UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

MRC Lifecourse Epidemiology Unit

Statistical approaches to the analysis of hierarchical data using simulations and real data from a study of musculoskeletal symptoms

by

Georgia Ntani

Thesis for the degree of Doctor of Philosophy

March 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Statistical epidemiology

Doctor of Philosophy

STATISTICAL APPROACHES TO THE ANALYSIS OF HIERARCHICAL DATA USING SIMULATIONS AND REAL DATA FROM A STUDY OF MUSCULOSKELETAL SYMPTOMS

by Georgia Ntani

Clustering of observations is a common phenomenon in epidemiological research. A first objective of this thesis was to explore the situations in which failure to account for clustering in statistical analysis could lead to erroneous conclusions. Using simulated data, I showed that effects estimated from a naïve regression model that ignored clustering were on average unbiased when the outcome was continuous, but were biased towards the null when the outcome was binary. The precision of effect estimates was overestimated when the outcome was binary, and also when both the outcome and explanatory variable were continuous. However, in linear regression with a binary explanatory variable, the precision of effects was somewhat underestimated by the naïve model. The magnitude of bias, both in point estimates and their precision, increased with greater clustering of the outcome variable, and was influenced also by clustering in the explanatory variable.

A second aim was to compare analytical approaches to clustering when synthesising results from multiple studies. Using real data from a large multicentre study, I showed that odds ratios (ORs) estimated from meta-analysis of summary results from component sub-studies were generally similar to those from multi-level modelling of pooled individual data. However, the precision of point estimates from meta-analysis was lower than that from multi-level analysis. Discrepancies between the two methods (including differences in ORs up to 27% and in precision up to 46%) were demonstrated when the outcome of interest was rare.

A third aim was to compare different methods for estimation of relative risks (RRs) when data are clustered. The random-intercept complementary log-log model produced estimates of effect and precision similar to those from the random-intercept log-binomial model (considered to be the best approach, but not always practical). Other models gave effect estimates close to those from the log-binomial model, but with less comparable precision. Contrary to the situation when RRs are being estimated in a set of independent (i.e. unclustered) observations, the random-intercept

Poisson model with robust variance produced less precise point estimates than those from the random-intercept log-binomial model.

Priorities for future work include exploration of: the consequences of ignoring clustering in the presence of effect modification and when marginal methods of analysis are used; situations in which meta-analytical estimates differ from those derived by pooled analysis; and specific situations in which the random-intercept Poisson model with robust variance is less likely to produce results similar to those from the random-intercept log-binomial model.

Contents

List of tables

List of figures

Abbreviations

CI	confidence interval
CLL	complementary log-log
GEE	generalised estimating equations
ICC	intraclass correlation coefficient
MSD	musculoskeletal disorders
OR	odds ratio
RI	random-intercept
RR	relative risk
SE	standard error

Declaration of authorship

I, Georgia Ntani declare that the thesis entitled “Statistical approaches to the analysis of hierarchical data using simulations and real data from a study of musculoskeletal symptoms” and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

this work was done wholly or mainly while in candidature for a research degree at this University;

where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

where I have consulted the published work of others, this is always clearly attributed;

where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

I have acknowledged all main sources of help;

where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

none of this work has been published before submission

Signed: ………………………………………………………………………..

Date:……………………………………………………………………………….

Acknowledgements

Here I am, at the end of my PhD that wouldn't have been possible if it wasn't for the help and support of several people who accompanied me in this journey. I am very grateful to both my supervisors David Coggon and Hazel Inskip for their ongoing support, generosity with time, practical help and constructive feedback whenever I needed it. I have greatly benefited and learnt a lot from their experience and their insight into the fields of epidemiology and statistics. I am also particularly grateful to David for his constant motivation and encouragement with the algebraic modelling developed in my PhD. With his exceptional mathematical skills and his clarity of mind he kept on giving me useful directions on how to complete this section.

The Colt Foundation provided financial support these four years. Most importantly, though, they have shown great interest in my work and have supported me in communicating parts of it in scientific meetings.

Holly Syddall has been an amazing office mate that never stopped showing interest in my thesis progress providing necessary chocolate supplies in times of need. My colleagues and friends Stefania D'Angelo and Elena Zaballa were always available for a chat and they never seemed to get tired with all my moaning when things were not working as planned. My dear friends Vasko, Maria, and Froso encouraged me and were always there for discussions in which little was said but much understood. The "heroes in white trousers" kept me sane all these years, with every single game being the best mechanism of stress-relief one would wish for.

Lastly, but most certainly not least, I want to thank my family. My Dad and my Mum for always being there for me, supporting me, and most importantly for helping me keep the right perspective in life. My brother (in law), Dionisis, and my sister, Poly, for their subtle and most effective ways of helping me get through moments of crisis and for always finding the time to listen to my most subtle concerns even in their busy days of early parenthood. Yianni, you have been my partner, mentor, friend, and my family these four very long years. You have supported me through all these roles, have put up with me, and have kept my feet on the ground like no one else would ever manage to. My deepest gratitude to you.

Στην ανιψιά μου

Chapter 1. Introduction

Musculoskeletal pain is that which affects the nerves, bones, tendons, and muscles. Pain resulting from musculoskeletal disorders (MSDs) is very common among people of working age and often results in morbidity and disability. MSDs are the leading cause of sickness absence in both European and non-European countries. In Europe, the cost associated with MSDs is estimated to be as high as 2% of Gross Domestic Product (GDP).

Several studies have investigated risk factors for MSDs. Most of these studies have focused on the risk that personal characteristics, such as sex and age, and occupational activities, such as lifting and bending, carry for the occurrence of musculoskeletal pain. A large body of evidence suggests that beyond these risk factors, psychological aspects of work, including job support, control, and satisfaction, also play an important role in MSDs. Additionally, many epidemiological studies have highlighted the impact of personal psychological factors, such as poor mental health and tendency to somatise, on the occurrence and persistence of musculoskeletal pain and related disability. However, even in combination, these established risk factors cannot adequately explain the large variation in the prevalence of disability from musculoskeletal pain that has been observed between populations and within some populations over time.

This gap in understanding prompted the hypothesis that culturally-determined health beliefs and expectations also have an important influence on common musculoskeletal complaints. To test this hypothesis, a multicentre study, the CUPID study was designed. In the course of this study, participants were recruited from 47 distinct occupational groups in 18 countries, and information was collected on a variety of musculoskeletal pain outcomes and potential risk factors, while local investigators also provided information on socio-economic group-specific factors, such as unemployment rates, social security provision for unemployment, and compensation for work-related MSDs.

The way in which participants were recruited and data were collected, gave a hierarchical form to the data from the CUPID study, individual participants being uniquely assigned to each of the 47 distinct occupational groups. This hierarchical structure provided a powerful and flexible resource with which to investigate the study hypothesis. However, it also posed a number of statistical challenges and raised several methodological questions that are the focus of this thesis.

A first challenge was to characterise the phenomenon of clustering, and to explore when it occurs and in what situations it is likely to affect statistical inference. In particular, how does the impact of failing to account for clustering depend on the nature and distribution of the main outcome and explanatory variables under investigation?

Beyond the importance of accounting for clustering, the methods by which clustered data can be analysed were a further interest. One frequently encountered situation in which clustering of data is likely to occur is meta-analysis. Usually, the results from different studies are combined by modelling based on the effect estimates from each component study and their standard errors. More rarely, however, individual data from each study are combined in a pooled analysis. It is important to know how well the results from these alternative methods of analysis agree. The hierarchical data in the CUPID study could be considered as coming from independently conducted studies in each occupational group, and this provided an opportunity to compare different methods of meta-analysis using a single dataset.

Finally, of great interest also was the optimal method of estimating relative risks for binary outcomes when data are clustered. The main outcomes in the CUPID study were all binary, and prevalence ratios would provide the most readily interpretable measure of association with such outcomes. However, while statistical models that can be used to estimate prevalence ratios have been widely explored for data that are not clustered, less is known about their relative merits where data are clustered.

In the first chapter of the thesis, I introduce the problem of musculoskeletal disorders with a brief description of epidemiological evidence on established risk factors. I then describe the hypothesis about cultural and psychosocial influences on common musculoskeletal complaints that led to the CUPID study, the methods of which I outline at the end of the chapter. In the second chapter, I describe the phenomenon of clustering and present approaches to the analysis of clustered data, leading to a description of the main research questions that this thesis aimed to answer. In the third chapter, I describe the approach that I followed to search for evidence that was already available from the published literature on the consequences of ignoring clustering. In the fourth and fifth chapter, I explore the consequences of ignoring clustering in linear and logistic regression, respectively, using data derived from Monte Carlo simulations. In each of these chapters, the consequences of ignoring clustering are explored separately for continuous and binary explanatory variables. Each case is introduced by first considering the simplest scenario in which observations come from only two clusters, before moving on to consider multiple clusters. For the case of a continuous outcome variable, algebraic calculations were also made to check generalisability of findings. In Chapter 6, I extend this exploration by fitting naïve (linear and logistic) regression models with dummy variables for the clusters in a subset of the simulated datasets used in previous chapters (Chapter 4 and Chapter 5). In Chapter 7 I use data from the CUPID study to compare risk estimates derived from meta-analysis of summary results with those derived from analytical methods that use all primary data (pooled analysis of individual data). Chapter 8 compares different methods for estimating prevalence ratios when data are hierarchically structured. For that, I used data from the CUPID study to fit various regression models and derive estimates which could be interpreted as relative risks. In Chapter 9, I review the main findings of the thesis, summarise the key conclusions that can be drawn from each chapter, and discuss directions for further research in the area of multilevel data analysis.

1.1 Musculoskeletal disorders

The term, musculoskeletal disorders (MSDs), refers to various conditions that affect the nerves, tendons, muscles, and supporting structures of the body. MSDs may affect many areas of the body including the neck, shoulders, wrists, upper or lower back, and knees, where they can cause symptoms ranging from mild discomfort to debilitating pain and disability. In some cases, symptoms arise from identifiable disease of tissues, but often there is no evidence of specific underlying pathology (1).

The burden from MSDs has been described in various studies. A large community-based postal survey of individuals of working age in the UK showed that 28% had upper limb pain lasting from 7 days to 6 months, 14% for >6 months, and 10% had pain characterised as disabling (2). Another large cross-sectional survey of a UK general practice population reported a 34% prevalence of neck pain in the past 12 months, 20% in the past week, and 11% prevalence of neck pain interfering with normal activities (3). A systematic review on shoulder pain reported that the 1-month prevalence in the studies reviewed varied from 19-31%, that of 1-year prevalence from 5-47%, and that of lifetime prevalence from 7-67% (4). Back pain alone accounted for 25% of pain regarded as troublesome in the past month in a sample of 4049 surveyed people in a cross-sectional study in the UK (5), while a higher prevalence (40%) was reported in a survey in Great Britain in which a broader definition of the symptom was used (6).

These symptoms can lead to disability and need of health care, with a parallel important loss of time from work. A systematic analysis for the Global Burden of Disease Study in 2010 showed that low back pain was in the top ten disorders impacting on disability-adjusted life years, with other MSDs also being very common in many of the 21 regions examined (7). In Europe, MSDs are the biggest cause of incapacity for work with costs reaching 2% of gross domestic product (8). Also, MSDs often result in long-term sickness absence, impacting on both employers and employees. Sickness absence leads to loss of productivity, which may also occur when people experiencing musculoskeletal symptoms are present at work but cannot perform to their normal capacity (9).

A substantial body of epidemiological research has shown that MSDs are associated with occupational physical activities, such as lifting, bending, forceful repetitive movements and work with sources of vibration (10-13), characterising them as work-related disorders. Added to that, further research has shown that psychosocial aspects of work play an important role in the occurrence of musculoskeletal pain. These include job demands, lack of support and control, dissatisfaction with work, and job insecurity. The association between these factors and MSDs has been widely studied in cross-sectional and longitudinal studies. A recent systematic review exploring the association between psychosocial stressors and musculoskeletal problems, gathered evidence from 45 longitudinal studies and presented the pooled effects of a number of different stressors on pain in the lower back, neck/shoulder, and upper and lower extremity (14). Most of the effects on the onset of musculoskeletal problems in this large review were small (odds ratios ranging from 1.15 to 1.66) but statistically significant.

Additionally, tendency to report and worry about common somatic symptoms has been linked to musculoskeletal pain. This association has been shown in many epidemiological studies. Most have been cross-sectional in design, and have shown strong associations with the prevalence of regional musculoskeletal symptoms (15-19) as well as that of pain in multiple sites of the body (2, 20, 21). Longitudinal studies have found that people who tend to somatise more are more likely to develop MSDs and associated disability (22-24). Additionally, poor mental health has been shown to be more common in people with musculoskeletal problems, in cross sectional studies (15, 25, 26). As for somatising tendency, longitudinal studies have indicated that people with poor mental health are more likely to develop new musculoskeletal pain (22, 24, 27).

1.2 Cultural and psychosocial influences

Musculoskeletal symptoms sometimes arise from identifiable pathology, but most often, underlying pathology cannot be established, in which case the symptoms are characterised as "non-specific". The frequent occurrence of musculoskeletal complaints in the absence of identifiable underlying pathology, reinforces the argument that psychological factors are a key element in understanding musculoskeletal disorders (28). Indeed, several studies have shown a robust association of psychosocial factors with reporting of symptoms and musculoskeletal health outcomes (29). Also, a recent study of patients investigated for possible carpal tunnel syndrome showed that psychosocial risk factors were more strongly associated with non-specific than specific musculoskeletal pain (30).

However, the identified occupational risk factors together with the known personal psychological risk factors cannot explain the major variations in the prevalence of disability from that have been observed between populations and within populations over time.

Health beliefs have also been shown to be associated with musculoskeletal outcomes. A systematic review of randomised control trials of medical advice regarding acute back pain showed that shifting people’s beliefs from a traditional perspective of bed rest to positive action that promotes physical activity can result in reduction of back pain prevalence and its impact (31). Similarly, an interventional study aiming to change beliefs about back pain and related disability, showed that a positive shift in beliefs was followed by a reduction in disability and workers’ compensation costs related to back pain (32).

Health beliefs and expectations are importantly driven by a person’s culture and environment. Culture may affect perception of health, how illness and pain are experienced, and understanding of what the possible causes of a specific illness might be (33). Furthermore, the cultural and societal environment that people experience may influence their perception of wellbeing (34).

The above considerations prompted the hypothesis that musculoskeletal illness and related disability which cannot be attributed to a detectable organic pathology, may be importantly influenced by health beliefs and expectations that are partly shaped by the individual’s characteristics and partly by the cultural environment that they experience (35).

1.3 The CUPID study

To examine the hypothesis that culturally-determined health beliefs and expectations have an important influence on common musculoskeletal complaints in addition to that of well-established risk factors, the CUPID (Cultural and Psychosocial Influences on Disability) study was designed. The objective of the CUPID study was to compare the prevalence of musculoskeletal pain and related disability among workers who were similarly exposed to biomechanical occupational risk factors but came from different cultural backgrounds. Additionally, the CUPID study aimed to explore risk factors for prevalence of symptoms and disability.

The methods of the CUPID study have been described in detail elsewhere (36). In brief, data were collected during 2006-2011 in 18 countries. In each country, participants were recruited from 1-4 occupations, providing a total of 47 occupational groups, with occupational group being defined by the combination of occupation and country (Table 1.1). Participants were aged between 20 and 59 years and had all been in their current job for at least 1 year. Information was collected through a questionnaire that was self-administered or administered at interview.

The questionnaire included sections about demographic characteristics; height; age that full-time education was completed; smoking habits; current occupation; pain at different anatomical sites and related disability; sickness absence in the past 12 months due to musculoskeletal problems or other illness; awareness of others with musculoskeletal pain; awareness of repetitive strain injury (RSI) or similar terms; adverse beliefs about musculoskeletal pain; and personal psychological factors.

Table 1.1. Countries and occupations from which participants were recruited within country.

Country	Abbreviation	Occupations from which participants were recruited within country
Brazil	BR	Office workers, nurses, sugar cane cutters
Ecuador	EC	Office workers, nurse assistants, flower plantation workers
Colombia	CO	Office workers
Costa Rica	CR	Office workers, nurses, telephone call centre workers
Nicaragua	NI	Office workers, nurses, machine operators
UK	UK	Office workers, nurses, mail sorters
Spain	SP	Office workers, nurses
Italy	IT	Nurses, assembly line workers
Greece	GR	Office workers, nurses, postal clerks
Estonia	EE	Office workers, nurses
Lebanon	LB	Office workers, nurses, food production workers
Iran	IR	Office workers, nurses
Pakistan	PK	Office workers, nurses, mail sorters
Sri Lanka	LK	Office workers, nurses, mail sorters (other workers 1), sewing machinists (other workers 2)
Japan	JP	Office workers, nurses, transportation operatives (other workers 1), sales workers (other workers 2)
South Africa	SA	Office workers, nurses
Australia	AU	Nurses
New Zealand	NZ	Office workers, nurses, mail sorters

Information about current occupation included number of hours worked per week, occupational physical activities, and psychosocial aspects of work. Participants were asked whether an average day involved: lifting weights of ≥25kg by hand; working with the hands above shoulder height for more than 1 hour; use of a computer keyboard or other repetitive movement of the wrist or fingers for longer than 4 hours; kneeling or squatting for longer than 1 hour; and repeated bending or straightening of the elbow for longer than 1 hour. The psychosocial aspects of work on which participants provided information were: time pressure, incentives, and job control, satisfaction, and security.

Questions about musculoskeletal pain focused on six anatomical sites. These were the low back, neck, shoulder, elbow, wrist/hand, and knee. For the last four sites information was provided for both the right and the left side. Participants were asked about pain in each of the above mentioned regions over the past month and the past 12 months. Related disability was assessed by questions asking about the level of difficulty that participants had experienced in doing specified everyday activities due to the site-specific pain. Along with pain and related disability, participants reported the number of days of sick leave taken over the past year due to pain in the specified region.

Adverse beliefs about musculoskeletal pain were assessed by questions adapted from the Fear Avoidance Beliefs Questionnaire (37). Through these, participants were considered to have adverse beliefs about: work-relatedness, if they believed that musculoskeletal pain is commonly caused by work; physical activity, if they believed that physical activity should be avoided and rest is needed for pain to get better; and prognosis, if they believed that neglecting such problems can cause permanent health problems and that such problems do not usually get better within three months.

Distress from somatic symptoms was ascertained using questions taken from the Brief Symptom Inventory (BSI) (38), which provided a measure of participants' tendency to somatise. Mental health was assessed through questions taken from the Short Form-36 (SF-36) questionnaire (39). These questions led to a score that was grouped into thirds of its distribution, creating three mental health categories: good, intermediate, and poor.

In addition to data collected from study participants through questionnaires, local collaborators provided information about group-level socio-economic factors specific to the occupational group. This included the local unemployment rate at the time of the survey, entitlement to sick pay in the first three months of absence, availability of social security support for the unemployed, financial support for ill health retirement, whether fees were paid for primary medical care and entitlement to compensation for work-related musculoskeletal disorders.

Chapter 2. Clustering of data

2.1 Introduction to problem

The design of the CUPID study, as described in section 1.3, meant that the data had a hierarchical structure. This was because observations for individuals within an occupational group were likely to be more similar to each other than those for individuals from different groups. That similarity could occur for several reasons. Firstly, individuals are not randomly assigned to occupations (40). They normally select their job, and the decision is often influenced by personal characteristics (e.g. personality, academic ability, social class and physical attributes such as strength, mobility, agility etc). Thus, specific occupations tend to gather individuals of similar profiles. Also, people within the same occupational environment interact and thus are likely to influence one another. One person's beliefs and expectations are liable to influence those of others in the same group. Of particular importance, health beliefs about musculoskeletal pain are likely to be shaped by the conditions and cultural environment that an individual experiences in the workplace (41). Additionally, some factors, such as local unemployment rate, will apply similarly to all individuals in the same group. Consequently many relevant variables are likely to be more similar among individuals belonging to the same occupational group than in individuals from different occupational groups, and particularly from different countries.

The greater similarity within as compared with between occupational groups might apply not only to risk factors of interest, but also to the pain outcomes under study. As mentioned earlier, musculoskeletal pain has been shown to be predicted by psychosocial factors such as low mood and tendency to somatise. As the profile of these psychosocial factors along with other unmeasured parameters (potentially associated with the outcome of interest) is likely to be more similar among people from the same occupational group, one would expect the same to apply to the musculoskeletal outcomes. For example, in an occupational group with a high prevalence of poor mental health and tendency to somatise, one might expect also a correspondingly high prevalence of musculoskeletal pain.

Such grouping of observations may have implications for statistical inference. This can be illustrated using the example of a study to explore the association between the occurrence of low back pain and the prevalence of adverse health beliefs about prognosis of musculoskeletal pain in the occupational group to which a person belongs (categorised as 'high' or 'low'). Suppose that this study uses a sample from 40 occupational groups, each comprising 100 individuals, and that 20 of the 40 occupational groups have a high prevalence of adverse beliefs about prognosis of pain, while the remainder have a low prevalence (Figure 2.1). That equates to 20 groups, or 2000 individuals, that are exposed to the risk factor and another 20 groups, or 2000 individuals that are not. In the first group of 2000 individuals (those with high group prevalence of adverse beliefs about prognosis) there are 300 people who have low back pain, while in the second group of individuals there are 100 people with low back pain. That implies a prevalence of 15% in 2000 exposed people as compared with 5% in 2000 unexposed individuals. It might be, however, that low back pain is very strongly clustered within occupational groups, and that the 300 exposed individuals with low back pain all come from just three occupational groups (each with 100% prevalence), while the 100 unexposed individuals with low back pain all come from a single occupational group (again with 100% prevalence). Then the comparison of exposed to unexposed individuals (300/2000 vs 100/2000) with low back pain essentially becomes a comparison of exposed to unexposed occupational groups (3/20 vs 1/20) in which individuals have low back pain. The comparison remains 15% (3/20 groups) versus 5% (1/20 groups), but the sample on which this comparison is made is 40 rather than 4000. So, if we wrongly consider that the comparison is between individuals rather than between groups of individuals (i.e. a sample size of 4000 as opposed to 40), our estimate of the strength of association will be spuriously precise.

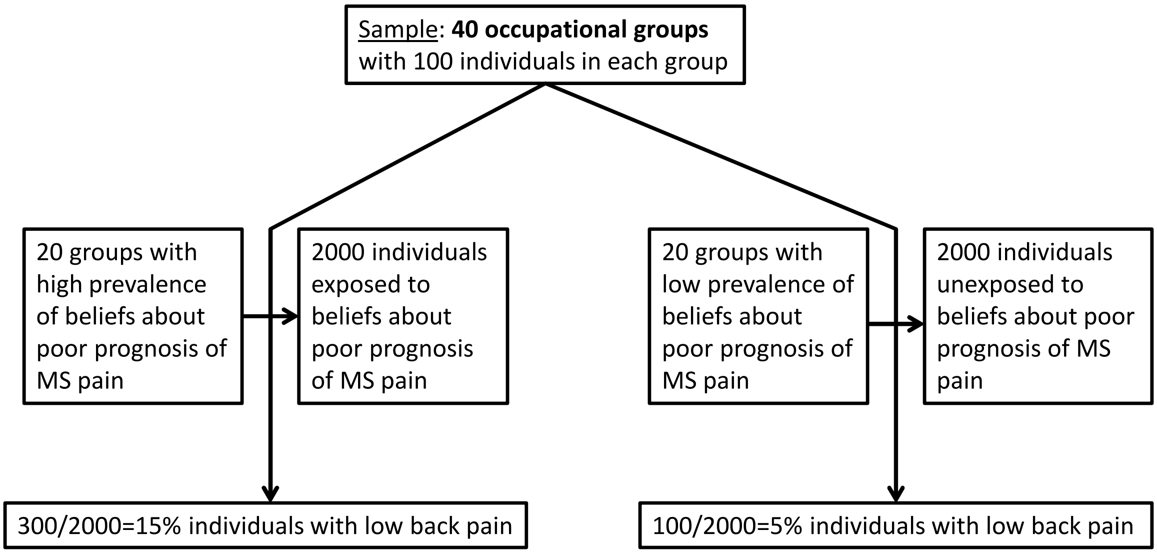

Sample: **40 occupational groups** with 100 individuals in each group
20 groups with high prevalence of beliefs about poor prognosis of MS pain
2000 individuals exposed to beliefs about poor prognosis of MS pain
20 groups with low prevalence of beliefs about poor prognosis of MS pain
2000 individuals unexposed to beliefs about poor prognosis of MS pain
300/2000=15% individuals with low back pain
100/2000=5% individuals with low back pain

Figure 2.1. Example of study exploring association between low back pain and group prevalence of adverse beliefs about prognosis of musculoskeletal pain

The above example is also formulated numerically in Table 2.1. A comparison between individuals would indicate a three-fold increased risk of low back pain when exposed to a high group prevalence of adverse health beliefs about the prognosis of musculoskeletal pain. In a sample of 4000 individuals, such a ratio would be highly significant statistically, with a narrow confidence interval (95% CI: 2.4, 3.7). A comparison between groups, however, even though it would yield the same risk estimate, would carry much greater statistical uncertainty (95% CI: 0.3, 26.5), reflecting the small number of data points (N=40).

Table 2.1. Results from hypothetical example on association between low back pain and group prevalence of adverse beliefs about prognosis of musculoskeletal pain

Low Back Pain cases / Exposed	Low Back Pain cases / Unexposed	Relative Risk (95% CI)
300/2000 individuals	100/2000 individuals	3 (2.4, 3.7)
3/20 groups	1/20 groups	3 (0.3, 26.5)

In another example, but in a similar context of 40 occupational groups, consider a study investigating the association between number of hours worked per week and age. Assume that there is a very tight similarity between observations within each of the occupational groups, such that each group comprises individuals very close in age and with very similar (but not the same) working patterns (in terms of number of hours worked per week). In such a scenario, the variance due to differences between occupational groups would be much larger than the variance within groups. Thus, ignoring grouping of observations in occupational groups, the variance due to differences in the higher level (groups), which was larger, would be mixed with variance between individuals within the same group, which was smaller. That would result in inappropriate uncertainty around the effect of age on number of hours worked per week. By acknowledging and separating the sources of variation (due to occupational groups, and due to individuals within occupational groups), one might get a more accurate estimate of precision for the association between the two variables.

2.2 Definition of clustering and scope

Clustering of observations within identifiable subsets of a population is a common phenomenon, and is encountered in many different disciplines. For example, in an educational context, students' scores for achievement are likely to be more similar between students who belong to the same class or the same school than between students from different schools (42). In a political context, voting behaviours tend to be more similar among individuals from the same region, and even more so among those who share the same local environment (43, 44). In medicine, patients registered with the same general practice are likely to be more similar to each other than patients under the care of different practices. In sociology, people from the same neighbourhood are more likely to exhibit the same social behaviour than people living in different neighbourhoods (45). In genetics, individuals who come from the same family are more likely to have a similar genetic profile than those from different families. In growth trajectories, repeated observations over time from one individual are more likely to be similar than observations from different individuals (46).

Where nesting of observations occurs, it may have implications for statistical inference, depending on the answers to two main questions. First is the extent to which the observed clustering is a feature of the population as a whole, and not simply of the particular study sample that has been drawn from that population. Second is whether the clustering of outcome variables can be explained by the predictor variables (if any) that have been measured.

Regarding the first consideration, the phenomenon of similarity can be considered a population characteristic if it would still be present in a sample consisting of all possible elements from the set of observations that could be made. In contrast to that, if similarity of observations is not a population characteristic, then possible similarities observed within groups (and parallel dissimilarities across groups) of observations in a study sample would be attributable to chance alone. If the clustering of the observations in the outcome variable of a study sample has occurred simply by chance, then it should not affect statistical inference. Otherwise clustering would occur in all studies since one could always define it a posteriori in a sample according to the observed distribution of the outcome variable.

When clustering does occur in the wider population, the observed similarities of the outcome variable will be attributable to explanatory variables, which may or may not be known. When relevant explanatory variables have been identified and measured, and subsequently are accounted for in the statistical analysis, then the clustering of observations is likely to be reduced or even completely eliminated, and where the clustering is eliminated, standard methods of statistical inference can again be applied .

There will, however, be implications for statistical inference if there is clustering of an outcome in the wider population that cannot be fully explained by measured explanatory variables. The extent to which clustering of an outcome variable impacts on statistical inference will depend on the degree of clustering after allowance for measured explanatory variables – i.e. the level of variation across groups compared to the within-group variation after adjustment for measured explanatory variables. Such variables may take the same value for all members of any given group, or they may vary between individuals within a group. For example, in a study of social behaviours, variables of the first type might be prevalence of low-educated residents in the neighbourhood (measured at the group level), and the presence or absence of resources such as reliable public transport, while variables of the second type could be age or sex (which are measured at the individual level). The residual clustering must depend on other relevant parameters that are unknown, or at least unmeasured. As for the measured parameters, these unmeasured parameters can be at the lower level of observation or the group level.

In studies where clustering occurs, it is common practice to partition observed variation into two levels – the group level and the individual level. High within-group similarity of observations can be thought of as lower variation of the outcome of interest within-groups (individual-level) compared with higher variation across groups (or group-level variation). For example, consider the parameter of social behaviour measured in groups of individuals from different neighbourhoods. Similarity of social behaviour between individuals living in the same neighbourhood can also be expressed as higher variation in social behaviour across neighbourhoods compared with variation between individuals within neighbourhoods.

In summary, when data come from a population comprising identifiable groups, such that after allowance for measured explanatory variables, values for an outcome variable show greater similarity within groups than across the whole population, they are said to be clustered or hierarchical, and the groups are termed clusters.

The magnitude of clustering is often quantified by a measure called intracluster correlation coefficient (ICC) (also referred to as intraclass correlation coefficient). This measure is defined as the ratio of the variance in the outcome variable between groups to the overall variance in the model (between and within groups). ICC can thus be expressed as the proportion of variance in the outcome measure that is explained by the group level. When most of the variation is found between individuals of the same cluster and the cluster means of the outcome variable vary little, the between-cluster variance will be low and thus the ratio of the ICC will be very close to zero. In that situation there is little clustering. On the other hand, when the variation is mostly observed at the group level, then the ICC will be close to one.

2.3 Approaches to the analysis of clustered data

Several approaches are used in the analysis of clustered data. These can be grouped into four broad categories: i) ignoring the clustering of observations ii) treating the clusters as independent observations (aggregation method), iii) fixed effects regression analysis using a dummy variable for each cluster, and iv) statistically accounting for clustering without estimating cluster effects. To illustrate the approaches to the analysis of clustered data, I consider the example of a study of students from a number of different schools. Here the lower unit of observation is the student and the higher level of observation is the school which the student attends. Let the outcome of interest be the student's score in maths and the main predictor variable be gender.

I) Ignoring clustering of observations

Ignoring clustering of observations is an approach that was widely used before multilevel analytical techniques became available. Researchers following this approach would consider observations at the lower level to be mutually independent and analyse the data disregarding their common element, the clusters, which in this example are the schools. In the case of a continuous outcome variable, a typical analysis technique would be an ordinary least squares (OLS) regression, and in the case of a binary outcome, ordinary logistic (OL) regression. With this approach, in the example of the students nested within schools, the score in maths would simply be regressed on gender using an OLS model.

This approach, often referred to as a disaggregation approach (47), has been widely discussed with regard to potential biases in estimates of parameters and their precision. These will be discussed in more detail in Chapter 4 and Chapter 5. In brief, if data are pooled from, say, N clusters with n observations in each cluster, to create a single large database and the inter-dependency of observations is ignored, parameters are estimated using the whole sample of $N \times n$ individuals. When observations are not independent and there is a strong within-cluster dependency, increasing the number of observations in a cluster will only increase the information about a given association minimally. However, the apparent statistical power will increase substantially. For example, consider an extreme case in which an outcome and an exposure variable are measured 100 times in each of 100 people and both of the variables are time invariant. That would result in clustered data in which the measurements were clustered within individuals. As both variables are time-invariant, all of the observations within any given cluster will be identical. Ignoring the clustering of observations would mean that the association between exposure and outcome was estimated on a sample of 10000 observations (100 measurements for each of the 100 clusters/individuals) whereas the association should be viewed more appropriately as being estimated from a sample of 100 measurements (one per person). Failure to account for the clustering would be likely to produce over-precise estimates of associations, and possibly false-positive results, depending on the level of clustering. It should be noted however, that such extreme examples are rare and while there is concern about over-precise estimates, depending on the degree of intraclass correlation, there may or may not be differences between analyses that account for clustering and those that do not.

II) Treating clusters as independent observations

This method is often referred to as aggregation (47). In aggregation, observations made at the lower level are used to derive summary measures for each cluster, commonly the mean of observations per cluster, which are then used as measures specific to the higher level (cluster level). Analysis is then conducted with clusters rather than individuals as the unit of analysis. This approach ensures independence of observations as they each result from different clusters, and thus standard analytical techniques can be applied.

In the example of the students who are clustered within schools, an average measure per school could be created for maths score and a prevalence measure of males could be created for the students in each school. Then the number of observation points would reduce to the number of schools, and using standard techniques, such as OLS linear regression, one could explore the association between maths score and prevalence of males.

By reducing the number of observations to the number of clusters, this approach means that one can confidently analyse data without violating assumptions of independence of observations. On the other hand, the sample size is reduced (sometimes markedly, depending on the number of observations per cluster in relation to the number of clusters) which may result in considerable loss of statistical power. Also, the method does not take into account the size of the clusters; some clusters may be larger in size and thus contribute more information than other clusters that are smaller in size. In aggregation methods, no weights are assigned to clusters, and unweighted analyses may result in inappropriate conclusions. Additionally, as information is summarised at the upper level of observation, information at the individual level is lost. Analysis using individual level data from each cluster may not give the same results as analysis of averaged data, a discrepancy called the ecological fallacy, which has been discussed widely in the context of ecological studies (48).

III) Regression analysis with dummy variables for clusters

This method uses conventional regression techniques that incorporate information on clustering. That is done simply by adding indicator variables for the clusters to which individuals belong in the regression model. In the above example, if the number of schools from which students were recruited were *N*, then, *N-1* dummy variables could be included in a statistical model that was fitted to estimate the association between maths score and gender. When the outcome of interest, in this case the maths score, is continuous and normally distributed, this method is equivalent to an Analysis of Covariance (ANCOVA). It has also been described as the least squares dummy variable (LSDV) approach (49).

With this approach, when clusters are fitted in the model as dummy variables, the analysis “controls for” cluster effects and estimates can be made of the main effect of the predictor on the outcome variable, which is assumed to be the same (fixed) across clusters. Even though this is a statistically valid method, it can only be applied when the explanatory variable of interest varies between individuals within clusters (individual-specific variable). Where the explanatory variable is specified at cluster level, its effect cannot be estimated because it cannot be discriminated from the effects of the cluster dummy variables (knowing the cluster determines the explanatory variable). To illustrate this point, consider an example in which we wish to investigate the effect of type of school (state vs private) on maths score in a sample of students recruited from 20 different schools, 10 of them being state schools and 10 private schools. A LSDV model would then be of the form:

$$Maths\ score = a + \sum_{i=2}^{20} \beta_i School_i + \beta_{21} Private + \varepsilon.$$

with a being the intercept, β_i with $i = 2, \ldots, 20$ being the estimated effect of each school i on the outcome as compared to the first school ($i = 1$), β_{21} being the estimated effect of private schools on the outcome as compared to state schools, and ε being the residual variation. In this model, information about the difference in maths scores between students in the two types of schools (β_{21}) is accounted for by the dummy variables for individual schools.

The confounding of effects between the cluster-level variables and the dummy variables for the clusters can be better explained through the algebraic representation of a linear regression model. Consider a sample of n individuals and a multiple linear regression model of the form

$$y = X\beta + e \qquad 2.1$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ x_{21} & \cdots & x_{2N} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nN} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

The columns of the matrix $\boldsymbol{X}$ are for the covariate variable vectors *N*. Then, based on the OLS method, the estimate of the unknown parameter β will be

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

In our case, the first *N-1* variables (x_i to x_{N-1}) will be cluster-indicator variables, and the last variable will be the cluster-level variable (x_N) the effect of which we wish to estimate. As the last variable is invariant within clusters, it will be a linear combination of the cluster-indicator variables. For that reason, the matrix X^T will have rows that correspond to linearly dependent row-vectors and the product of the two matrices $X^T X$ will be a matrix with rows that correspond to linearly dependent row-vectors. It follows that the determinant of $X^T X$ will be 0. Thus $X^T X$ will not be invertible and the effect of the last variable cannot be estimated.

Where the predictor under investigation is measured at the lower level of observation, the LSDV method can be a statistically valid approach and is often preferred when the number of clusters is small. However, as the number of clusters increases, fitting dummy variables to account for cluster differences may lead to loss of statistical power as many degrees of freedom are used.

IV) Accounting for clustering without estimating cluster effects

In a simple situation in which the interest is in testing the null hypothesis that there are no differences between groups of observations in the presence of clustering, conventional statistical tests have been adapted to account for within-cluster correlation of observations. Such tests depend on the distribution of the outcome variable. For continuous and binary outcomes the adapted conventional t-test and χ^2 test have been described by Gonen et al (50). For continuous data, that are not normally distributed, the methods used to test for differences between two or more groups of observations are the Wilcoxon rank-sum test and the Kruskal-Wallis test. In the context of clustered data these have also been adapted separately for cases in which the predictor variable is cluster-specific (51, 52) (i.e. a cluster-level variable) and individual-specific (53) (i.e. an individual-level variable). Similar adaptations have also been developed for paired observations that are clustered (54, 55) (for example, the right and left shoulders of individuals who are clustered within countries).

Beyond simple statistical tests, statistical modelling that accounts for clustering effects can be used when more complex analysis of data is required (i.e. estimation of adjusted associations). Such statistical modelling includes clustered robust estimation, generalised estimating equations (56) (GEEs), and random effects (RE) modelling (57, 58).

Clustered robust estimation works in the same way as an OLS model. The difference is that the method relaxes the assumptions of independence of observations. In other words, the observations are assumed to be independent across clusters, but not necessarily within clusters. Using this method, point estimates of parameters are unaffected, but their standard errors are modified depending on the level of within-cluster correlation of observations and the number of observations within each cluster.

GEEs are marginal methods. In GEEs the mean response and the within-cluster correlations are modelled separately, considering the former as the primary focus and accounting for the latter to produce valid conclusions. GEEs are most commonly used to analyse longitudinal data (59) in which there are repeated measures within individuals and the objective is to describe the marginal expectation of the response variable as a function of the explanatory variables, while taking into consideration the within-cluster correlation of the observations.

Random effects modelling is a conditional method, in that it estimates the effect of the explanatory variable of interest conditional on the cluster. This model also relaxes the assumptions about independence of observations and distinguishes between the two sources of variation in the outcome variable – i.e. within and between clusters. Applying this approach, we can allow the average value of the outcome to vary across the clusters, assuming a constant effect of the explanatory variable on the outcome variable (a random-intercepts (RI) model), or we can allow both the average value of the outcome and the effect of the explanatory on the outcome variable to vary across clusters (a random-effects (RE) model). These models, however, are subject to assumptions concerning the distributions of the random intercepts and the random effects that one has to consider when implementing them. Despite these assumptions, however, random effects modelling is often preferred to dummy variable regression analysis when the number of clusters is large due to the lower number of degrees of freedom used.

Several studies have assessed the similarities and dissimilarities of the RE and GEE methods (60, 61) and differences in the effect estimates that they generate (62-65). The conclusion from these studies has been that researchers must be cautious when choosing between marginal and conditional methods as the two may answer different questions. Consider the example of maths scores in relation to gender among students from different schools. The marginal model (GEE) would estimate the effect of being a randomly selected male on maths score compared with a randomly selected female, independent of school. The conditional model (RE), on the other hand, would describe the effect of being a randomly selected male on maths score compared with a randomly selected female from the same school.

The above mentioned methods can be applied whether the outcome variable is continuous or categorical (binary, ordinal, or nominal). When the outcome under investigation is binary (0/1), then another approach to the analysis of clustered data is the conditional logistic regression model (66), which can be conceived as an extension of the RI logistic regression model. In the case of a continuous outcome, when the cluster-specific mean of the response variable is subtracted from the response variable, the cluster-mean-centred variable is generated and a model that uses the latter as an outcome variable corresponds to conditional maximum likelihood estimation. In the case of a binary outcome variable, the cluster-specific intercepts can be eliminated from the model by constructing a likelihood that is conditional on the number of positive responses (=1) in the cluster. That will correspond to the conditional maximum likelihood estimation known as conditional logistic regression. The drawback of this method is that only effects of explanatory variables that vary within clusters can be estimated.

2.4 Research questions

This thesis focuses on the analysis of clustered data. A major objective was to explore the situations in which clustering of observations needs consideration in the analysis stage of a study, the consequences of failing to account for clustering, and how that is likely to affect statistical inference. The thesis then compares two methods that are commonly applied in meta-analysis of data that are clustered by study population. Finally, it compares different methods for the estimation of relative risk (which is a frequently used measure of association when the outcome of interest is binary) when data are clustered.

The three major research questions on which this thesis will focus are: i) what are the consequences for statistical inference of ignoring clustering of observations, ii) how do estimates of association and related precision that are derived using meta-analysis of summary results compare to those from pooled analysis of individual data, and iii) how can relative risks be estimated when observations are clustered. These research questions are described in more detail below.

2.4.1 Consequences of ignoring clustering

As described in the previous section, clustered data are commonly encountered in epidemiological research. An increasing number of scientists have identified the phenomenon in existing data, while new designs of studies are being used in which data are hierarchically structured. Thus, it is not surprising that researchers are now increasingly aware of the importance of accounting for clustering, both in the statistical analysis of the data and in the interpretation of the results. However, there is still a large number of recently published studies in which clustering has been ignored (67-71). Possible reasons may be the complexity of models that account for clustering, and under-appreciation of the importance of the problem.

Several studies have addressed research questions around the problem of clustering in relation to statistical inference. These have clearly distinguished the effects of clustering according to whether the main explanatory variable is constant or varying within clusters (72). However, in practice, even where an explanatory variable varies within cluster, its variation across clusters may be much larger than that within clusters. Little is known about the effects of clustering is such a situation. Also, most of the methodological studies have considered explanatory variables that were continuous, and very few have addressed binary explanatory variables. No study so far has investigated the consequences of ignoring clustering according to different scenarios for the distribution of the explanatory variable. In addition, even though the implications of failing to account for clustering for the estimation of standard errors have been well described, effects on point estimates have received little attention.

The first objective of the thesis was to explore consequences of ignoring clustering in statistical inference using simulated data. Two different distributions of the outcome variable were examined: continuous and binary, using linear and logistic regression, respectively. Also, in each of the two cases (linear and logistic regression), both continuous and binary explanatory variables were considered, as shown in the table below. The investigation focused on how the distribution of the explanatory variable in relation to the cluster variable affects point estimates of risk and their precision, assuming different levels of within-cluster similarity of observations of the outcome variable. In addition, it explored coverage rates of true effect estimates by 95% confidence intervals, and the extent to which analyses that ignore clustering can give misleading estimates of type I error.

		Outcome	
		Continuous	**Binary**
Exposure	**Continuous**	Linear regression (section 4.3)	Logistic regression (section 5.3)
	Binary	Linear regression (section 4.5)	Logistic regression (section 5.5)

2.4.2 Meta-analysis and pooled analysis

Meta-analysis is a systematic approach to synthesis of results from independent studies that have explored a specified research question. The aim is to integrate the separate effect estimates reported from the independent studies into a combined "summary effect estimate". For example, the research interest might be the effect of lifting on the occurrence of low back pain. In a meta-analysis, effect estimates for lifting in relation to low back pain would be obtained from independent studies and synthesised to give an overall effect estimate for lifting. In a meta-analysis, the greater the similarity of the studies in terms of conditions under which they have been carried out, ascertainment of the main exposures and outcomes under investigation, and analytical approaches followed, the more valid the overall effect estimate will be considered. In contrast, when studies included in a meta-analysis have been conducted under different conditions, using different methods to ascertain exposures and outcomes, the overall meta-estimate of effect may be called into question.

Meta-analysis can be viewed as a special case of hierarchically structured data analysis. The independent studies can be considered as different clusters in a hierarchical dataset, made up of the individual participants in each study.

When information is available about the individuals participating in the studies one wishes to combine, it is also possible to analyse pooled individual data using various statistical techniques including fixed-effects models, which may or may not account for the study in which the individual took part, or random-effects models (if the number of studies is sufficiently large).

Previous research has suggested that pooled analysis of individual data is preferable to meta-analysis of summary results from each component study. Also, pooled analysis of individual data has been considered by some researchers as the gold standard for combining evidence from multiple studies (73). However, these conclusions were based on only a small number of studies that employed both statistical approaches, of which most pooled data from randomised control trials, and very few from observational studies. No study so far has compared results from meta-analysis and pooled analysis of individual data from multiple studies in which procedures for data collection and definitions of variables were standardised.

The CUPID study, as described in section 1.3, is a multicentre study in which participants were grouped within 47 distinct occupational groups. The 47 groups can be regarded as 47 independent studies in which data collection procedures and ascertainment of musculoskeletal outcomes and potential risk factors were standardised. In each of these groups it is thus possible to explore the same research question using identical statistical methods. Then the whole sample can be used to explore the same research question when applying pooled analysis of individual data.

The second aim of this thesis was to compare estimates produced when results from the separate studies were meta-analysed with estimates from models fitted to the individual-level data. This was done using data from the CUPID study covering musculoskeletal pain outcomes with high and low overall prevalence. The comparison of estimates derived from the two methods was further explored in different scenarios of correlations between the explanatory variables.

2.4.3 Estimation of relative risks from clustered data

Binary outcomes are regularly encountered in epidemiological research. They are commonly analysed by logistic regression modelling, which estimates odds ratios. These are a good approximation to relative risks, but are further from the null value of one, and the difference becomes greater as the prevalence of the outcome increases, notably when it exceeds 10% (74).

Relative risks can be estimated by fitting log binomial regression models. However, this model often has convergence problems during the iterative estimation procedure that it uses, and alternatives have been suggested for the estimation of relative risks (74). For example, Cox Proportional Hazards (PH), and Poisson regression models with robust confidence intervals can be used to estimate relative risks when the log-binomial model fails to converge.

Estimation of relative risks can also be a problem when observations are grouped within clusters, and thus cannot be regarded as independent; and in these circumstances, a similar approach (based on Poisson regression with robust standard errors) has been suggested (75). Since the publication in 2004 of the paper that originally proposed Poisson regression with robust standard errors (otherwise referred to as a modified Poisson model) in the context of clustered data (76), many studies have applied this method to estimate relative risks, but only one study has attempted to test its validity (77). Moreover, that study considered only one of the two main modelling approaches (modified Poisson regression modelling) to account for clustered data. No study to date has explored the performance of the Cox PH model in the estimation of relative risks where clustering is present.

The third aim of the thesis was to review available statistical models for the estimation of relative risks when data are clustered. These models were then fitted to data from the CUPID study, using outcomes of high and low prevalence, and the derived risk estimates were compared with those from log-binomial models.

2.5 Analytical model to account for clustering in this thesis

As reported in previous section, the two main approaches for modelling clustered data are the marginal (or GEE) and the conditional (or random effects) models. The choice between the two approaches depends on the research interest. That said, when one is interested in clustering effects then a conditional model should be employed. A marginal model does not produce direct estimates of cluster variance, but instead treats this variance as a nuisance. Effect estimates from the two methods are thus interpreted differently. The point estimate derived from the marginal model is interpreted as change in the outcome variable per unit increase in the explanatory variable, with the corresponding precision of the point estimate being corrected for clustering effects. In contrast to that, the point estimate derived from the conditional model is interpreted as change in the outcome variable per unit increase in the explanatory variable for individuals grouped into the same cluster. That is why the marginal methods are also termed population-averaged methods while the conditional methods are also termed cluster-specific methods. Results from marginal models are interpreted on the higher cluster level making them less suitable for prediction at the individual level (78). In this thesis, the interest is to explore how estimates of effect and related precision at the individual level are influenced by clustering effects, and modelling these effects (i.e. quantifying the variance at the higher level) is a crucial part of description in this investigation. Additionally, as clustering effects are not explicitly modelled in marginal models, these models do not allow the estimates of effect to vary across clusters, which is a potential extension for future work. For the above reasons, to account for clustering effects in (simulated and real) data presented in this thesis, I employed conditional methods.

Chapter 3. Background literature

A major focus of the research for this thesis was the consequences of failing to account for clustering in regression analysis, where there is a single outcome and a single explanatory variable. To find out what was already known about this focus of my research, I first carried out a preliminary search in Google Scholar to establish the nature and scope of the relevant literature. For that, I used keywords and phrases relating to "consequences of ignoring clustering", and tried to identify key papers in the area, focusing on those that considered both a naïve and a multilevel modelling approach, giving priority to those that applied RIs, but excluding those that used RIs only in conjunction with random coefficients. Through this preliminary search, I identified a small number of relevant papers, which I augmented by a "snowballing" approach (i.e. by identifying further articles from reference lists). This yielded a total of 29 published papers that had explored the same or similar research questions to mine.

Recognising that these were unlikely to represent all of the relevant literature, I then tried to carry out a more systematic search, of the sort described in a paper by Denison et al (79). It was already clear that evidence was available from a wide range of scientific fields, so following advice from a librarian at the Hartley library, University of Southampton, the systematic search was carried out using the Scopus database. The search strategy is summarised in the following steps:

Step 1: Identify publications which included key terms relevant to **ignoring or failing to account** for clustering
Step 2: Identify publications which included key terms relevant to **clustered data**
Step 3: Identify publications which included key terms relevant to **consequences** of ignoring clustering
Step 4: Identify the overlap of publications from the previous three steps
Step 5: From the overlapping publications exclude those which the title or abstract indicated had used statistical techniques that were beyond the scope of my thesis (for example generalised estimating equations) and those that were not in the English language or published only as abstracts
Step 6: Exclude duplicates

This search strategy initially identified more than 50,000 potentially relevant publications, and the number was only minimally reduced when the exclusions were made (i.e. of articles in languages other than English, conference abstracts, and duplicates). To address the challenge posed by this large number of publications, which could not all be reviewed, I sought advice from two librarians (Health Services Library, Southampton General Hospital) who suggested ways in which the search strategy might be refined. However, even after applying those methods, the number of references was still approximately 35,000. Moreover, they did not include some of the main articles on consequences of ignoring clustering of which I was aware from my preliminary literature search. After further discussions with the two librarians and with colleagues in the MRC Lifecourse Epidemiology Unit, I then modified my search strategy in the following ways:

- I decreased the number of key terms in each of the first three steps described above
- I restricted the search of the key terms to the titles and/or abstracts of potentially relevant articles
- I limited the potentially relevant publications to those that were in journals which were more likely to publish research on statistical methods in areas relevant to my thesis (for example excluding journals in the field of astronomy).

The specific steps along with the key words used, and the number of returns at each step are described in detail in Table 3.1.

This search strategy identified 1,655 publications (step 10, Table 3.1), the titles of which were screened to exclude those that clearly were not relevant to my research question. When the title was not sufficiently informative to determine the relevance of the article, I also checked the abstract. After that screening process, and after excluding duplicates, I ended up with 113 articles (step 12, Table 3.1) in which information regarding my research questions was likely to be included in the main text. The full text of these articles was then scanned, and they were classified into three groups: consequences of ignoring clustering in linear regression; consequences of ignoring clustering in logistic regression; and other. Most articles fell into the last category for the following reasons: that they did not provide sufficient information in the manuscript; that the regression models fitted were not of the type in which I was interested, and that the explanatory variable of interest was defined at cluster-level with no consideration of explanatory variables distributed at individual-level. After also excluding articles because it was not possible to compare directly the statistical modelling approach that was used with those to be considered in my thesis, I ended up with 30 publications.

I next checked the reference lists of the 30 publications that met my criteria for review, and this identified a further 19 publications for inclusion. 5 of these had also been found in my preliminary scoping search and 14 were new. At the same time, 16 of the relevant papers that had emerged from the scoping search were not picked up by the more systematic search. None of the newly identified publications materially altered the conclusions that could be drawn from those that had already been formed in the preliminary scoping exercise.

Table 3.1. Steps followed in the systematic approach for searching for evidence on consequences of ignoring clustering and number of documents returned

Step	Search	Number of documents returned
1	TITLE (incorrect* OR account* OR fail* OR *appropriate OR ignor* OR "consequences" OR "neglect" OR "adjust" OR "unadjusted" OR "avoid" OR "OLS" OR "ordinary least squares" OR "conventional" OR "traditional") OR ABS (incorrect* OR account* OR fail* OR *appropriate OR ignor* OR "consequences" OR "neglect" OR "adjust" OR "unadjusted" OR "avoid" OR "OLS" OR "ordinary least squares" OR "conventional" OR "traditional")	7,044,397
2	TITLE ("multilevel" OR "hierarchical" OR "clustered" OR "cluster" OR "clusters" OR "clustering" OR "nested data" OR "nested observations" OR "nesting" OR "nested structure" OR "correlated data" OR cluster* OR "grouped data" OR "structured data" OR "structure of data" OR "dependent observations" OR "dependency of observations" OR "dependent data" OR "dependency of data") OR ABS ("multilevel" OR "hierarchical" OR "clustered" OR "cluster" OR "clusters" OR "clustering" OR "nested data" OR "nested observations" OR "nesting" OR "nested structure" OR "correlated data" OR cluster* OR "grouped data" OR "structured data" OR "structure of data" OR "dependent observations" OR "dependency of observations" OR "dependent data" OR "dependency of data")	985,101
3	TITLE (statistic* OR *estimat* OR "estimation" OR "type-I error" OR "type 1 error" OR "type I error" OR "level effect" OR "level-effects" OR "effect estimate" OR "point estimate" OR "regression coefficient" OR "effect estimates" OR "point estimates" OR "regression coefficients" OR "odds" OR "standard error" OR "standard errors" OR "variance estimate" OR "variance estimates") OR ABS (statistic* OR *estimat* OR "estimation" OR "type-I error" OR "type 1 error" OR "type I error" OR "level effect" OR "level-effects" OR "effect estimate" OR "point estimate" OR "regression coefficient" OR "effect estimates" OR "point estimates" OR "regression coefficients" OR "odds" OR "standard error" OR "standard errors" OR "variance estimate" OR "variance estimates")	4,615,494
4	#1 AND #2 AND #3	34,866
5	TITLE ("Generalised estimating equations" OR "Generalized estimating equations" OR "Generalised estimating equation" OR "Generalized estimating equation" OR "Bayesian") OR ABS ("Generalised estimating equations" OR "Generalized estimating equations" OR "Generalised estimating equation" OR "Generalized estimating equation" OR "Bayesian")	117,619
6	#4 AND NOT #5	32,141
7	#6 LIMITED TO ENGLISH LANGUAGE	30,712
8	#7 LIMITED TO JOURNALS AND REPORTS	25,201
9	#8 LIMITED TO ARTICLES, REVIEWS, CONFERENCE PAPERS, AND REPORTS	24,554
10	#9 LIMITED TO JOURNALS LIKELY TO HAVE RELEVANT PUBLICATIONS TO THE RESEARCH QUESTION	1,655
11	POTENTIALLY RELEVANT ARTICLES BASED ON TITLE AND ABSTRACT SCREENING	115
12	EXCLUDING 2 DUPLICATES	113
13	RELEVANT PAPERS BASED ON FULL ARTICLE	30

Limitations of methods

My approach to searching the literature on the consequences of ignoring clustering in regression analysis had certain limitations. The final search was based only on keywords encountered in the titles and abstracts of articles, rather than in all possible fields. Moreover, the references identified were limited to articles, reviews, conference papers, and reports published in journals that were likely to carry material relevant to my research question. This was the only way that a systematic approach to searching the literature across such a broad field would return a manageable number of publications. However, it would be surprising if the yield were complete, especially given the failure of the systematic search to pick up 16 relevant papers found in the preliminary scoping exercise. On the other hand, the new articles that were identified by the systematic search did not materially alter the conclusions that could be drawn from the scoping exercise, which gives some reassurance that important evidence was not missed.

Summary of evidence found

In brief, the consequences of ignoring clustering (in terms of Type I error, bias in effect estimates and effects on their precision) have been explored by a number of investigators over the past three decades (80). The earliest reports focused on linear regression, and concluded that when analytical approaches failed to account for clustering, estimates of effect were unbiased (44, 81-85), but their precision would be underestimated when the explanatory variable of interest was constant for all observations grouped under the same cluster (making it a cluster-level variable), and overestimated when it varied within the cluster (i.e. it was an individual-level variable) (81, 86, 87). Furthermore, because of the effects on precision, Type I error rates would increase or decrease according to whether the explanatory variable under investigation was a cluster-level or individual-level variable.

Following on from those early papers, there has been extensive discussion of the consequences of ignoring clustering where an explanatory variable is defined at cluster-level. However, less attention has been given to situations in which the explanatory variable of interest is defined at the individual level, and an erroneous belief emerged that independently of whether an explanatory variable is cluster-level or individual-level, when analysis does not account for clustering it produces over-precise estimates of effect (44, 88-92). More recent publications have challenged this idea by focusing on individual-level explanatory variables too and comparing the precision of effect estimates when clustering is and is not accounted for in the analysis (82, 83, 93-97). A few of the papers describe simulation studies (82, 85, 93, 94, 98), but most have been based on examples of real data (99-106). Despite clear evidence from several analyses of real data that effect estimates for individual-level explanatory variables from naïve models can be less precise than those from corresponding RI models, which contradicts a widely held belief, results from other studies have shown that the effect of clustering on the precision of effect estimates for individual-level explanatory variables can be in either direction (99-102, 105, 106). Rates of coverage by 95% confidence intervals were reported in two studies, one of which focused on continuous (96) and the other on binary (94) explanatory variables. They both reported coverage rates very close to the nominal value of 95%. Also, increased Type I error rates were reported in two studies (83, 107).

Most of the articles that were identified, and especially those that adopted a theoretical approach, focused on continuous outcome variables. Articles concerning the consequences of ignoring clustering in logistic regression were more recent, and mainly compared naïve to multilevel models using real data rather than simulation studies. Unlike for linear regression, most studies agree regarding the effect of failing to account for clustering on the precision of effect estimates: effect estimates (expressed as log odds ratios) derived from naïve logistic regression were closer to the null than those from multi-level logistic regression, and their precision was greater (78, 108-117). However, a few articles have reported that effect estimates derived from the two models were very similar (118), or substantially different (119). I found only a few studies that examined effects on interval coverage rates, and these indicated that rates were lower than the nominal value when OL was used (108, 116, 120). No studies were identified that explored Type I error rates for effects of individual-levels variables when OL was used.

Overall, very few of the articles that I identified discussed Type I error rates and coverage by 95% confidence intervals, and none of them explored how the precision of effect estimates, either from linear or logistic regression, is influenced by the relative distribution of the explanatory variable within and between clusters. In addition, to the best of my knowledge, no attempt has been made to explore whether the well-discussed effects of clustering may differ depending on whether the explanatory variable of interest is binary or continuous in its distribution. These questions formed part of my investigation of the consequences of failing to adjust for clustering, and are discussed in the next two chapters.

The evidence that was found on consequences of ignoring clustering in linear and logistic regression is described in more detail in the introductory sections of the chapters that follow.

Chapter 4. Consequences of ignoring clustering in linear regression

4.1 Introduction

As described in Chapter 2, multilevel modelling is an analytical technique commonly used to account for grouping of observations within clusters. Use of this modelling approach has increased a lot over recent years. A review paper on the use of multilevel modelling by public health researchers, noted that the increase in the use of these modelling techniques was evidenced by numerous published articles, commentaries and books on the subject (80). However, despite the wide use that multilevel modelling has received, there are still numerous cases in which analytical techniques to account for grouping of observations have been ignored, as indicated by reviews (67-71). Recent systematic reviews have reported that clustering was accounted for in only 21.5% of multicentre trials (121) and 47% of cluster randomised trials (122). This may be partly due to computational/statistical complexities (123), but authors omit to discuss the limitations of their chosen analytical technique, possibly because of a lack of clarity about the effects of ignoring clustering. As outlined in Chapter 2, the first aim of this thesis, was to explore the consequences of failing to account for clustering in statistical inference.

The consequences of ignoring clustering in the simple case of a continuous outcome variable and a single explanatory variable X were first described in the early 80's. Scott and Holt (81) showed, from an algebraic perspective, that when the size of the clusters is fixed to n for all clusters, the precision of the effect estimate is inflated by a factor of $\left[1 + (n-1)\rho_x\rho_y\right]$, with ρ_x being the intra-cluster correlation of the explanatory variable X, and ρ_y the intra-cluster correlation of the outcome variable after adjustment for X. As highlighted later by Donner (86), the inflation identified by Scott and Holt (81) can result in an upward bias when the estimate of ρ_x is negative, and a downward bias otherwise ($\rho_x > 0$). A negative ρ_x occurs when the values of X within clusters vary considerably more than the values of X across clusters (86). Negative values of ρ_x can be interpreted as very low intraclass correlation (124) of the explanatory variable which can only occur for individual-level variables, while the highest possible value of ρ_x $(= 1)$ is possible only when all values within cluster are the same (i.e. it is a cluster-level variable). The case in which $\rho_x = 1$ was also explored by Kloek et al who showed that in this situation, the variance of effect estimates from a simple linear model is severely underestimated (87).

Since the early exploration of consequences for the precision of effect estimates when analytical methods do not account for clustering, the impact of clustering in statistical inference has been widely discussed. However, most investigation has focused on the case of a single explanatory variable that is cluster-specific (i.e. a cluster-level variable). Numerous studies have shown that in the case of cluster-specific explanatory variables, the effect estimates from a naïve model that fails to adjust for clustering are over-precise leading to rates of Type I error that are higher than the nominal value of 5%. This message has been widely disseminated in the epidemiological literature, and has led to the erroneous belief that failing to adjust for clustering causes underestimation of standard errors and increases in Type I error, whatever the type of explanatory variable (96). Several researchers have reported this direction of bias without specifying the conditions under which it occurs (44, 88-92).

More recent studies have challenged the notion that standard errors are always underestimated by reporting counter-examples (93, 95-97) or explaining in more detail how the precision of effect estimates can be biased in either direction (83, 94, 125). Chuang et al (94) have demonstrated that where variation from cluster differences is augmented by variation at the lower level (i.e. within clusters variation), inflation of standard errors, spuriously large p-values and false-negative results for individual-level explanatory variables can result. On the other hand, when the explanatory variable is cluster-specific (i.e. it takes the same value for all individuals in any given cluster), the degrees of freedom are inflated, as estimates of effect are related to the number of individuals whereas in reality it is the number of clusters that determines the level of statistical uncertainty. This leads to spuriously precise estimates of effect. To illustrate these opposite effects on precision, Sainani presented two simple examples of clustered data, and focused on errors associated with failure to account for clustering, according to whether the explanatory variable was cluster-specific or individual-specific, demonstrating two directions of bias (125).

Several studies have compared estimation from traditional models and multilevel models using simulated or real data. These have found that when the main explanatory variable was measured at the cluster-level and traditional regression models were used, the standard error of the estimate of its effect was biased downwards, and the opposite occurred for individual-level variables (82, 93, 94, 98, 103, 104). However, some of the studies reviewed (especially those using real data), showed that even for individual-level variables in the absence of adjustment for clustering, standard errors of the effect were either higher or lower compared to those estimated from the multilevel models (85, 99-102, 105, 106). Others have reported that ignoring clustering has very small effect on the precision of effect estimates (91, 126-128). While very small differences in precision can be explained by small intraclass correlation, the reasons why some individual-level effects have higher and others have lower standard errors when estimated by traditional analytical techniques is unclear, and has not been discussed.

Beyond effects on standard errors, estimates of effect size from traditional as compared with multilevel regression techniques are also of interest. Some investigators have reported that when traditional regression methods are applied where multilevel modelling methods would be more appropriate, effect estimates are unbiased but inefficient (44, 81-85). However, studies comparing naïve methods with multilevel modelling techniques using real data show that effect estimates can differ considerably, with the direction of differences being variable (96, 101-103, 126). The errors in the estimation of effects are not well described in the literature.

To date, no study has investigated the extent to which bias can occur in effect estimates when clustering is ignored, the determinants of that bias, or the exact consequences for the precision of estimates according to different distributions of the explanatory variable and, in particular, the extent to which the explanatory variable varies within as compared with between clusters.

The aim of this chapter is to assess the detailed implications of failing to account for clustering effects on the effect estimate (regression coefficient), its precision (characterised by the standard error (SE)), and coverage by 95% confidence intervals, when exploring the relation of a continuous outcome variable to an explanatory variable that is either continuous or binary.

To assist understanding, the question is first considered where the observations are grouped into two clusters, and it is then expanded to the situation of multiple clusters. This approach is followed separately for a continuous and then a binary explanatory variable. For the continuous explanatory variable, differences between estimates derived from models that account for clustering and those that do not are examined according to the ratio of between to within cluster dispersion of the explanatory variable. For the binary explanatory variable, differences in estimates are examined in relation to variation in the prevalence of the explanatory variable across the clusters.

In addition, for each of the two types of explanatory variable, the extent to which analyses that ignore clustering can give misleading estimates of type I error is considered, when the real effect of the explanatory variable is set to zero.

4.2 Two clusters – Continuous explanatory variable

Where both the outcome and the explanatory variable are continuous and individual level data are clustered within only two groups, the effect of clustering is best taken into account by fitting a dummy variable for cluster as a covariate in a linear (typically OLS) regression model (part III in section 2.3). In this case, the comparison between estimates derived from the RI and the OLS models equates to a comparison between estimates from an OLS model that adjusts for the effect of cluster and an OLS model that does not adjust for the clustering effect.

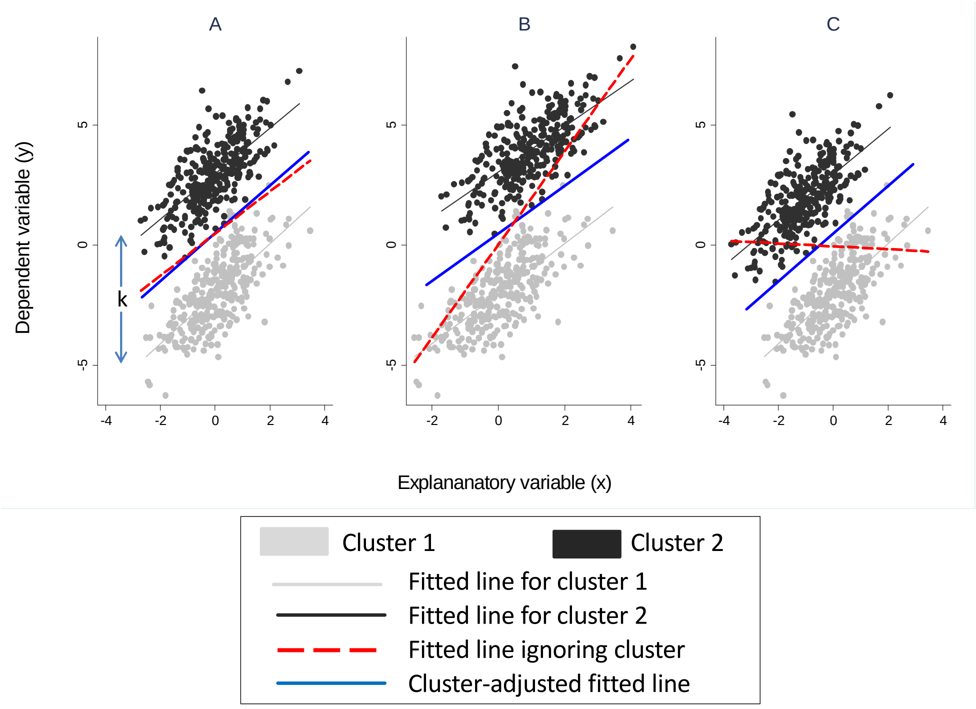


Figure 4.1. Scatter plot of continuous outcome y and continuous explanatory variable x, with observations grouped in two clusters. A. $\bar{x} = \bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$; B. $\bar{x}_{cluster\ 1} < \bar{x}_{cluster\ 2}$; C. $\bar{x}_{cluster\ 1} > \bar{x}_{cluster\ 2}$

To illustrate the problem, simulated data were generated for a continuous outcome and a continuous explanatory variable, with observations grouped in two clusters. The outcome and the explanatory variables were plotted against each other, and are presented in Figure 4.1, in which the two colours correspond to the observations grouped within the two clusters. The corresponding regression lines are depicted in the same colours as the observations for the cluster and are parallel to each other. The cluster-specific intercepts of the regression lines differ by a constant value, k. The red dashed line is that fitted without accounting for clustering, while the blue solid line is the fitted line from a regression model that included adjustment for clustering. The three different subplots of the figure (A, B, and C) correspond to three examples of shifts in the distribution of x between the first and second clusters. In subplot A, the cluster-specific mean values of the explanatory variable are identical ($\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$). In subplots B and C, the cluster-specific mean values of the explanatory variable for the second cluster are shifted to the right ($\bar{x}_{cluster\ 1} < \bar{x}_{cluster\ 2}$) and left ($\bar{x}_{cluster\ 1} > \bar{x}_{cluster\ 2}$), respectively, of those for the first cluster. When the cluster-specific distributions of x are overlaid ($\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$), subplot A, the regression line from the cluster-adjusted analysis is approximately the same as that from the cluster-unadjusted analysis. However, when $\bar{x}_{cluster\ 1} < \bar{x}_{cluster\ 2}$ (subplot B), the cluster-unadjusted regression line is steeper than the cluster-adjusted regression line. The opposite is observed when $\bar{x}_{cluster\ 1} > \bar{x}_{cluster\ 2}$ (subplot C).

The estimated regression coefficients and the corresponding standard errors from the two models were further explored by considering varying shifts in the distribution of the explanatory variable x for the second cluster from that of the first cluster, varying distances between the cluster-specific intercepts (k's), and different dispersions of the error term in the linear regression model. This was done using Monte Carlo simulated data, as described below.

4.2.1 Methods

In the simplest case, in which there is a single explanatory variable, the OLS linear regression is specified by a model of the form:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{4.1}$$

where, for individual i, y_i and x_i are the values of the outcome and explanatory variables respectively, β_0 and β_1 are the intercept and the slope of the regression line (the latter interpreted as the effect of the explanatory variable on the outcome), and e_i is the residual or error term, with $e_i|x_i \sim N(0, SD_e^2)$.

The OLS regression model that accounts for the cluster in which observations are grouped takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 cluster + e_i \tag{4.2}$$

where β_2 is the effect of the second cluster on the outcome y_i as compared with that for the first cluster, keeping the rest of the parameters in the prediction part of the equation constant.

The consequences of failing to adjust for the effect of cluster (equation 4.1) in the estimated regression coefficients and corresponding SEs were explored by simulating datasets with observations nested within two clusters. For each Monte Carlo simulation, the number of observations in each cluster was set to 200 resulting in samples of 400 observations overall. For simplicity, the size of the effect of x_i on y_i was arbitrarily set to 1 ($\beta_1 = 1$), and the average value of y_i when $x_i = 0$ across the two clusters was arbitrarily set to 0 ($\beta_0 = 0$).

Cluster specific intercepts were initially generated (u_1 and u_2 for the first and the second cluster respectively) with u_1=$u_2 + k$, where k was the distance between the two intercepts, and a random variable for the error term e_i from a normal distribution of mean zero and standard deviation SD_e. To set values for the continuous explanatory variable, I initially generated a variable x_{0i} from the standard normal distribution $\left(x_{0i} \sim N(0,1)\right)$. For that variable the average value of x_{0i} for the first cluster was the same as that for the second cluster. I then generated a further 400 explanatory variables x_{mi}, with $m = 1, \ldots, 400$. For each of these 400 variables, $x_{mi} = x_{0i}$ for the first cluster and $x_{mi} = x_{0i} + shift_m$ for the second cluster, with $shift_m$ being the mth value of a vector variable $shift$, the values of which were drawn from a normal distribution of mean zero and standard deviation $SD_{shift} = 10$. As such, the 400 explanatory variables x_{mi}, with $m = 1, \dots, 400$, had varying distances between the cluster-specific means (Figure 4.2). The corresponding values for the outcome variable were generated as $y_{ij} = x_{ij} + e_{ij} + u_j$. Thus, the 400 different explanatory variables yielded 400 outcome variables.

The above process was repeated for varying distances ($k = \{0.5, 1, 1.5\}$) between the cluster-specific intercepts (u's), and standard deviations of the error of $y_i |\, x_i$ ($SD_e = \{0.5, 1, 1.5\}$).

For each set (y_i, x_i), two models were fitted; an OLS model that adjusted for the cluster and another that did not. From the estimates generated from the two models, I calculated the difference in the regression coefficients (difference=$\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) and the ratio of the corresponding SEs (ratio=$SE^{cluster-adjusted}/SE^{cluster-unadjusted}$). These were plotted against values for the shift of the distribution of the explanatory variable in the second cluster and summarised by means and SDs. Coverage by 95% CIs was calculated for varying distances (k) between the cluster-specific intercepts (u's), standard deviations of $y_i |\, x_i$ (SD_e), and levels of the shift of the distribution of the explanatory variable in the second cluster.

To explore the implications of ignoring the effect of clustering in hypothesis testing, simulations were repeated assuming no association between the explanatory and the outcome variable, with the size of the effect of x_i on y_i being set to 0 ($\beta_1 = 0$). The p-values of the cluster-adjusted and cluster-unadjusted associations were saved, the percentages of p-values <0.05 were calculated, and as for coverage by 95% CIs, were summarised across values of k, u, SD_e, and levels of the shift of the distribution of the explanatory variable in the second cluster.

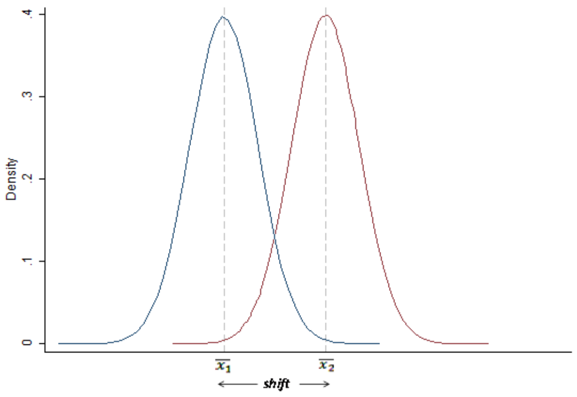
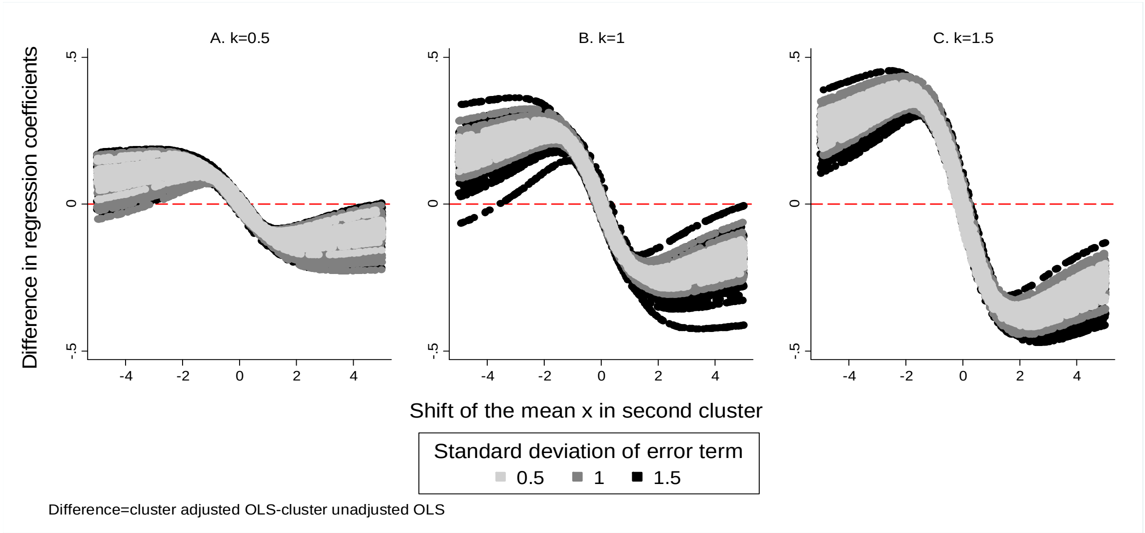


Figure 4.2. Shift of mean value of explanatory variable for second cluster. Blue line shows the distribution of x for the first cluster and red line shows the shifted distribution of x for the second cluster

4.2.2 Results

Difference in regression coefficients

Figure 4.3 plots the differences in regression coefficients ($\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) according to the shift in the distribution of the explanatory variable x for the second cluster from the distribution of x for the first cluster. The three subplots of the figure are for the different distances between the cluster-specific intercepts examined (A. $k = 0.5$; B. $k = 1$; C. $k = 1.5$). In each subplot, results for different SD_e are represented by different shades of grey, with the lighter colours corresponding to lower values of SD_e.



Difference=cluster adjusted OLS-cluster unadjusted OLS

Figure 4.3. Scatter plot of difference between regression coefficients estimated from OLS that adjusted and OLS that did not adjust for cluster against shift in the mean value of x for the second cluster from the mean value of x for the first cluster

When the two distributions were overlaid (shift of distribution of x for the second cluster = 0), the regression coefficients estimated from the two models were approximately the same (difference between regression coefficients=0). When the distribution of x for the second cluster was shifted to the right of the distribution of x for the first cluster, the estimated regression coefficient from the model that adjusted for the cluster was lower than that estimated from the model that did not adjust for cluster. The opposite occurred when the distribution of x in the second cluster was shifted to the left of the distribution of x in the first cluster. The difference in the regression coefficients became positive, meaning that $\beta_1^{Cluster-adjusted}$>$\beta_1^{Cluster-unadjusted}$. As the shift increased, the difference in the regression coefficients increased, reaching a maximum value, after which further increase in the distance between the two distributions (distance between $\bar{x}_{cluster\ 1}$

and $\bar{x}_{cluster\ 2}$) caused the differences between the regression coefficients to decrease, moving back towards zero. For constant SD_e (comparison of subplots A, B and C of Figure 4.3), increasing distance between the cluster-specific intercepts (k's), increased the maximum difference between the regression coefficients estimated from the two models. For given distance between the cluster-specific intercepts (k's) (comparison of the three shades within each subplot in Figure 4.3), increasing dispersion of the error term (SD_e) increased the dispersion of the differences in regression coefficients.

Table 4.1. Mean (SD) of differences between regression coefficients estimated from OLS modelling that adjusted, and OLS modelling that did not adjust for cluster, for varying SD_e and selected shifts of the mean value of x for the second cluster

		Difference in$\boldsymbol{\beta_1}$'s		
		$\boldsymbol{k = 0.5}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 1.5}$
$\boldsymbol{shift = 0}$	$\boldsymbol{SD_e = 0.5}$	-0.00(0.01)	0.00(0.02)	-0.01(0.04)
	$\boldsymbol{SD_e = 1}$	-0.00(0.01)	-0.00(0.02)	0.01(0.04)
	$\boldsymbol{SD_e = 1.5}$	0.00(0.01)	0.00(0.03)	0.01(0.04)
1.95≤$\boldsymbol{shift}$≤$\boldsymbol{2.05}$	$\boldsymbol{SD_e = 0.5}$	-0.12(0.01)	-0.25(0.01)	-0.38(0.01)
	$\boldsymbol{SD_e = 1}$	-0.13(0.03)	-0.26(0.03)	-0.38(0.03)
	$\boldsymbol{SD_e = 1.5}$	-0.12(0.03)	-0.25(0.04)	-0.39(0.03)
-2.05≤$\boldsymbol{shift}$≤$\boldsymbol{-1.95}$	$\boldsymbol{SD_e = 0.5}$	0.13(0.01)	0.25(0.01)	0.37(0.02)
	$\boldsymbol{SD_e = 1}$	0.12(0.03)	0.24(0.02)	0.38(0.02)
	$\boldsymbol{SD_e = 1.5}$	0.13(0.03)	0.24(0.04)	0.37(0.03)
3.95≤$\boldsymbol{shift}$≤$\boldsymbol{4.05}$	$\boldsymbol{SD_e = 0.5}$	-0.10(0.02)	-0.20(0.02)	-0.30(0.02)
	$\boldsymbol{SD_e = 1}$	-0.12(0.05)	-0.21(0.04)	-0.31(0.04)
	$\boldsymbol{SD_e = 1.5}$	-0.09(0.05)	-0.22(0.07)	-0.32(0.06)
-4.05≤$\boldsymbol{shift}$≤$\boldsymbol{-3.95}$	$\boldsymbol{SD_e = 0.5}$	0.10(0.02)	0.20(0.02)	0.30(0.03)
	$\boldsymbol{SD_e = 1}$	0.10(0.04)	0.20(0.04)	0.30(0.04)
	$\boldsymbol{SD_e = 1.5}$	0.10(0.05)	0.19(0.05)	0.28(0.06)

The differences between the regression coefficients estimated from the two models are summarised for different SD_e and for selected shifts of the mean value of x for the second cluster in relation to the first cluster in Table 4.1. The differences in the regression coefficients were all approximately equal to zero for zero shift of the distribution of x in the second cluster. When the shift was approximately equal to 2 (i.e. twice the within-cluster SD of x), the average differences of the regression coefficients were larger than those for a shift approximately equal to 4. For fixed SD_e, increasing k increased the absolute mean differences between regression coefficients. For any given k, increasing SD_e increased the SD of the differences while the mean values remained approximately the same.

Ratio of standard errors

The ratios of the SEs of the regression coefficients estimated from the two models ($SE^{cluster-adjusted}/SE^{cluster-unadjusted}$) were plotted against the shift in the distribution of x for the second cluster from the distribution of x for the first cluster (Figure 4.4). As in Figure 4.3, the three subplots are for the different distances between the cluster-specific intercepts examined (A. $k = 0.5$; B. $k = 1$; C. $k = 1.5$), while in each subplot, the three different shades of grey correspond to different values of SD_e.

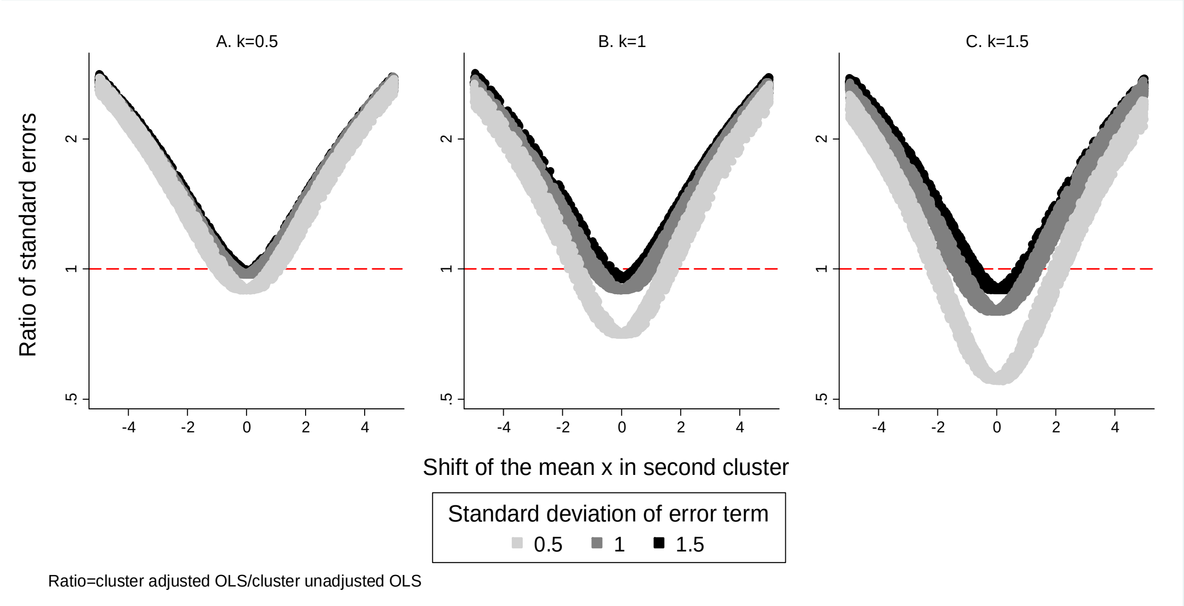


Figure 4.4. Ratios of standard errors estimated from OLS regression that adjusted and OLS regression that did not adjust for cluster, against shift in the mean value of x for the second cluster from the mean value of x for the first cluster.

Overall, when shift=0, the ratio was at its minimum value while shifting the distribution of x for the second cluster to either the right or the left of the distribution of x for the first cluster resulted in higher ratios. In most cases explored here, the ratio was >1, meaning that the SEs estimated from the cluster-adjusted model were larger than those estimated from the cluster-unadjusted model. However, for given SD_e, increasing k (comparison of same shade across different subplots of the figure), resulted in lower minimum values of the ratio (when shift=0). In other words, as the distance between the cluster-specific intercepts increased while the cluster-specific distributions of the explanatory variables were very similar ($\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$), the SEs estimated from the OLS models that adjusted for clustering of the observations were lower than those estimated from the OLS model that did not adjust for clustering. For constant k, increasing SD_e (comparison of the three shades within each subplot of the figure) increased the minimum value of the ratio.

The minimum values of the ratio of the SEs for the different combinations of SD_e and k, and for zero shift of the distribution of the explanatory variable x for the second cluster are presented inTable 4.2. When SD_e was held constant, an increase in k was associated with a lower minimum value of the ratio. For $SD_e = 0.5$ and $k = 1.5$, the minimum ratio of the SEs was 0.55 meaning that the estimated SE from the cluster-adjusted OLS model was approximately half of the SE estimated from the cluster-unadjusted OLS model.

Table 4.2. Minimum value of the ratio of SEs of the regression coefficients estimated from OLS that adjusted, and OLS that did not adjust for cluster, for varying SD_e and for zero shift of the mean value of x for the second cluster

		Ratios of SEs		
		$\boldsymbol{k = 0.5}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 1.5}$
$\boldsymbol{shift = 0}$	$\boldsymbol{SD_e = 0.5}$	0.892	0.703	0.551
	$\boldsymbol{SD_e = 1}$	0.971	0.893	0.797
	$\boldsymbol{SD_e = 1.5}$	0.987	0.949	0.894

As also seen in Figure 4.4, as SD_e increased the minimum ratio of SEs increased approaching 1, when the distance between the cluster-specific intercepts was kept constant. When SD_e was held constant, increasing k decreased the minimum value of the ratio. For $SD_e = 0.5$ and $k = 1.5$, the minimum ratio of the SEs was 0.55, meaning that the estimated SE from the cluster-adjusted OLS model was approximately half of the SE estimated from the cluster-unadjusted OLS model.

Coverage by 95% confidence intervals

Coverage by 95% CIs of the simulated effect $\beta_1^{Cluster-adjusted}$=1, was calculated for the cluster-adjusted and the cluster-unadjusted models and is presented in Table 4.3. With the cluster-adjusted model, coverage was very close to the nominal value of 95%; it had an average of 93% and it ranged from 88% to 98% across the different levels of distance between the cluster-specific intercepts k, and values of SD_e. In contrast to the cluster-unadjusted model, coverage by 95% CIs from the cluster-unadjusted model was considerably lower, with an average value of 11.9%. It also varied quite importantly from 5.3% up to 31.6%. Coverage for the cluster-unadjusted model was lower for decreasing k, and for higher values of SD_e.

To explore how coverage by 95% CIs is influenced by the shift of the distribution of x in the second cluster from that in the first cluster, I considered the absolute shift of the distribution of x in cluster 2 from that in cluster 1, and then quarters of the distribution of the absolute shifts. I found that with the cluster-unadjusted OLS model, almost none of the simulated datasets falling in the 2nd, 3rd, or 4th quarter of the distribution of absolute shifts included the simulated effect, essentially producing a coverage of approximately 0% in that area of high distance between cluster-specific mean values of x. In contrast to that observation, with the cluster-adjusted OLS model, coverage varied very little from 95% across the different categories of absolute shifts.

Table 4.3. Coverage (%) by 95% confidence intervals of simulated effect $\beta_1^{Cluster-adjusted}$=1 from the cluster-unadjusted and cluster-adjusted models according to distance between the cluster-specific intercepts k, and values of SD_e

		$\boldsymbol{k = 0.5}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 1.5}$
Cluster-unadjusted OLS model	$\boldsymbol{SD_e = 0.5}$	9.1	18.6	31.6
	$\boldsymbol{SD_e = 1}$	5.3	9.2	13.5
	$\boldsymbol{SD_e = 1.5}$	4.6	6.5	8.8
Cluster-adjusted OLS model	$\boldsymbol{SD_e = 0.5}$	92.3	88.0	98.0
	$\boldsymbol{SD_e = 1}$	94.0	95.7	91.8
	$\boldsymbol{SD_e = 1.5}$	87.7	91.7	96.1

Type I error

Table 4.4 shows the percentage of the simulated datasets for which the null hypothesis was rejected, when in fact there was no association between the explanatory and the outcome variable (frequency of type I error). Results are presented separately for varying cluster-specific distances k and varying SD_e. The frequency of type I error was very high when clustering was not taken into account, while, with adjustment for clustering, it was much closer to the 5% that would be expected. Decreasing k, while keeping SD_e constant, increased type I error in models that did not account for clustering. Also, increasing SD_e and keeping a constant k, increased type I error.

Table 4.4. Proportion (%) of simulated datasets for which the null hypothesis was rejected, when true effect size of the explanatory on the outcome variable was assumed to be 0

		$k = 0.5$	$k = 1$	$k = 1.5$
Cluster-unadjusted OLS model	$SD_e = 0.5$	90.9	81.6	66.5
	$SD_e = 1$	94.0	90.7	86.6
	$SD_e = 1.5$	94.9	92.9	91.1
Cluster-adjusted OLS model	$SD_e = 0.5$	9.7	2.0	5.7
	$SD_e = 1$	5.8	11.8	4.0
	$SD_e = 1.5$	12.2	2.0	3.8

To explore the effect of shift in the distribution of x for the second cluster, I categorised the absolute values of the shifts into quarters of its distribution. I found that with the cluster-unadjusted OLS model, almost all simulated datasets falling into the 2nd, 3rd, and 4th quarter of the distribution of absolute shifts produced statistically significant results (Type I error rate $\approx 100\%$), while those falling into the first quarter produced rates that ranged from 19% up to 80% for different values of k and SD_e.

4.3 Multiple clusters – Continuous explanatory variable

4.3.1 Methods

As described in the previous section, in the simplest case, in which there is a single explanatory variable, the OLS linear regression is specified by a model of the form presented in equation 4.1.

For a continuous outcome and a continuous explanatory variable, the RI multi-level model can be viewed as an extension of the OLS model, and is specified as:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + e_{ij} \\ &= \beta_0 + \beta_1 x_{ij} + e_{ij} + u_j \end{aligned} \qquad 4.3$$

where the index i refers to the individual and the index j to the cluster, and $\beta_{0j} = \beta_0 + u_j$, the estimate of the intercept for cluster j. The term u_j represents the error for cluster j around the fixed intercept value of β_0, and is assumed to be normally distributed with $u_j|x_{ij} \sim N(0, SD_u^2)$. The term e_{ij} represents the additional error within the cluster, also referred to as the individual level error term, with $e_{ij}|x_{ij}, u_j \sim N(0, SD_e^2)$.

An important parameter in hierarchical data is the intraclass correlation coefficient (ICC) (also discussed in section 2.2), which characterises the extent to which the outcome variable y_{ij} is similar within clusters, given the distribution of the explanatory variable x_{ij} (129). For a continuous outcome variable, a continuous explanatory variable, and with the nomenclature used in equation 4.3, the ICC is defined as ICC $= \frac{SD_u^2}{SD_u^2+SD_e^2}$ (66).

To explore the study questions, simulated datasets were generated according to the assumptions of the RI model. For each Monte Carlo simulation, both the number of clusters and the number of observations per cluster were set to 100. For simplicity, the size of the effect of x_{ij} on y_{ij} was arbitrarily set to 1 ($\beta_1 = 1$), and the average value of y_{ij} when x_{ij}= 0 was arbitrarily set to 0 ($\beta_0 = 0$).

To set values x_{ij} for the explanatory variable in a cluster j, an individual level variable was generated as $x_{0ij} \sim N\left(0, SD_{x_{ij}}^2\right)$, with $SD_{x_{ij}} = 1$, and a cluster-specific variable as $shift_j \sim N\left(0, {SD_{shift}}^2\right)$. The individual level variable was then added to the cluster-specific shift, so that $x_{ij} = x_{0ij} + shift_j$. The corresponding values for the outcome variable y_{ij} were generated according to equation 4.3. For this purpose, the individual-level error terms were drawn from a random normal distribution with mean zero and variance $SD_{e_{ij}}^2$, and the cluster-level error terms were drawn from a random normal distribution with mean zero and variance $SD_{u_j}^2$.

Simulated data were generated for two different scenarios of dispersion of the error term, $SD_{e_{ij}} = 1$ and $SD_{e_{ij}} = 2$, and for combinations of values for SD_{u_j} (0.0316, 0.05485, 0.1005, 0.1759, 0.3333 and 0.6547 when $SD_{e_{ij}} = 1$, and 0.0633, 0.1097, 0.2010, 0.3517, 0.667, and 1.309 when $SD_{e_{ij}} = 2$ (chosen to give expected values for the ICC of 0.001, 0.003, 0.01, 0.03, 0.1 and 0.3 respectively, for the two different values of $SD_{e_{ij}}$), with $SD_{shift} \sim U[a, b]$, the parameters a and b being arbitrarily chosen to be 0 and 15, respectively. For each combination of expected ICC and $SD_{e_{ij}}$, 1000 simulated datasets were produced.

For each simulated dataset, two linear regression models were fitted; an OLS model which ignored the clustering (equation 4.1), and a RI multi-level model which allowed for clustering effects (equation 4.3). For each of the models, the regression coefficient (i.e. the effect of the explanatory variable on the outcome variable) and its standard error (SE) were estimated. To compare results from the two models, the difference between the estimated regression coefficients ($\beta_1^{RI} - \beta_1^{OLS}$), and the ratio of their SEs (SE^{RI}/SE^{OLS}) were calculated.

To assess how the comparison between the two models was affected by the distribution of x_{ij} within and between clusters, these two measures were plotted against the dispersion of the x_{ij} between clusters (dispersion of $shift_j$), which equates to the relative dispersion of x_{ij} between to within clusters when $SD_{x_{ij}} = 1$, and is described in what follows simply as 'the relative dispersion of the x_{ij}'). In addition, descriptive statistics were produced for the distributions of the two measures across simulated samples, according to values for expected ICC and $SD_{e_{ij}}$.

The accuracy of the 95% confidence intervals for the regression coefficient β_1 from the two methods was assessed by calculating the proportion of the estimated confidence intervals that included the true value that had been used in the simulations. A method was considered to have appropriate coverage if 95% of the 95% confidence intervals included the value of the effect β_1 used in the simulations. Deviations from this ideal could reflect bias in the estimates of effect, unsatisfactory standard errors (130), or both.

To assess impacts on type I error, the simulations were repeated assuming no association between x_{ij} and y_{ij} (i.e. $\beta_1 = 0$), and the proportions of datasets for which the null hypothesis was rejected at a 5% significance level (i.e. the absolute magnitude the estimated effect exceeded 1.96 times its standard error) in OLS and RI modelling were compared according to ICC.

Due to random sampling variation the estimated ICC values were within given ranges of the target levels of ICC. These ranges were 0.0005-0.0014, 0.0025-0.0034, 0.005-0.014, 0.025-0.034, 0.05-0.14, and 0.25-0.34 for target levels of 0.001, 0.003, 0.01, 0.03, 0.1 and 0.3 respectively. Simulations resulting in estimated ICC values outside of these ranges were discarded and not used further. In the description of the results that follows ICC values are labelled according to the target levels.

4.3.2 Results

Difference in regression coefficients

Differences in regression coefficients ($\beta_1^{RI} - \beta_1^{OLS}$) estimated from the two linear models (described in equations 4.1 and 4.3) were explored in relation to the relative dispersion of the x_{ij}. The results are illustrated in Figure 4.5. The two different subplots of the figure (A, and B) correspond to the two different values of $SD_{e_{ij}}$ described in the methods. In each scatter plot, the different levels of within-cluster similarity of observations (ICCs) are depicted by different shades of grey with darker shades corresponding to simulated results for higher ICCs. On average, for small relative dispersion of x_{ij}, differences were more narrowly spread for small ICCs and more widely spread for large ICCs. For each value of ICC, increasing the relative dispersion of x_{ij} resulted in larger differences in regression coefficients up to a relative dispersion of $x_{ij} = 1$ (i.e. same dispersion of x_{ij} between and within clusters). Beyond that point, further increase in the relative dispersion of x_{ij} resulted in smaller differences in regression coefficients from the two methods, approaching a difference of zero. Comparison of the different subplots of Figure 4.5 indicates that higher dispersion of the error term resulted in greater differences of the regression coefficients from the two models, even for small relative dispersion of x_{ij}; for high ICC (=0.3) the range of differences for $SD_{e_{ij}} = 1$ was approximately -0.2 to 0.2, corresponding to a 20% difference in the regression coefficients from the two methods, and this range decreased to approximately -0.1 to 0.1 for $SD_{e_{ij}} = 2$.

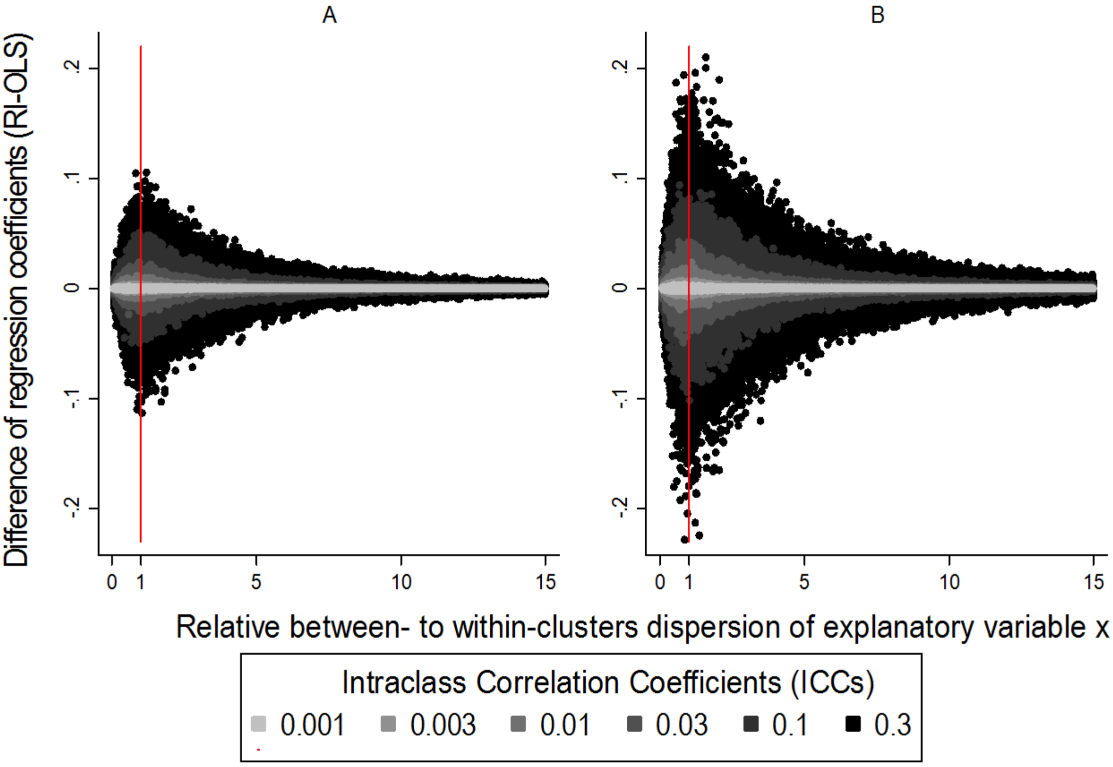


Figure 4.5. Difference between regression coefficients estimated from RI and OLS models ($\beta_1^{RI} - \beta_1^{OLS}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend). Figure A: $SD_{e_{ij}} = 1$. Figure B: $SD_{e_{ij}} = 2$

Table 4.5 summarises the differences between regression coefficients from the RI and OLS models for varying ICCs and for different dispersions of the error term. The estimated regression coefficients from the two methods were on average very similar (β_1^{RI}, $\beta_1^{OLS} \cong 1$). For the narrowest dispersion of e_{ij} that was assumed, and for intraclass correlation coefficients of about 0.001, differences in the regression coefficients ranged up to 0.003. For the same dispersion of e_{ij} (i.e. in any specific subplot of Figure 4.5), increasing ICCs tended to increase the spread of the differences in estimated regression coefficients from the two methods reaching approximately 0.20. For constant ICC, differences in regression coefficients increased as the dispersion of e_{ij} increased. For example, for ICC=0.30, differences in regression coefficients ranged up to 0.21 for $SD_{e_{ij}} = 2$ and were as low as 0.11 for $SD_{e_{ij}} = 1$. Overall, differences were minimal for small ICCs and small dispersion of e_{ij}, and they increased up to 0.2 for high ICCs and high dispersion of e_{ij}.

Table 4.5. Range of differences between regression coefficients estimated from RI and OLS models ($\beta_1^{RI} - \beta_1^{OLS}$) according to ICC and dispersion of e_{ij}

ICC	**Within-cluster $SD_{e_{ij}}$=1**	**Within-cluster $SD_{e_{ij}}$=2**
0.001	-0.001 to 0.001	-0.003 to 0.003
0.003	-0.004 to 0.004	-0.007 to 0.007
0.01	-0.009 to 0.011	-0.019 to 0.023
0.03	-0.024 to 0.023	-0.049 to 0.042
0.1	-0.050 to 0.050	-0.102 to 0.094
0.3	-0.113 to 0.105	-0.228 to 0.209

Ratio of standard errors

The ratios of SEs derived from the RI and OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) were examined in relation to the relative dispersion of x_{ij}, and are presented in Figure 4.6. As in Figure 4.5, the different levels of ICCs are represented by different shades of grey, with lighter shades corresponding to lower ICCs and darker shades to higher ICCs. The ratio took its minimum value for the smallest relative dispersion of x_{ij} and increased as the relative dispersion of x_{ij} increased, tending asymptotically to a maximum value. The minimum and maximum values of the ratio of the SEs (the latter also corresponding to its asymptote) were ICC-dependent, higher ICCs (darker colours) resulting in lower minimum and higher maximum values for the ratio. The relative dispersion of x_{ij} at which the ratio of SEs approached its asymptote was also ICC-dependent, being higher for larger ICCs. For very small values of relative dispersion of x_{ij}, the minimum value of the ratio of the SEs was approximately one for small levels of ICC and was less than one for higher ICCs. Particularly for small values of relative dispersion and ICC ≅ 0.10 or 0.30, the ratio of SEs was <1, meaning that SEs from RI models were smaller than from OLS models.

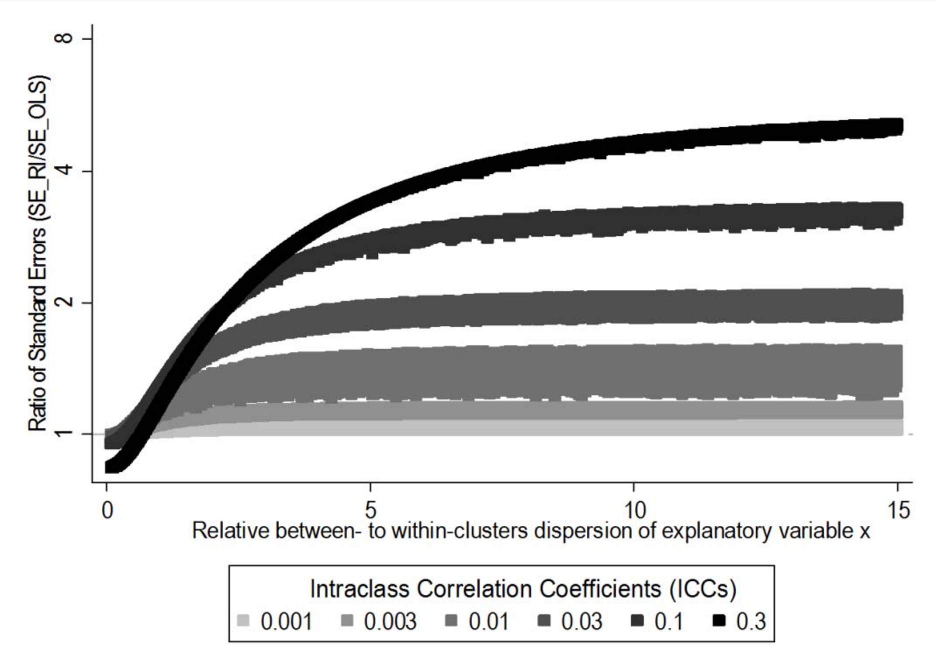


Figure 4.6. Ratios of standard errors estimated from RI and OLS models ($\mathrm{SE}_{\beta_1^{\mathrm{RI}}}/\mathrm{SE}_{\beta_1^{\mathrm{OLS}}}$) plotted against relative between- to within-clusters dispersion of explanatory variable x_{ij}

Table 4.6 summarises the range of the ratios of the SEs of the regression coefficients estimated from the two models and shows the relative dispersions of x_{ij} for which the SEs from the two models were approximately equal (ratio of SEs ≅ 1). These are described separately for the different levels of ICC. For the lowest ICC examined, the ratios of SEs from the two models varied from 1 to 1.07 and became equal to 1 for a relative dispersion of x_{ij} of 0.1. Increasing ICC increased the range of the ratio of SEs, and also the value of the relative dispersion of x_{ij} at which the two estimated SEs were approximately equal. For the highest ICC examined (ICC ≅ 0.30), the minimum value of the ratio of SEs was 0.84 and the maximum value was 5.32, while the relative dispersion for which $\mathrm{SE}_{\beta_1^{\mathrm{RI}}}$=$\mathrm{SE}_{\beta_1^{\mathrm{OLS}}}$ was 0.66.

Unlike the differences between regression coefficients, the ratios of SEs for a given relative dispersion of x_{ij} and ICC level were very similar for the different values of $SD_{e_{ij}}$ (data not shown), and therefore results are not presented separately.

Table 4.6. Range of ratios of SEs of the regression coefficients estimated from RI and OLS models ($\mathrm{SE}_{\beta_1^{\mathrm{RI}}}/\mathrm{SE}_{\beta_1^{\mathrm{OLS}}}$) and relative dispersions of explanatory variable x ($SD\_\bar{x}_j/within\ cluster\ SD\_x_{ij}$) for which SEs were equal (ratio=1)

ICC	**Range of ratios of SEs of the regression coefficients estimated from the RI and OLS models**		$\boldsymbol{SD\_\bar{x}_j/within\ cluster\ SD\_x_{ij}}$ **for** $\mathbf{SE}_{\beta_1^{\mathrm{RI}}}\mathbf{=SE}_{\beta_1^{\mathrm{OLS}}}$
	Minimum ratio	**Maximum ratio**	
0.001	1.00	1.07	0.104
0.003	1.00	1.16	0.114
0.01	1.00	1.57	0.137
0.03	0.99	2.10	0.204
0.1	0.95	3.35	0.349
0.3	0.84	5.32	0.663

Coverage of 95% confidence intervals

Table 4.7 shows the extent to which 95% confidence intervals covered the simulated effect of the explanatory variable on the outcome (β_1=1), when derived from the two statistical models, for different levels of ICC, and for fifths of the distribution of the relative dispersion of the explanatory variable x_{ij}.

Table 4.7. Coverage (%) by 95% confidence intervals of simulated effect β_1=1 under the RI and OLS models according to fifths of the distribution of the relative between- to within-cluster dispersion of explanatory variable x_{ij}

ICC	**Bottom fifth=1**		**2**		**3**		**4**		**Top fifth=5**		**Total**	
	RI	**OLS**	**RI**	**OLS**	**RI**	**OLS**	**RI**	**OLS**	**RI**	**OLS**	**RI**	**OLS**
0.001	95.07	94.39	95.01	93.86	95.17	94.04	95.07	93.98	95.10	94.03	95.08	94.07
0.003	95.13	92.84	95.32	92.03	95.37	91.94	95.37	91.98	95.38	91.97	95.31	92.16
0.01	94.85	87.26	94.71	83.79	94.59	83.67	94.70	83.67	94.80	83.53	94.73	84.34
0.03	94.78	74.80	94.81	67.67	94.79	67.50	94.90	67.28	94.96	67.63	94.85	68.92
0.1	94.73	54.82	94.72	44.94	94.61	44.84	94.95	44.72	94.83	44.35	94.77	46.74
0.3	94.64	37.94	94.94	27.74	94.67	27.46	94.94	27.28	94.77	27.22	94.79	29.43

Irrespective of ICC, coverage with the RI model was approximately 95%. Coverage for the OLS model was close to 95% for very low ICC and decreased for increasing levels of ICC. For the highest ICC level examined (ICC=0.3), OLS showed a notably poor coverage of 29%. For a given ICC, coverage of 95% confidence intervals did not vary much by relative between- to within-cluster dispersion of the explanatory variable x_{ij}, although they were somewhat higher in the bottom fifth as compared to the 2nd, 3rd, 4th, and 5th fifth of the distribution of relative dispersion of explanatory variable x_{ij}.

Type I error

To assess the frequency of type I error, defined as incorrect rejection of a true null hypothesis, under the OLS and the RI multi-level models, simulations were repeated assuming no association between the explanatory variable x_{ij} and the outcome variable y_{ij} ($\beta_1^{RI} = \beta_1^{OLS} = 0$).

Figure 4.7 shows the proportion of datasets for which the null hypothesis was rejected at a 5% significance level for varying levels of ICC. Using the RI multi-level model, the association between x_{ij} and y_{ij} was statistically significant (i.e. the effect differed from zero by more than 1.96*SE) in approximately 5% of the datasets for all ICCs. However, using the OLS models, type I error varied with ICC. For a very small ICC, type I error was very close to that under the RI model ($\sim$6%) but increased rapidly as the ICC increased, reaching $\sim 70\%$ for ICC≅0.30.

To explore reasons for the increased rates of Type 1 error with OLS regression, I focused on the simulated data sets for an expected ICC of 0.30 that gave a significant association when using OLS regression, but not when using multi-level modelling (n= 130,632 simulated datasets). A significant association between the outcome and the explanatory variable could be a result of either a strong effect or a small SE. Assuming that the random-intercept multilevel model was the appropriate method to use for such a data structure, for each data set, I tested whether $\beta_1^{RI} > 1.96 * SE^{OLS}$ and whether $\beta_1^{OLS} > 1.96 * SE^{RI}$. A higher count of the first inequality compared to the second one would mean that the observed high rates of Type 1 error were mostly due to underestimated SEs. Conversely, a higher count of the second inequality compared to the first one would indicate that the observed high rates of Type 1 error were mostly due to biases in the estimated effect under the OLS model. I found that in many of the simulated datasets $\beta_1^{OLS} > 1.96 * SE^{RI}$ (n = 28,013), while for 30,910 datasets neither inequality was satisfied. For the majority of the datasets (n=71,709), $\beta_1^{RI} > 1.96 * SE^{OLS}$ and $\beta_1^{OLS} < 1.96 * SE^{RI}$. This indicated that the significance was driven mostly by errors in the SE of the OLS estimate of effect.

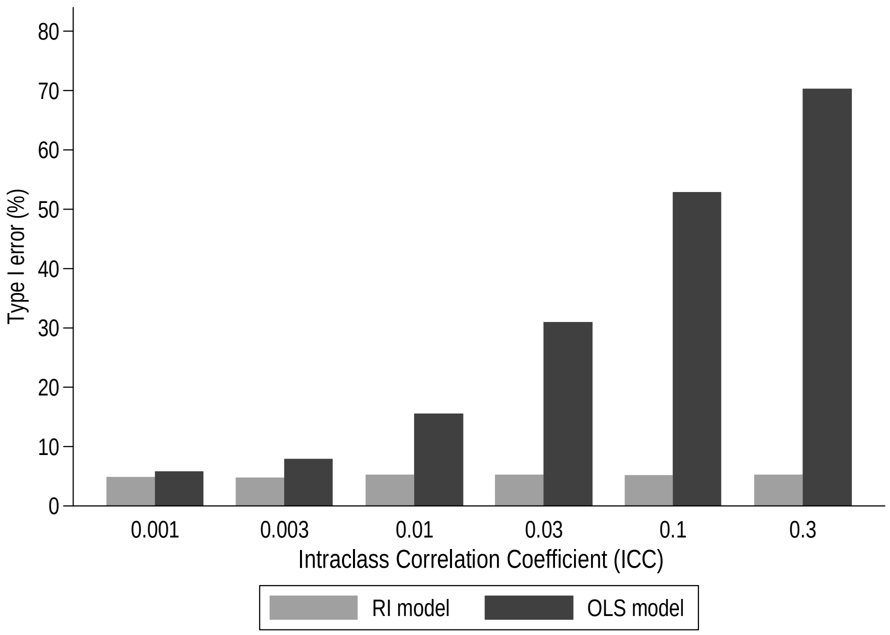


Figure 4.7. Proportion (%) of datasets for which the null hypothesis was rejected according to level of ICC when $\beta_1^{RI} = 0$

4.4 Two clusters – Binary explanatory variable

The implications of ignoring clustering effects in linear regression were next investigated in the case of a binary explanatory variable. As in the previous section, I initially investigated the situation in which observations were grouped within two clusters. As explained in previous section, the comparison between estimates derived from RI and OLS models, in the case of two clusters, becomes a comparison between an OLS model that adjusts for a cluster effect by using a dummy variable for cluster as a covariate and an OLS model that does not adjust for cluster.

To illustrate the case of two clusters when the explanatory variable is binary, data were simulated for a continuous outcome and a binary explanatory variable, with observations grouped in two clusters, as shown in Figure 4.8. The three subplots of the figure (A, B, C) correspond to three different scenarios of cluster-specific prevalence of the explanatory variable x. In subplot A, the prevalence of x is the same in both clusters (20%), in subplot B, the prevalence of x is higher in cluster 1 than in cluster 2 (20% v 5%) while in subplot C, the prevalence of x is lower in cluster 1 than in cluster 2 (20% v 35%). In each subplot, the two shades (light and dark grey) correspond to observations from the first and the second clusters, and the solid lines of the same shade are the fitted lines for the two separate clusters. The red dashed line is the fitted line for all observations when clustering is ignored, while the blue solid line is the overall fitted regression line between the y- and x-adjusted for clustering.

When the prevalence of the explanatory variable was the same in cluster 1 and cluster 2, the two fitted lines (adjusted for cluster (blue solid line), and unadjusted for cluster (red dashed line)) were approximately the same. This indicates that when prevalence was the same in the two clusters, adjusting for the clustering did not affect the estimated effect of the explanatory variable on the outcome variable. In contrast, when the prevalence of x differed between the clusters, the estimated effect differed according to whether the relationship between the outcome and the explanatory variable was adjusted for the cluster effects (subplots B and C in Figure 4.8). Specifically, when the prevalence of x in the first cluster was higher than in the second cluster, the slope of the regression line was steeper after adjusting for clustering than when no adjustment for clustering was made (subplot B, Figure 4.8). Conversely, when the prevalence in the first cluster was lower than that in the second cluster, the slope of the regression line with adjustment for cluster was less than that of the unadjusted regression line.

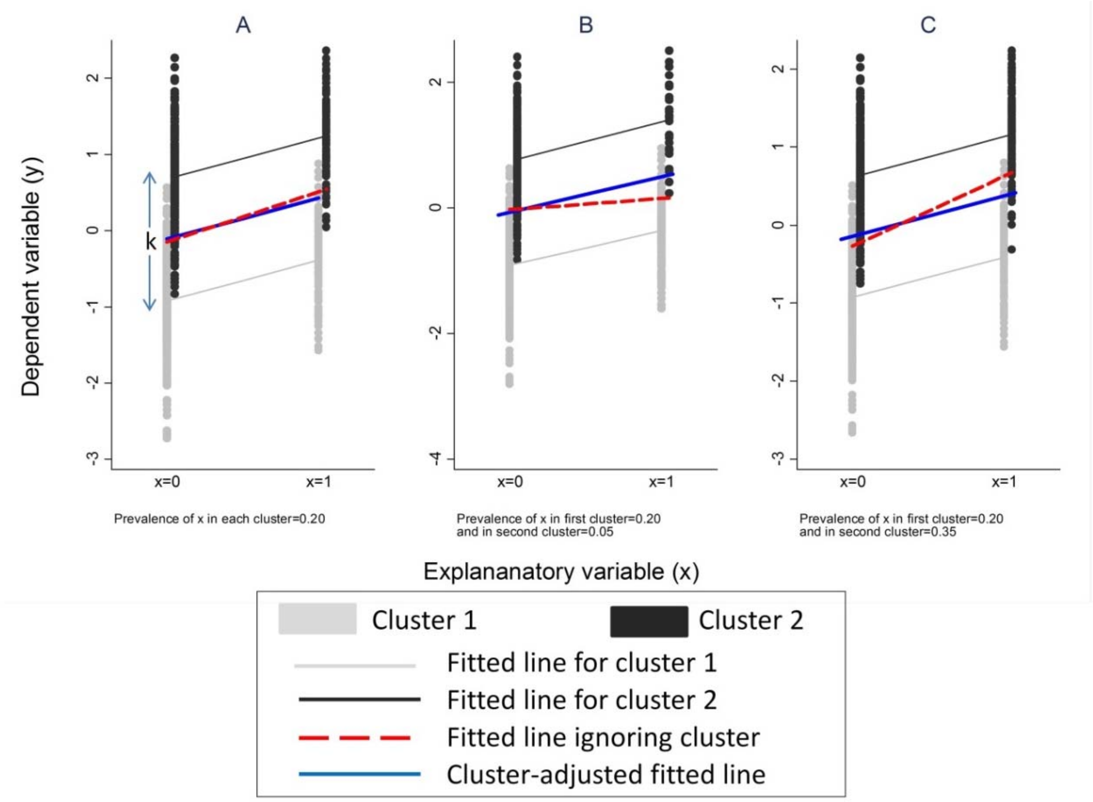


Figure 4.8. Scatter plot of continuous outcome y and binary explanatory variable x, with observations grouped in two clusters. The two shades (light and dark grey) correspond to the two clusters in which the observations are grouped, and k is the distance between the cluster-specific intercepts.

The differences in effect estimates (characterised by regression coefficients) and their corresponding SEs from a cluster-adjusted OLS and a cluster-unadjusted OLS model were further explored by considering different scenarios of clustering of the continuous outcome and binary explanatory variable. Specifically, regression coefficients and their SEs were estimated for varying differences in the prevalence of x in the two clusters (by varying the prevalence of x for the second cluster while keeping that in the first cluster fixed), distances between the cluster-specific intercepts, and dispersions of the error term, as described below.

4.4.1 Methods

For this investigation, the methods followed were similar to those described in section 4.2.1. In brief, I generated simulated data grouped in two clusters with 200 observations per cluster. In each simulation I generated a binary explanatory variable x_0 of given prevalence, which was the same for the two clusters. I then generated another 400 samples, each comprising two clusters of 200 observations, such that the prevalence in the first cluster was again that of the explanatory variable x_0, but that in the second cluster was different. The differences in prevalence rates between the two clusters across the 400 samples were values derived from a random normal distribution of mean zero and standard deviation arbitrarily set to 10. I then generated cluster-specific intercepts (u_1 and u_2 for the first and the second cluster, respectively) of given separation k (u_1=$u_2 + k$) and an error term of $y|x$ (e), the latter drawn from a normal distribution with mean zero and SD = 1. The values of the continuous outcome were generated as $y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij} + u_j$, with the regression coefficient β_1 arbitrarily set to 1 and the overall intercept β_0 arbitrarily set to 0. Simulated data were generated for combinations of the prevalence of x in the first cluster (0.05, 0.1, 0.2, and 0.4), and distance between the cluster-specific intercepts k (0.5, 1, 1.5). Each set of simulations was repeated 50 times.

To explore the effect of the error term e, simulations were repeated with the error term drawn from a normal distribution with mean zero and SD = 0.5, and SD = 1.5.

For each simulated dataset, a cluster-unadjusted (as described in equation 4.1) and a cluster-adjusted (equation 4.2) OLS model were fitted, and the effect estimates with their SEs were saved. From the estimates generated from the two models, I calculated the difference between the regression coefficients ($\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) and the ratio of the corresponding SEs ($SE^{cluster-adjusted}/SE^{cluster-unadjusted}$). These were plotted against varying values for the prevalence of the explanatory variable x in the second cluster while keeping the prevalence of x in the first cluster constant. Coverage by 95% CIs was summarised according to distances (k) between the cluster-specific intercepts (u's), standard deviations of $y_i|\, x_i$ (SD_e), and levels of the shift in the prevalence of the explanatory variable in the second cluster.

As in the case of observations grouped in two clusters with a continuous explanatory variable (section 4.2), to explore the consequences of ignoring cluster effects for type I error, simulations were repeated assuming an effect of x_i on y_i equal to 0 ($\beta_1 = 0$). The proportion of the simulated datasets in which the derived association between outcome and explanatory variable was significant at a 5% significance level was calculated and summarised for values of k, SD_e, and shift in the prevalence of the explanatory variable in the second cluster.

4.4.2 Results

Difference in regression coefficients

The differences between regression coefficients estimated from the cluster-adjusted and cluster-unadjusted OLS models ($\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) were plotted against the prevalence of the independent variable x in the second cluster while the prevalence of x in the first cluster remained constant (Figure 4.9). The four subplots of the figure correspond to the four fixed prevalence rates of x for the first cluster (A. 0.05; B. 0.1; C. 0.2; D. 0.4). Simulated results for the three different distances of the cluster-specific intercepts (k's) are represented by the three different shades of grey in each subplot of the figure. In all simulated data presented in this figure, the error term assumed to have a mean value of zero and SD=1.

When the prevalence of x in the second cluster was zero while that in the first cluster was higher, differences between the regression coefficients were positive, meaning that $\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted} > 0$. As the prevalence of x in the second cluster increased, this difference between the regression coefficients decreased, reaching a value of zero when the prevalence of x in the two clusters was the same. Further increase of the prevalence of x in the second cluster, resulted in the difference between the regression coefficients becoming increasingly more negative. The rate of decrease of the absolute differences was higher when the absolute difference in the prevalence of x between the two clusters was smaller, and lower when the absolute difference in the prevalence of x between the two clusters was larger.

The range of differences in the regression coefficients was dependent on the distance between the cluster-specific intercepts (k's) (Figure 4.9); the range of differences was smaller for smaller k and larger for larger k. The level of the fixed prevalence of x in the first cluster did not affect the pattern of the differences in the regression coefficients in relation to the prevalence of x in the second cluster, nor the range of the differences for varying distance k (comparison of subplots A, B, C, and D, Figure 4.9).

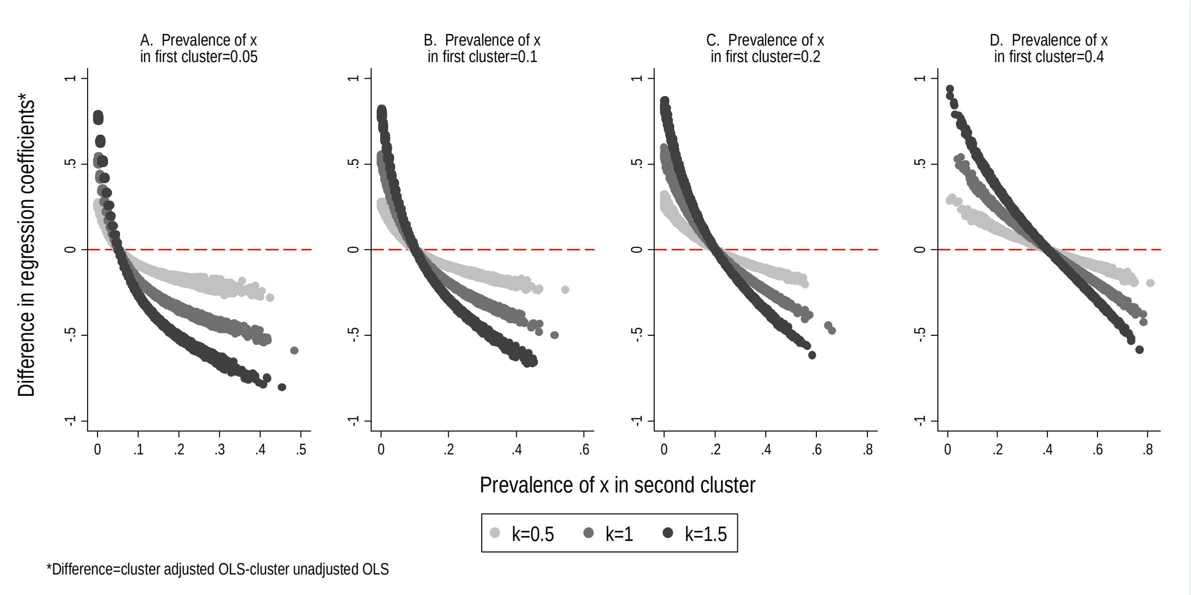


Figure 4.9. Differences in regression coefficients for varying prevalence of explanatory variable x in second cluster keeping prevalence of x in first cluster constant to A) 0.05, B) 0.1, C) 0.2, and D) 0.4

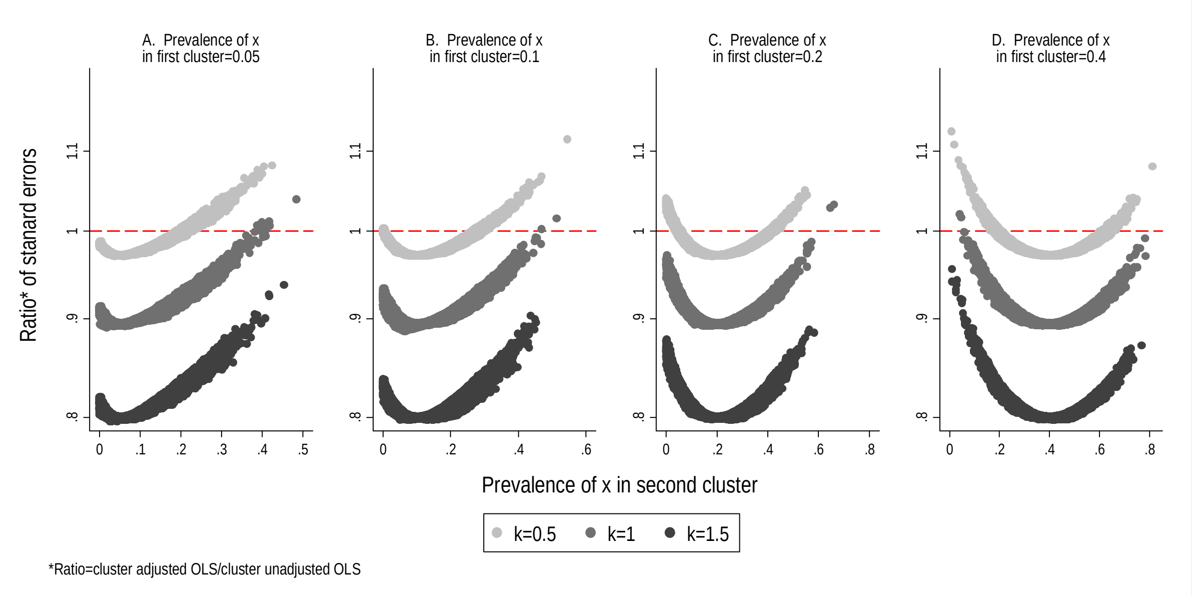
Ratio of standard errors

The ratios of the SEs of the regression coefficients from the two models ($SE^{cluster-adjusted}/SE^{cluster-unadjusted}$) were plotted against the prevalence of the independent variable x in the second cluster while the prevalence of x in the first cluster remained constant (Figure 4.10). As in Figure 4.9, the four different subplots correspond to the four different levels of the fixed prevalence of x in the first cluster (A. 0.05; B. 0.1; C. 0.2; D. 0.4) and the three different shades of grey in each subplot of the figure correspond to the three different distances between the cluster-specific intercepts examined ($k = \{0.5, 1, 1.5\}$). The error term was drawn from the standard normal distribution.

Overall, the ratios of the SEs were almost all below one, meaning that when there was no allowance for clustering, the SEs of the regression coefficients that were not adjusted for clustering were higher than the corresponding values derived from the OLS model that did adjusted for it. The ratio reached its minimum value when the prevalence of x in the second cluster was the same as that in the first cluster. Increasing differences in the prevalence of x between the two clusters (with prevalence of x in the second cluster either lower or higher than that in the first cluster) increased the ratio of the SEs, such that there was a U-shaped variation in the ratios across different prevalence rates of x in the second cluster for fixed prevalence of x in the first cluster.

The values of the ratios of the SEs were dependent on the distance between the cluster-specific intercepts (k). For smaller k, the ratios of the SEs were closer to 1 (minimum ratio ≈0.97 for $k = 0.5$) while increasing k resulted in lower values for the ratios (minimum ratio ≈0.9 and 0.8 for $k = 1$ and 1.5, respectively).

As observed for the differences in regression coefficients, neither the pattern of the ratios of SEs in relation to varying prevalence of x in the second cluster, nor the range of the values of the ratios were affected by the level of the fixed prevalence of x for the first cluster (comparison of subplots A, B, C, and D, Figure 4.10).



*Ratio=cluster adjusted OLS/cluster unadjusted OLS

Figure 4.10. Ratio of SEs of regression coefficients for varying prevalence of explanatory variable x in second cluster, keeping prevalence of x in first cluster constant to A) 0.05, B) 0.1, C) 0.2, and D) 0.4

Effect of error term

To explore the effect of the standard deviation of the error term e (SD_e) on the differences in the regression coefficients and the ratios of the corresponding SEs estimated from the two OLS models, simulations were repeated for SD_e=0.5 and SD_e =1.5 with the distance between the cluster-specific intercepts $k = 1$ and the prevalence of x in the first cluster set to 0.2.

As seen in subplot A of Figure 4.11, neither the pattern of the differences in the regression coefficients in relation to increasing prevalence of x in the second cluster nor the range of the differences were affected by increasing the dispersion of the error term. However, the dispersion of the differences was slightly greater as might be expected. In contrast to the regression coefficients, the ratio of the SEs was influenced substantially by changing SD_e (subplot B of Figure 4.11) with the ratio taking values closer to 1 for $SD_e = 1.5$ compared with $SD_e = 0.5$.

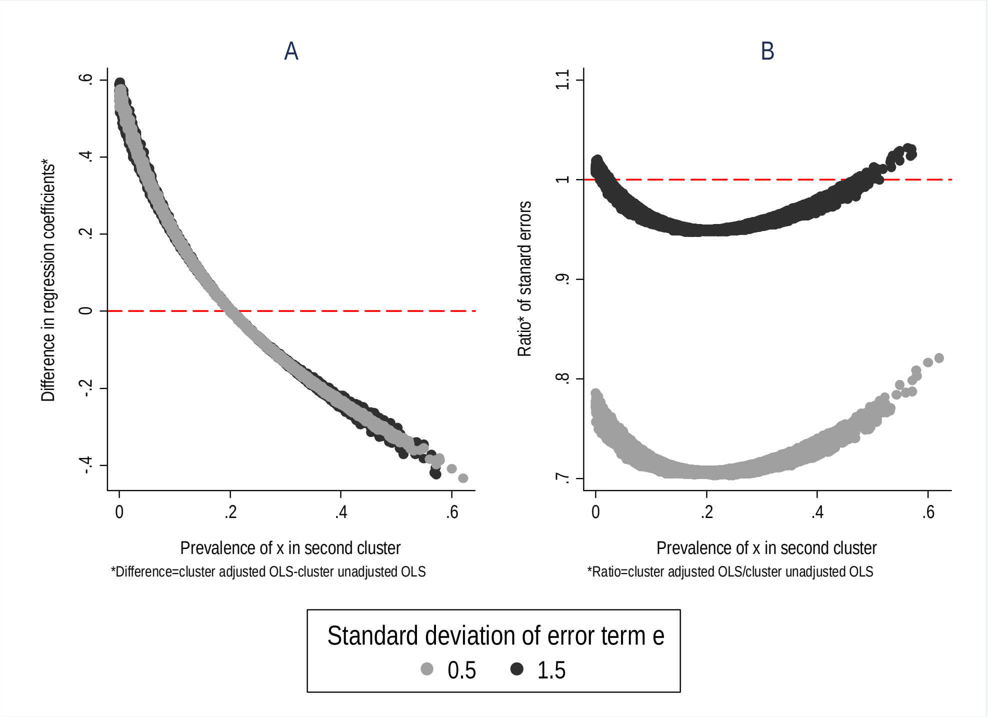


Figure 4.11. Differences between regression coefficients and ratios of the corresponding standard errors derived from cluster-adjusted and cluster-unadjusted OLS models according to prevalence of explanatory variable x in the second cluster for $SD_e = 1.5$ (dark grey) and $SD_e = 0.5$ (light grey), when distance between the cluster-specific intercepts $k = 1$ and prevalence of x in the first cluster was 0.2

Coverage by 95% confidence intervals

Coverage by 95% CIs of the simulated effect $\beta_1^{Cluster-adjusted}$=1 from the cluster-adjusted and cluster-unadjusted models was calculated for different values of k and SD_e. To explore how the relative prevalence of x in the two clusters influenced coverage by 95% CIs, I calculated the absolute difference in the prevalence rates ($|prevalence\ of\ x\ in\ cluster\ 1 - prevalence\ of\ x\ in\ cluster\ 2|$) and I considered quarters of its distribution. Coverage for the cluster-adjusted model was very close to the nominal value of 95% (average of 95.3%). It varied from 94.7% to 96.0% and from 94.6% to 96.5% for the different values of k and SD_e, respectively. Coverage also varied very little across quarters of the distribution of the differences in the two prevalence rates (second column of Table 4.8).

Table 4.8. Coverage by 95% confidence intervals under the cluster-adjusted and cluster-unadjusted model across quarters of the distribution of the absolute difference in prevalence rates of x in the two clusters

Quarters	**Cluster-adjusted model**	**Cluster-unadjusted model**
1st	95.1	97.0
2nd	95.4	92.6
3rd	95.5	81.2
4th	95.2	56.6
All	95.3	81.9

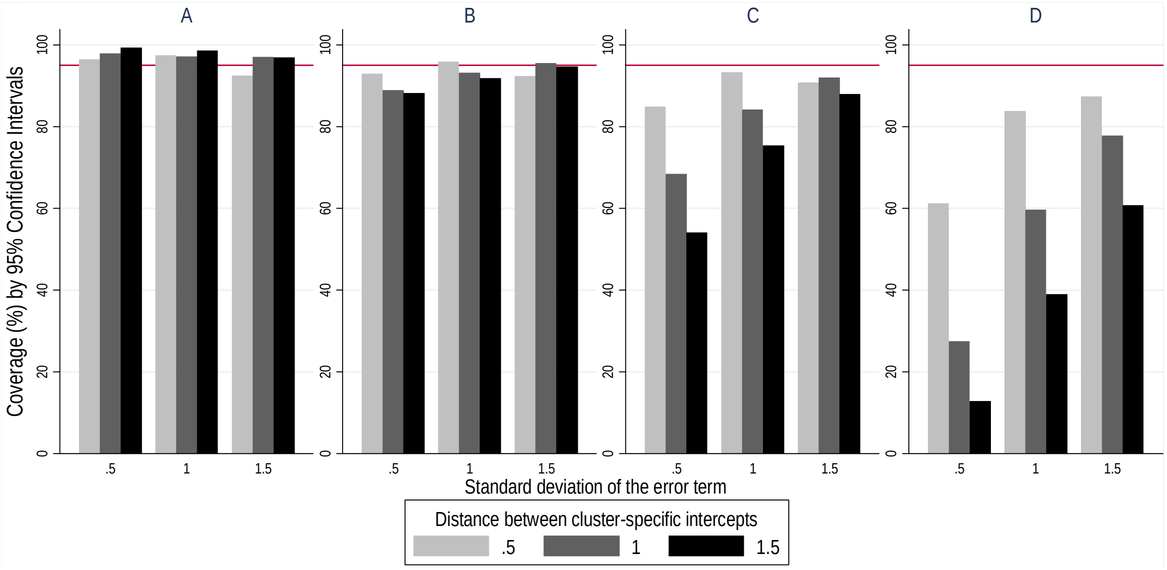


Figure 4.12. Coverage by 95% confidence intervals for the simulated effect $\beta_1^{Cluster-adjusted}$=1 from the cluster-unadjusted model for different values of distance between the cluster-specific intercepts k, standard deviation of error term (SD_e), and quarters of the distribution of the absolute difference in prevalence rates of x in the two clusters depicted in the 4 sub-plots (A) 1st, B) 2nd, C) 3rd, and D) 4th). The red horizontal line in the graphs represents the nominal value of 95% for coverage.

Unlike coverage with the cluster-adjusted model, that for the cluster-unadjusted model varied considerably across values of k and SD_e and quarters of the distribution of absolute differences in prevalence rates of the two clusters. The levels of coverage from the cluster-unadjusted model are illustrated in Figure 4.12. Coverage was above the nominal value of 95% when the difference between the cluster prevalence rates was small, but it fell below that value as the difference increased (right column of Table 4.8). Increasing values of k, for constant SD_e, decreased coverage. Coverage also decreased with decreasing SD_e, for any given value of k. These trends were more apparent when there were bigger differences between prevalence rates in the two clusters (Figure 4.12).

Type I error

To explore type I error in the cluster-unadjusted and cluster-adjusted OLS models, the effect size of the explanatory variable on the outcome variable was set to zero. As for the coverage by 95% CIs, type I error rates were explored in relation to values of k and SD_e and quarters of the distribution of absolute differences in prevalence rates of the two clusters. Overall, 22% of the simulated datasets showed a significant association at the 5% significance level when a cluster-unadjusted OLS model was fitted, while the corresponding proportion when a cluster-adjusted OLS model was fitted was 5%. Type I error rates from the cluster-adjusted model varied very little with difference in prevalence rates of x in the two clusters (Table 4.9) and by values of k and SD_e (range: 4.5% to 5.7%).

In contrast to the cluster-adjusted OLS model, when clustering was ignored, Type I error rates were lower than 5% when the difference in prevalence rates between the two clusters was small (Table 4.9 and sub-plot A in Figure 4.13) and increased to values higher than the nominal value of 5% when differences were larger. Type I error increased with decreasing SD_e, and with increasing distance between cluster-specific intercepts. The highest type I error rate was 92.9% and it was observed for the highest value of k, lowest value of SD_e, and in the highest quarter of the distribution of the absolute difference between prevalence rates in the two clusters (sub-plot D in Figure 4.13).

Table 4.9. Type I error rates for the cluster-adjusted and cluster-unadjusted models across quarters of the distribution of the absolute difference in prevalence rates of x in the two clusters

Quarters	**Cluster-adjusted model**	**Cluster-unadjusted model**
1st	5.1	3.3
2nd	5.2	11.1
3rd	4.8	25.0
4th	4.8	46.8
All	5.0	21.5

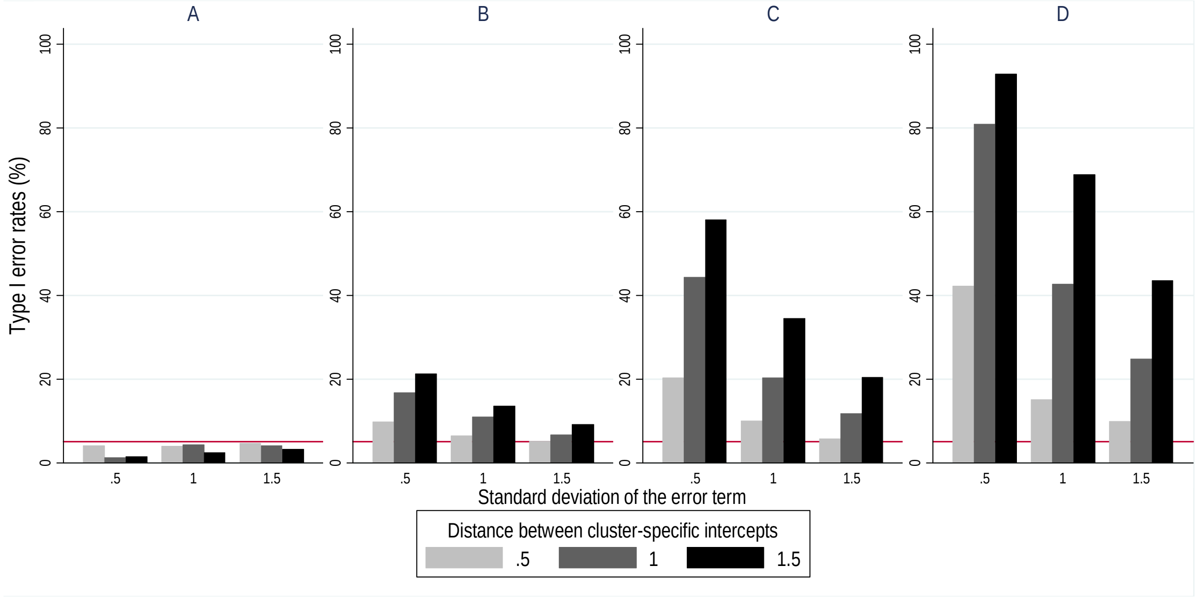


Figure 4.13. Type I error rates for the cluster-unadjusted model for different values of distance between the cluster-specific intercepts k, standard deviation of error term (SD_e), and quarters of the distribution of the absolute difference in prevalence rates of x in the two clusters depicted in the 4 sub-plots (A) 1st, B) 2nd, C) 3rd, and D) 4th). The red horizontal line in the graphs represents the nominal value of 5% for type I error.

4.5 Multiple clusters – Binary explanatory variable

4.5.1 Methods

In the case of a single binary explanatory variable, the OLS model that ignores clustering effects is described by the model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

which was set out earlier in equation 4.1, and the RI model that adjusts for clustering effects by the model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij} + u_j$$

(as presented in equation 4.3), in which x_{ij} follows a binomial distribution.

The methods used to explore the consequences of ignoring clustering were similar to those described in section 4.3.1. In brief, I generated simulated datasets of 100 clusters with 100 observations per cluster and I arbitrarily set the size of the effect of x_{ij} on y_{ij} to 1 ($\beta_1 = 1$), and the average value of y_{ij} when x_{ij}= 0 to 0 ($\beta_0 = 0$). I then set the prevalence of x_{ij} in each cluster to be the sum of a constant (the same in all clusters) and a cluster-specific variable $shift_j \sim N\left(0, SD_{shift}{}^2\right)$. The corresponding values for the outcome variable y_{ij} were generated according to equation 4.3. The individual-level error terms were drawn from a random standard normal distribution ($N(0,1)$), and the cluster-level error terms were drawn from a random normal distribution with mean zero and variance SD_u^2.

Simulated data were generated for combinations of the values for SD_u described in section 4.3.1 with the overall prevalence of x_{ij} set to 0.05, 0.1, 0.2, and 0.4, while $SD_{shift} \sim U[a, b]$ with the parameters a and b arbitrarily chosen to be 0 and 0.05, respectively. For each combination of expected ICC and $SD_{x_{ij}}$, 100 simulated datasets were produced, with ICC values being those selected for the simulation process for linear regression with a continuous explanatory variable described in section 4.3.1.

For each simulated dataset, an OLS and RI model were fitted and the regression coefficients were estimated with corresponding SEs. In accordance with previous analyses, I calculated the difference between the estimated regression coefficients ($\beta_1^{RI} - \beta_1^{OLS}$), and the ratio of their SEs ($SE_{\beta_1^{RI}} / SE_{\beta_1^{OLS}}$). Then, these two measures were plotted against the dispersion of the prevalence of the explanatory variable x_{ij} across the clusters.

The coverage of the 95% confidence interval was calculated as the proportion of simulated confidence intervals that included the true value, and was used as an indication of errors arising from problems in the estimation of effect or of the corresponding standard error.

The simulations were repeated assuming zero effect of the explanatory on the outcome variable (i.e. $\beta_1 = 0$). The proportions of datasets for which the null hypothesis was rejected at a 5% significance level in OLS and RI modelling were calculated to asses impacts on type I error according to levels of ICC.

4.5.2 Results

Difference in regression coefficients

Differences in regression coefficients ($\beta_1^{RI} - \beta_1^{OLS}$) estimated from the two linear models (described in equations 4.1 and 4.3) according to the dispersion of the prevalence of the binary explanatory variable x_{ij} across the clusters are presented in Figure 4.14. The four different subplots of the figure (A, B, C, and D) correspond to the four different overall prevalence rates of x_{ij}. In each scatter plot, the different levels of within-cluster similarity of observations (ICCs) are depicted in different shades of grey with darker shades corresponding to simulated results for higher ICCs.

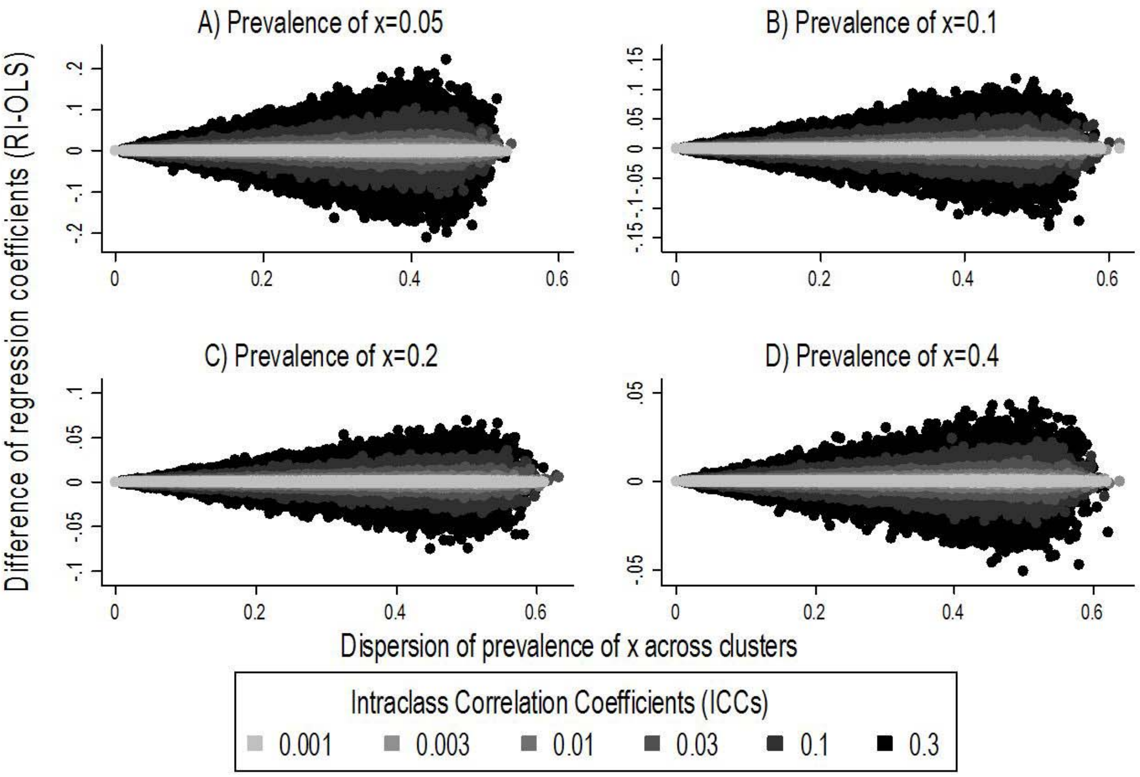


Figure 4.14. Difference between regression coefficients estimated from RI and OLS models ($\beta_1^{RI} - \beta_1^{OLS}$) according to dispersion of prevalence of explanatory variable x_{ij} across clusters. Figure A: Overall prevalence of x_{ij} 0.05. Figure B: Overall prevalence of x_{ij} 0.1. Figure C: Overall prevalence of x_{ij} 0.2. Figure D: Overall prevalence of x_{ij} 0.4

On average, for small dispersion of the cluster-specific prevalence of x_{ij}, differences were more narrowly spread for small ICCs and more widely spread for large ICCs. For each value of ICC, small dispersion of cluster-specific prevalence of x_{ij} resulted in small differences between the regression coefficients. However, increasing the dispersion of cluster-specific prevalence of x_{ij}, resulted in larger differences between the regression coefficients from the two methods. Comparing the different subplots of Figure 4.14 (note the different scales on the y-axes), higher overall prevalence of x_{ij} resulted in regression coefficients from the two models being more similar even for large dispersion of the prevalence of x_{ij} across clusters; for ICC=0.3, differences ranged from -0.2 to 0.2, corresponding to a 20% difference in the regression coefficients from the two methods, when the overall prevalence of x_{ij} was 0.05, and this range decreased to approximately -0.05 to 0.05 for an overall prevalence of x_{ij} of 0.4.

The ranges of differences in regression coefficients estimated from the RI and the OLS models for varying overall prevalence of the independent variable x_{ij} and ICC's are presented in Table 4.10.

Table 4.10. Range of difference in regression coefficients estimated from the RI and the OLS models ($\beta_1^{RI} - \beta_1^{OLS}$) for varying ICC and overall prevalence of the independent variable x_{ij}

	Overall prevalence of x_{ij}			
ICC	**0.05**	**0.1**	**0.2**	**0.4**
0.001	-0.003 to 0.004	-0.002 to 0.002	-0.001 to 0.001	-0.001 to 0.001
0.003	-0.007 to 0.007	-0.005 to 0.006	-0.003 to 0.003	-0.002 to 0.002
0.01	-0.022 to 0.026	-0.012 to 0.013	-0.007 to 0.008	-0.004 to 0.004
0.03	-0.045 to 0.042	-0.027 to 0.031	-0.018 to 0.017	-0.011 to 0.012
0.1	-0.105 to 0.103	-0.058 to 0.054	-0.034 to 0.035	-0.023 to 0.024
0.3	-0.211 to 0.222	-0.130 to 0.118	-0.075 to 0.069	-0.051 to 0.045

The estimated regression coefficients were on average very similar. The ranges of differences were very small for small ICC and large overall prevalence of x_{ij}, being up to 0.004. The ranges of the differences increased with increasing ICC and with decreasing overall prevalence of x_{ij}. For example, when the ICC was 0.3, differences ranged up to 0.2 when the overall prevalence of x_{ij} was 0.05, but only up to 0.13, 0.075, and 0.05 when the overall prevalence of x_{ij} was 0.1, 0.2, and 0.4, respectively.

Ratio of standard errors

The ratios of the SEs of the regression coefficients estimated from the RI and OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) were investigated in relation to the dispersion of the prevalence of the explanatory variable x_{ij} across the clusters. The results are presented in Figure 4.15. As in Figure 4.14, the four different subplots of Figure 4.15 correspond to the four different levels of underlying prevalence of the explanatory variable x_{ij}. In each subplot, the different levels of ICCs are represented by different shades of grey, with darker shades corresponding to higher ICCs. Overall the ratios of the SEs were below one for most of the situations examined, indicating that the SEs of the regression coefficients estimated from the RI model were smaller than those under the OLS model. The ratio of the SEs achieved its minimum value for the smallest dispersion of the prevalence of x_{ij} across the clusters, and increased progressively with increasing dispersion of x across clusters. For small ICCs (<0.1), the SEs from the two models were very similar. However, increasing the ICC to 0.1 or higher led to the ratio of the SEs decreasing to values much lower than 1. For constant ICC, comparison of subplots A, B, C, and D, shows that the rate of increase of the ratio of the SEs was higher for lower underlying prevalence rates of the x_{ij}.

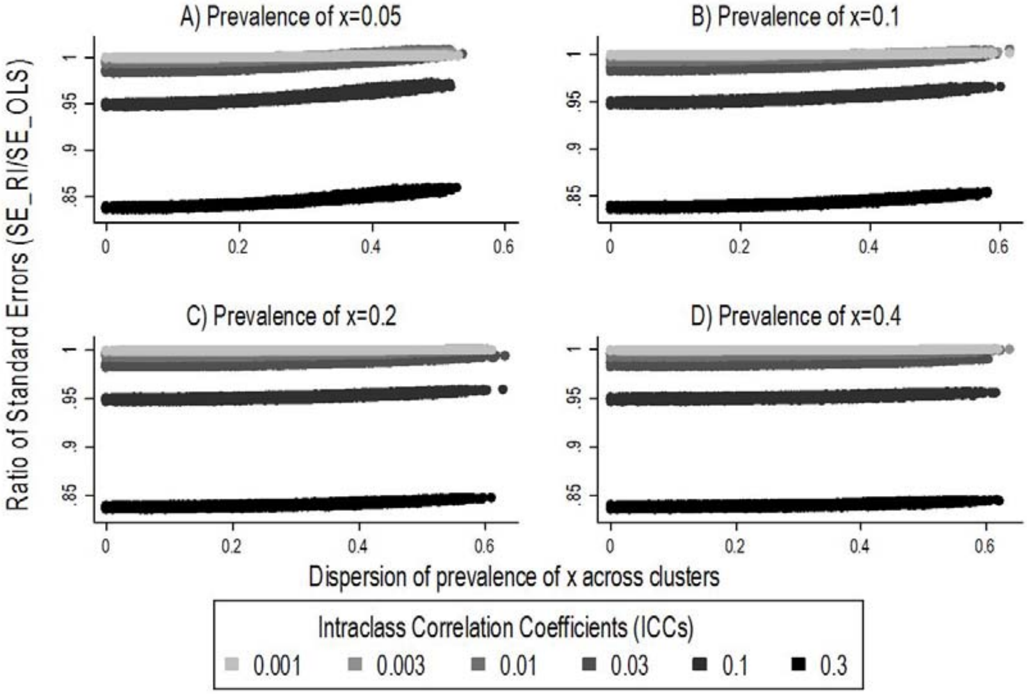


Figure 4.15. Ratios of standard errors of regression coefficients estimated from RI and OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) according to dispersion of prevalence of independent variable x_{ij} across clusters. Figure A: Overall prevalence of x_{ij} 0.05. Figure B: Overall prevalence of x_{ij} 0.1. Figure C: Overall prevalence of x_{ij} 0.2. Figure D: Overall prevalence of x_{ij} 0.4

Table 4.11. Range of the ratios of SEs estimated from the RI and the OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) for varying ICC and overall prevalence of the independent variable x_{ij}

ICC	Overall prevalence of x_{ij}=0.05	Overall prevalence of x_{ij}=0.1	Overall prevalence of x_{ij}=0.2	Overall prevalence of x_{ij}=0.4
0.001	0.999 to 1.003	0.999 to 1.002	0.999 to 1.001	0.999 to 1.000
0.003	0.998 to 1.005	0.998 to 1.003	0.998 to 1.001	0.998 to 1.000
0.01	0.993 to 1.008	0.993 to 1.005	0.993 to 1.001	0.993 to 1.000
0.03	0.983 to 1.006	0.983 to 1.001	0.983 to 0.994	0.983 to 0.992
0.1	0.946 to 0.972	0.947 to 0.966	0.947 to 0.960	0.946 to 0.956
0.3	0.836 to 0.860	0.835 to 0.854	0.835 to 0.848	0.835 to 0.845

The ranges of the ratios of SEs estimated from the RI and the OLS models for overall prevalence of the independent variable x_{ij} of 0.05, 0.1, 0.2, and 0.4 and for varying ICCs are summarised in

Table 4.11. As shown in Figure 4.15, almost all ratios were <1. The minimum value of the ratio was very close to 1 for small ICCs, while increasing ICC resulted in lower minimum values of the ratio. For fixed ICC, the maximum value of the ratio was higher for small overall prevalence of x_{ij}. For example, for ICC=0.3 the maximum value of the ratio was 0.860 for an overall prevalence of the explanatory variable 0.05, 0.854 for a prevalence 0.1, 0.848 for a prevalence of 0.2, and 0.845 for a prevalence of 0.4.

Coverage of 95% confidence intervals

The proportion of the estimated 95% confidence intervals for the association between the explanatory and the outcome variable that included the true value of 1, when derived from the RI model was 95%. The OLS model also produced 95% confidence intervals 95% of which included the true value but only for ICC≤0.03. As ICC increased, coverage from the OLS model deviated from the nominal value of 95%. As shown in Figure 4.16, when ICC was 0.1 or 0.3, coverage was on average lower for lower prevalence of x_{ij}; it fell below the nominal value of 95% for 0.05 prevalence of x_{ij} and it increased to values higher than 95% for 0.40 prevalence of x_{ij} (comparison of the four sub-plots of the figure). Also, for any given prevalence of x_{ij}, coverage was lower for increasing dispersion of prevalence of x_{ij} across clusters. Variation of the average coverage by categories of prevalence rates of x_{ij} and overall prevalence of x_{ij} was higher when ICC was higher (ICC=0.3) than when it was lower (ICC=0.1). The smallest and the largest values of coverage were 87% and 98% and they were observed when overall prevalence of x_{ij} was 0.05 and in the bottom and top thirds respectively of the distribution of dispersion of prevalence of x_{ij} across clusters. Coverage as high as 98% was also seen in the bottom third of the distribution of dispersion of prevalence of x_{ij} across clusters for the other prevalence rates (0.10, 0.20, and 0.40) explored.

Type I error

The frequency of type I error rates with the two analytical approaches, was assessed for different values of ICC, overall prevalence of x, and dispersion of prevalence of x across clusters. With the RI model, type I error rates for the different conditions had an average of 5% and varied very little (from 4.6% to 5.4%). The same observation was made for type I error rates under the OLS model when ICC values were lower than 0.1; the average value was 5% and varied from 4.6% to 5.6% for different ICC values (<0.1), overall prevalence rates of x, and dispersion of prevalence of x across clusters. However, for ICC values of 0.1 and 0.3, type I error rates diverged from 5%. The variation of rates in those cases is illustrated in Figure 4.17 for the four prevalence rates of x (subplots A, B, C, and D of the figure), and for thirds of the distribution of dispersion of prevalence of x across clusters. For small dispersion of prevalence rates of x (bottom third of the distribution), type I error was lower than 5%, and it increased as dispersion increased. This trend was more prominent for lower values of overall prevalence of x, and for ICC=0.3 compared to ICC=0.1. The smallest and the largest values of type I error were 2% and 12% and they were observed when overall prevalence of x was 0.05 and in the bottom and top thirds respectively of the distribution of dispersion of prevalence of x across clusters.

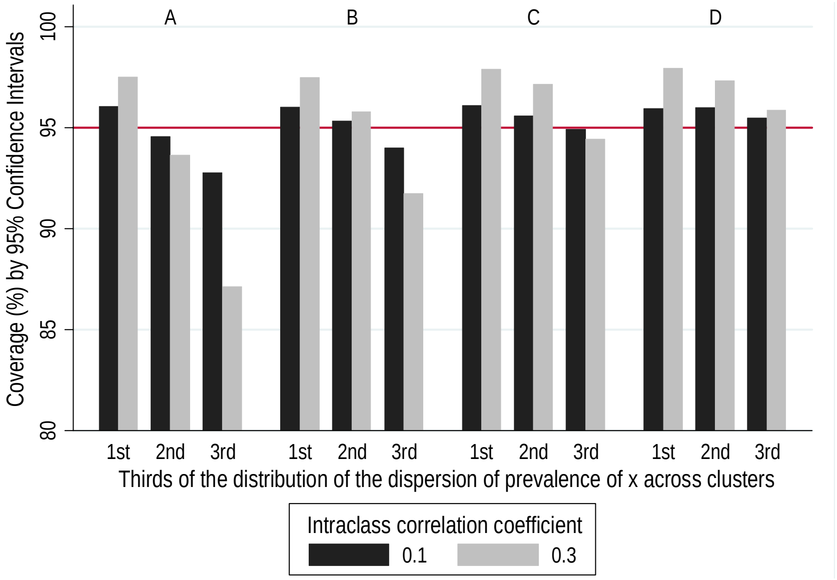


Figure 4.16. Coverage (%) by 95% confidence intervals from the OLS model for ICC=0.1 and 0.3, by overall prevalence rates of x (A) 0.05, B) 0.10, C) 0.20, and D) 0.40), and thirds of the distribution of the dispersion of prevalence of x across clusters

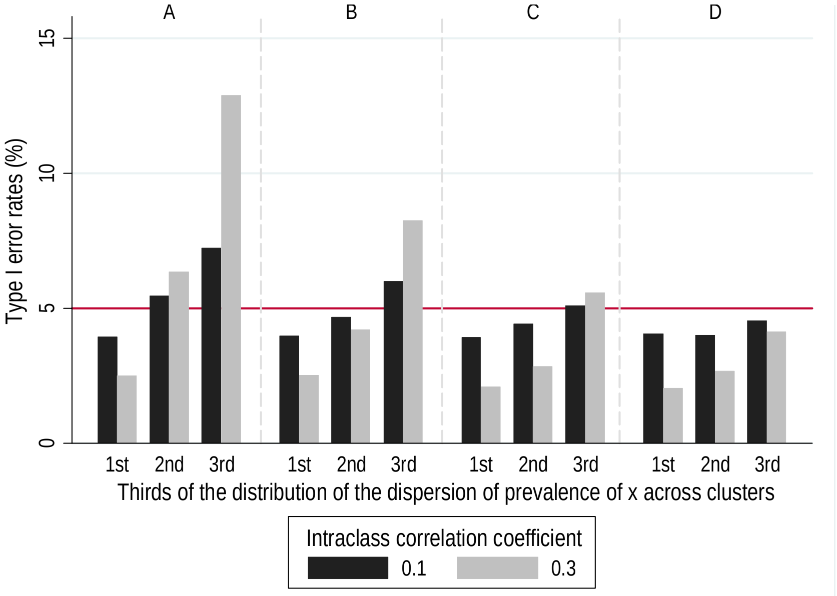


Figure 4.17. Type I error rates (%) from the OLS model for ICC=0.1 and 0.3, by overall prevalence rates of x (A) 0.05, B) 0.10, C) 0.20, and D) 0.40), and thirds of the distribution of the dispersion of prevalence of x across clusters

4.6 Summary of results

4.6.1 Continuous outcome - continuous predictor – 2 clusters

The difference between $\beta_1^{Cluster-adjusted}$ and $\beta_1^{Cluster-unadjusted}$ was very small (≈ 0) when the cluster-specific distributions of the explanatory variable x were overlaid. The difference between the two estimates increased as the distribution of x for the second cluster was shifted away from that of the first cluster, up to a shift of 2 (twice the within-cluster SD of x). Beyond that point, the difference in the point estimates produced from the two models started decreasing, approaching a value of zero again. Differences were larger when the distance between the cluster-specific intercepts k was larger. Higher values of variance of the error term led to increased dispersion of differences between $\beta_1^{Cluster-adjusted}$ and $\beta_1^{Cluster-unadjusted}$, while the mean values remained approximately the same.

SEs from the cluster-adjusted model were lower than those from the cluster-unadjusted model, when the distributions x in the two clusters were overlaid. Shifting the distribution of x in the second cluster away from that of the first cluster increased SEs from the cluster-adjusted model more compared to those from the cluster-unadjusted model resulting in ratios $SE^{cluster-adjusted}/SE^{cluster-unadjusted}$ >1. When the cluster-specific distributions of x were overlaid, the ratio of SEs depended on both the distance between the cluster-specific intercepts k and the variance of the error term. Increasing k, or the variance of the error term decreased the ratio of SEs to values <1. When values of x for the second cluster were shifted from those of the first cluster, k and the variance of the error term had no effect on ratios of SEs.

Coverage by 95% CIs was at the nominal level of 95% when the model adjusted for clustering. With the cluster-unadjusted model, coverage was severely biased downwards. Coverage by 95% CIs decreased for smaller values of k, higher values of variance of the error term, and larger shifts in the distribution of x in the second cluster from that in the first cluster.

Type I error rates were close to the nominal value of 5% when clustering was accounted for. However, with the cluster-unadjusted OLS model, error rates increased to considerably higher values. This increase was observed with increasing shifts of the cluster-specific distributions of x, increasing variance of the error term, and decreasing values of k.

These observations are summarised in Table 4.12.

Table 4.12. Summary of simulation results when the outcome variable was continuous, the explanatory variable was continuous, and observations were grouped in 2 clusters

$\lvert\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}\rvert$ ↑ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ approximately < 2 within-cluster SD of x ↓ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ approximately > 2 within-cluster SD of x ↑ when k ↑
$SE^{cluster-adjusted} < SE^{cluster-unadjusted}$ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert \approx 0$ Minimum of $SE^{cluster-adjusted}/SE^{cluster-unadjusted}$ ↓ when k ↑ ↓ when SD_e ↑ $SE^{cluster-adjusted}/SE^{cluster-unadjusted}$ ↑ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑
Coverage by 95% CIs from the cluster-adjusted model: on average 95% Coverage by 95% CIs from the cluster-unadjusted model was < 95% and it ↓ when k ↓ ↓ when SD_e ↑ ↓ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑
Type I error from the cluster-adjusted model: on average 5% Type I error from the cluster-unadjusted model was >5% and it ↓ when k ↑ ↓ when SD_e ↓ ↓ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑

4.6.2 Continuous outcome - continuous predictor – multiple clusters

Differences between β_1^{RI} and β_1^{OLS} were on average very small (≈ 0) for all values of ICC, and relative dispersion of x_{ij} across clusters. Dispersion of differences, however, increased as ICC increased, and for higher values of $SD_{e_{ij}}$. It also increased as relative dispersion of x_{ij} increased, up to a value of ≈ 1, while further increase in the relative dispersion of x_{ij} causing a decrease in the dispersion of $\beta_1^{RI} - \beta_1^{OLS}$.

In most cases, $SE^{RI} > SE^{OLS}$. However, the ratio, SE^{RI}/SE^{OLS}, was less than 1 when ICC was high, and when the relative dispersion of x_{ij} was small. The ratios of SEs were also very close to 1 for lower ICC values and when the relative dispersion of x_{ij} was small. Increasing relative dispersion of x_{ij}, and increasing values of ICC, were associated with increasing ratio of SEs.

Coverage by 95% CIs was at the nominal level of 95% with the RI model. Under the naïve OLS model, coverage was lower than 95%. It decreased as ICC increased, but it was not affected by increasing relative dispersion of x_{ij}.

Type I error from the RI model was very close to 5%. However, with the OLS model, Type I error increased with increasing ICC values, although it was not influenced by increasing values of relative dispersion of x_{ij}.

These observations are summarised in Table 4.13.

Table 4.13. Summary of simulation results when the outcome variable was continuous, the explanatory variable was continuous, and observations were grouped in 100 clusters

$\beta_1^{RI} - \beta_1^{OLS}$: on average ≈ 0 for any ICC and relative dispersion of x_{ij} Range of $\beta_1^{RI} - \beta_1^{OLS}$: ↑ when ICC ↑ ↑ when $SD_{e_{ij}}$↑ ↑ when relative dispersion of x_{ij} <1 ↓ when relative dispersion of x_{ij} >1
$SE^{RI} > SE^{OLS}$ in most cases SE^{RI}/SE^{OLS} < 1 for high ICC and small relative dispersion of x_{ij} SE^{RI}/SE^{OLS} ≈ 1 for lower ICC and small relative dispersion of x_{ij} SE^{RI}/SE^{OLS} ↑ when ICC ↑ ↑ when relative dispersion of x_{ij} ↑
Coverage by 95% CIs from the RI model: on average 95% Coverage by 95% CIs from the OLS model was < 95% and it ↓ when ICC ↑
Type I error from the RI model: on average 5% Type I error from the OLS model was >5% and it ↑ when ICC ↑

4.6.3 Continuous outcome - binary predictor – 2 clusters

The estimate of effect from the cluster-adjusted model was the same as that from the cluster-unadjusted model when the prevalence of the explanatory variable x was the same in both clusters. Increasing absolute difference in the prevalence rates of x in the two clusters increased the absolute difference of $\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$. Also, increasing the distance between the cluster-specific intercepts k, increased the difference in effects estimated from the two models.

SEs derived from the cluster-adjusted model were always lower than those from the cluster-unadjusted model resulting in ratios $SE^{cluster-adjusted}/SE^{cluster-unadjusted}<1$. The ratio of SEs had a minimum value when prevalence rates of x in the two clusters were the same, and it increased towards 1 as the difference between cluster-specific prevalence rates of x increased. Ratios of SEs were generally lower for increasing values of k, and for lower values of variance of the error term.

Coverage by 95% CIs from the cluster-adjusted model was on average 95% and varied very little across values of k and for difference in the prevalence rates of x in the two clusters. However, coverage under the cluster-unadjusted model was biased downwards; it decreased with decreasing variance of the error term, with increasing values of k, and with increasing difference between the prevalence rates of x in the two clusters.

Type I error rates calculated from the cluster-adjusted model were approximately 5%. In contrast, the cluster-unadjusted model produced error rates that were higher than 5%; they increased with decreasing variance of the error term, with increasing values of k, and with increasing difference between the prevalence rates of x in the two clusters.

These observations are summarised in Table 4.14.

Table 4.14. Summary of simulation results when the outcome variable was continuous, the explanatory variable was binary, and observations were grouped in two clusters

$\lvert\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}\rvert$	≈ 0 when $\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$ ↑ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑ ↑ when k ↑
$SE^{cluster-adjusted} < SE^{cluster-unadjusted}$	
Minimum of $SE^{cluster-adjusted}/SE^{cluster-unadjusted}$	when $\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$ ↓ when k ↑ ↑ when SD_e ↑
$SE^{cluster-adjusted}/SE^{cluster-unadjusted}$	↑ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑
Coverage by 95% CIs from the cluster-adjusted model:	on average 95%
Coverage by 95% CIs from the cluster-unadjusted model was < 95% and it	↓ when k ↑ ↓ when SD_e ↓ ↓ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑
Type I error rates from the cluster-adjusted model:	on average 5%

Type I error rates from the cluster-unadjusted model were > 5% and it

↓ when k ↑

↓ when SD_e ↓

↓ when $|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}|$ ↑

4.6.4 Binary outcome - binary predictor – multiple clusters

The estimates of effect from the RI and the OLS models were on average the same. However, differences in effect estimates ($\beta_1^{RI} - \beta_1^{OLS}$) varied more for higher values of ICC, higher dispersion of prevalence of x_{ij} across clusters, and for lower values of overall prevalence of x_{ij}.

SEs from the RI model were consistently lower than those from the OLS model. Ratios of the SEs ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) were lower when ICC was higher, and increased slightly with increasing dispersion of prevalence of x_{ij} across clusters. Overall prevalence of x_{ij} did not affect patterns of ratios in relation to ICC values and dispersion of prevalence of x_{ij} across clusters.

Coverage was very close to 95% under the RI. It was also close to 95% under the OLS model for ICC≤0.03 but varied by dispersion of prevalence of x_{ij} across clusters and by overall prevalence of x_{ij} when ICC was 0.1 or 0.3. Bigger variation of coverage around the value of 95% was seen for ICC=0.3.

Type I error rates were very close to 5% under the RI model for any ICC value. Error rates were also very close to the nominal value of 5% under the OLS model but only when the ICC value was <0.1. When ICC was equal to 0.1 or 0.3, type I error rates varied from lower to higher than 5%, depending on the overall prevalence of x_{ij} and on the dispersion of prevalence of x_{ij} across clusters. For small dispersion, type I error was <5% and it increased with increasing dispersion of prevalence of x_{ij}. For lower overall prevalence of x_{ij}, type I error rates increased up to 12% for large dispersion of prevalence of x_{ij} across clusters.

These observations are summarised in Table 4.15.

Table 4.15. Summary of simulation results when the outcome variable is continuous, the explanatory variable is binary, and observations are grouped in 100 clusters

$\beta_1^{RI} - \beta_1^{OLS}$: on average ≈ 0 for any ICC and relative dispersion of x_{ij}

Range of $\beta_1^{RI} - \beta_1^{OLS}$: ↑ when ICC ↑

↑ when relative dispersion of x_{ij} ↑

$SE^{RI} < SE^{OLS}$ in all cases SE^{RI}/SE^{OLS} ↓ when ICC ↑ ↑ slightly when relative dispersion of x_{ij} ↑
Coverage by 95% CIs from the RI model: on average 95% Coverage by 95% CIs from the OLS model: on average 95% when ICC≤0.03 Coverage by 95% CIs from the OLS model varied from <95% to >95% for ICC=0.1 or 0.3 ↓ when dispersion of prevalence of x_{ij} across clusters ↑ ↓ when overall prevalence of x_{ij} ↓
Type I error rates from the RI model: on average 5% Type I error rates from the OLS model: on average 5% when ICC≤0.03 Type I error rates from the OLS model varied from <5% to values >5% for ICC=0.1 or 0.3 ↑ when dispersion of prevalence of x_{ij} across clusters ↑ ↑ when overall prevalence of x_{ij} ↓

4.7 Algebraic approach

The consequences of ignoring clustering in linear regression were investigated in the previous sections of this chapter using simulated data. In this section, the problem is approached algebraically using the same setting of varying parameters as before. Initially the general formulae for difference in regression coefficients and ratio of SEs are derived, and then the special case of two clusters is considered as an illustration.

In a simple linear regression which models the relationship between a dependent variable y and an explanatory variable x by a line of the form $y = a + \beta x$, the best estimate of β based on a sample of n data points (x_i, y_i) is given by

$$\hat{\beta} = \frac{E(x_i - \bar{x})(y_i - \bar{y})}{E(x_i - \bar{x})^2}$$

with an estimated variance of

$$Var(\hat{\beta}) = \frac{E(y_i - \bar{y})^2 - \hat{\beta}^2 E(x_i - \bar{x})^2}{(n-2)E(x_i - \bar{x})^2}$$

Consider a situation of N different clusters, each of size n, with observations denoted as (x_{ij}, y_{ij}), where i is the index for the observation in the cluster, and j is the index of the cluster ($i = 1,2,..,n$, and $j = 1,2,\ldots,N$) . Within each cluster j, the variance of observations x_{ij}, will be calculated as

$$E\left(x_{ij}-\bar{x}_j\right)^2=\sigma_{x_j}^2$$

while between the j clusters, the variance will be

$$E\left(\bar{x}_j-\bar{x}\right)^2=\sigma_{\bar{x}_j}^2$$

Following the specification of the RI model, each of the different clusters has a specific intercept u_j, added to the overall intercept a. The variance of the intercepts will be

$$Var\left(u_j\right)=E\left(u_j-\bar{u}\right)^2=\sigma_u^2$$

Then, the cluster-specific point estimates of effect, $\hat{\beta}_j$, and their variances can be calculated from the formulae above as

$$\hat{\beta}_j=\frac{E\left(x_{ij}-\bar{x}_j\right)\left(y_{ij}-\bar{y}_j\right)}{E\left(x_{ij}-\bar{x}_j\right)^2}$$

and

$$Var\left(\hat{\beta}_j\right)=\frac{E\left(y_{ij}-\bar{y}_j\right)^2-\hat{\beta}_j{}^2E\left(x_{ij}-\bar{x}_j\right)^2}{(n-2)E\left(x_{ij}-\bar{x}_j\right)^2}$$

The overall adjusted estimate of β across all N clusters will be denoted as $\hat{\beta}_a$, which is calculated as

$$\hat{\beta}_a=\frac{\sum_{j=1}^{N}\left(\hat{\beta}_j/Var\left(\hat{\beta}_j\right)\right)}{\sum_{j=1}^{N}\left(1/Var\left(\hat{\beta}_j\right)\right)}$$

and has variance

$$Var\left(\hat{\beta}_a\right)=\frac{1}{\sum_{j=1}^{N}\left(1/Var\left(\hat{\beta}_j\right)\right)}$$

For a given cluster, j, the estimate $\hat{\beta}_j$ will be

$$\hat{\beta}_j=\frac{E\left(x_{ij}-\bar{x}_j\right)\left(y_{ij}-\bar{y}_j\right)}{E\left(x_{ij}-\bar{x}_j\right)^2}=\frac{E\left(x_{ij}-\bar{x}_j\right)\left(a+\beta x_{ij}+u_j+\varepsilon_{ij}-a-\beta\bar{x}_j-u_j\right)}{E\left(x_{ij}-\bar{x}_j\right)^2}$$

$$=\frac{E\left(x_{ij}-\bar{x}_j\right)\left(\beta\left(x_{ij}-\bar{x}_j\right)+\varepsilon_{ij}\right)}{E\left(x_{ij}-\bar{x}_j\right)^2}=\frac{\beta E\left(x_{ij}-\bar{x}_j\right)^2}{E\left(x_{ij}-\bar{x}_j\right)^2}=\beta$$

So, the cluster-specific $\hat{\beta}_j$s are unbiased.

Within each cluster, the variance of $\hat{\beta}_j$ will be

$$Var(\hat{\beta}_j\ ) = \frac{E\left(y_{ij}-\bar{y}_j\right)^2 - \hat{\beta}_j{}^2 E\left(x_{ij}-\bar{x}_j\right)^2}{(n-2)E\left(x_{ij}-\bar{x}_j\right)^2}$$

$$= \frac{E\left(a+\beta x_{ij}+u_j+\varepsilon_{ij}-a-\beta\bar{x}_j-u_j\right)^2 - \beta^2 E\left(x_{ij}-\bar{x}_j\right)^2}{(n-2)E\left(x_{ij}-\bar{x}_j\right)^2}$$

$$= \frac{E\left(\beta\left(x_{ij}-\bar{x}_j\right)+\varepsilon_{ij}\right)^2 - \beta^2 E\left(x_{ij}-\bar{x}_j\right)^2}{(n-2)E\left(x_{ij}-\bar{x}_j\right)^2}$$

$$= \frac{E\left(\beta^2\left(x_{ij}-\bar{x}_j\right)^2\right) + 2\beta E\left(\left(x_{ij}-\bar{x}_j\right)\varepsilon_{ij}\right) + E\left(\varepsilon_{ij}{}^2\right) - \beta^2 E\left(x_{ij}-\bar{x}_j\right)^2}{(n-2)E\left(x_{ij}-\bar{x}_j\right)^2}$$

$$= \frac{\sigma_\varepsilon^2}{(n-2)E\left(x_{ij}-\bar{x}_j\right)^2} = \frac{\sigma_\varepsilon^2}{(n-2)\sigma_{x_j}^2}$$

since x_{ij} and ε_{ij} are independent, and $E\left(\varepsilon_{ij}\right) = 0$.

The overall $\hat{\beta}$ across all clusters will therefore be

$$\hat{\beta}_a = \frac{\sum_{j=1}^{N}\left(\hat{\beta}_j/Var\left(\hat{\beta}_j\right)\right)}{\sum_{j=1}^{N}\left(1/Var\left(\hat{\beta}_j\right)\right)} = \frac{\sum_{j=1}^{N}\left(\beta\Big/\left(\frac{\sigma_\varepsilon^2}{(n-2)\sigma_{x_j}^2}\right)\right)}{\sum_{j=1}^{N}\left(1\Big/\left(\frac{\sigma_\varepsilon^2}{(n-2)\sigma_{x_j}^2}\right)\right)} = \beta\frac{\sum_{j=1}^{N}\left(1\Big/\left(\frac{\sigma_\varepsilon^2}{(n-2)\sigma_{x_j}^2}\right)\right)}{\sum_{j=1}^{N}\left(1\Big/\left(\frac{\sigma_\varepsilon^2}{(n-2)\sigma_{x_j}^2}\right)\right)} = \beta$$

with an overall variance of

$$Var\left(\hat{\beta}_a\right) = \frac{1}{\sum_{j=1}^{N}\left(1/Var\left(\hat{\beta}_j\right)\right)} = \frac{1}{\sum_{j=1}^{N}\left(1\Big/\left(\frac{\sigma_\varepsilon^2}{(n-2)\sigma_{x_j}^2}\right)\right)} = \frac{\sigma_\varepsilon^2}{N(n-2)E(\sigma_{x_j}^2)}$$

If clustering is ignored, the cluster-unadjusted point estimate of β, $\hat{\beta}_u$ will be calculated as

$$\hat{\beta}_u = \frac{E\left(x_{ij}-\bar{x}\right)\left(y_{ij}-\bar{y}\right)}{E\left(x_{ij}-\bar{x}\right)^2} = \frac{E\left(x_{ij}-\bar{x}\right)\left(a+\beta x_{ij}+u_j+\varepsilon_{ij}-a-\beta\bar{x}-\bar{u}\right)}{E\left(x_{ij}-\bar{x}\right)^2}$$

$$= \frac{E\left[\left(x_{ij}-\bar{x}\right)\left(\beta\left(x_{ij}-\bar{x}\right)+\left(u_j-\bar{u}\right)+\varepsilon_{ij}\right)\right]}{E\left(x_{ij}-\bar{x}\right)^2}$$

$$= \frac{\beta E\left(x_{ij} - \bar{x}\right)^2 + E\left(x_{ij} - \bar{x}\right)\left(u_j - \bar{u}\right) + E\left(x_{ij} - \bar{x}\right)\varepsilon_{ij}}{E\left(x_{ij} - \bar{x}\right)^2} = \beta + \frac{E\left(x_{ij} - \bar{x}\right)\left(u_j - \bar{u}\right)}{E\left(x_{ij} - \bar{x}\right)^2}$$

since x_{ij} and ε_{ij} are independent, and $E\left(\varepsilon_{ij}\right) = 0$.

However,

$$E\left(x_{ij} - \bar{x}\right)\left(u_j - \bar{u}\right) = E\left(x_{ij}u_j\right) - E\left(x_{ij}\bar{u}\right) - E\left(\bar{x}u_j\right) + E(\bar{x}\bar{u})$$

$$= E\left(\bar{x}_j u_j\right) - \bar{x}\bar{u} = Cov\left(\bar{x}_j, u_j\right)$$

as $E\left(x_{ij}u_j\right) = E\left(\bar{x}_j u_j\right)$.

Also,

$$E\left(x_{ij} - \bar{x}\right)^2 = E\left(x_{ij} - \bar{x}_j + \bar{x}_j - \bar{x}\right)^2 = E\left(\left(x_{ij} - \bar{x}_j\right) + \left(\bar{x}_j - \bar{x}\right)\right)^2$$

$$= E\left(x_{ij} - \bar{x}_j\right)^2 + E\left(\bar{x}_j - \bar{x}\right)^2 + 2E\left(x_{ij} - \bar{x}_j\right)\left(\bar{x}_j - \bar{x}\right) = E(\sigma_{x_j}^2) + \sigma_{\bar{x}_j}^2$$

as $E\left(x_{ij} - \bar{x}_j\right)\left(\bar{x}_j - \bar{x}\right) = Cov\left(x_{ij}, \bar{x}_j\right) = 0$

Thus, the estimate $\hat{\beta}_u$ takes the form

$$\hat{\beta}_u = \beta + \frac{E\left(x_{ij} - \bar{x}\right)\left(u_j - \bar{u}\right)}{E\left(x_{ij} - \bar{x}\right)^2} = \beta + \frac{Cov\left(\bar{x}_j, u_j\right)}{E(\sigma_{x_j}^2) + \sigma_{\bar{x}_j}^2}$$

So, bias in the point estimate will be

$$\boxed{bias = \hat{\beta}_a - \hat{\beta}_u = \beta - \beta - \frac{Cov\left(\bar{x}_j, u_j\right)}{E(\sigma_{x_j}^2) + \sigma_{\bar{x}_j}^2} = -\frac{Cov\left(\bar{x}_j, u_j\right)}{E(\sigma_{x_j}^2) + \sigma_{\bar{x}_j}^2}}$$

The variance of the cluster-unadjusted point estimate of β, $\hat{\beta}_u$, will be calculated as

$$Var\left(\hat{\beta}_u\right) = \frac{E\left(y_{ij} - \bar{y}\right)^2 - \hat{\beta}_u{}^2 E\left(x_{ij} - \bar{x}\right)^2}{(nN - 2)E\left(x_{ij} - \bar{x}\right)^2}$$

$$= \frac{E\left(a + \beta x_{ij} + u_j + \varepsilon_{ij} - a - \beta\bar{x} - \bar{u}\right)^2 - \hat{\beta}_u{}^2 E\left(x_{ij} - \bar{x}\right)^2}{(nN - 2)E\left(x_{ij} - \bar{x}\right)^2}$$

$$= \frac{E\left(\beta\left(x_{ij}-\bar{x}\right)+\left(u_j-\bar{u}+\varepsilon_{ij}\right)\right)^2-\hat{\beta}_u{}^2E\left(x_{ij}-\bar{x}\right)^2}{(nN-2)E\left(x_{ij}-\bar{x}\right)^2}$$

$$= \frac{\beta^2E\left(x_{ij}-\bar{x}\right)^2+E\left(u_j-\bar{u}+\varepsilon_{ij}\right)^2+2\beta E\left(\left(x_{ij}-\bar{x}\right)\left(u_j-\bar{u}+\varepsilon_{ij}\right)\right)-\hat{\beta}_u{}^2E\left(x_{ij}-\bar{x}\right)^2}{(nN-2)E\left(x_{ij}-\bar{x}\right)^2}$$

$$= \frac{\begin{gathered}(\beta^2E\left(x_{ij}-\bar{x}\right)^2+E\left(u_j-\bar{u}\right)^2+2E\left(\left(u_j-\bar{u}\right)\varepsilon_{ij}\right)+E\left(\varepsilon_{ij}\right)^2\\ +2\beta E\left(\left(x_{ij}-\bar{x}\right)\left(u_j-\bar{u}\right)\right)+2\beta E\left(\left(x_{ij}-\bar{x}\right)\varepsilon_{ij}\right)-\hat{\beta}_u{}^2E\left(x_{ij}-\bar{x}\right)^2)\end{gathered}}{(nN-2)E\left(x_{ij}-\bar{x}\right)^2}$$

$$= \frac{\left(\beta^2-\hat{\beta}_u{}^2\right)E\left(x_{ij}-\bar{x}\right)^2+E\left(u_j-\bar{u}\right)^2+E\left(\varepsilon_{ij}\right)^2+2\beta E\left(\left(x_{ij}-\bar{x}\right)\left(u_j-\bar{u}\right)\right)}{(nN-2)E\left(x_{ij}-\bar{x}\right)^2}$$

$$= \frac{\beta^2-\hat{\beta}_u{}^2}{(nN-2)}+\frac{\sigma_u^2+\sigma_\varepsilon^2+2\beta Cov\left(\bar{x}_j,u_j\right)}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)}$$

$$= \frac{\beta^2-\left(\beta+\dfrac{Cov\left(\bar{x}_j,u_j\right)}{E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2}\right)^2}{(nN-2)}+\frac{\sigma_u^2+\sigma_\varepsilon^2+2\beta Cov\left(\bar{x}_j,u_j\right)}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)}$$

$$= \frac{\beta^2-\beta^2-\left(\dfrac{Cov\left(\bar{x}_j,u_j\right)}{E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2}\right)^2-2\beta\dfrac{Cov\left(\bar{x}_j,u_j\right)}{E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2}}{(nN-2)}+\frac{\sigma_u^2+\sigma_\varepsilon^2+2\beta Cov\left(\bar{x}_j,u_j\right)}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)}$$

$$= \frac{-Cov\left(\bar{x}_j,u_j\right)^2}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)^2}-\frac{2\beta Cov\left(\bar{x}_j,u_j\right)}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)}+\frac{\sigma_u^2+\sigma_\varepsilon^2+2\beta Cov\left(\bar{x}_j,u_j\right)}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)}$$

$$= \frac{(\sigma_u^2+\sigma_\varepsilon^2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)-Cov\left(\bar{x}_j,u_j\right)^2}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)^2}$$

Thus, the ratio of the SEs of the cluster-adjusted to that of the cluster-unadjusted point estimates will be

$$Ratio_{SEs} = \frac{SE\left(\hat{\beta}_a\right)}{SE\left(\hat{\beta}_u\right)} = \frac{\sqrt{\dfrac{\sigma_\varepsilon^2}{N(n-2)E(\sigma_{x_j}^2)}}}{\sqrt{\dfrac{(\sigma_u^2+\sigma_\varepsilon^2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)-Cov\left(\bar{x}_j,u_j\right)^2}{(nN-2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)^2}}}$$

$$\boxed{Ratio_{SEs} = \sqrt{\frac{(nN-2)\sigma_\varepsilon^2\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)^2}{N(n-2)E(\sigma_{x_j}^2)\left((\sigma_u^2+\sigma_\varepsilon^2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)-Cov\left(\bar{x}_j,u_j\right)^2\right)}}}$$

As an illustration, the above formulae for bias in the effect estimate and for the ratios of SEs can then be applied to the simplified case of two clusters, with equal numbers of observations per cluster described in sections 4.2 for continuous and 4.4 for binary explanatory variable.

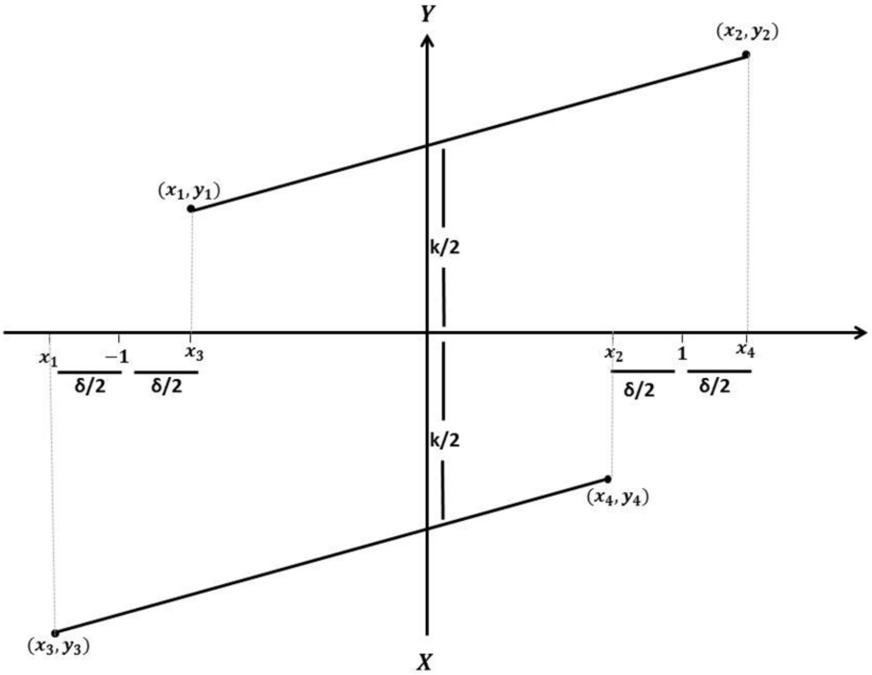


Figure 4.18. Case of two clusters of equal size with continuous explanatory variable x_{ij}, in which the first cluster, comprising observations (x_1, y_1) and (x_2, y_2), is shifted to the right of its original distribution and the second cluster, comprising observations (x_3, y_3) and (x_4, y_4), is shifted to the left of its original distribution by a value $\delta/2$.

I first consider the case of a continuous explanatory variable x_{ij}, in which the cluster-specific distributions are overlaid ($\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2} = 0$), and their variance is equal to 1 ($\sigma_{x_j}^2 = 1$, for $j = 1,2$). Then the distribution of x_{ij} in the first cluster is shifted to the right of its original distribution by a value $\delta/2$ and the distribution of x_{ij} in the second cluster is shifted to the left of its original distribution by a value $\delta/2$, resulting in a difference between the cluster-specific mean values of x_{ij} of δ (Figure 4.18). As such, the variance of x_{ij} between the two clusters is

$$\sigma^2_{\bar{x}_j} = \frac{1}{2}\sum_j \left(\bar{x}_j - \bar{x}\right)^2 = \frac{1}{2}((\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2) = \frac{1}{2}((\delta/2)^2 + (-\delta/2)^2) = \delta^2/4$$

while that within clusters remains 1 ($\sigma^2_{x_j} = 1$). The cluster-specific intercepts are centred around 0 and with a distance k between them (Figure 4.18), with the cluster-specific intercepts being $u_1 = k/2$ and $u_2 = -k/2$ for the first and the second cluster, respectively. The variance of the two intercepts will then be

$$\sigma^2_u = \frac{1}{2}\sum_j \left(u_j - \bar{u}\right)^2 = \frac{1}{2}((u_1 - \bar{u})^2 + (u_2 - \bar{u})^2)$$

$$= \frac{1}{2}((u_1)^2 + (u_2)^2) = \frac{1}{2}((k/2)^2 + (-k/2)^2) = k^2/4$$

The covariance between $\bar{x}_j$ and u_j is calculated as

$$Cov\left(\bar{x}_j, u_j\right) = \frac{1}{2}\sum_j \left(\bar{x}_j - \bar{x}\right)\left(u_j - \bar{u}\right) = \frac{1}{2}\sum_j \left(\bar{x}_j u_j\right) - \bar{x}\bar{u} = \frac{1}{2}\sum_j \left(\bar{x}_j u_j\right)$$

$$= \frac{(k/2)(\delta/2) + (-k/2)(-\delta/2)}{2} = \frac{k\delta}{4}$$

Then, the bias in the effect estimate, when clustering is not accounted for in the regression model is calculated as

$$bias = -\frac{Cov\left(\bar{x}_j, u_j\right)}{E(\sigma^2_{x_j}) + \sigma^2_{\bar{x}_j}} = -\frac{k\delta}{4\left(1 + \frac{\delta^2}{4}\right)} = -\frac{k\delta}{4 + \delta^2}$$

The ratios of SEs is calculated as

$$Ratio_{SEs} = \frac{SE\left(\hat{\beta}_a\right)}{SE\left(\hat{\beta}_u\right)} = \sqrt{\frac{(nN-2)\sigma^2_\varepsilon\left(E(\sigma^2_{x_j}) + \sigma^2_{\bar{x}_j}\right)^2}{N(n-2)E(\sigma^2_{x_j})\left((\sigma^2_u + \sigma^2_\varepsilon)\left(E(\sigma^2_{x_j}) + \sigma^2_{\bar{x}_j}\right) - Cov\left(\bar{x}_j, u_j\right)^2\right)}}$$

$$\cong \sqrt{\frac{(1 + \delta^2/4)^2}{(k^2/4 + 1)(1 + \delta^2/4) - (k\delta/4)^2}} = \sqrt{\frac{(1 + \delta^2/4)^2}{k^2/4 + k^2\delta^2/16 + 1 + \delta^2/4 - (k\delta/4)^2}}$$

$$= \sqrt{\frac{(1+\delta^2/4)^2}{k^2/4+1+\delta^2/4}}$$

Bias of effect estimate was plotted against varying values of distance between the cluster-specific mean values of x_{ij} (i.e. shift of mean value of x_{ij} in the second cluster from that in the first cluster), and for the three values of distance between the cluster-specific intercepts k that were examined in section 4.2, in Figure 4.19.

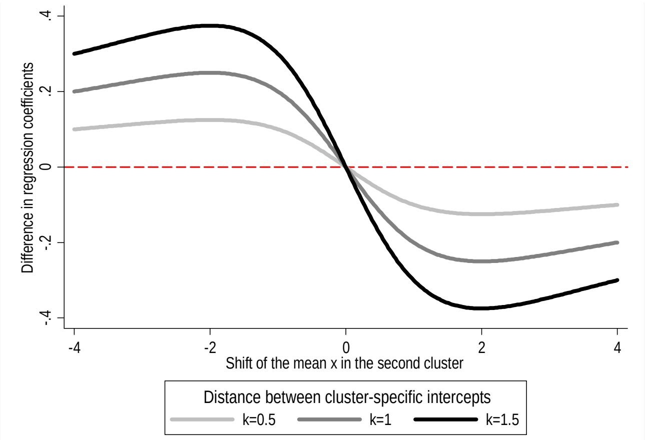


Figure 4.19. Difference between cluster-adjusted and cluster-unadjusted regression coefficients against shift in the mean value of x_{ij} for the second cluster from the mean value of x_{ij} for the first cluster

Like bias of effect estimates, ratios of SEs were plotted against values of distance between the cluster-specific mean values of x_{ij}, and for the three values of distance between the cluster-specific intercepts k that were examined in section 4.2, and presented in Figure 4.20.

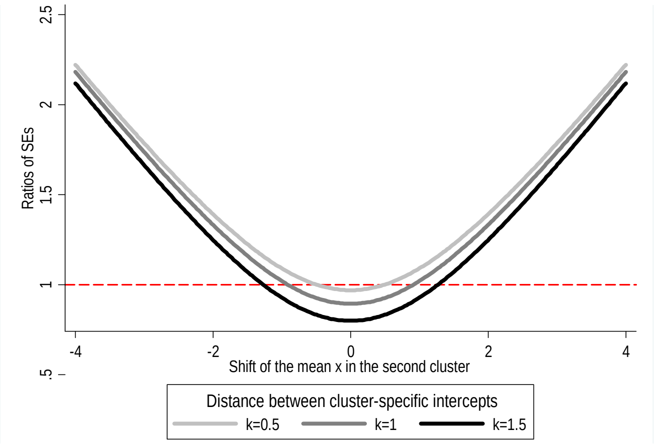


Figure 4.20. Ratios of SEs of cluster-adjusted and cluster-unadjusted regression coefficients against shift in the mean value of x_{ij} for the second cluster from the mean value of x_{ij} for the first cluster

In the case of two clusters, when the explanatory variable x_{ij} was binary, the prevalence of x_{ij} in the first cluster was set to p, while that in the second cluster was set to q. As in the case of two clusters with a continuous explanatory variable, the distance between the cluster-specific intercepts was equal to k (Figure 4.21).

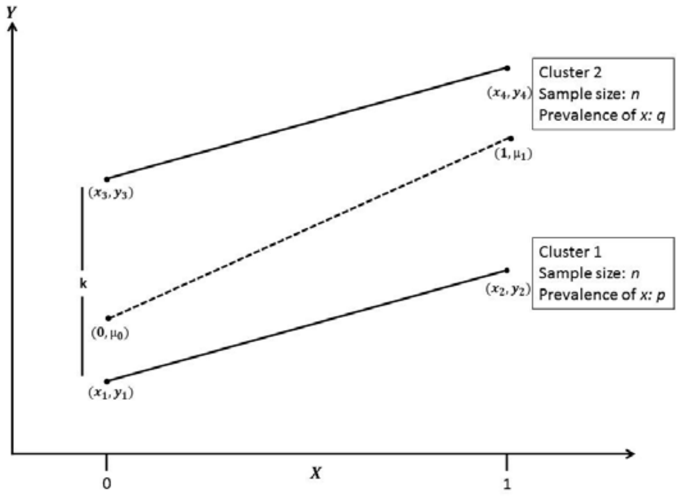


Figure 4.21. Case of two clusters of equal size with binary explanatory variable x_{ij}, the prevalence of which in the first cluster is p, and and in the second cluster is q.

In the case described above, the average value of all x_{ij} is $\bar{x} = \frac{p+q}{2}$, and the variance of x_{ij} between clusters is

$$\sigma^2_{\bar{x}_j} = \frac{1}{2}\sum_j \left(\bar{x}_j - \bar{x}\right)^2 = \frac{1}{2}\left(\left(p - \frac{p+q}{2}\right)^2 + \left(q - \frac{p+q}{2}\right)^2\right)$$

$$= \frac{1}{2}\left(p^2 - p(p+q) + \frac{(p+q)^2}{4} + q^2 - q(p+q) + \frac{(p+q)^2}{4}\right)$$

$$= \frac{1}{2}\left(-2pq + \frac{(p+q)^2}{2}\right) = \frac{1}{4}(p-q)^2$$

The mean of the variance within clusters is

$$E(\sigma^2_{x_j}) = \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{2}\left(x_{ij} - \bar{x}_j\right)^2 = \frac{1}{2n}n\left(p(1-\bar{x}_1)^2 + (1-p)\bar{x}_1{}^2 + q(1-\bar{x}_2)^2 + (1-q)\bar{x}_2{}^2\right)$$

$$= \frac{1}{2}(p(1-p)^2 + (1-p)p^2 + q(1-q)^2 + (1-q)q^2)$$

$$= \frac{1}{2}\big((1-p)(p-p^2+p^2)+(1-q)(q-q^2+q^2)\big) = \frac{1}{2}\big(p(1-p)+q(1-q)\big)$$

Then

$$E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2 = \frac{1}{2}\big(p(1-p)+q(1-q)\big)+\frac{1}{4}(p-q)^2$$

$$= \frac{1}{4}(p^2+q^2-2pq+2p-2p^2+2q-2q^2) = \frac{1}{4}(p+q)(2-p-q)$$

The covariance between $\bar{x}_j$ and u_j is calculated as

$$Cov(\bar{x}_j,u_j) = \frac{1}{2}\sum_j(\bar{x}_j-\bar{x})(u_j-\bar{u}) = \frac{1}{2}\sum_j(\bar{x}_j u_j)-\bar{x}\bar{u} = \frac{1}{2}\sum_j(\bar{x}_j u_j)$$

$$= \frac{(-k/2)(p)+(k/2)(q)}{2} = \frac{(k/2)(q-p)}{2} = \frac{(q-p)k}{4}$$

The variance of the two intercepts is equal to the one calculated above for the case of continuous x_{ij} (i.e. $\sigma_u^2 = k^2/4$).

Bias in the effect estimate when clustering is ignored, and the ratio of SE of the cluster-adjusted to that of the cluster-unadjusted estimate are calculated as follows:

$$bias = -\frac{Cov(\bar{x}_j,u_j)}{E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2} = -\frac{\frac{(q-p)k}{4}}{\frac{1}{4}(p+q)(2-p-q)} = \frac{(q-p)k}{(p+q)(2-p-q)}$$

$$Ratio_{SEs} = \frac{SE(\hat{\beta}_a)}{SE(\hat{\beta}_u)} = \sqrt{\frac{(nN-2)\sigma_\varepsilon^2\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)^2}{N(n-2)E(\sigma_{x_j}^2)\left((\sigma_u^2+\sigma_\varepsilon^2)\left(E(\sigma_{x_j}^2)+\sigma_{\bar{x}_j}^2\right)-Cov(\bar{x}_j,u_j)^2\right)}}$$

$$\cong \sqrt{\frac{\left(\frac{1}{4}(p+q)(2-p-q)\right)^2}{\left(\frac{1}{2}\big(p(1-p)+q(1-q)\big)\right)\left(\left(\frac{k^2}{4}+1\right)\left(\frac{1}{4}(p+q)(2-p-q)\right)-\left(\frac{(q-p)k}{4}\right)^2\right)}}$$

Assuming a constant prevalence of x_{ij} in the first cluster of $p = 0.2$, bias in estimated effects against varying prevalence of x_{ij} in the second cluster, was plotted in Figure 4.22

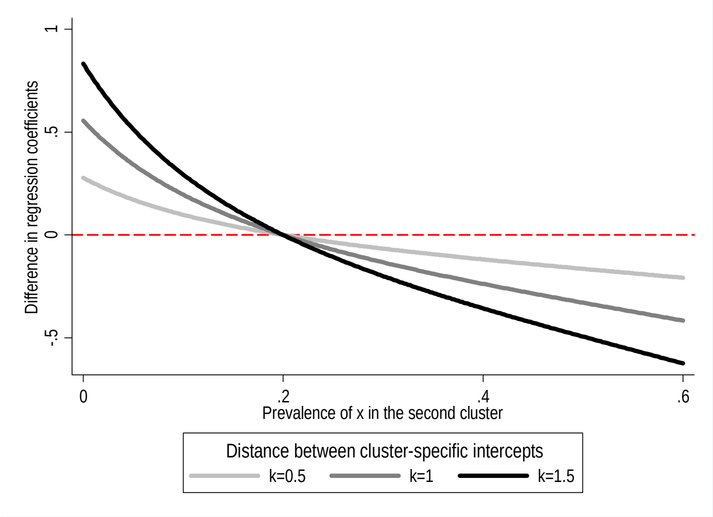


Figure 4.22. Difference between cluster-adjusted and cluster-unadjusted regression coefficients against prevalence of x_{ij} in the second cluster when prevalence of x_{ij} in the first cluster is fixed to $p = 0.2$

The values of ratios of SEs were then plotted against varying prevalence of x_{ij} in the second cluster when prevalence of x_{ij} in the first cluster was fixed to $p = 0.2$ (Figure 4.23).

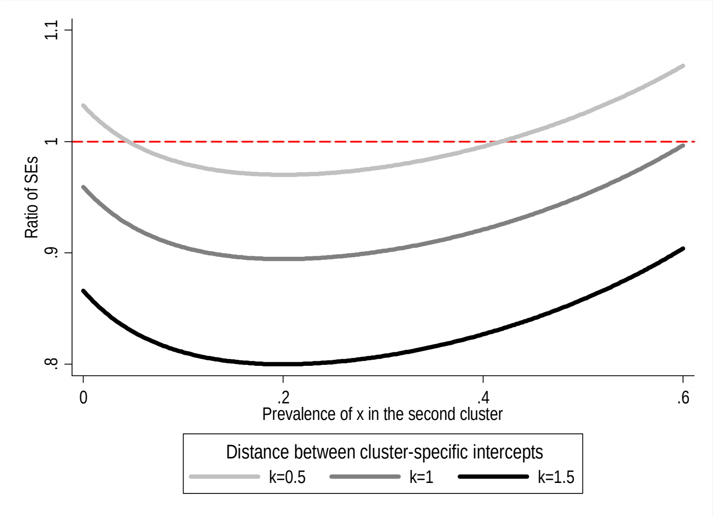


Figure 4.23. Ratios of SEs of cluster-adjusted and cluster-unadjusted regression coefficients against prevalence of x_{ij} in the second cluster when prevalence of x_{ij} in the first cluster is fixed to $p = 0.2$

Observations regarding differences in regression coefficients and ratios of SEs for the case of two clusters, as calculated algebraically in this section, match those from the simulated data described in previous sections (4.2 and 4.4). Specifically, equations plotted in figures Figure 4.19, Figure

4.20, Figure 4.22, and Figure 4.23 are a replicate of those presented in figures Figure 4.3Figure 4.4Figure 4.9Figure 4.10, respectively.

These findings support the validity of my earlier findings based on simulations. They confirm that (under the assumption that clusters are of equal size and that cluster-specific intercepts and individual error terms are independent) the bias in estimates of effect when clustering is ignored in linear regression depends on the variance of the explanatory variable within and between clusters and the covariance of cluster-specific intercepts with the mean values of the explanatory variable in each cluster, while bias in the precision of such estimates depends also on the variance of cluster-specific intercepts, and of the random error in the outcome variable after adjustment for cluster and the explanatory variable. However, bias is not contingent on other assumptions that were made in the simulations.

4.8 Discussion

In this chapter, attention was restricted to the implications of ignoring clustering in statistical inference regarding the relationship between a continuous outcome and a single explanatory variable. Two different types of explanatory variable were considered – continuous and binary. For each of the two categories of explanatory variable, the implications for statistical inference of failing to account for clustering were explored by comparison of estimated effects and related precision, assessment of the coverage by 95% confidence intervals, and estimation of the frequency of type I error. As the main interest of this thesis more generally, and of the current chapter more specifically, is the grouping of observations within multiple clusters, the discussion will be developed around results from the analysis of multiple clusters. The findings when observations were grouped into two clusters (sections 4.2 and 4.4) served merely as an introduction to the more general case of multiple clusters and will not be discussed further.

In the cases both of a continuous and a binary explanatory variable, where the true slope of the regression line was non-zero, I found that the cluster-unadjusted OLS and RI models gave on average very similar estimates of effect for any level of ICC. However, despite the average value of difference in point estimates from the two methods being zero, differences occurred in both directions and varied more when the level of ICC increased. The largest differences in estimates of effect between OLS and multi-level RI regression modelling were only about 20% of the true value and they occurred where the ICC was high (0.3). For a continuous explanatory variable, the largest errors in the differences of estimated effects occurred when the dispersion of the explanatory variable within clusters was approximately the same as that between clusters (relative dispersion approximately equal to 1), while, for a binary explanatory variable, differences increased with increasing dispersion of prevalence of the explanatory variable across clusters.

Conclusions drawn from comparison of SEs estimated from cluster-unadjusted OLS and RI models are different for continuous as compared with binary explanatory variables. When the explanatory variable of interest was continuous, the SEs of regression coefficients were generally larger for the multi-level RI model than for the cluster-unadjusted OLS model, their ratio being highest (>4) for a high ICC (0.3) and where most of the dispersion of the explanatory variable was between rather than within clusters. However, contrary to what is widely stated, the spuriously greater precision of OLS method was not universal. Where most of the dispersion of the continuous explanatory variable was within rather than between clusters, OLS regression gave larger SEs than multi-level modelling. When the explanatory variable was binary, SEs estimated from the RI model, were generally lower than those from the cluster-unadjusted OLS model. The SEs differed by up to 15% for the highest ICC value (ICC=0.3).

The rates of coverage of 95% confidence intervals for estimates of effect, whether of a continuous or a binary explanatory variable, when derived from a RI model were at the nominal level of 95%, irrespective of other parameters (i.e. ICC, relative dispersion of the continuous explanatory variable, or dispersion of the prevalence of the binary explanatory variable across clusters). When the explanatory variable was binary, the cluster-unadjusted OLS model also resulted in a good coverage of the 95% confidence intervals when ICC was low (≤ 0.01). However, for higher values of ICC, coverage varied slightly (range: 87% - 98%) around the nominal value of 95% depending on the overall prevalence and the dispersion of the cluster-specific prevalence rates of x_{ij}. In contrast, when the explanatory variable was continuous, the model that failed to account for clustering resulted in poor coverage rates, especially as ICC increased, reaching a rate as low as 30% for ICC=0.3.

Setting the effect of the explanatory variable on the outcome variable to zero allowed exploration of the frequency of type I error. With the RI model, in all of the scenarios explored, type I error was very close to 5%. When the explanatory variable of interest was continuous, I found that failure to allow for clustering increased rates of Type I error, and that the inflation of type I error was particularly pronounced (up to 70%) when the degree of clustering, was high (ICC=0.3). In contrast to this, when the explanatory variable was binary, type I error under the naïve OLS model was close to the expected value of 5% for low levels of clustering (ICC<0.1). However, when ICC was high (0.1 or 0.3), I found that type I error rates varied more around 5%, with values as low as 2% (when overall prevalence of the explanatory variable was low and the dispersion of its prevalence across clusters was small) and as high as 12% (when overall prevalence of the explanatory variable was low and the dispersion of its prevalence across clusters was large).

The analysis for each specification of parameters (expected ICC, relative dispersion of a continuous explanatory variable, overall prevalence or dispersion of prevalence rates across clusters of a binary explanatory variable) was based on 1,000 simulated samples of 10,000 observations grouped in 100 clusters, each of 100 individuals. By using such a large sample size (larger than in most epidemiological investigations), I reduced random sampling variation, making it easier to characterise any systematic differences between the two methods of analysis. However, the approach may have led to underestimation of the maximum differences between estimates of effect that could arise from OLS as compared with multi-level modelling. Additionally, the number of observations per cluster was the same in all simulations, making it impossible to draw conclusions about effects of ignoring clustering for varying cluster sizes. It should also be noted that all samples were generated by sampling from normal distributions, with data structured according the assumptions of RI multi-level modelling – i.e. that the cluster level residuals (u_j in equation 4.3) were normally distributed and uncorrelated, individual level residuals (e_{ij} in equation 4.3) were uncorrelated, and cluster level and individual level residuals were uncorrelated. Given that these assumptions were met, RI methods were expected to function as intended. However, previous research on violation of assumptions in multi-level modelling has shown that assuming a non-normal distribution of the cluster level error term (u_j), has no or very little impact on the estimation of parameters (90). In support of that, based on the formulae for bias and relative precision when clustering is ignored, I showed that it is unlikely that results would have been different if another distribution of the cluster level error term was assumed. Algebraic calculations for derivations of these formulae however, did assume independence of cluster level and individual level residuals. For that, if the assumption regarding correlation of the u_j's and e_{ij}'s residuals were ignored, the multi-level model would be incorrectly specified and comparisons with the cluster-unadjusted OLS modelling approach might differ from those that were seen. Also, data were simulated following the specification of the RI regression model rather than that of the RE model described in section 2.3. That was done because the RI model is more frequently used, especially when there is no a priori expectation of differential effects of the explanatory on the outcome variables across the different clusters. Simulating data following the specification of the RE model would have added complexity to the algorithm used for simulation, and the (already long) computational time required. In addition, increasing values of the relative between- to within-cluster dispersion were generated by increasing the between-cluster dispersion of the cluster-specific mean values of the explanatory variable, keeping the within-cluster dispersion constant. One could also achieve increasing values of relative dispersion by altering the within-cluster dispersion of the explanatory variable. However, there is no reason to expect results to differ if the latter approach were chosen, especially given the findings from algebraic calculations.

The effect of clustering when a cluster-unadjusted model is fitted could also have been assessed by calculating bias as [(estimated effect – true effect)/true effect], as defined in earlier studies

(85). Instead, I defined bias by the difference in the effect estimates derived from the two analytical models. The data were simulated following the model specification of RI linear regression, which is one of the most well established and frequently chosen analytical approaches to account for clustering. As such, given that all resulting effect estimates were positive, deviations of the difference in regression coefficients from the value of zero can only represent deficiencies of the OLS model, provided that the assumptions of the RI model are met. Therefore, there is no reason to expect that the conclusions one would draw from an alternative definition of bias would be more reliable, provided the conditions under which data were simulated and the models fitted were the same.

When multilevel RI modelling was applied to the simulated clustered datasets with a continuous or a binary explanatory variable, the rate of Type I error was 5%, and the coverage by 95% CIs was 95%, as would be expected, given the method by which the simulated samples were generated. In comparison, when cluster-unadjusted models were fitted to clustered data with a continuous explanatory variable, rates of Type I error were higher, particularly when the ICC was high. For the highest level of ICC examined (0.3), type I errors were as frequent as 70%. However, even with an ICC of only 0.01, rates of Type 1 error were more than 10%. Consistent with this, coverage by 95% confidence intervals was considerably lower than the nominal value for higher ICC levels. The lowest coverage of 29% was for the highest ICC level. In contrast to these results Huang et al (96) have reported values of coverage very close to 95% from the OLS model for a continuous explanatory variable. Differences between findings presented in this chapter and those presented by Huang et al (96) can be explained by zero clustering in the explanatory variable assumed in the latter. Sensitivity analysis restricting the simulated datasets only to those in which clustering in the explanatory variable was not meaningful showed that interval coverage rates were very close to 95% independent of clustering in the outcome variable (results not shown). When the explanatory variable was binary, both the interval coverage and Type I error rates varied little around the nominal values of 5% and 95%, and only for ICC values higher than 0.01. Overall coverage rates were higher for higher ICC and decreased for increasing dispersion of the cluster-specific prevalence rates of x_{ij} across clusters and for decreasing overall prevalence of the x_{ij}. A similar observation of small variation of interval coverage around 95% for higher ICC values has been made before (82). Type I error when the explanatory variable was binary and its prevalence was low, varied around 5% with values falling below 5% for small dispersion of prevalence of the explanatory variable, and above 5% for large dispersion. For larger overall prevalence of the explanatory variable, Type I error rates fell below 5%. In accordance with these findings, Galbraith et al (107) have shown that cluster-unadjusted models resulted in relatively conservative Type I error. Also, in a context of individually randomised trials, Kahan et al (83) have shown that Type I error increased with increasing ICC and increasing difference in the probability of assignment of patients to treatment arms in randomised controlled trials.

It has been widely stated that when data are clustered, effects estimated by OLS regression are unbiased but inefficient (44, 81-85). My results confirm that for data of the type that I simulated, coefficients from OLS regression were on average very similar to those from RI multi-level modelling. Previous studies based on simulation data have shown similar results (82, 85, 90, 96). However, for individual simulated samples, the estimates may differ, and the potential magnitude of the differences depends on the level of within-cluster similarity of the outcome variable. For an ICC of 0.3, the estimates of effect from the two analytical methods could differ by up to 20%. In addition, when the explanatory variable is continuous, the error in estimates of the regression coefficient is larger when the between-cluster dispersion of the explanatory variable is similar to that within-cluster. When the explanatory variable is binary, the error increases as the dispersion of the prevalence rates across clusters increases, and when the overall prevalence rate across all clusters is lower (<10%). These errors in the estimated effect indicate that in an individual study, failure of regression analysis to account for clustering of observations could result in considerably higher or lower estimates of effect than those derived from multilevel analysis. This has been illustrated in numerous published papers of real data, which have shown that estimates from the two analytical methods can differ to a lesser or greater extent (67, 91, 101, 102, 126). However, in those publications, no or very limited information is provided to establish whether the error observed is due to dispersion of the cluster-specific mean values of the continuous explanatory variables, or dispersion of prevalence rates for the binary explanatory variable across clusters.

Most often it is stated that regression coefficients are spuriously precise when clustering is not taken into account in regression models. However, in several reports, authors have failed to specify the conditions under which this applies (88-92, 131). Other authors have pointed out that when the explanatory variable is defined at the level of the cluster, and a cluster-unadjusted approach is followed, SEs tend to be spuriously low, and that the opposite occurs when the explanatory variable varies within clusters (83, 94, 96, 125). Bias in SEs for effects of cluster-varying explanatory variables has been shown in results from real data when both models were fitted (101, 126, 132). However, others have reported contradictory results in which SEs of effects of individual-level explanatory variables from OL regression were very similar to, or lower than, those from a multi-level model (102, 103, 105, 106). It should be noted that the dichotomy between cluster- and individual-level variables is not clear-cut. There can be varying degrees of clustering in the explanatory variable, with the extremes being variables for which the values are completely unclustered (zero differences between clusters), and variables for which the values are the same within each cluster. However, in real data, an explanatory variable can lie anywhere in between. An early report focused on this issue by considering the level of clustering in the explanatory variable as the main driver for the expected bias of the precision of the effect estimates (81), rather than the absolute distinction between cluster-constant and cluster-varying explanatory variables. The authors reported that as clustering in the explanatory variable decreases, the bias in SEs from a cluster-unadjusted model is expected to be upwards, and the opposite is expected when clustering in the explanatory variable increases. Taking into consideration clustering in the explanatory variable (ρ_x) as well as in the outcome variable (ρ_y), a later study using simulated data showed that for a given level of ρ_y, increasing ρ_x resulted in increasing the ratio of estimated SEs ($SE_\beta^{RI}/SE_\beta^{OLS}$) from values <1 to values ≈ 1 (93). My results for continuous explanatory variables agree with this, with ratios of SEs ($SE_\beta^{RI}/SE_\beta^{OLS}$) moving from values <1 to values >1, as clustering in the explanatory variable, expressed as relative dispersion, increased.

Bias in the precision of effect estimates for binary explanatory variables when clustering is ignored has received very limited attention in the published literature. Several of the reported studies have used real data to compare naïve and multi-level models, using both continuous and binary individual-level explanatory variables (102, 126). For the majority of binary explanatory variables used in the models fitted in these studies, SEs derived from the OLS model were larger than those derived from the multi-level model. The same conclusion was drawn from a study using simulated data (82). However, none of the studies using real data has explored the level of bias in relation to variation in the prevalence of the binary explanatory variable, and the study of simulated data assumed constant prevalence of the explanatory variable in all clusters. Simulation results presented in this chapter show that, irrespective of the dispersion of prevalence of the explanatory variable across clusters and the overall prevalence in all clusters, SEs from multi-level model are always lower than those from the OLS model, and the bias is higher for higher ICC values.

In this chapter, the problem of ignoring clustering was further explored algebraically, using the same setting with regard to parameters that remained fixed and those that varied. Following mathematical calculations, I derived a formula for i) bias in the point estimate when clustering was not accounted for in the analytical approach and ii) the ratio of the SEs of cluster-adjusted to cluster-unadjusted point estimates. Using the formulae derived, I applied the specifications of the simplified case in which observations were grouped within two clusters. Results from the algebraic approach perfectly matched those from the simulated data. This algebraic proof can thus be regarded as a validation of the findings presented in this chapter. Additionally, as general formulae were derived, one can directly calculate bias in the regression coefficient and corresponding precision when clustering is not accounted for in the regression model, using different conditions from those explored in this chapter. That can be done assuming different values of: i) within- and/or between-cluster variance of the explanatory variable, ii) covariance between the cluster-specific intercepts and mean values of x_{ij}, iii) variance of the error term, and iv) number and/or size of clusters. Lastly, another important aspect of the derived formulae for bias and relative precision is their confirmation that the assumption of the RI model about the normality of the cluster-level error term is not critical to its function.

In this chapter, the focus was on the association between a continuous outcome and an explanatory variable that was defined at the individual level (cluster-varying explanatory variable). I showed that when the explanatory variable was continuous, and most of the variation occurred within rather than between clusters, the cluster-unadjusted OLS model gave larger SEs for the regression coefficient than multi-level modelling. This is consistent with reports in which ignoring clustering resulted in spuriously high SEs when the explanatory variable varied within cluster. The reverse occurred when most of the dispersion of the explanatory variable was between rather than within clusters. In this situation the explanatory variable approaches the characteristics of a cluster specific variable. I further showed that there was a variation in the threshold value above which SEs estimated from OLS models that did not adjust for clustering, were higher than those from multi-level modelling. That threshold value depended on the ICC (Table 4.6). I additionally showed that when the explanatory variable under investigation was binary, ignoring clustering in statistical modelling resulted in higher SEs for the estimated effect than those derived from the random-intercept model. The SEs differed more for higher ICCs but not with the overall prevalence of the explanatory variable, nor with the dispersion of its prevalence across clusters (Figure 4.15). Unlike SEs, the point estimates were unbiased for either continuous or binary explanatory variables (Figure 4.5 and Figure 4.14).

In conclusion, my results support the use of multi-level modelling to account for clustering effects in linear regression analyses of data that are hierarchically structured, especially where ICCs might exceed 0.01. Failure to do so is likely to result in incorrect estimates of effect (either too high or too low) with spurious precision in the case of continuous explanatory variables or with overestimated precision in the case of binary explanatory variables, and may lead to incorrect inferences. The errors in estimates of effect of a continuous explanatory variable will be smaller when most of the dispersion of the explanatory variable is between rather than within clusters – i.e. the variable comes closer to being cluster-specific. Similarly, when the explanatory variable is binary, smaller differences in the effect estimates occur when the dispersion of the prevalence of the explanatory variable across clusters is small, or when its overall prevalence across clusters is high.

Additionally, I identified situations in which a naïve analytical approach is more likely to importantly affect statistical inference, i.e. when rates of Type I error and interval coverage deviate more from the nominal values of 5% and 95% respectively. These situations are when the explanatory variable is continuous, and ICC levels are greater than 0.01. It is then that Type I error rates are higher than 10% and interval coverage rates are lower than 80%. On the other hand, statistical inference when a naïve regression model is fitted is less likely to be of concern when the explanatory variable is binary, as the error and coverage rates deviate very little from the nominal values. However, even for a binary explanatory variable, error rates can be higher than 10%, and corresponding interval coverage rates can be lower than 90% (but possibly not lower than 80%). This occurs when ICC is high, the overall prevalence of the explanatory variable is low (approximately 5%), and the dispersion of the cluster-specific rates is large. In all circumstances in which the ICC is very small, clustering is minimal and there is little difference between RI and OLS regression.

The investigation of implications of ignoring clustering in statistical inference when the outcome of interest is continuous is extended in the next chapter to binary outcomes. As in this chapter, effect estimates and their precision will be considered separately for continuous and binary explanatory variables.

Chapter 5. Consequences of ignoring clustering in logistic regression

5.1 Introduction

Clustered data are frequently encountered in epidemiological research, particularly with the increasing use of multicentre and longitudinal studies. In many of these studies the outcome of interest follows a binomial distribution. In those cases, logistic regression is often the regression model of choice to explore the risk of occurrence of the outcome under investigation in relation to risk factors. Thus it is of interest to know how results from naïve logistic models compare with those from corresponding multi-level logistic regression models when observations are clustered.

A number of researchers have tried to address this question previously. Unlike the case of linear regression, the picture drawn from published literature on consequences of ignoring clustering in logistic regression is more coherent. Several studies that have analysed real data have shown that effect estimates (expressed as log-odds ratios) derived from naïve logistic regression, and also their precision, were lower than those from multi-level logistic regression (78, 109-111, 113-115, 117). These findings have been supported by analyses of simulated data, which have demonstrated varying levels of bias towards the null (depending on the level of clustering) when clustering was not taken into account (108, 112, 116). In contrast to these studies, a few others have reported slightly different findings, with estimates of effect and precision being very similar (118), or with point estimates derived from the two analytical models being substantially different (119). However, none of these studies has investigated the extent to which bias in the estimates of effect and their precision varies according to different distributions of the explanatory variable across clusters.

Less attention has been given to coverage by 95% confidence intervals and rates of type I error when clustering is not accounted for in logistic regression models. Abo-Zaid et al showed that coverage was significantly lower than the nominal value, especially when the explanatory variable was binary (108). A similar but more moderate under-coverage problem for the effect of a binary explanatory variable has also been reported by Xu et al, who additionally showed that when the explanatory variable was continuous, coverage was very close to the nominal value (116). Interestingly, another report, which also focused on binary explanatory variables, showed that coverage reduced considerably as clustering of the explanatory variable increased (120).

In this chapter, I extend the investigation of consequences of ignoring clustering in statistical inference to the situation in which the outcome under investigation follows a binomial distribution. As in the previous chapter, two types of explanatory variable were examined: continuous and binary. In each case, the implications of ignoring clustering were assessed in relation to the estimate of effect of the explanatory variable on the outcome variable, quantified by the log-odds ratio, its precision, and coverage by 95% confidence intervals.

As previously, the research question is introduced by exploring how effect estimates and related precision compared between the two models when observations were grouped into two clusters, and it is then expanded to the case of grouping into multiple clusters. The conditions under which the comparisons were made were the same as those described in the chapter on linear regression. When the explanatory variable was continuous, differences in estimates derived from the two models were examined according to the ratio of between- to within-cluster dispersion of the explanatory variable. For the binary explanatory variable, differences in estimates were explored for varying dispersion of prevalence rates of the explanatory variable across the clusters. In addition, differences in estimates were explored for outcome variables of low and high overall prevalence. Also, type I error rates were assessed for both the naïve logistic and the multi-level logistic regression model when the effect size was set to zero.

5.2 Two clusters – Continuous explanatory variable

To describe the problem of clustering of observations into two clusters, when the outcome of interest is binary and the explanatory variable is continuous, I initially generated as examples, three simulated datasets based on the ordinary logistic regression model

$$\ln\left(\frac{p}{1-p}\right) = a + \beta_1 x + \beta_2 cluster$$

with p being the probability of occurrence of the outcome of interest. In the first, the mean values of the explanatory variable were the same in the two clusters, while in the second and third case, the cluster-specific mean values of the explanatory variable differed by a constant value of 2 and 3, respectively (Table 5.1). In all three simulations, β_1 was set to 1.099 (OR=3), $p = 0.35$, while the distance between the cluster-specific intercepts was kept constant and equal to $k = 1.5$.

When the mean value of the explanatory variable was very similar in the two clusters (case 1 in Table 5.1), the cluster-unadjusted point estimate of effect of x on y was slightly lower than that derived from the cluster-adjusted logistic regression model. When the cluster-specific mean values of x differed by a constant value of 2 (case 2 in Table 5.1), the cluster-unadjusted OR was approximately 1.5-fold higher than the cluster-adjusted OR. Further increase in the distance between the cluster-specific mean values of x (case3 in Table 5.1) increased the difference between the estimates of effect from the two models a bit further.

Table 5.1. Example of observations grouped in two clusters, when the outcome variable was binary and the explanatory variable was continuous

Dataset	Prevalence of y	Mean value of x	Unadjusted OR	Adjusted OR
Case 1	38.0%	1st cluster: 0.01	2.52	2.83
		2nd cluster: -0.01		
Case 2	41.3%	1st cluster: -0.99	4.54	2.95
		2nd cluster: 0.99		
Case 3	44.1%	1st cluster: -1.49	4.82	3.03
		2nd cluster: 1.49		

These observations were extended, using simulated datasets with varying distances between the cluster-specific mean values of the explanatory variable, distances between the cluster-specific intercepts, and overall prevalence rates of the outcome variable as described in the following sections.

5.2.1 Methods

In the simple case of a binary outcome variable (typically coded as 0 or 1) and a single explanatory variable, the expectation of the outcome variable is the probability that the outcome is 1 (conditional on the explanatory variable).

$$E(y_i|x_i) = \Pr(y_i = 1|x_i)$$

When this is modelled using the ordinary logistic (OL) regression, the link function used is the $logit$ function as described below

$$logit(p_i) = logit(\Pr(y_i = 1|x_i)) = \beta_0 + \beta_1 x_i \qquad 5.1$$

$$y_i \sim Bernoulli(p_i)$$

where, for individual i, y_i and x_i are the values of the outcome and the explanatory variables respectively, and p_i is the probability of occurrence of the outcome of interest. The $logit$ transformation is defined as the logged odds, i.e. the logged probability of presence as compared to absence of the outcome (equation 5.2).

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \ln(odds) = \ln\left(\frac{probability\ of\ presence\ of\ the\ outcome}{probability\ of\ absence\ of\ the\ outcome}\right) \qquad 5.2$$

The parameter β_0 in equation 5.1 is the log odds of the outcome for individuals with $x_i = 0$, and β_1 is the log odds ratio of the outcome for every unit increase in the explanatory variable x_i.

The OL model that accounts for the cluster in which the observations are grouped is the one specified in equation 5.1, with the addition of an extra term for the cluster in the right hand side of the equation, i.e.

$$logit(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 cluster \qquad 5.3$$

where β_2 is the log odds ratio of the outcome when an observation is grouped in the second cluster as compared to the first, keeping x_i constant.

As in previous chapter, the consequences of ignoring grouping of observations in the two clusters for estimation of the log odds ratio and the related SE were explored through Monte Carlo simulations. Each simulated dataset consisted of 400 observations which were nested in two clusters (200 observations per cluster). For simplicity, the effect of x_i on y_i (β_1) was arbitrarily set to $\ln(2)$, and the log odds of the outcome for individuals with $x_i = 0$ (β_0) was set to $\ln\left(\frac{Prevalence_y}{1-Prevalence_y}\right)$, for a given prevalence of the outcome variable.

In the simulation, initially, two intercepts, u_1 and u_2, were generated for the first and the second cluster, respectively, with $u_2 = u_1 + k$, where k was the distance between the two intercepts. The values of the continuous explanatory variable x_i were generated by first generating a variable x_{0i} from the standard normal distribution ($x_{0i} \sim N(0,1)$) and then adding a value $m/2$ to x_{0i} when individual i was grouped in the first cluster, and subtracting the same quantity ($m/2$) from x_{0i} when i was grouped in the second cluster. Next, the predicted probabilities of the outcome variable were calculated from equation 5.3 solving for p_i, as follows

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 x_i + u_j\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i + u_j\right)} \qquad 5.4$$

with $j = 1$ or $j = 2$, depending on whether observation i was grouped in the first or second cluster, respectively. Then the binary outcome variable was generated using a random variable l from the uniform distribution $U(0,1)$; for every individual i, the outcome was positive (value of 1) if the predicted probability from equation 5.4 was higher than the value of l for that individual, and negative (value of 0) otherwise.

The above process was repeated for: varying distances ($k = \{0.5, 1, 1.5\}$) between the cluster-specific intercepts (u's); different values m for the distance between the cluster-specific mean values of x_i ($m = -5, -4.9, -4.8, -4.7, \dots, 0, 4.7, 4.8, 4.9, 5$); and different population prevalence rates of the outcome variable (10% and 30%).

For each combination of the above parameters, 25 simulated datasets were generated. For each simulated dataset, two models were fitted; an OL model that adjusted for cluster (equation 5.3) and another that did not (equation 5.1). From the fitted models, I generated the difference in the estimated log odds ratios (difference=$\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) and the ratio of the corresponding SEs (ratio=$SE^{cluster-adjusted}/SE^{cluster-unadjusted}$). These were plotted against the distance between the cluster-specific mean values of x_i and summarised by means and SDs. To assist illustration of patterns of differences in effect estimates and ratios of SEs by differences of cluster-specific mean values of x_{ij}, local polynomial curves were also used. These are a scatterplot smoothing approach in which the outcome (values plotted on the vertical axis) is fitted to a polynomial form of the regressor (values on the horizontal axis) via locally weighted least squares, without making any assumption regarding the functional form for the response value given the regressor (133).

Also, the coverage of the 95% confidence intervals was explored by calculating the proportion of the simulated datasets for which the 95% CIs of the cluster-adjusted and the cluster-unadjusted models included the simulated effect (OR=2).

To explore type I error when clustering is ignored, simulations were repeated assuming no association between x_i and y_i, with the effect being set to 0 ($\beta_1 = 0$). The p-values from the cluster-adjusted and cluster-unadjusted associations were saved and the percentages of p-values <0.05 were calculated and summarised.

5.2.2 Results

Difference in regression coefficient

Figure 5.1 plots differences in the log odds ratios estimated from the cluster-adjusted and cluster-unadjusted logistic regression models ($\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) against the shift in the distribution of the explanatory variable x for the second cluster from that for the first cluster. The top two sub-plots of the figure present simulated data, while the bottom two present the confidence intervals resulting from polynomial smooth plots. The two sub-plots in the left and the right part of the figure are for population prevalence of the outcome variable y of 10% and 30%, respectively. Varying distances between the cluster-specific intercepts ($k = 0.5, 1, 1.5$) are indicated by different shades of grey.

When the cluster-specific distributions of x were overlaid, the differences in the log odds ratios were very close to zero. Increasing the distance between the distributions of x in the two clusters (moving to the right or the left of 0 in the x-axis of the figure) increased the differences between the estimated log odds ratios, up to a maximum negative value when the distance between $\bar{x}_{cluster\ 1}$ and $\bar{x}_{cluster\ 2}$ was ~ 2 times the SD of x). Beyond that, further increase in the displacement of the cluster-specific distributions of x caused the differences in the log odds ratios to reduce (towards zero). In almost all cases (all of which assumed an OR=2) the differences were negative indicating that the log odds ratios from the cluster-unadjusted models were higher than those from the cluster-adjusted models.

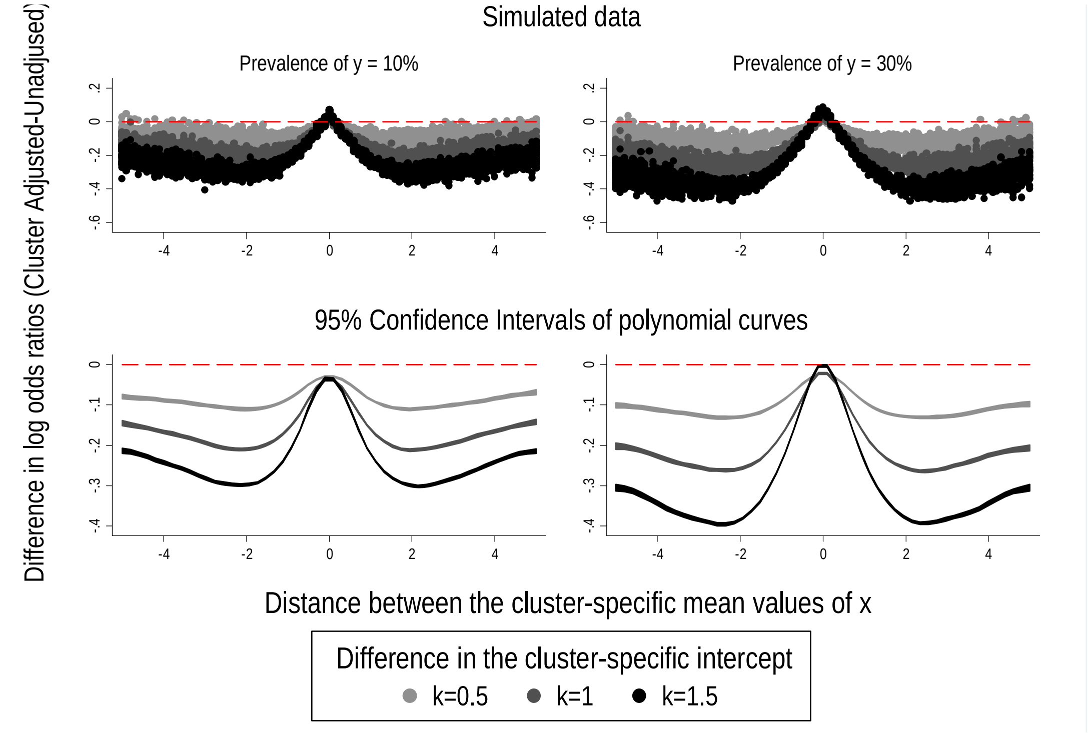


Figure 5.1. Scatter plot of the differences between log odds ratios estimated from OL regression that adjusted and OL that did not adjust for cluster against distance between the cluster-specific mean values of x, for varying distances between the cluster-specific intercepts and prevalence of the outcome variable in the population

Differences in the log odds ratios estimated from the two methods are summarised in Table 5.2 separately for the two population prevalence rates of the outcome variable, the three values of the cluster-specific intercepts considered in these simulations ($k's$), and for selected values of the distance between the cluster-specific distributions of x.

Table 5.2. Mean (SD) of differences between log odds ratios estimated from OL regression that adjusted and OL regression that did not adjust for cluster, for varying distances between the cluster-specific intercepts and prevalence of the outcome variable in the population.

		Difference in log odds ratios		
		$\boldsymbol{k = 0.5}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 1.5}$
Prevalence of y =10%	$\boldsymbol{-4.05 \leq shift \leq -3.95}$	-0.090 (0.040)	-0.152 (0.039)	-0.253 (0.029)
	$\boldsymbol{-2.05 \leq shift \leq -1.95}$	-0.116 (0.030)	-0.210 (0.031)	-0.297 (0.024)
	$\boldsymbol{-0.05 \leq shift \leq 0.05}$	0.006 (0.005)	0.018 (0.008)	0.047 (0.017)
	$\boldsymbol{1.95 \leq shift \leq 2.05}$	-0.120 (0.033)	-0.223 (0.027)	-0.305 (0.026)
	$\boldsymbol{3.95 \leq shift \leq 4.05}$	-0.073 (0.033)	-0.170 (0.029)	-0.256 (0.025)
Prevalence of y =30%	$\boldsymbol{-4.05 \leq shift \leq -3.95}$	-0.110 (0.041)	-0.228 (0.032)	-0.342 (0.052)
	$\boldsymbol{-2.05 \leq shift \leq -1.95}$	-0.127 (0.024)	-0.267 (0.025)	-0.381 (0.024)
	$\boldsymbol{-0.05 \leq shift \leq 0.05}$	0.008 (0.005)	0.034 (0.011)	0.067 (0.013)
	$\boldsymbol{1.95 \leq shift \leq 2.05}$	-0.128 (0.029)	-0.259 (0.017)	-0.390 (0.025)
	$\boldsymbol{3.95 \leq shift \leq 4.05}$	-0.107 (0.040)	-0.226 (0.039)	-0.345 (0.036)

As also seen in Figure 5.1, when $\bar{x}_{cluster\ 1} \approx \bar{x}_{cluster\ 2}$, differences in the log odds ratios were very small. As the absolute difference $|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}|$ increased, differences in the log odds ratios increased, with $\beta_1^{Cluster-adjusted} < \beta_1^{Cluster-unadjusted}$. The maximum difference in the log odds ratios occurred when $|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}| \approx 2\ SDs$ of x. The absolute maximum value of the differences was different for the two population prevalence rates of y; the maximum difference in the estimated log odds ratios was somewhat higher when the population prevalence of y was 30% compared to 10%. For a given prevalence of y, when $\bar{x}_{cluster\ 1} \approx \bar{x}_{cluster\ 2}$, the average difference in the log odds ratio was positive and it increased with increasing k. More generally, for a given prevalence of y, and distance between $\bar{x}_{cluster\ 1}$ and $\bar{x}_{cluster\ 2}$, increasing k increased differences of log odds ratios estimated from the two methods. Overall, estimates from cluster-adjusted models were lower than those from cluster-unadjusted models, the differences being highest when $|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}| \approx 2\ SDs$ of x, and higher for higher cluster-specific intercepts, and higher population prevalence of the outcome variable y.

Ratios of standard errors

The ratios of the SEs of the log odds ratios estimated from the two methods ($SE^{cluster-adjusted}$/ $SE^{cluster-unadjusted}$) were plotted against the distance between the cluster-specific mean values of x (Figure 5.2). As in the corresponding plot for the difference in the log odds ratios (Figure 5.1), the top two sub-plots show simulated data, while the bottom two show the 95% CIs of the polynomial curves for the simulated data. The two sub-plots at the left of the figure show results for a population prevalence of the outcome variable y of 10%, while the corresponding two sub-plots at the right are for population prevalence of y of 30%. As previously, the lighter shades of grey correspond to a lower value of distance between the cluster-specific intercepts.

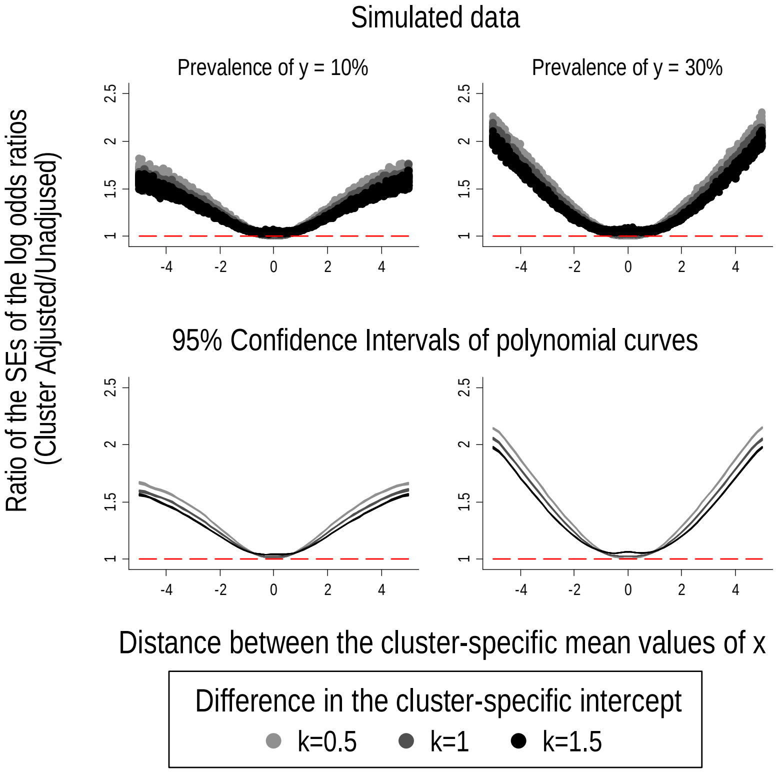


Figure 5.2. Scatter plot of ratios of the SEs of the log odds ratios estimated from OL regression that adjusted to those from OL that did not adjust for cluster against distance between the cluster-specific mean values of x, for varying differences in the cluster-specific intercepts and prevalence of the outcome variable in the population

When the two cluster-specific distributions of x were overlaid, the ratio of the SEs of the log odds ratio took a minimum value. Increasing the distance between the cluster-specific mean values of x made the ratio of the SEs increase too, meaning that as the distance between $\bar{x}_{cluster\ 1}$ and $\bar{x}_{cluster\ 2}$ increased, the SEs from the cluster-adjusted logistic regression model were higher than those from the cluster-unadjusted. For $k = 0.5$, the minimum value of the ratio of the SEs was very close to 1, meaning that for $\bar{x}_{cluster\ 1} = \bar{x}_{cluster\ 2}$ the SEs from the two methods were approximately equal (Table 5.3). The minimum value of the ratio of the SEs increased as the distance between the cluster-specific intercepts increased (Table 5.3). Also, for any given value of k, the minimum value of the ratio of the SEs was lower for lower population prevalence of the outcome variable y.

Table 5.3. Minimum value of the ratios of SEs of the log odds ratios estimated from OL regression that adjusted to those from OL regression that did not adjust for cluster for varying distances between the cluster-specific intercepts and prevalence of the outcome variable in the population.

	Minimum ratio of the SEs of the log odds ratios $(\mathbf{SE}_{\beta_1^{Cluster\ adjusted}}/\mathbf{SE}_{\beta_1^{Cluster\ unadjusted}})$		
	$\boldsymbol{k = 0.5}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 1.5}$
Prevalence of y = 10%	1.001	1.012	1.030
Prevalence of y = 30%	1.004	1.021	1.049

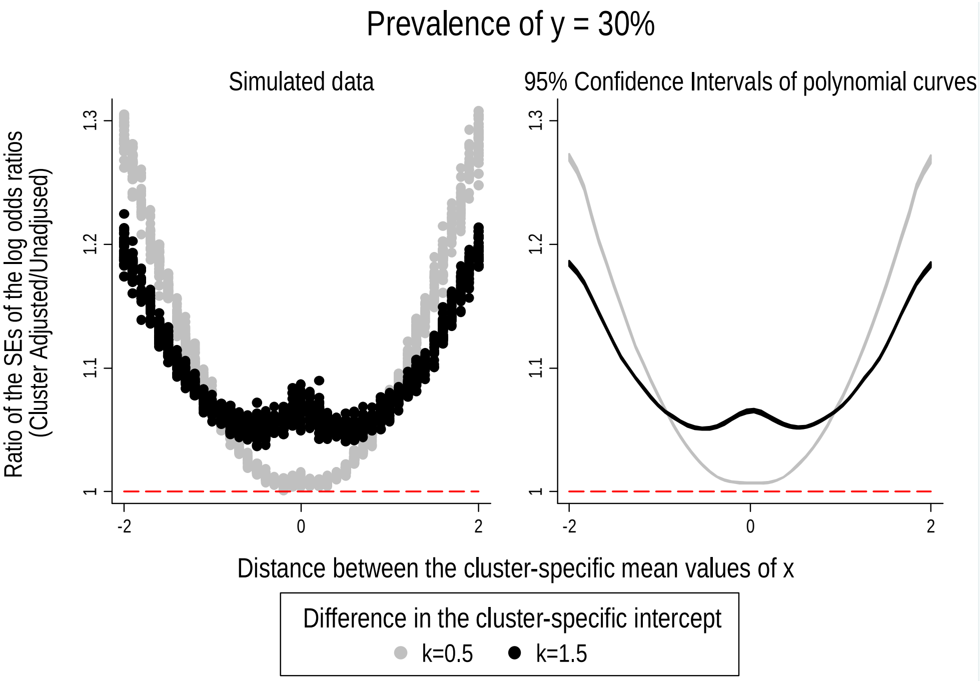


Figure 5.3. Scatter plot (sub-figure on the left) and 95% CIs of the local polynomial curves (sub-figure on the right) of ratios of the SEs of the log odds ratios estimated from OL regression that adjusted to those from OL that did not adjust for cluster against distance between the cluster-specific mean values of x, for distance between the cluster-specific intercepts $k = 0.5$ and $k = 1.5$.

The pattern of change of the ratios of the SEs of the log odds ratios for increasing difference between $\bar{x}_{cluster\ 1}$ and $\bar{x}_{cluster\ 2}$ was somewhat different for higher values of distance between the cluster-specific intercepts k. Figure 5.3, as Figure 5.2, shows the ratios of the SEs for increasing differences between $\bar{x}_{cluster\ 1}$ and $\bar{x}_{cluster\ 2}$, but focusing on a shift of $\bar{x}_{cluster\ 1}$ from $\bar{x}_{cluster\ 2}$ up to 2 times the SD of x, and only for population prevalence of y of 30% and for the lowest and the highest values of k (0.5 and 1.5). When $k = 0.5$, increasing the shift of $\bar{x}_{cluster\ 1}$ from $\bar{x}_{cluster\ 2}$ made the ratios of the SEs increase smoothly. However, when $k = 1.5$, increasing the shift of $\bar{x}_{cluster\ 1}$ from $\bar{x}_{cluster\ 2}$, made the ratios of the SEs decrease before they started increasing to higher values, giving a W-shaped pattern.

Coverage by 95% Confidence Intervals

Table 4.5 shows the proportion of the simulated datasets for which the 95% CIs of the cluster-adjusted and the cluster-unadjusted models included the simulated effect (OR=2), according to the distance between the cluster-specific intercepts k and population prevalence of the outcome variable y. For any k and any prevalence of y, the coverage for the cluster-adjusted model was very close to the nominal value of 95%. However, coverage for the cluster-unadjusted model became very low as the distance between the cluster-specific intercepts k increased, and for higher prevalence of y. For $k = 1.5$, and for 30% prevalence of y, the coverage of the 95% CIs was as low as 7.5%.

Table 5.4. Coverage (%) by 95% confidence intervals of simulated effect OR=2 for the cluster-unadjusted and cluster-adjusted models, according to distance between the cluster-specific intercepts k, and population prevalence of the outcome variable y.

	Prevalence of y = 10%		**Prevalence of y = 30%**	
	95% CI coverage cluster-unadjusted model	**95% CI coverage cluster-adjusted model**	**95% CI coverage cluster-unadjusted model**	**95% CI coverage cluster-adjusted model**
k=0.5	34.8	95.0	15.4	95.3
k=1	11.1	95.0	8.9	94.5
k=1.5	8.2	94.9	7.5	95.2

Type I error

Type I error was assessed by calculating the proportion of the simulated datasets that resulted in a significant association between the outcome and the explanatory variable in the cluster-unadjusted and the cluster-adjusted model, when the log odds ratio was set to 0 (i.e. OR=1). Results are presented by distance between the cluster-specific intercepts (k) and population prevalence of y in Table 5.5. Type I error from the model that statistically adjusted for clustering was at the nominal value of 5% for any distance k examined. However, increasing distance k increased type I error from the cluster-unadjusted model. The rate of increase was steeper when the population prevalence of the outcome variable was higher. The highest type I error was 87% and it occurred when estimates were not adjusted for clustering and for the larger distance between cluster-specific intercepts assumed in these simulations.

Table 5.5. Proportion (%) of simulated datasets for which the null hypothesis was rejected, when the true effect size (log odds ratio) of the explanatory on the outcome variable was set to 0

	Prevalence of y = 10%		**Prevalence of y = 30%**	
	Type I error in the cluster-unadjusted model	**Type I error in the cluster-adjusted model**	**Type I error in the cluster-unadjusted model**	**Type I error in the cluster-adjusted model**
k=0.5	19.2	4.4	35.8	4.7
k=1	53.6	5.1	77.7	5.4
k=1.5	76.0	4.1	86.6	4.2

5.3 Multiple clusters – Continuous explanatory variable

5.3.1 Methods

In the simplest case of a single explanatory variable, the OL regression model is described in equation 5.1. For a binary outcome and a continuous explanatory variable, the RI logistic regression model is defined as in equation 5.1 with the addition of a random intercept u_j in the linear prediction, as follows:

$$logit\big(\Pr(y_{ij} = 1|x_{ij}, u_j)\big) = \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_{0j} + \beta_1 x_{ij} = \beta_0 + \beta_1 x_{ij} + u_j \qquad 5.5$$

where the index i refers to the individual and the index j to the cluster in which observations are grouped, and β_{0j} the estimate of the log odds for $x_{ij} = 0$. The term u_j represents the effect of being in group j on the log odds that $y_{ij} = 1$ when $x_{ij} = 0$. It is also referred to as the cluster-level residual and it follows a normal distribution with mean 0 and standard deviation SD_u ($u_j \sim N(0, SD_u^2)$). The term β_1 is the effect on log odds of a unit increase in x_{ij} for individuals in the same cluster (same value of u_j); it is often referred to as cluster-specific effect of x_{ij}. The exponential of β_1, $\exp(\beta_1)$, is an odds ratio (OR) comparing the odds for individuals differing by one unit for x_{ij} and in the same cluster j.

As discussed in the previous chapter, the ICC is a value that expresses the level of similarity within clusters, and it is based on the distinction between variances at the two levels (individual and cluster level). The ICC value depends on the variance of the mean values of the continuous outcome variable (corrected for the explanatory variables in the regression model) across clusters (SD_u^2) and the variance within clusters (SD_e^2). In the case of a binary outcome variable, the individual level values are 0 and 1 and the variance at the lower level is defined as $p_i(1 - p_i)$. It is thus defined on a probability scale in contrast to the cluster-level variance which is defined on a linear scale. The problem of lack of comparability of the variances at the two levels in the case of a binary outcome has been identified and well described before, and alternative approaches have been suggested (134, 135). The best-described approach is the latent variable method in which the lower level variance is described on a logistic scale. With this approach, underlying the observed binary outcome y_{ij}, there is a latent continuous outcome y_{ij}^*, representing the propensity for $y_{ij} = 1$. This unobserved individual variable follows a logistic distribution with mean zero and variance $\pi^2/_3$. Based on this approach, the ICC for RI logistic regression models is defined as $ICC = \frac{SD_u^2}{SD_u^2 + \pi^2/_3}$ (136).

To explore the study questions, in a similar way to the previous chapter, datasets were generated from Monte Carlo simulations based on the assumptions of the RI model. Each simulated dataset had 10000 observations with 100 observations nested within each of 100 clusters. For simplicity, the size of the effect of x_{ij} on y_{ij} was arbitrarily set to $\ln(2)$ (i.e. OR=2), and the log odds of the outcome for individuals with $x_i = 0$ (β_0) was set to $\ln\left(\frac{Prevalence_y}{1 - Prevalence_y}\right)$, for a given prevalence of the outcome variable.

As in the simulation of datasets for continuous outcome and explanatory variable (section 4.3.1), the values of the explanatory variable were generated as $x_{ij} = x_{0ij} + shift_j$, where x_{0ij} was an individual-level variable from the standard normal distribution ($\sim N(0,1)$), and $shift_j$ a cluster-level variable from the normal distribution with mean zero and standard deviation SD_{shift}. Also, the cluster-level error term u_j was generated from a random normal distribution with mean zero and standard deviation SD_{u_j}. Then, I generated the predicted probabilities of the outcome variable from equation 5.5 solving for p_{ij}, as follows:

$$p_{ij} = \frac{\exp\left(\beta_0 + \beta_1 x_{ij} + u_j\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{ij} + u_j\right)} \qquad 5.6$$

The binary outcome variable was subsequently generated through the use of a random variable l from the uniform distribution $U(0,1)$; for each individual i, the outcome was positive (value of 1) if the predicted probability from equation 5.6 was higher than the value of l for that individual, and negative (value of 0) otherwise.

Simulated data were generated for different values of SD_{u_j} (0.0574, 0.0995, 0.1823, 0.3190, 0.6046, 1.1874) chosen to give expected values of ICC of 0.001, 0.003, 0.01, 0.03, 0.1, and 0.3 respectively (following the definition of ICC for logistic regression models). Also, different values of SD_{shift} were selected from a random uniform distribution $U[a, b]$, with the parameters a and b arbitrarily chosen to be 0 and 15 respectively. Moreover, simulated data were generated for low (10%) and high (30%) population prevalence of the outcome variable. For each combination of values of SD_{u_j}, SD_{shift}, and population prevalence of y_{ij}, 50 datasets were simulated.

For each simulated dataset, an OL and a RI logistic regression model were fitted and the effect estimates (β_1^{OL} and β_1^{RI}), along with the corresponding SEs ($SE_{\beta_1^{OL}}$ and $SE_{\beta_1^{RI}}$), were saved. To compare results from the two models, I calculated the difference in regression coefficients ($\beta_1^{RI} - \beta_1^{OL}$) and the ratio of their SEs ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OL}}$). These were described for varying between-to within-cluster dispersions of the explanatory variable x_{ij}, different population prevalence rates of the outcome variable, and different levels of ICC, through summary statistics and scatterplots. Similarly to the previous section, local polynomial curves were used to assist illustration of patterns.

The coverage by 95% confidence intervals was assessed by calculating the proportions of estimated confidence intervals that included the true value (simulated effect). To assess type I error, I repeated the simulations with the effect of x_{ij} on y_{ij} set to 0 ($\beta_1 = 0, i.e. OR = 1$). Type I error was then described as the proportion of datasets for which the null hypothesis was rejected at a 5% level under the OL and the RI logistic regression models, for increasing ICC values.

The research question was further explored by analysis of real data from the CUPID study. This is described in the section that follows the description of results of the simulation studies.

5.3.2 Results

Differences in log odds ratios

Differences in log odds ratios ($\beta_1^{RI} - \beta_1^{OL}$) estimated from the RI and the OL models were plotted against the relative between- to within-cluster dispersion of explanatory variable x_{ij}, separately for high and low population prevalence of the outcome variable y_{ij}, and are presented in Figure 5.4. The top two sub-plots of the figure present differences in log odds ratios from simulated data, while the bottom two sub-plots of the figure show the 95% CIs of the corresponding polynomial curves. In each of the sub-plots, the shades depict different levels of ICC with darker shades of grey corresponding to higher ICC values. Differences are also summarised numerically in Table 5.6, according to ICC values and for the two levels of population prevalence of the outcome variable, using means (SDs), medians (IQRs), and ranges.

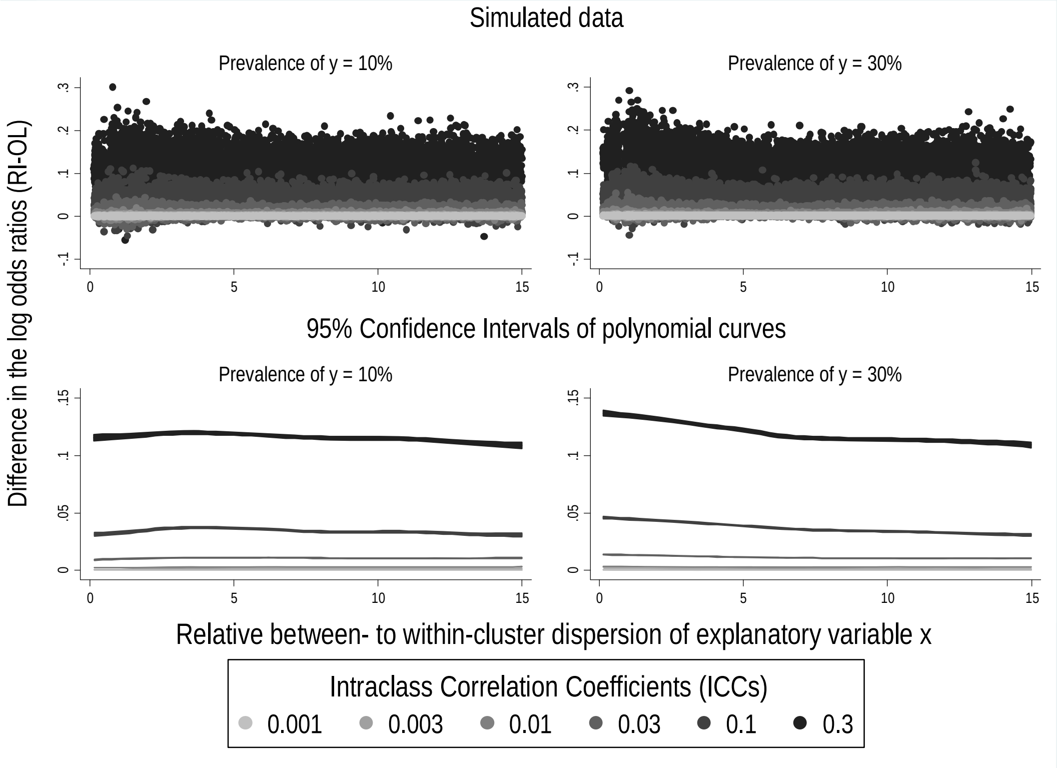


Figure 5.4. Difference between log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OL}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend) and for the two different population prevalence rates assumed.

Table 5.6. Descriptive statistics for differences in log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OL}$) according to ICC

Population prevalence of outcome	Intraclass Correlation Coefficient	Mean (SD)	Median (IQR)	Minimum	Maximum
10%	0.001	0 (0)	0 (0-0)	-0.00	0.00
	0.003	0 (0)	0 (0-0)	-0.00	0.01
	0.01	0 (0)	0 (0-0)	-0.01	0.02
	0.03	0.01 (0.01)	0.01 (0.01-0.02)	-0.03	0.05
	0.1	0.03 (0.02)	0.03 (0.02-0.05)	-0.05	0.11
	0.3	0.12 (0.04)	0.12 (0.09-0.14)	-0.06	0.30
30%	0.001	0 (0)	0 (0-0)	-0.00	0.00
	0.003	0 (0)	0 (0-0)	-0.00	0.01
	0.01	0 (0)	0 (0-0)	-0.01	0.02
	0.03	0.01 (0.01)	0.01 (0.01-0.02)	-0.02	0.06
	0.1	0.04 (0.02)	0.04 (0.02-0.05)	-0.04	0.12
	0.3	0.12 (0.04)	0.12 (0.10-0.14)	-0.01	0.29

When ICC was low (≤0.03) differences between the estimated log odds ratios were very small (mean difference ≈0). However, for larger values of ICC, the average difference increased. For low population prevalence of the outcome variable and for ICC=0.1, the average value of difference was 0.03 and for ICC=0.3, the average difference was 0.12. In the first case (low prevalence of y_{ij} and ICC=0.1), the average value of the estimated log odds ratio from the RI logistic regression model was 2.0, while that from the OL model was 1.93. For ICC=0.3, the corresponding average of estimated ORs was 2.0 from the RI model and 1.79 from the OL model. These results indicate that as the within cluster similarity of observations increased, the bias towards the null of the effect estimate from the model that ignored clustering increased too. Furthermore, differences in the log odds ratios seemed to have a slow rate of decrease for increasing relative dispersion of x_{ij} (Figure 5.4). However, the average difference in the log odds ratios did not appear to be associated with the population prevalence of y_{ij}, although the range of differences was somewhat smaller for higher prevalence of y_{ij}.

Differences in the log odds ratios varied more for higher ICC values. They ranged from -0.02 to 0.06 for ICC=0.03, and from -0.01 to 0.30 for ICC=0.3, for both high and low population prevalence of the outcome variable (Table 5.6). However, this might have been a consequence of the wider range of ICC estimates within the ICC=0.3 category (range of ICC values from 0.25 to 0.34) as compared to lower ICC categories. To explore this further, differences in the log odds ratios were plotted (simulated data and 95% CIs of the corresponding polynomial curves) against relative dispersion of x_{ij} for different levels of ICC within the category of ICC=0.3 (Figure 5.5) and are summarised numerically in Table 5.7.

As shown in Figure 5.5, the average value of differences between estimates from the two methods increased with increasing ICC. However, the range of differences was very similar across subcategories of ICC, meaning that for equally narrow windows of ICC estimates, higher average of ICC values yielded wider ranges of differences in the log odds ratios from cluster-adjusted and cluster-unadjusted models. The polynomial curves presented in Figure 5.5 (centre of the 95% CIs presented in the figure) also show that as the relative dispersion of the cluster-mean values of x_{ij} increased, the difference in the log odds ratios decreased. However, for any given ICC level, the difference in log odds ratios for the smallest relative dispersion was larger than the difference in the log odds ratios for the largest relative dispersion by only 0.02 on average.

Table 5.7. Descriptive statistics for differences in log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OL}$) according to ICC estimates within the category of ICC=0.3

Intraclass Correlation Coefficient	**Mean (SD)**	**Minimum**	**Maximum**
0.25-0.27	0.10 (0.03)	-0.02	0.20
0.27-0.29	0.11 (0.03)	-0.05	0.25
0.29-0.30	0.12 (0.03)	-0.01	0.24
0.30-0.31	0.12 (0.04)	-0.03	0.27
0.31-0.32	0.13 (0.04)	-0.06	0.27
0.32-0.34	0.13 (0.04)	-0.01	0.30

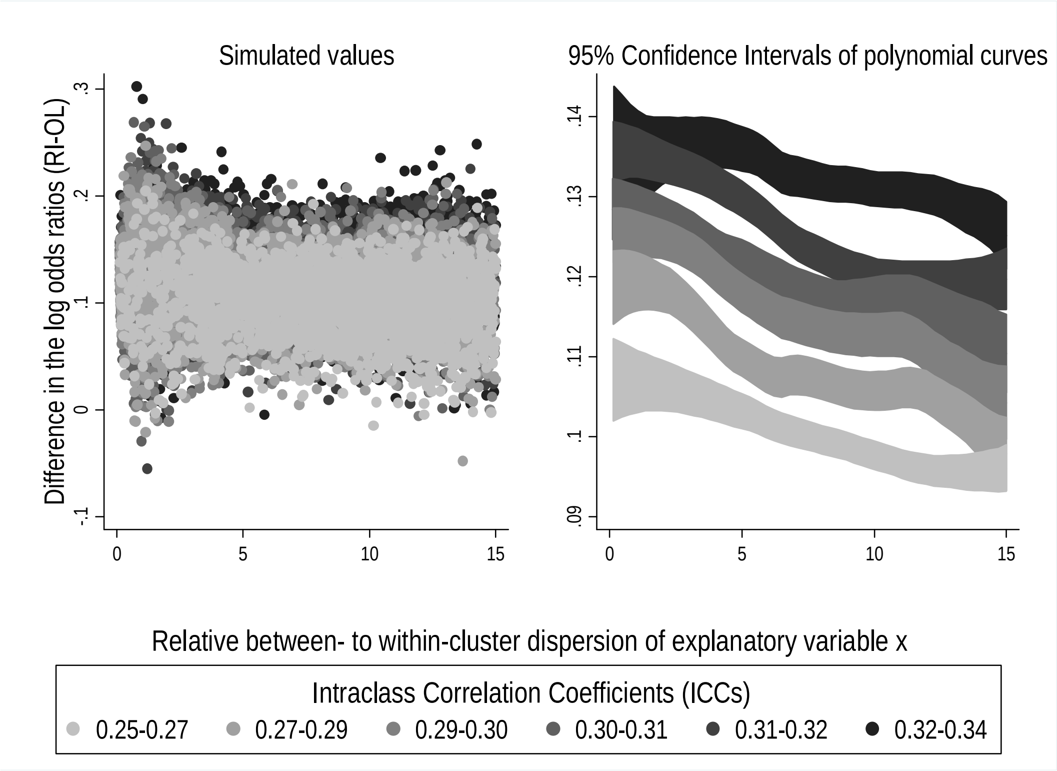


Figure 5.5. Differences between log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OL}$) plotted against relative between- to within-clusters dispersion of explanatory variable x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend) within the range 0.25-0.34

Ratio of standard errors

Similarly to the differences in the log odds ratios derived from the two methods, the ratios of their SEs ($\mathrm{SE}_{\beta_1^{RI}}/\mathrm{SE}_{\beta_1^{\mathrm{OL}}}$) were explored in relation to the relative dispersion of x_{ij}. Results are illustrated in Figure 5.6. As in Figure 5.4, the top two sub-plots in Figure 5.6 show the simulated data, while the bottom two sub-plots depict the 95% CIs of the corresponding polynomial curves. The two sub-plots at the left and the right are for low and high population prevalence of y_{ij}, respectively. The different shades within the figure depict different levels of ICC as indicated in the legend. Descriptive statistics for the ratios of SEs are additionally presented in Table 5.8.

Irrespective of the ICC level, the ratio had a minimum value which occurred at the smallest relative dispersion of x_{ij} and increased as the relative dispersion of x_{ij} increased, reaching a maximum value. The ratio of the SEs then remained relatively steady with further increases in the relative dispersion of x_{ij}, but with a small drop for the highest values. The minimum value of the ratio depended on the ICC; for the lowest ICC examined (0.001), the minimum ratio of SEs was very close to 1, while it increased to more than 2 when the ICC was 0.3. Similarly, the maximum value of the ratio was ICC-dependent; higher ICC values resulted in a higher maximum for the ratio. As well as the average values of the ratios of SEs, their range varied with ICC. The range of ratios was narrower when the ICC was low, and wider when ICC increased.

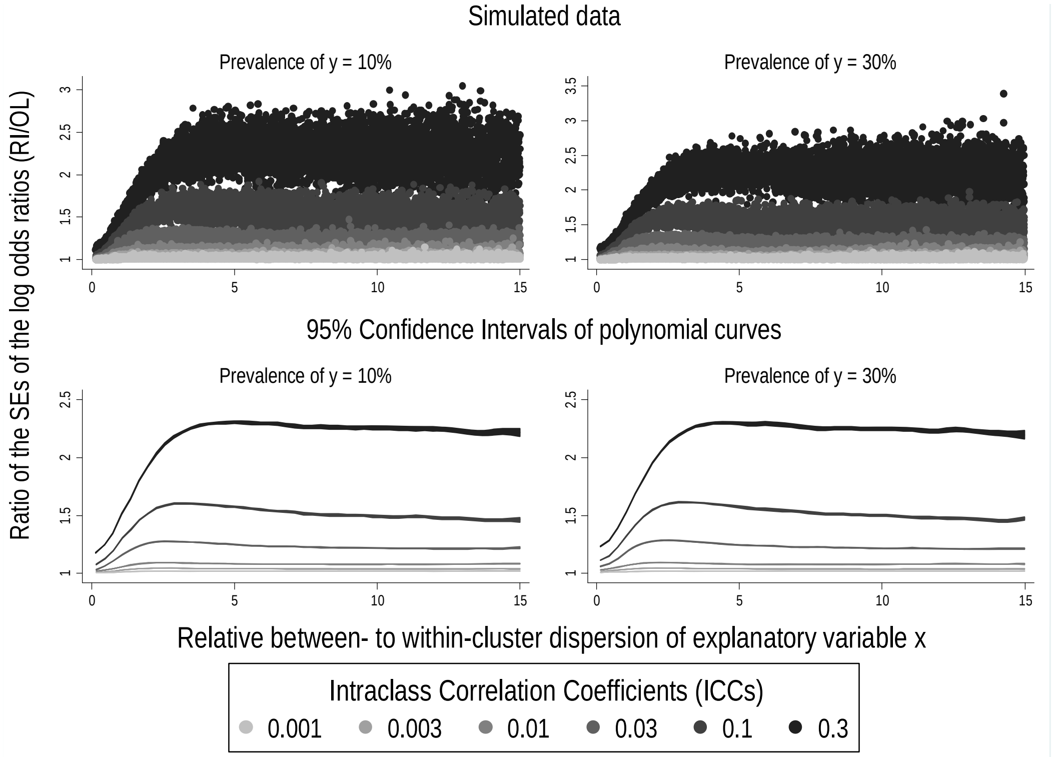


Figure 5.6. Ratios of standard errors of log odds ratios estimated from RI and OL models ($\mathrm{SE}_{\beta_1^{RI}}/\mathrm{SE}_{\beta_1^{\mathrm{OL}}}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend)

Table 5.8. Descriptive statistics for ratios of the standard errors of the log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OLS}$) according to ICC

	ICC	**Mean (SD)**	**Median (IQR)**	**Minimum**	**Maximum**
Population prevalence of y_{ij}=10%	0.001	1.02 (0.01)	1.02 (1.01-1.02)	1.00	1.15
	0.003	1.04 (0.01)	1.04 (1.03-1.04)	1.00	1.14
	0.01	1.08 (0.03)	1.08 (1.06-1.09)	1.00	1.25
	0.03	1.22 (0.06)	1.23 (1.20-1.26)	1.01	1.47
	0.1	1.49 (0.15)	1.51 (1.41-1.60)	1.03	1.92
	0.3	2.15 (0.33)	2.21 (2.05-2.36)	1.11	3.05

	ICC	Mean (SD)	Median (IQR)	Minimum	Maximum
Population prevalence of y_{ij}=30%	0.001	1.02 (0.01)	1.02 (1.01-1.02)	1.00	1.11
	0.003	1.04 (0.01)	1.04 (1.03-1.04)	1.00	1.13
	0.01	1.08 (0.03)	1.08 (1.06-1.09)	1.00	1.25
	0.03	1.23 (0.06)	1.23 (1.20-1.26)	1.02	1.43
	0.1	1.50 (0.14)	1.52 (1.42-1.60)	1.04	1.97
	0.3	2.14 (0.31)	2.20 (2.03-2.34)	1.14	3.39

To explore whether the last observation was due to a wider range of ICC estimates within higher ICC categories (for the ICC=0.3 category, ICC estimates varied from 0.25 to 0.34, while they varied only from 0.025 to 0.034 in the ICC=0.03 category), ratios of the SEs of the log odds ratios were plotted (simulated data and 95% CIs of the corresponding polynomial curves) against relative dispersion of x_{ij} for different levels of ICC within the category of ICC=0.3 (Figure 5.7).

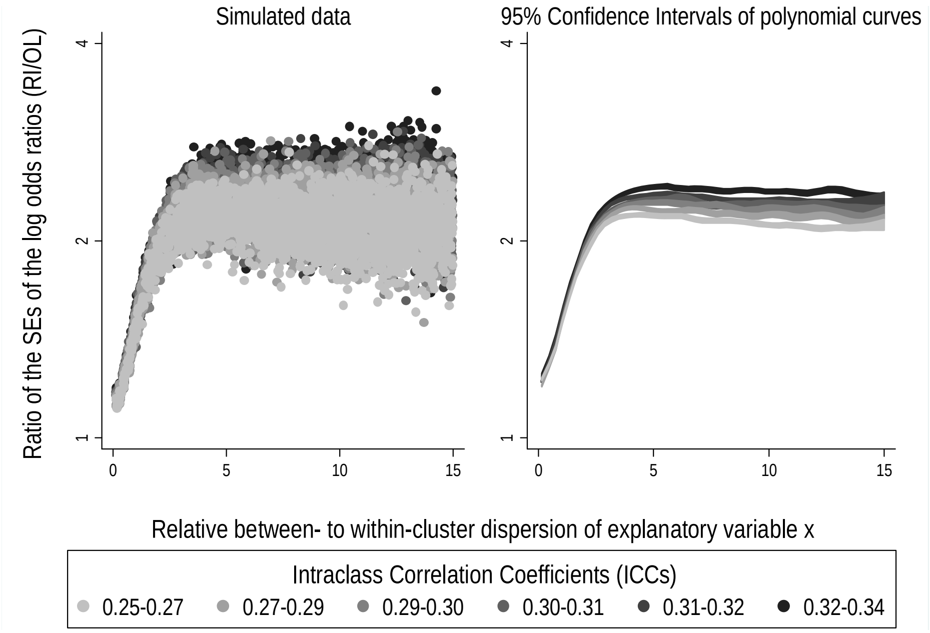


Figure 5.7. Ratios of standard errors of the log odds ratios estimated from RI and OL models $(\mathrm{SE}_{\beta_1^{\mathrm{ML}}}/\mathrm{SE}_{\beta_1^{\mathrm{OLS}}})$ plotted against relative between- to within-clusters dispersion of explanatory variable x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend) within the range 0.25-0.34

Table 5.9. Descriptive statistics for ratios of SEs of the log odds ratios estimated from RI and OL models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OL}}$) according to ICC estimates within the category of ICC=0.3

Intraclass Correlation Coefficient	**Mean (SD)**	**Minimum**	**Maximum**
0.25-0.27	2.09 (0.22)	1.11	2.79
0.27-0.28	2.09 (0.31)	1.11	2.84
0.28-0.30	2.08 (0.34)	1.12	2.93
0.30-0.31	2.10 (0.36)	1.12	2.87
0.31-0.32	2.19 (0.32)	1.12	2.93
0.32-0.34	2.33 (0.27)	1.13	3.39

The results are also summarised in Table 5.9. Even though the average ratio of SEs increased with increasing ICC, the range of values increased very little (from 2.09 for 0.25≤ICC≤0.27 to 2.33 for 0.32≤ICC≤0.34).

Coverage of 95% CIs

Coverage of the simulated OR=2 by 95% CIs estimated from the RI model was very close to the nominal value of 95%. The average coverage was 95.4% and varied from 94.1% to 96.8% across different ICC values, quarters of the distribution of the relative dispersion of x_{ij}, and population prevalence rates of the outcome variable. Coverage from the OL model was also very close to 95% when ICC was low (≤0.003) for both low and high population prevalence of y_{ij}, and for all parts of the distribution of relative dispersion of x_{ij} (Figure 5.8). When ICC increased (>0.003), coverage decreased and it became particularly poor for ICC=0.3. Also, coverage by 95% CIs was lower when the population prevalence of y_{ij} was higher. Moreover, coverage varied with increasing relative dispersion of x_{ij}; for any given ICC level, coverage was higher in the first quarter and lower in the second quarter, and it increased gradually in the third and top quarter of the distribution of the relative dispersion of x_{ij} (Figure 5.9). The lowest value for coverage by 95% CIs was 1.3% and it occurred in the second quarter of the distribution of relative dispersion of x_{ij}, for ICC=0.3, and high population prevalence of y_{ij}.

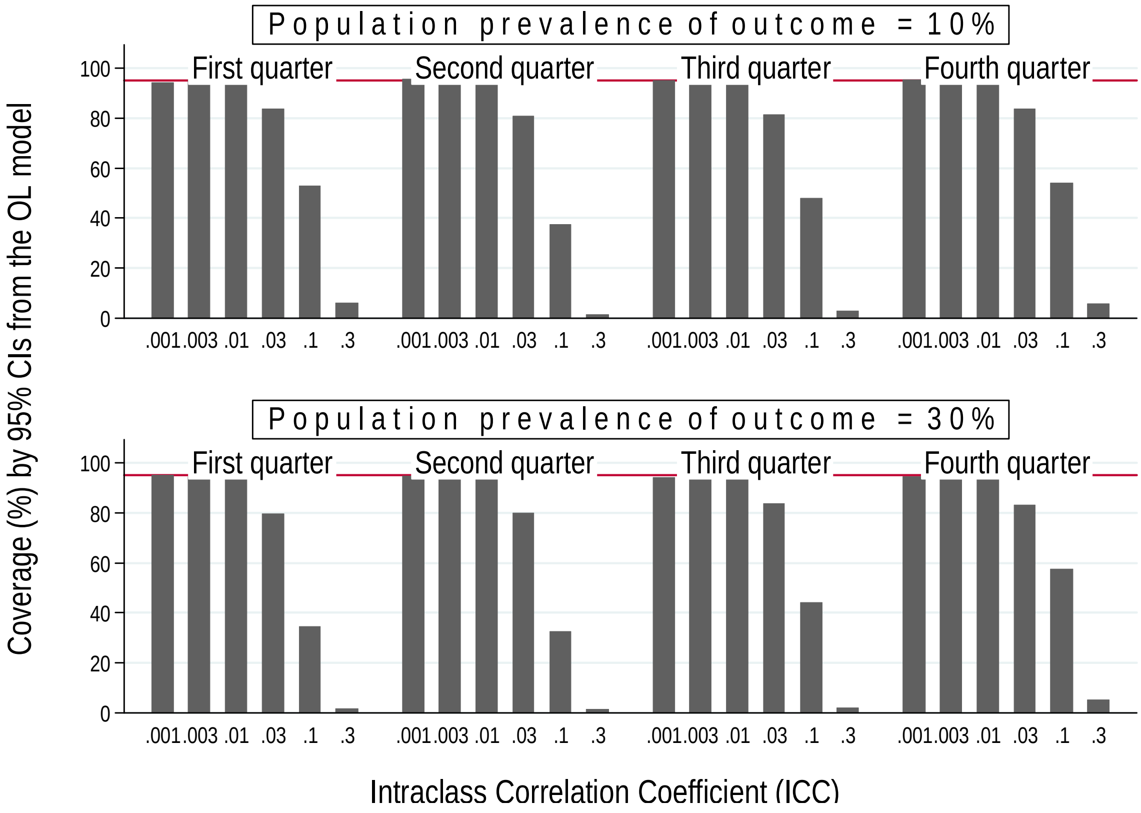


Figure 5.8. Coverage (%) by 95% confidence intervals of simulated effect OR=2 for the OL model, according to quarters of the distribution of the relative between- to within-cluster dispersion of explanatory variable x_{ij}, population prevalence of outcome variable y_{ij}, and ICC. The horizontal red line is for the nominal value of 95%.

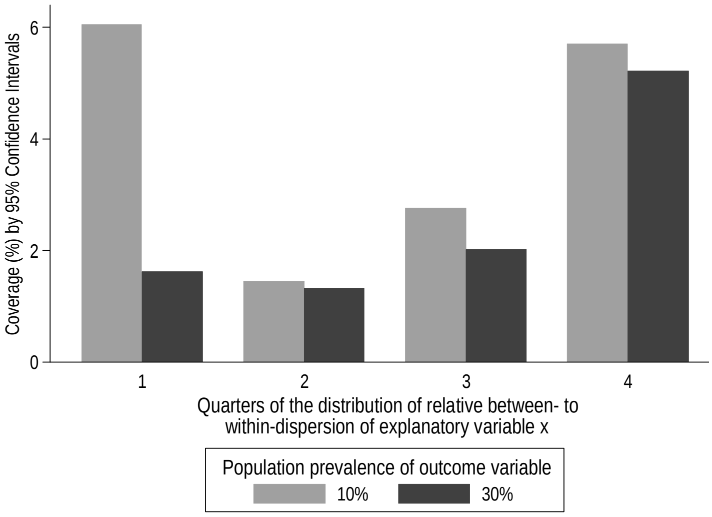


Figure 5.9. Coverage (%) by 95% confidence intervals of simulated effect OR=2 under the OL model according to quarters of the distribution of the relative between- to within-cluster dispersion of explanatory variable x_{ij}, and population prevalence of outcome variable y_{ij}, when ICC=0.3

Type I error

Type I error under the RI model was on average 5%, and it ranged from 3.7 to 5.7 across the different levels of relative dispersion of the explanatory variable, x_{ij}, ICC, and population prevalence of the outcome variable. No consistent pattern of the variation of error rates from the RI model was observed with any of these parameters.

Error rates from the OL model were close to the nominal value of 5% when the ICC was low (≤0.01) and also the prevalence of the outcome variable (Figure 5.10). Increasing ICC, prevalence of outcome variable, and relative dispersion of x_{ij}, all increased type I error (Figure 5.10). The maximum type I error rate was 67.8%, and it occurred at the top quarter of the distribution of relative dispersion of x_{ij}, when the ICC was 0.3 and the population prevalence of $y_{ij}$was high.

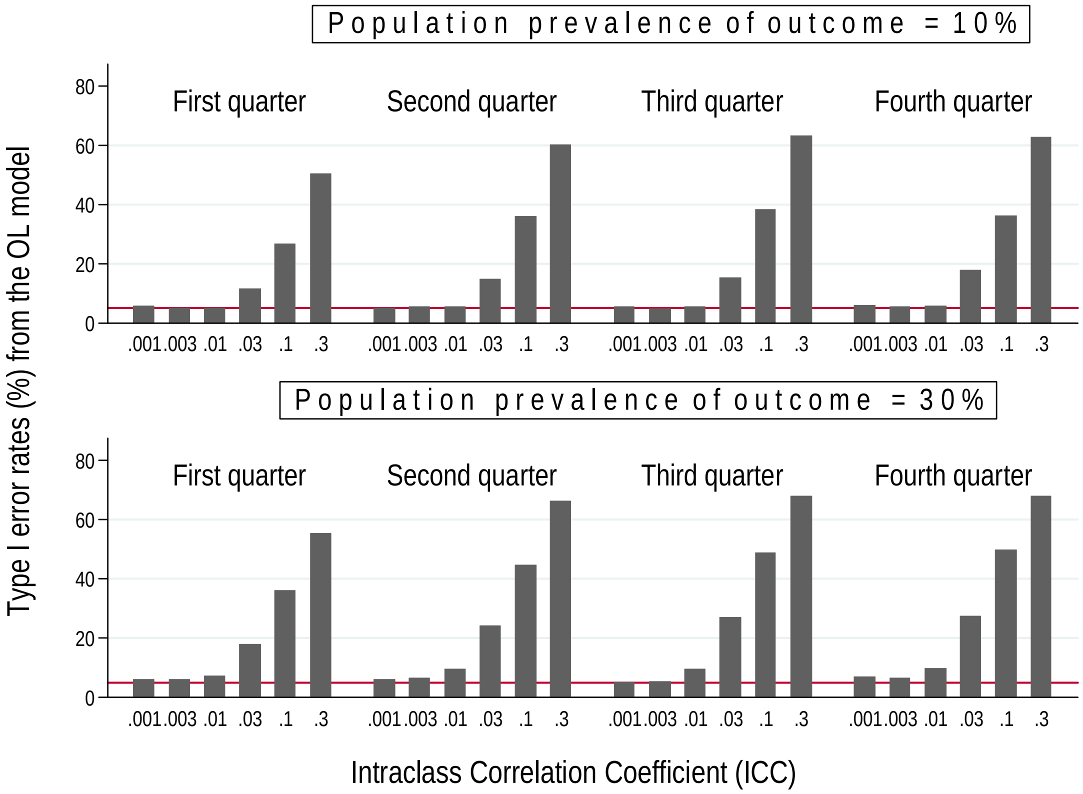


Figure 5.10. Proportion (%) of datasets for which the null hypothesis was rejected under the OL logistic regression model, according to quarters of the distribution of the relative between- to within-cluster dispersion of explanatory variable x_{ij}, population prevalence of outcome variable y_{ij}, and ICC. The horizontal red line is for the nominal value of 5%.

5.3.3 Analysis of CUPID data

Most variables in the CUPID study have an ordinal or binary form, with only two measures on a continuous scale (age of participant at recruitment and hours worked per week). To explore the consequences of failing to account statistically for clustering when using real data, all ordinal variables were dichotomised (by collapsing all but the null category together). That resulted in 115 binary variables overall with prevalence rates that ranged from 1.7% to 97.8%. These variables were used as outcome variables (without considering what these variables represented) in 230 logistic regression models, half of which used age and the other half hours worked per week, as explanatory variables. Each of the regression models was fitted without accounting for clustering (OL regression model) and after accounting for clustering using RIs. As estimates of log odds ratios in the simulation studies presented earlier in this chapter were all positive, to enable direct comparison of results, from the 230 models only estimates that were either both positive (N=100) or both negative (N=56) from the two analytical methods were compared. When for a given association estimates from both the OL and the RI model were negative (N=56), they were both turned into positive.

The estimated ICCs across the RI logistic regression models fitted ranged from 0.06 to 0.73, with a median value of 0.15. The distribution of the estimated ICC values is shown in Figure 5.11.

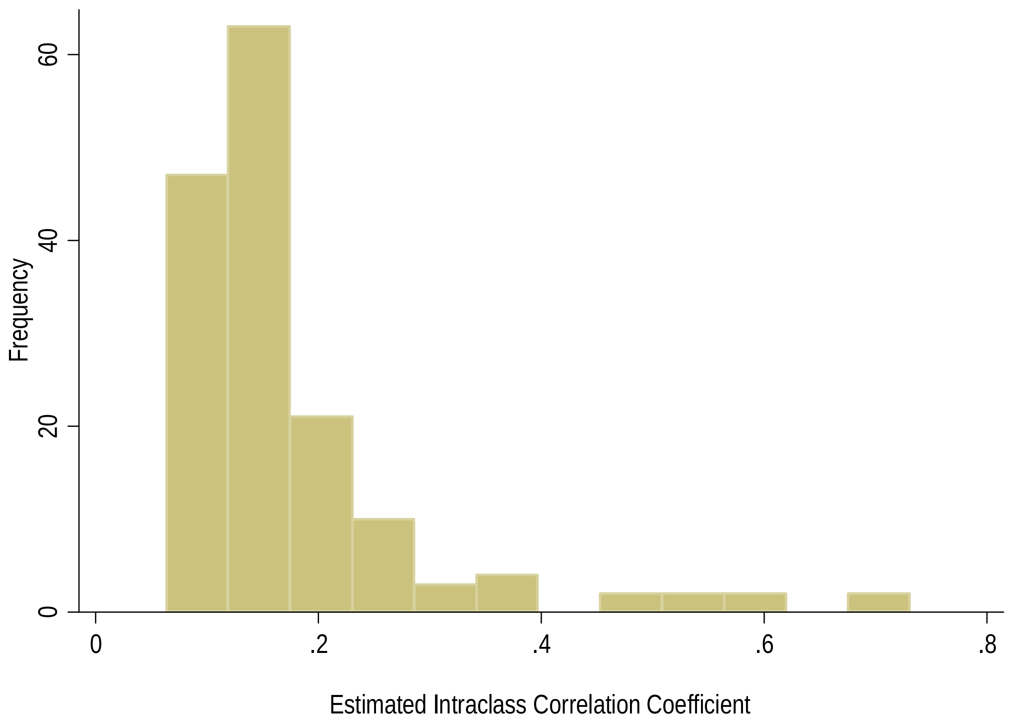


Figure 5.11. Distribution of estimated unadjusted intraclass correlation coefficients using data from the CUPID study

The log odds ratios estimated from the two methods are plotted against each other in Figure 5.12 (left hand side plot), and so are the corresponding SEs of the log odds ratios (right hand side plot). The data points are coloured in different shades of grey with the shading corresponding to quarters of the distribution of the estimated ICC values from the RI logistic regression models. The dashed line in each of the two plots of the figure corresponds to the $y = x$ line. Agreement between log odds ratios from the two methods was good (estimates very close to the $y = x$ line) when they were further from the null value of zero, but when they were closer to the null value (log odds ratios < 0.02) there was more dispersion without any clear pattern. SEs were consistently higher when derived from the RI than when derived from the OL model (Figure 5.12).

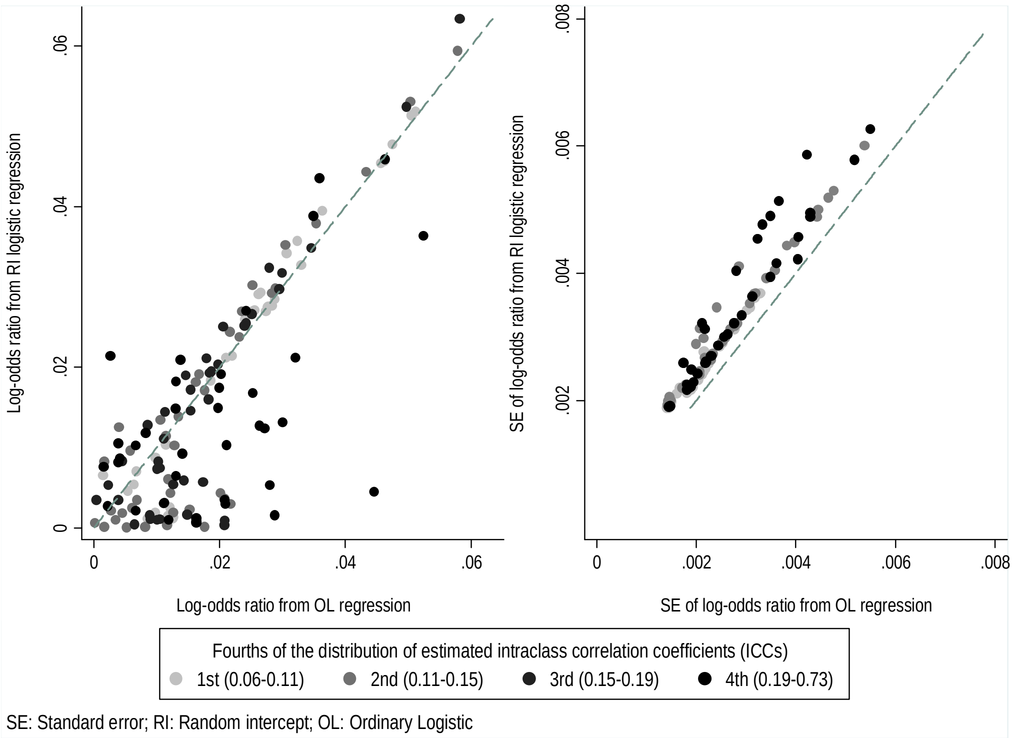
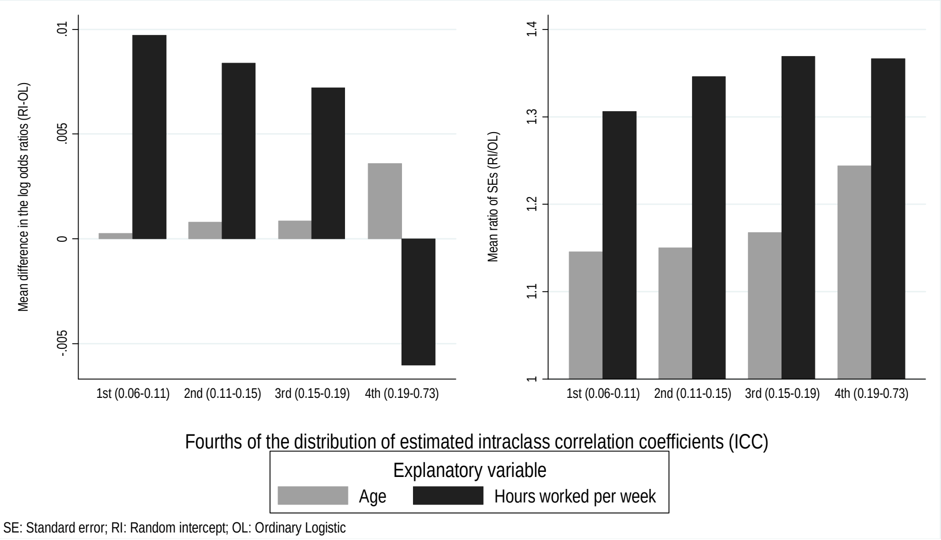


Figure 5.12. Scatter plots of the log odds ratios (left) and the corresponding standard errors (right) estimated from RI logistic and OL regression models using data from the CUPID study, for different levels of estimated ICC

I next explored the effect of the relative dispersion (between- to within-clusters) of the explanatory variables on the differences in the estimated log odds ratios and the ratio of the corresponding SEs. To calculate relative dispersion, for each of the explanatory variables, I divided the dispersion of the cluster-specific mean values (between-cluster SD) of the variable, by the average dispersion of the variable in each of the different clusters (i.e. occupational groups). The calculated relative dispersion of age was 0.640, while that of hours worked per week was 0.972. Figure 5.13 shows the mean values of differences in the log odds ratios from the two methods (left), and of the ratios of the corresponding SEs (right), by quarters of the distribution of the estimated ICC values, and by explanatory variable.

The average difference in log odds ratios was much lower when age was used as the explanatory variable compared to hours worked per week. Also, increasing categories of estimated ICC increased differences in log odds ratios for age, but not for hours worked per week. These observations agree partly with those from the simulated data. In section 5.3.2, it was shown that increasing ICC increased differences in log odds ratios. That was seen in the analysis of data only when age was used as an explanatory variable.

With regard to ratios of SEs, the mean values deviated much less from the null value of one when the explanatory variable was age than when it was hours worked per week. In addition, for both explanatory variables, increasing ICC levels increased the ratios of SEs further from 1. These observations agree with those from the analysis of simulated data, in which it was shown that increasing ICC increased ratios of SEs, and that for any given ICC, ratios of SEs were higher when relative dispersion was larger (for values of relative dispersion approximately up to 3).



SE: Standard error; RI: Random intercept; OL: Ordinary Logistic

Figure 5.13. Bar chart of mean values of the differences in the log odds ratios (RI-OL) (left) and the ratios of the corresponding standard errors (RI/OL) (right) using data from the CUPID study, according to the explanatory variable used in the models

There was no pattern of differences in log odds ratios from the two models, nor of the ratios of their corresponding SEs, for increasing prevalence of the outcome variable (data not shown).

5.4 Two clusters – Binary explanatory variable

As in previous sections, to describe the problem in the case of a binary outcome and a binary explanatory variable, with observations nested within two clusters, I simulated data, as shown in Table 5.10. I considered three situations. In the first situation, the prevalence of the explanatory variable was the same in both clusters. In the second case, the prevalence of the explanatory variable in the first cluster was lower than that in the second cluster, and in the third case, higher. In all three cases, the prevalence of the explanatory variable was set to 0.2, while the prevalence of the outcome variable was different in the two clusters (leading to clustering of the outcome variable in all three cases).

Table 5.10. Example of observations grouped in two clusters, when both the outcome and the explanatory variables were binary

			All		Cluster 1		Cluster 2		OR	
			x		x		x		Unadjusted	Adjusted
1st case			**0**	**1**	**0**	**1**	**0**	**1**		
	y	**0**	283	66	149	37	134	29	1.62	1.65
		1	37	14	11	3	26	11		
			x		x		x			
2nd case			**0**	**1**	**0**	**1**	**0**	**1**		
	y	**0**	252	94	149	37	103	57	2.49	2.01
		1	28	26	11	3	17	23		
			x		x		x			
3rd case			**0**	**1**	**0**	**1**	**0**	**1**		
	y	**0**	298	55	149	37	149	18	0.65	0.77
		1	42	5	11	3	31	2		

As shown in Table 5.10, when the prevalence of the explanatory variable was the same in the two clusters (1st case), the adjusted OR was very similar to the unadjusted. However, when the prevalence of the explanatory variable was lower in the first cluster than in the second (2nd case), the cluster-adjusted OR was larger than the unadjusted. Also, when the prevalence of the explanatory variable was higher in the first cluster than in the second (3rd case), the cluster-adjusted OR was lower than the unadjusted.

Differences in effect estimates and their corresponding SEs in the cluster-unadjusted and cluster-adjusted logistic regression models in the case of two clusters were explored in more detail, as described in the following sections.

5.4.1 Methods

The methods followed for this investigation were very similar to those described in section 5.2.1. In brief, simulated datasets of 400 observations, evenly nested within two clusters, were generated. In each simulation, I generated a binary explanatory variable x_i of a given prevalence in each of the two clusters. I then generated cluster-specific intercepts u_1 and u_2 such that $u_2 = u_1 + k$, where k was the distance between the two intercepts. The predicted probabilities for the outcome variable were calculated from equation 5.4. In this equation, the log odds of the outcome for individuals with $x_i = 0$ (β_0) was set to ln $\left(\frac{Prevalence_y}{1-Prevalence_y}\right)$, for a given prevalence of the outcome variable, while the effect β_1 was set to $\ln(2)$. The predicted probabilities were then used to derive the binary outcome variable y_i by first generating a random variable l from the uniform distribution $U(0,1)$ and then for each individual i assigning values 0 or 1 to y_i depending on whether the predicted probability for that individual was lower or higher than the corresponding value of the random variable l.

Simulated data were generated for different combinations of distance between the cluster-specific intercepts ($k = \{0.5, 1, 1.5\}$), population rates of the outcome variable (10% and 30%), and prevalence rates of the explanatory variable in each of the two clusters. The prevalence rates of x_i in the first cluster that were examined were 0.05, 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9, and 0.95; those for the second cluster varied from 0 to 1 with a step size of 0.01. For each combination of the above parameters, 25 simulated datasets were generated.

As in the section 5.2.1, here also, in each set of data, a cluster adjusted (equation 5.3) and a cluster-unadjusted (equation 5.1) model were fitted. I then calculated the difference between the effects estimated from the two models (difference=$\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) and the ratio of the corresponding SEs (ratio=$SE^{cluster-adjusted}/SE^{cluster-unadjusted}$). The difference in the effect estimates, the ratios of the SEs, and the coverage by estimated 95% confidence intervals were described for different combinations of k, prevalence rates of x_i, and for high and low population prevalence of y_i. To enable a better illustration of patterns, scatter plots and local polynomial curves were used as in previous sections.

To explore type I error, the above simulation was repeated assuming an effect $\beta_1 = 0$. The p-values from the two models were saved and the percentages of p-values <0.05 were calculated and summarised.

5.4.2 Results

Difference in the log odds ratios

The differences between log odds ratios estimated from the cluster-adjusted and the cluster-unadjusted models ($\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$) were plotted against increasing prevalence of the explanatory variable x in the second cluster, while that in the first cluster remained constant (Figure 5.14 and Figure 5.15). Figure 5.14 plots the differences in the log odds ratios when the population prevalence of y was simulated to be 10%, and Figure 5.15 when the prevalence of y was simulated to be 30%. Different prevalence rates of x in the first cluster were examined (0.05, 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9, and 0.95) as shown in the two figures. The sub-plots at the top of each of the two figures show simulated data, while those at the bottom of the figures show the 95% CIs of the local polynomial smooth plots. As in previous sections, three cases of differences between the cluster-specific intercepts k ($k = 0.5, 1, 1.5$) were explored; data points and corresponding curves in lighter shades of grey depict lower values of k, and darker shades depict higher values of k.

When the prevalence of x in the second cluster was lower than that in the first cluster, the differences in the log odds ratios were positive. As the prevalence of x in the second cluster increased approaching the prevalence of x in the first cluster, differences decreased, reaching a value of zero when the prevalence of x was the same in the two clusters. Further increase in the prevalence of x in the second cluster (prevalence of $x_{cluster\ 2}$ > prevalence of $x_{cluster\ 1}$), made the difference in the effect estimates from the two models increase. When prevalence of $x_{cluster\ 2}$ > prevalence of $x_{cluster\ 1}$, the differences in the log odds ratios were negative, indicating that $\beta_1^{Cluster-adjusted}<\beta_1^{Cluster-unadjusted}$. The differences seen in point estimates for different prevalence rates of the explanatory variable in the two clusters were due to deviations of $\beta_1^{Cluster-unadjusted}$ from the simulated effect of $ln(2)$ (OR=2), while $\beta_1^{Cluster-adjusted} \approx \ln(2)$. For example, when prevalence of $y_{ij} = 0.3$, $k = 1$, and prevalence of $x_{cluster\ 2}$ > prevalence of $x_{cluster\ 1}$, the average value of $\beta_1^{Cluster-unadjusted}$ was 1.073, while when prevalence of

$x_{cluster\ 2}$ < prevalence of $x_{cluster\ 1}$ the average value of $\beta_1^{Cluster-unadjusted}$ was 0.257. In both cases, the average value of $\beta_1^{Cluster-adjusted}$ was 0.7.

For a given difference between the prevalence rates of x_{ij} in the two clusters, and prevalence of the outcome variable, y, differences in effect estimates from the two methods were larger for larger differences in the cluster-specific intercepts k. Also, the rate of change of differences in log odds ratios for increasing prevalence of x in the second cluster, when keeping the prevalence of x in the first cluster constant, was steeper for larger values of k (observation made from polynomial curves).

The direction of differences between the log odds ratios, and the pattern of change for increasing difference in the prevalence of x in the two clusters were very similar in the two scenarios of population prevalence of y examined (comparison of Figure 5.14 and Figure 5.15 for a given combination of prevalence of x_{ij} in each cluster and value of k).

Ratio of standard errors

The ratios of the SEs of the log odds ratios estimated from the two methods were plotted against increasing prevalence of the explanatory variable x in the second cluster, for constant values of prevalence of x in the first cluster, in graphs very similar to those described above for differences in the effect estimates (Figure 5.16 and Figure 5.17). In all cases, the ratio of the SEs of the log odds ratio was higher than one, meaning that in all simulated data and for any scenario of difference between the cluster-specific intercepts k, prevalence of x, and prevalence of the outcome variable y, the SEs of $\beta_1^{Cluster-adjusted}$ were larger than those of $\beta_1^{Cluster-unadjusted}$. When the prevalence rates of x in the two clusters were similar, the ratios of the SEs were close to 1. Increasing the difference between the cluster-specific prevalence rates of x (or increasing/decreasing the prevalence of x in the second cluster from that of the first cluster), increased the ratio of the SEs giving a U-shaped pattern. The maximum ratio of SEs was ~3, indicating that the $SE^{cluster-adjusted}$ was 3-fold higher than the $SE^{cluster-unadjusted}$, and this occurred for the maximum difference in the prevalence rates of x in the two clusters.

The pattern of change in the ratio of SEs (by increasing difference in the cluster prevalence rates of x), and in the ratios of the SEs, did not depend on the population prevalence of y, nor on the distances between the cluster-specific intercepts k.

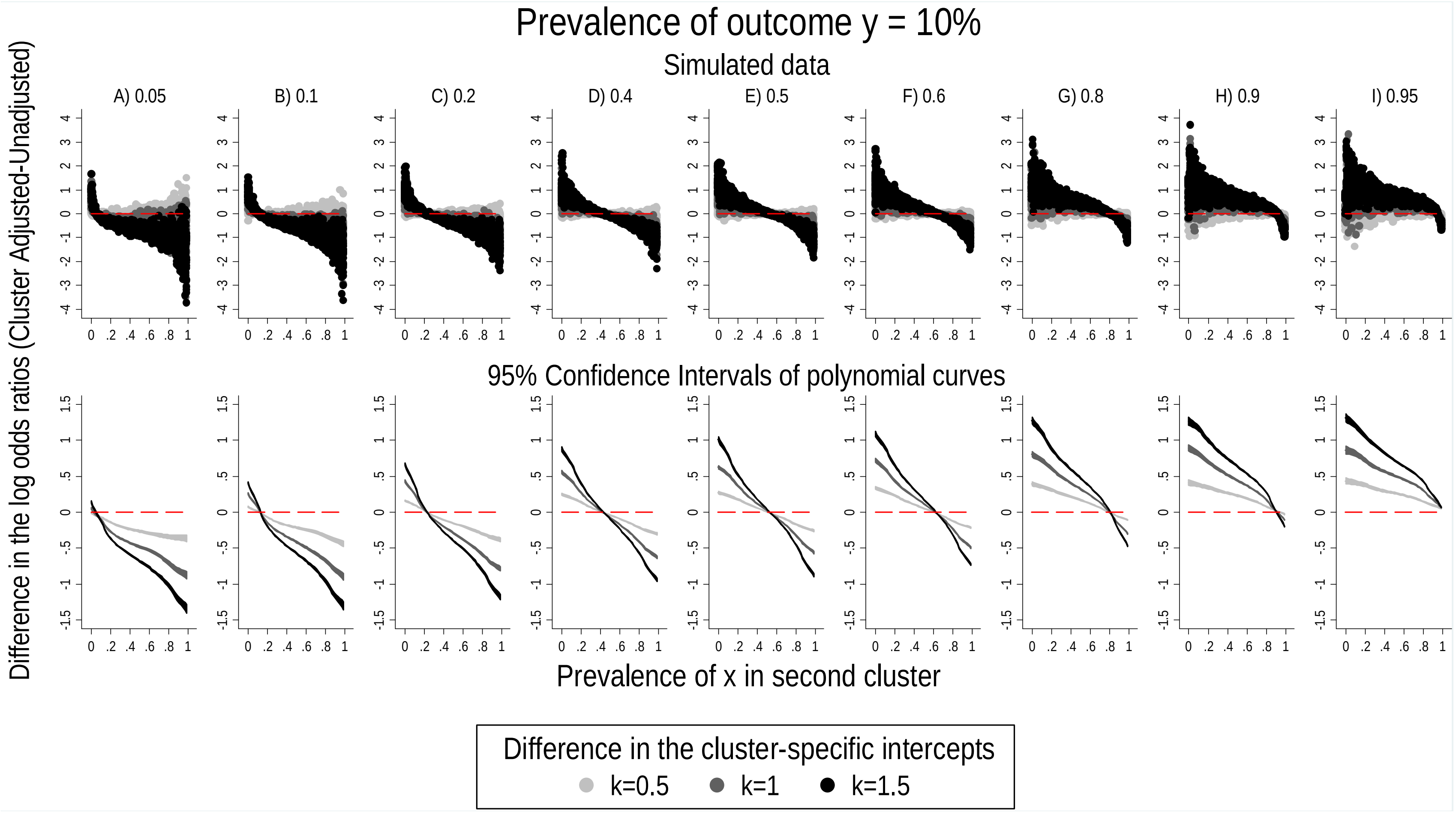


Figure 5.14. Differences in log odds ratios estimated from cluster-adjusted and cluster-unadjusted logistic regression models for varying prevalence of explanatory variable x in the second cluster keeping the prevalence of x in first cluster constant to A) 0.05, B) 0.1, C) 0. 2, D) 0.4, E) 0.5, F) 0.6, G) 0.8, H) 0.9, and I) 0.95, when the population prevalence of the outcome variable y was assumed to be 10%

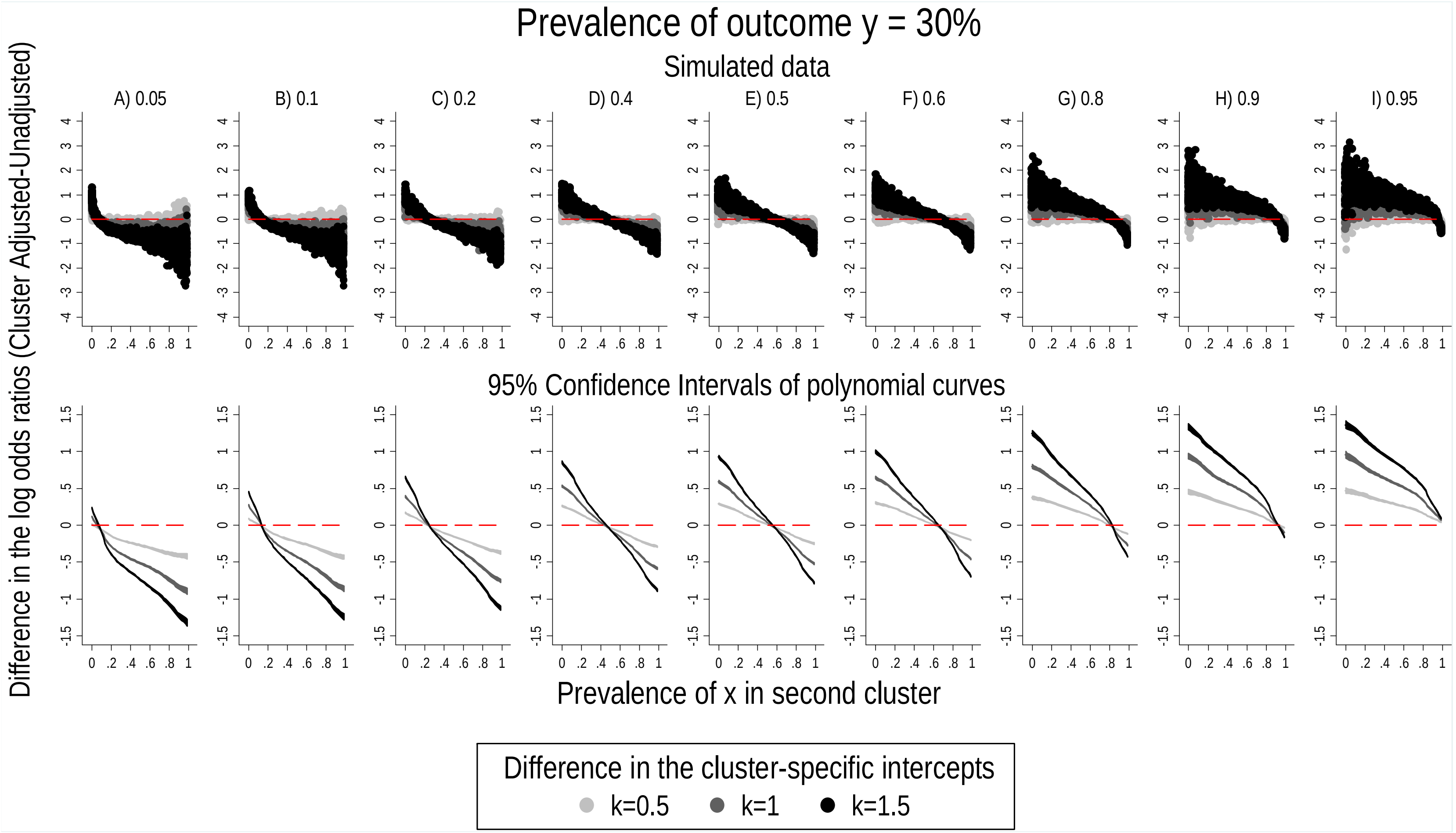


Figure 5.15. Differences in log odds ratios estimated from cluster-adjusted and cluster-unadjusted logistic regression models for varying prevalence of explanatory variable x in the second cluster keeping the prevalence of x in first cluster constant to A) 0.05, B) 0.1, C) 0. 2, D) 0.4, E) 0.5, F) 0.6, G) 0.8, H) 0.9, and I) 0.95 when the population prevalence of the outcome variable y was assumed to be 30%

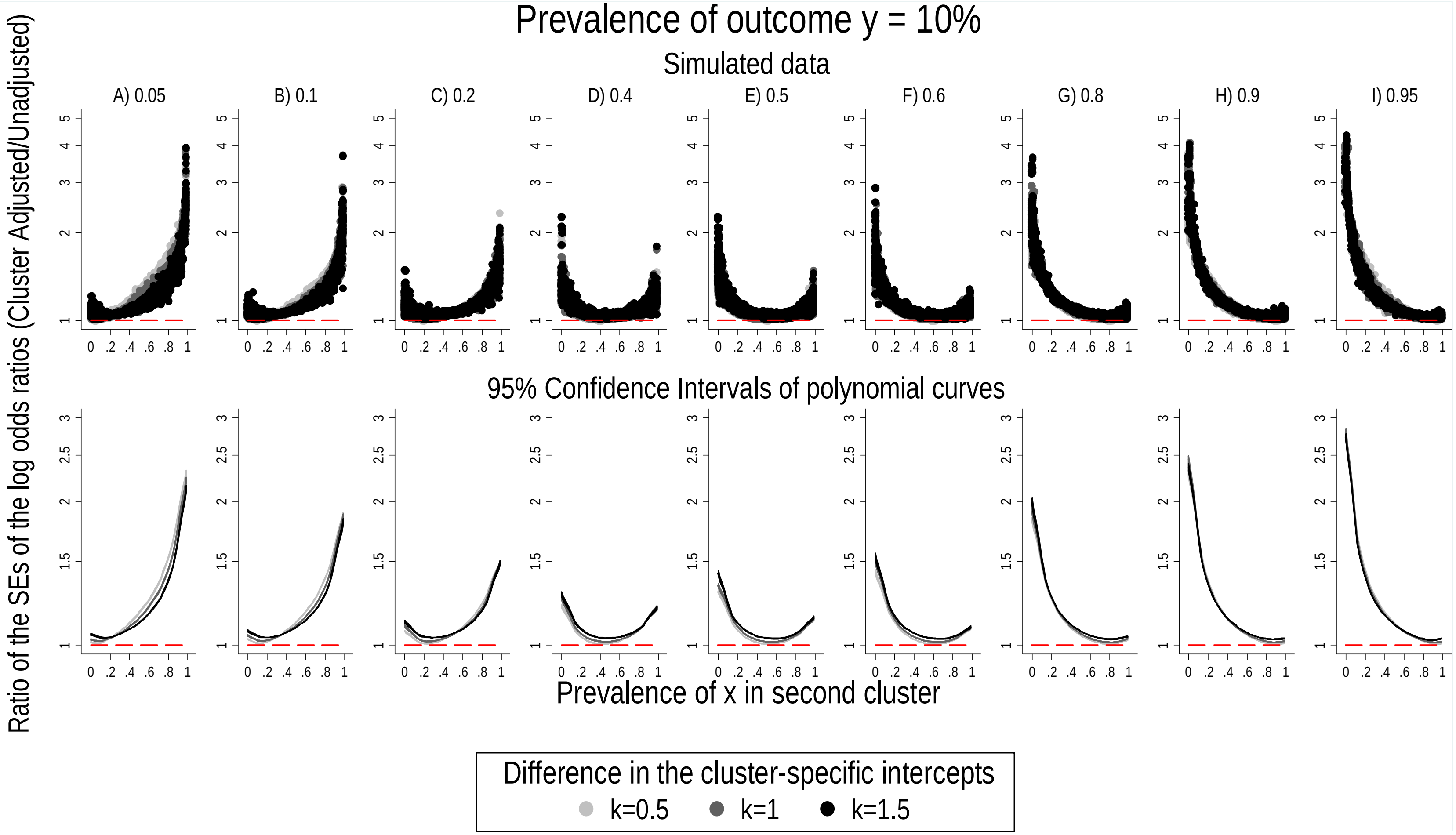


Figure 5.16. Ratios of the standard errors of the log odds ratios estimated from cluster-adjusted to those from cluster-unadjusted logistic regression models for varying prevalence of explanatory variable x in the second cluster keeping the prevalence of x in first cluster constant to A) 0.05, B) 0.1, C) 0. 2, D) 0.4, E) 0.5, F) 0.6, G) 0.8, H) 0.9, and I) 0.95 when the population prevalence of the outcome variable y was assumed to be 10%

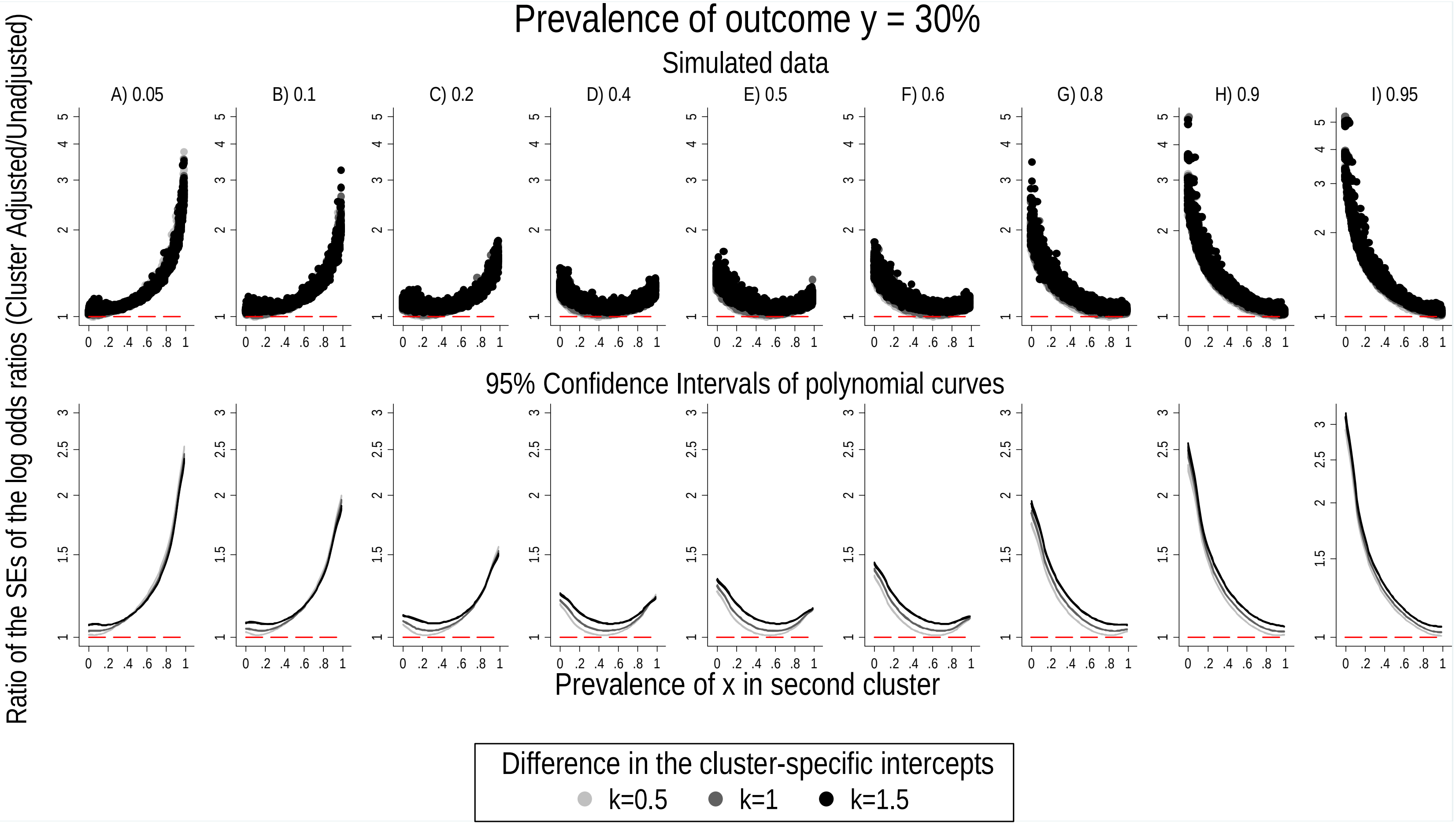


Figure 5.17. Ratios of the standard errors of the log odds ratios estimated from cluster-adjusted to those from cluster-unadjusted logistic regression models for varying prevalence of explanatory variable x in the second cluster keeping the prevalence of x in first cluster constant to A) 0.05, B) 0.1, C) 0. 2, D) 0.4, E) 0.5, F) 0.6, G) 0.8, H) 0.9, and I) 0.95 when the population prevalence of the outcome variable y was assumed to be 30%

Coverage of 95% Confidence Intervals

Coverage by 95% CIs from the cluster-unadjusted method when OR=2, according to different values for the cluster-specific intercepts k, prevalence rates of x in the first cluster, and fifths of the distribution of x in the second cluster, is illustrated in Figure 5.18. Numerical values of the coverage for a selected number of prevalence rates of x in the first cluster are also presented in Table 5.11.

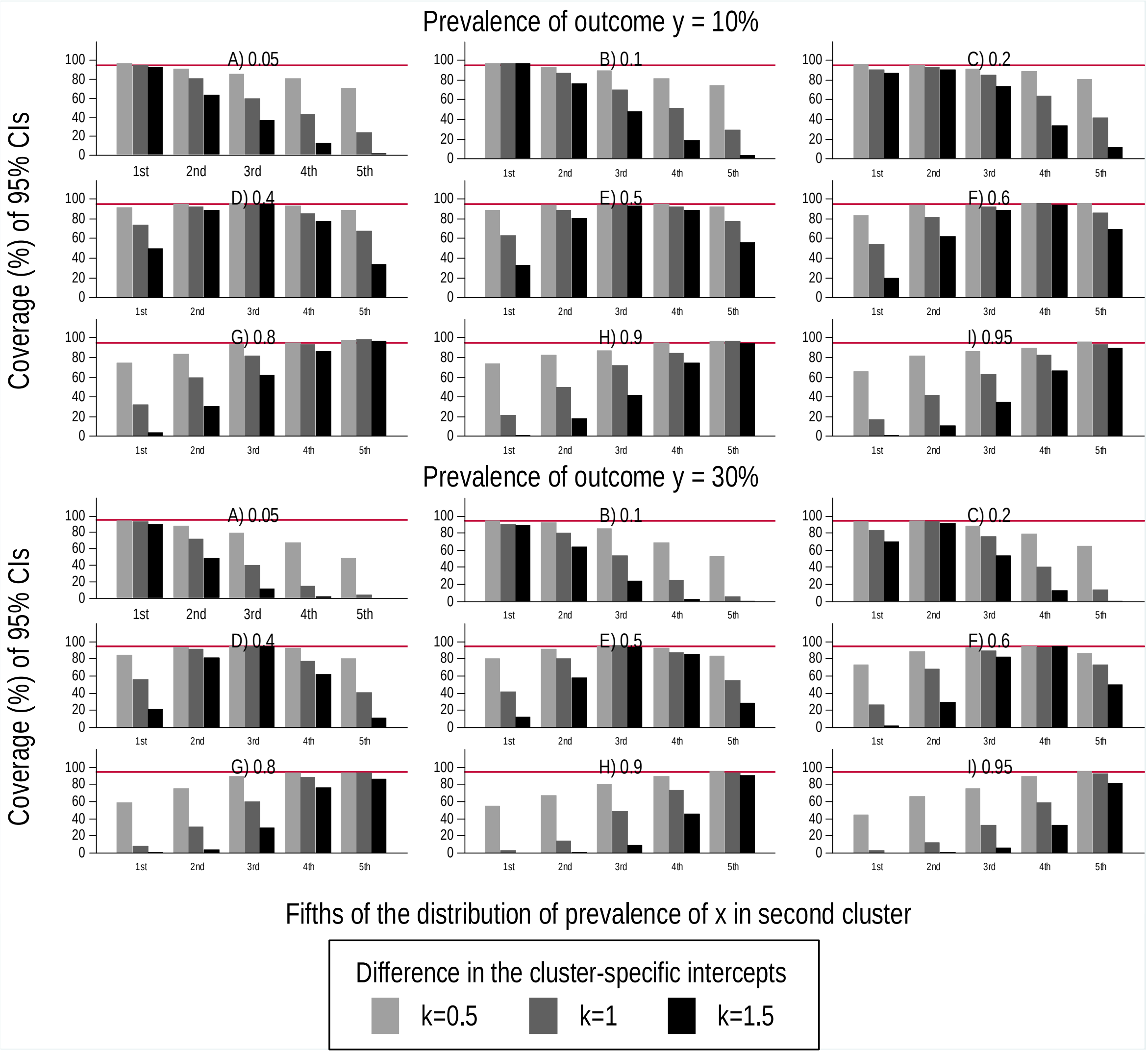


Figure 5.18. Coverage (%) by 95% confidence intervals of simulated effect OR=2 for the cluster-unadjusted and cluster-adjusted models according to difference between the cluster-specific intercepts k, prevalence rates of x in the first cluster (A) 0.05, B) 0.1, C) 0. 2, D) 0.4, E) 0.5, F) 0.6, G) 0.8, H) 0.9, and I) 0.95), and fifths of the distribution of x in the second cluster. The horizontal red line corresponds to the nominal value of 95%.

Table 5.11. Coverage (%) by 95% confidence intervals of simulated effect OR=2 for the cluster-unadjusted and cluster-adjusted models according to difference between the cluster-specific intercepts k, selected prevalence rates of x in the first cluster, and fifths of the distribution of x in the second cluster

	Prevalence of x in the 1st cluster	**Difference in cluster-specific intercepts**	**Fifths of the prevalence of x in the 2nd cluster**				
			1st	**2nd**	**3rd**	**4th**	**5th**
Prevalence of outcome 10%	**0.05**	$\boldsymbol{k = 0.5}$	96.5	90.8	85.3	80.8	70.6
	0.05	$\boldsymbol{k = 1}$	94.9	80.8	59.4	43.2	23.8
	0.05	$\boldsymbol{k = 1.5}$	92.7	63.4	37.1	13.2	2.0
	0.5	$\boldsymbol{k = 0.5}$	88.8	93.6	94.3	94.8	92.0
	0.5	$\boldsymbol{k = 1}$	63.2	88.8	93.9	92.0	77.0
	0.5	$\boldsymbol{k = 1.5}$	33.3	81.2	93.3	88.6	56.3
	0.95	$\boldsymbol{k = 0.5}$	66.1	82.2	86.5	90.0	95.4
	0.95	$\boldsymbol{k = 1}$	17.6	41.6	62.9	82.4	92.9
	0.95	$\boldsymbol{k = 1.5}$	1.2	10.8	34.9	66.6	90.0
Prevalence of outcome 30%	**0.05**	$\boldsymbol{k = 0.5}$	94.5	88.0	79.6	67.8	48.0
	0.05	$\boldsymbol{k = 1}$	93.0	71.4	40.4	14.4	4.2
	0.05	$\boldsymbol{k = 1.5}$	90.0	49.0	11.4	1.6	0.0
	0.5	$\boldsymbol{k = 0.5}$	80.6	91.0	94.5	93.0	83.2
	0.5	$\boldsymbol{k = 1}$	41.3	80.6	95.2	87.2	54.8
	0.5	$\boldsymbol{k = 1.5}$	11.8	57.6	93.3	85.4	28.8
	0.95	$\boldsymbol{k = 0.5}$	44.4	66.2	75.4	89.6	95.4
	0.95	$\boldsymbol{k = 1}$	2.5	12.2	32.4	59.0	92.6
	0.95	$\boldsymbol{k = 1.5}$	0.0	0.2	6.3	32.0	81.5

Overall, coverage from the method that did not adjust for clustering was poor when the difference between the prevalence rates of the explanatory variable in the two clusters was high (either low prevalence of x in the first cluster and top fifth of the distribution of x in the second cluster, or high prevalence of x in the first cluster and bottom fifth of the distribution of x in the second cluster), and for large differences in the cluster-specific intercepts, k. Coverage was particularly poor (<30%) for combinations of prevalence of x in the first cluster of 0.05-0.20 and in the top two fifths of the distribution of prevalence of x in the second cluster, and for combinations of prevalence of x in the first cluster of 0.60-0.95 and the bottom two fifths of the distribution of prevalence of x in the second cluster. As the difference between cluster-specific prevalence rates of x decreased, coverage approached the nominal level of 95%. For any given combination of prevalence rates of x in the two clusters and k, coverage was poorer in the simulated datasets with a population prevalence rate of y of 30%. Also, keeping all other parameters constant, coverage was lower for higher values of k. By contrast, coverage by 95% CIs from the model that adjusted for clustering was close to the nominal level of 95%, irrespective of any difference between prevalence rates of x in the two clusters, prevalence of y, and k. The average coverage for all combinations of the above parameters, was 95.3% and varied little (range: 93% to 97.9%). Data on coverage from the model that adjusted for clustering are not shown.

Type I error

To explore type I error, the simulated log odds ratio was set to zero (i.e. OR=1). Type I errors from the cluster-unadjusted model for different values of k, prevalence rates of x in the first cluster, and fifths of the distribution of prevalence of x in the second cluster, are illustrated in Figure 5.19. Table 5.12 presents the findings numerically, but only for selected values of prevalence of x in the first cluster (0.05, 0.5, and 0.95). Type I error from the model that did not adjust for clustering was approximately at the level of 5% when prevalence rates of x were similar in the two clusters and $k = 0.5$. Increasing the difference in prevalence rates of x increased type I error, which reached a level of 100% for the combinations of low/high prevalence of x in cluster 1 and top/bottom fifth of the distribution of prevalence of x in cluster 2, respectively. Keeping a constant difference between clusters in the prevalence of x, and a constant prevalence of the outcome variable, y, increasing difference between cluster-specific intercepts (k) increased type I error. Moreover, type I error was somewhat higher for higher prevalence of y, for any given combination of prevalence rates of x, and value of k.

Table 5.12. Proportion (%) of datasets for which the null hypothesis was rejected under the cluster-unadjusted model when true effect size (log odds ratio) of the explanatory on the outcome variable was assumed to be 0 (or OR=1), according to cluster-specific intercept k, for selected prevalence rates of x in the first cluster and fifths of the distribution of x in the second cluster

	Prevalence of x in the first cluster	**Difference in cluster-specific intercepts**	**Fifths of the prevalence of x in the second cluster**				
			1st	**2nd**	**3rd**	**4th**	**5th**
Prevalence of outcome 10%	**0.05**	$\mathbf{k = 0.5}$	5.1	6.4	11.4	17.5	26.4
	0.05	$\mathbf{k = 1}$	4.6	17.0	31.0	50.7	70.9
	0.05	$\mathbf{k = 1.5}$	7.9	31.8	55.6	76.2	96.2
	0.5	$\mathbf{k = 0.5}$	9.7	7.2	5.2	6.9	11.2
	0.5	$\mathbf{k = 1}$	27.8	7.6	5.0	9.9	28.0
	0.5	$\mathbf{k = 1.5}$	51.7	14.0	5.6	13.9	49.1
	0.95	$\mathbf{k = 0.5}$	24.1	19.4	11.2	9.3	5.5
	0.95	$\mathbf{k = 1}$	74.1	51.0	26.8	16.6	6.5
	0.95	$\mathbf{k = 1.5}$	95.7	79.8	53.4	28.2	7.9
Prevalence of outcome 30%	**0.05**	$\mathbf{k = 0.5}$	6.3	15.0	21.0	32.0	51.6
	0.05	$\mathbf{k = 1}$	9.1	32.6	64.8	83.8	96.4
	0.05	$\mathbf{k = 1.5}$	15.9	58.4	92.6	98.1	100.0
	0.5	$\mathbf{k = 0.5}$	14.9	5.8	5.2	6.3	15.8
	0.5	$\mathbf{k = 1}$	56.4	12.6	6.6	16.8	50.4
	0.5	$\mathbf{k = 1.5}$	81.1	23.8	4.8	28.2	81.4
	0.95	$\mathbf{k = 0.5}$	51.0	32.2	20.6	10.7	7.8
	0.95	$\mathbf{k = 1}$	96.9	84.8	64.6	34.1	7.4
	0.95	$\mathbf{k = 1.5}$	100.0	98.4	89.0	55.4	11.6

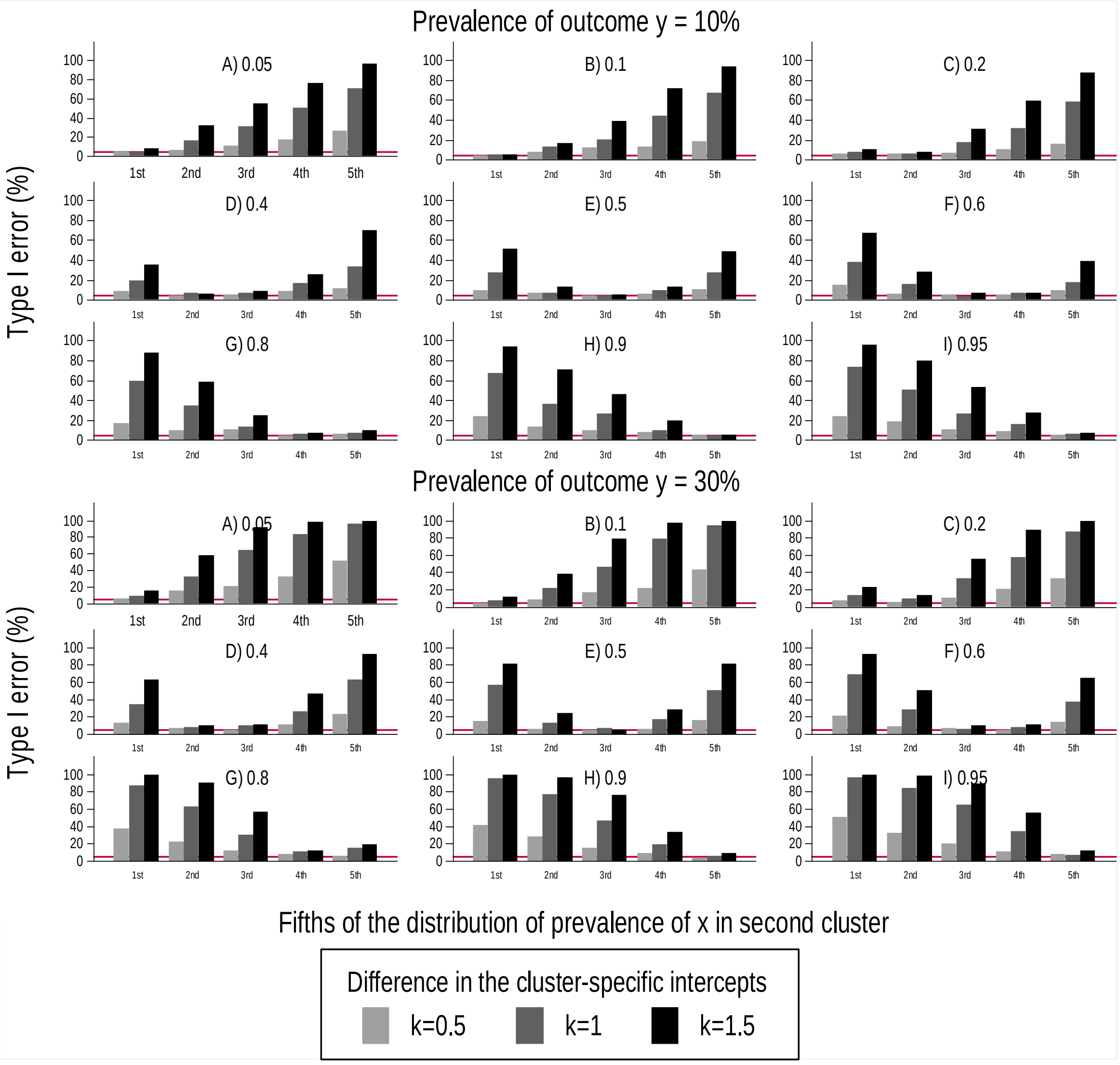


Figure 5.19. Proportion (%) of simulated datasets for which the null hypothesis under the cluster-unadjusted model was rejected, when the true effect size (log odds ratio) of the explanatory on the outcome variable was assumed to be 0 (or OR=1), by prevalence of x in the first cluster (A) 0.05, B) 0.1, C) 0. 2, D) 0.4, E) 0.5, F) 0.6, G) 0.8, H) 0.9, and I) 0.95), fifths of the distribution of the prevalence rates of x in the second cluster, and for two prevalence rates of the outcome variable y. The horizontal red line corresponds to the nominal value of 5%.

Type I error calculated from the cluster-adjusted models was very close to 5% (average value 4.7% across all combinations of prevalence rates of x, y, and values of k) and varied from 2.4 to 7.6. No consistent pattern of variation in type I error from the cluster-adjusted model was apparent according to prevalence rates of x, y, and values of k (data not shown).

5.5 Multiple clusters – Binary explanatory variable

5.5.1 Methods

In the case of a single binary explanatory variable, the ordinary logistic model that ignores clustering is described in equation 5.1, while the random intercept logistic regression model is given in equation 5.5, with x_{ij} a variable that follows a binomial distribution.

The methods used for the case of a binary outcome and a binary explanatory variable were similar to those described in section 5.3.1. Each simulated dataset consisted of 10000 observations nested evenly in 100 clusters. Initially, the prevalence of the binary explanatory variable x_{ij} in each cluster was defined as the sum of a constant (overall prevalence across clusters) and a cluster-specific variable $shift_j \sim N(0, SD_{shift}^2)$. Values of 0 and 1 were then assigned to each individual based on the prevalence of x_{ij} in the cluster. Then the cluster-level error term u_j was generated from a standard normal distribution of mean zero and variance SD_u^2. With the effect of x_{ij} on y_{ij} arbitrarily set to $\ln(2)$ (i.e. OR=2), and the log odds β_0 to $\ln\left(\frac{Prevalence_y}{1-Prevalence_y}\right)$, for a given prevalence of the outcome variable, I calculated the predicted probability of the outcome variable for each individual as per equation 5.6. A random variable l was then drawn from the uniform distribution $U(0,1)$. For each individual i, the outcome was positive (value of 1) if the predicted probability was higher than the value of l for that individual, and negative (value of 0) otherwise.

The simulations were repeated for different values of: overall prevalence of the explanatory variable, SD_{shift}, and SD_u. As in section 5.3.1, the values for overall prevalence of x_{ij} were 0.05, 0.1, 0.2, and 0.4. The values for SD_{shift} were drawn from a random uniform distribution $U[a,b]$, with the parameters a and b arbitrarily set to be 0 and 0.3, respectively. The values for SD_u were such, that the targeted values (0.001, 0.003, 0.01, 0.03, 0.1, and 0.3) of ICC were generated (based on the ICC definition for the logistic regression model). Simulated data were produced for low (10%) and high (30%) population prevalence of y_{ij}. For each combination of the above, 50 simulated datasets were generated.

For each simulated dataset, the estimates of effect and the corresponding SEs from the two regression methods were saved. The difference in the effect estimates ($\beta_1^{RI} - \beta_1^{OL}$), and the ratio of their SEs ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OL}}$) were described in relation to ICC values, population prevalence of y_{ij}, overall prevalence of the explanatory variable, and dispersion of the prevalence of the explanatory variable across clusters. Also, the coverage by 95% confidence intervals was summarised for different combinations of these parameters.

To explore effects on type I error, simulations were repeated after setting the effect estimate to zero ($\beta_1 = 0$, i.e. OR=1), and then calculating the proportion of simulated datasets for which the null hypothesis was rejected at a 5% significance level.

5.5.2 Results

Differences in log odds ratios

Figure 5.20 plots differences in the log odds ratios ($\beta_1^{RI} - \beta_1^{OL}$) estimated from the RI logistic regression and the OL regression models that did not adjust for clustering against the dispersion of the prevalence of the explanatory variable x_{ij} across clusters. The sub-plots in the top two rows of Figure 5.20 (top half of the figure) present data for an assumed population prevalence of the outcome variable y_{ij} of 10%, and the sub-plots in the bottom two rows of the figure (bottom half) for population prevalence of y_{ij} of 30%. For each of the two cases of prevalence of y_{ij}, different prevalence rates of the explanatory variable x_{ij} were explored (A) 0.05, B) 0.1, C) 0.2, and D) 0.4). For each combination of prevalence rates of y_{ij} and x_{ij}, both the simulated data and the 95% CIs of the polynomial curves are presented. As in previous chapters, in each sub-plot of the figure, the different levels of within-cluster similarity of observations (ICCs) are shown in different shades of grey, with darker shades depicting higher values of ICC. The average and the range of differences between the log odds ratios estimated from the two methods according to ICC, overall prevalence of x_{ij}, and population prevalence of y_{ij} are also presented in Table 5.13.

On average, differences were small for low values of ICC (≤0.03), for any prevalence of y_{ij} and x_{ij}, and dispersion of the cluster-specific prevalence of x_{ij}. Increasing ICC and population prevalence of the outcome variable increased the average difference in the log odds ratios. Also, differences were somewhat larger for higher prevalence of the outcome variable for any given ICC and prevalence of x_{ij}. For prevalence of y_{ij}=10% and ICC=0.3, the average difference in the log odds was 0.12, reflecting an average estimated OR of 2.01 under the RI model and 1.79 under the OL model. The corresponding average estimates for prevalence of y_{ij}=30% and ICC=0.3 were OR=2 from the RI model and OR=1.73 from the OL model, resulting in an average difference in the log odds ratios of 0.15. Unlike the average difference in the log odds ratios, which remained constant for any given ICC and prevalence of x_{ij} and y_{ij} (approximately flat polynomial curves), the dispersion of the differences increased with increasing dispersion of the cluster-specific prevalence of x_{ij} across the clusters (depicted in the wider 95% CIs of the polynomial curves). Comparison of results for different prevalence rates of x_{ij} (Table 5.13), when the rest of the parameters (prevalence of y_{ij}, and ICC) remained constant, showed that increasing the overall prevalence of x_{ij} did not change the average difference in the log odds ratios. However, increasing the overall prevalence of x_{ij} decreased the range of differences. This pattern was more apparent for higher values of ICC; for ICC=0.3, differences ranged from -0.39 to 0.59 when overall prevalence of x_{ij}=0.05, while they ranged from -0.24 to 0.45 when overall prevalence of x_{ij}=0.40.

Ratio of standard errors

As for the differences in the log odds ratios estimated from the two methods, the ratios of their SEs ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OL}}$) were examined in relation to the prevalence of y_{ij} and x_{ij}, dispersion of the prevalence of x_{ij} across clusters, and levels of ICC. The ratios of the SEs are plotted against dispersion of the prevalence of x_{ij} in Figure 5.21. As in the corresponding figure for the differences in the log odds ratios (Figure 5.20), simulated data and the 95% CIs of the polynomial curves are presented separately for the four different prevalence rates of x_{ij} and the two different population prevalence rates of y_{ij} (in different sub-plots), and for the different levels of ICC (shown in different shades within each sub-plot).

Overall the ratios of the SEs were above one, meaning that for all prevalence rates of the two main variables and their relative distributions across clusters, the SEs from the RI models were higher than the SEs from the OL models. The ratio of the SEs was associated with the ICC level; for constant prevalence rates of x_{ij} and y_{ij}, and dispersion of the prevalence of x_{ij} across clusters, higher ICCs resulted in higher ratios of SEs of the log odds ratios (comparison of different shades of grey in any sub-plot of Figure 5.21). This observation indicates that as within-cluster similarity of observations increased, imprecision of β_1^{RI} increased more than imprecision of β_1^{OL}. For any constant ICC, and combination of prevalence rates of x_{ij} and y_{ij}, the ratios of SEs increased as the dispersion of prevalence of x_{ij} across clusters increased.

The rate of increase depended on the ICC level; ratios of SEs increased faster with increasing dispersion of prevalence of x_{ij} across clusters for higher ICCs than for lower ICCs (fan-shaped pattern in Figure 5.21). Keeping ICC and the overall prevalence of x_{ij} constant, higher population prevalence of y_{ij} (comparison of top with bottom half of Figure 5.21) resulted in higher ratios of SEs.

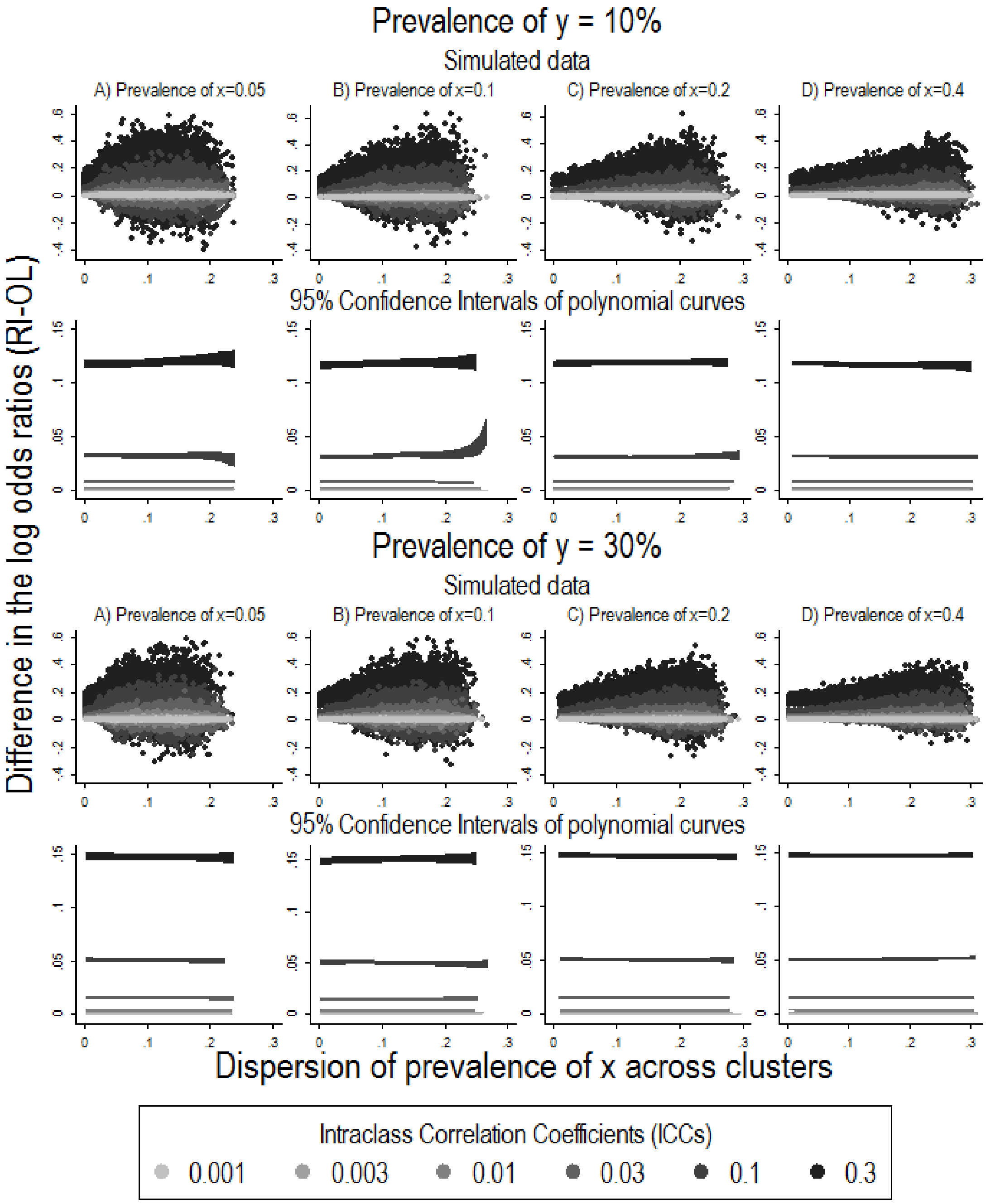


Figure 5.20. Difference between log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OLS}$) according to dispersion of prevalence of explanatory variable x_{ij} across clusters, for different levels of intraclass correlation (shades of grey as indicated in the legend) and for the two different population prevalence rates of the outcome variable

Table 5.13. Descriptive statistics for differences in log odds ratios estimated from RI and OL models ($\beta_1^{RI} - \beta_1^{OLS}$) according to ICC, overall prevalence of x_{ij}, and population prevalence of y_{ij}

		Overall prevalence of x_{ij}=0.05		**Overall prevalence of x_{ij}=0.1**		**Overall prevalence of x_{ij}=0.2**		**Overall prevalence of x_{ij}=0.4**	
	ICC	**Mean**	**Range**	**Mean**	**Range**	**Mean**	**Range**	**Mean**	**Range**
Prevalence of y_{ij} = 10%	0.001	0.000	-0.006 to 0.006	0.000	-0.006 to 0.006	0.000	-0.004 to 0.004	0.000	-0.004 to 0.004
	0.003	0.001	-0.014 to 0.017	0.001	-0.013 to 0.013	0.001	-0.011 to 0.013	0.001	-0.008 to 0.010
	0.01	0.002	-0.048 to 0.042	0.002	-0.043 to 0.049	0.002	-0.029 to 0.040	0.002	-0.027 to 0.036
	0.03	0.009	-0.116 to 0.111	0.008	-0.108 to 0.100	0.009	-0.077 to 0.101	0.008	-0.060 to 0.093
	0.1	0.032	-0.287 to 0.321	0.033	-0.197 to 0.313	0.031	-0.178 to 0.271	0.032	-0.136 to 0.209
	0.3	0.119	-0.394 to 0.589	0.118	-0.374 to 0.636	0.118	-0.329 to 0.623	0.116	-0.244 to 0.446
Prevalence of y_{ij} = 30%	0.001	0.001	-0.007 to 0.009	0.001	-0.006 to 0.008	0.001	-0.005 to 0.009	0.001	-0.005 to 0.006
	0.003	0.002	-0.019 to 0.022	0.002	-0.015 to 0.021	0.002	-0.014 to 0.017	0.002	-0.009 to 0.015
	0.01	0.004	-0.047 to 0.065	0.004	-0.055 to 0.058	0.004	-0.037 to 0.050	0.004	-0.028 to 0.044
	0.03	0.016	-0.112 to 0.207	0.016	-0.103 to 0.150	0.016	-0.093 to 0.109	0.016	-0.087 to 0.109
	0.1	0.051	-0.260 to 0.359	0.050	-0.224 to 0.272	0.051	-0.144 to 0.244	0.052	-0.118 to 0.229
	0.3	0.147	-0.306 to 0.589	0.151	-0.331 to 0.591	0.147	-0.271 to 0.536	0.148	-0.169 to 0.427

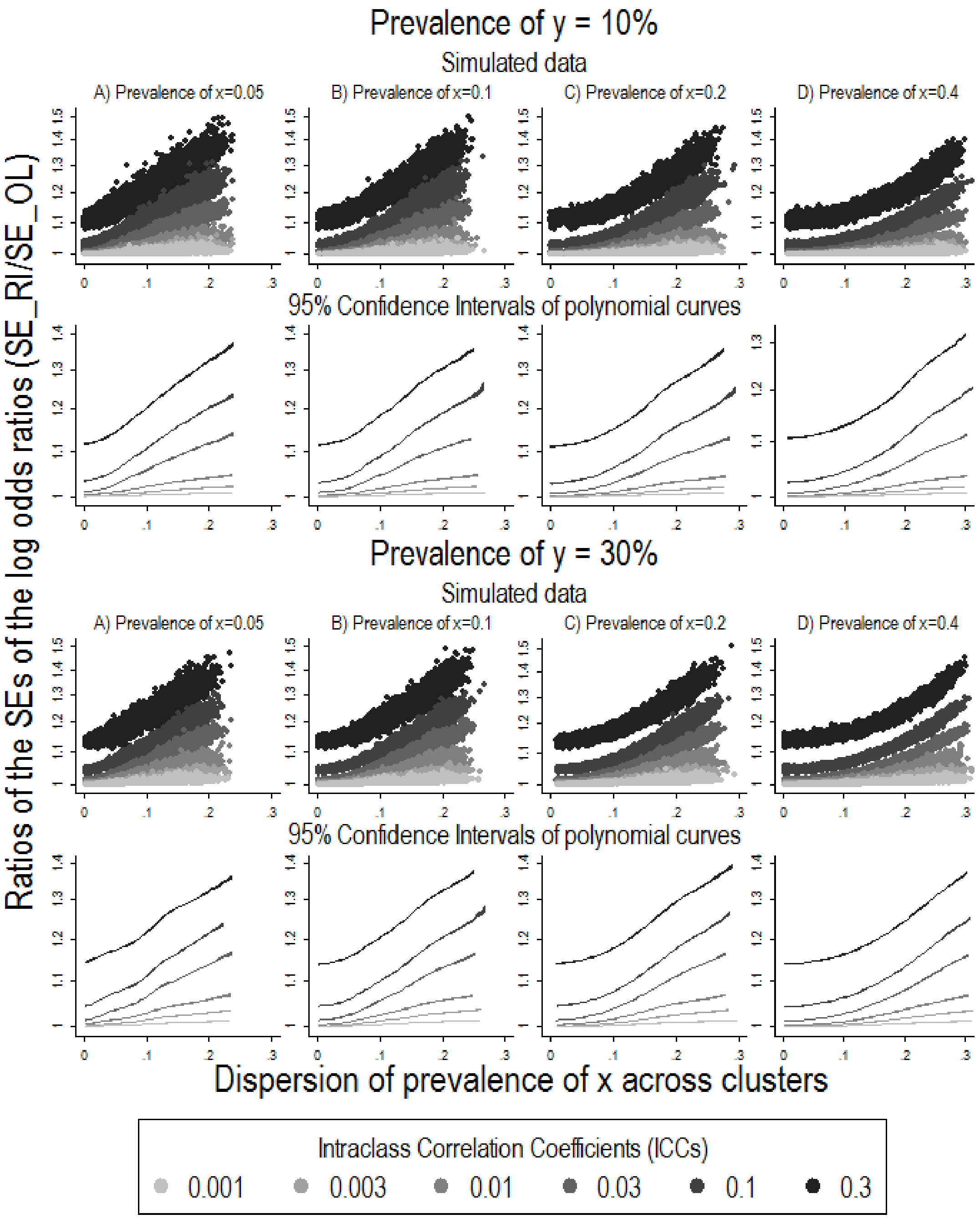


Figure 5.21. Ratio of the standard errors of the log odds ratios estimated from RI and OL models ($\beta_1^{RI}/\beta_1^{OLS}$) according to dispersion of prevalence of explanatory variable x_{ij} across clusters, for different levels of intraclass correlation (shades of grey as indicated in the legend) and for the two different population prevalence rates of the outcome variable

To explore the effect of increasing the overall prevalence of x_{ij} for constant ICC and population prevalence of y_{ij}, the ratios of SEs were plotted against the dispersion of the prevalence of x_{ij} across clusters for ICC=0.3 and for population prevalence of y_{ij}=10% (Figure 5.22). In the left-hand sub-plot of Figure 5.22, the simulated data are plotted out, while the right-hand sub-plot shows the 95% CIs of the polynomial curves. The ratios of the SEs were very similar for the different cluster-specific prevalence rates of x_{ij} when the dispersion of the prevalence rates across clusters was small. However, they increased more rapidly with increasing dispersion of prevalence rates of x_{ij} when the overall prevalence rates of x_{ij} were lower than when they were higher.

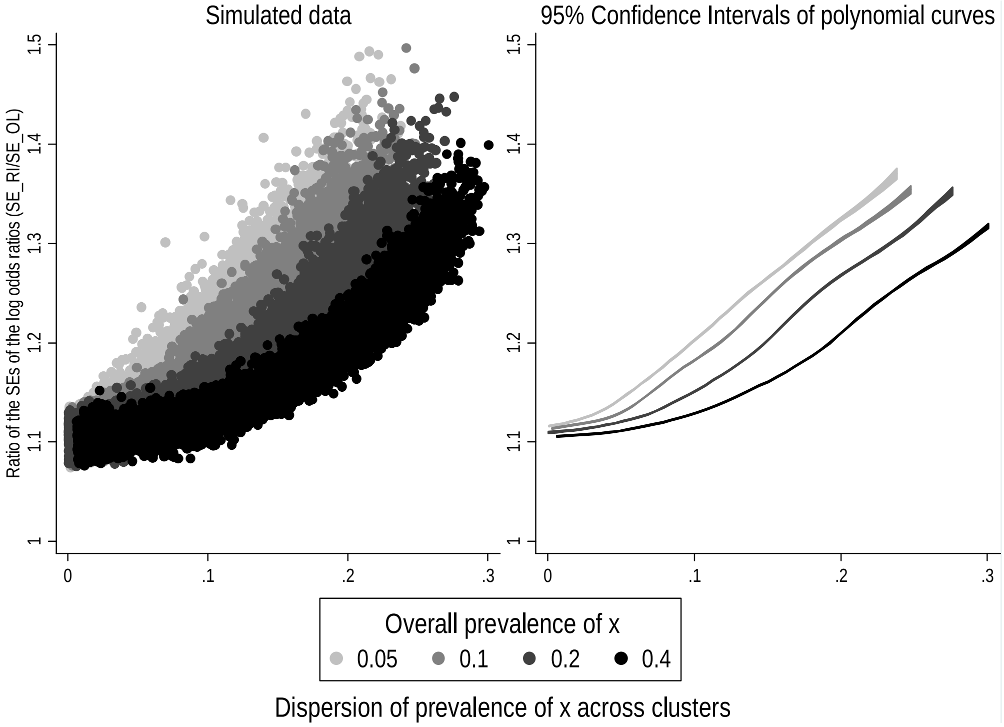


Figure 5.22. Ratio of the standard errors of the log odds ratios estimated from RI and OL models ($\beta_1^{RI}/\beta_1^{OLS}$) according to dispersion of prevalence of explanatory variable x_{ij} across clusters, for different prevalence rates of explanatory variable x_{ij}, for ICC=0.3 and population prevalence of y_{ij}=10%

Coverage by 95% Confidence Intervals

Table 5.14 describes coverage of the simulated OR=2 by the 95% CIs estimated from the OL regression models, according to quarters of the distribution of dispersion of cluster-specific prevalence rates of x_{ij}, ICC levels, overall prevalence of x_{ij}, and population prevalence of y_{ij}.

Table 5.14. Coverage (%) by 95% confidence intervals of simulated effect OR=2 from the OL model according to quarters of the distribution of the dispersion of prevalence of x_{ij} across clusters, ICC, overall prevalence of x_{ij} and population prevalence rate of y_{ij}

	ICC	**Quarters of the distribution of dispersion of x_{ij} across clusters**							
		First	**Second**	**Third**	**Forth**	**First**	**Second**	**Third**	**Forth**
		Overall prevalence of x_{ij} = 0.05				**Overall prevalence of x_{ij} = 0.1**			
Population prevalence of y_{ij}=10%	0.001	94.4	94.9	95.0	95.6	94.1	95.5	95.9	94.6
	0.003	95.1	94.9	95.1	95.9	95.6	94.7	94.7	95.0
	0.01	95.0	95.4	95.8	96.0	95.6	95.4	95.5	95.0
	0.03	94.9	93.7	91.4	89.1	94.5	92.8	91.8	90.2
	0.1	92.6	87.8	82.2	78.8	92.7	88.1	83.4	77.7
	0.3	78.9	63.3	52.0	48.6	69.0	58.5	54.3	47.1
		Overall prevalence of x_{ij} = 0.2				**Overall prevalence of x_{ij} = 0.4**			
	0.001	95.2	95.6	93.9	95.3	94.3	95.3	95.2	95.1
	0.003	95.6	94.7	94.5	95.4	93.8	95.3	95.2	94.9
	0.01	94.3	94.3	96.2	95.0	94.9	95.0	94.5	95.2
	0.03	95.0	94.4	92.9	92.0	95.8	93.6	94.1	92.0
	0.1	93.0	89.7	85.7	80.7	90.8	89.1	86.2	81.5
	0.3	52.5	51.6	50.3	44.6	39.6	39.9	43.0	43.6
		Overall prevalence of x_{ij} = 0.05				**Overall prevalence of x_{ij} = 0.1**			
Population prevalence of y_{ij}=30%	0.001	95.2	94.6	94.5	96.4	95.5	94.4	94.9	94.5
	0.003	94.9	94.9	94.7	93.5	94.9	94.8	94.3	94.6
	0.01	94.8	94.5	94.1	92.6	94.4	95.3	94.1	94.3
	0.03	94.2	92.6	90.3	86.4	94.9	91.4	89.4	87.6
	0.1	88.6	80.2	74.4	65.0	88.8	80.2	71.8	66.4
	0.3	60.4	49.7	40.7	37.4	40.6	40.4	37.1	33.6
		Overall prevalence of x_{ij} = 0.2				**Overall prevalence of x_{ij} = 0.4**			
	0.001	94.1	95.2	96.3	94.7	95.6	94.5	95.1	94.8
	0.003	95.7	94.9	94.9	93.6	94.6	95.4	94.6	94.8
	0.01	95.3	95.2	93.6	93.5	94.3	94.0	95.0	93.9
	0.03	94.8	91.3	90.3	86.7	93.5	93.5	91.1	87.6
	0.1	83.6	79.0	75.8	68.0	79.2	74.5	71.6	63.8
	0.3	17.0	23.2	30.3	30.4	4.8	9.2	14.2	23.8

When the OL model was fitted, coverage was approximately 95% when the ICC was low (ICC<0.01), for any part of the distribution of prevalence of x_{ij} across clusters, and all prevalence rates of x_{ij} and y_{ij}. For any combination of prevalence rates x_{ij} and y_{ij}, and quarter of the distribution of dispersion of cluster-specific rates of x_{ij}, increasing ICC decreased coverage by 95% CIs. The rate of decrease in coverage depended on the overall prevalence of x_{ij} and the population prevalence of y_{ij}, as well as the dispersion of prevalence of x_{ij} across clusters. For any prevalence of the outcome variable, the decrease in coverage for increasing ICC was smaller when the overall prevalence of x_{ij} was lower. Moreover, for constant quarter of the distribution of dispersion of the cluster-specific prevalence of x_{ij}, higher compared to lower prevalence of y_{ij} resulted in considerably poorer coverage, especially when the overall prevalence of x_{ij} was high, and when the ICC was 0.3. Interestingly, for low prevalence of y_{ij} and overall prevalence of x_{ij} = 0.4, the rate of decrease in coverage (for increasing ICC) was higher for the bottom quarter of the distribution of dispersion of prevalence rates of x_{ij} than it was at the top quarter.

The same was observed when the prevalence of y_{ij} was high and the overall prevalence of x_{ij} was 0.2 or 0.4 (Figure 5.23). In all other combinations of prevalence rates of the outcome and the explanatory variables, coverage decreased more with increasing ICC for the top quarter of the distribution of dispersion of the cluster-specific prevalence of x_{ij} than for the bottom. The lowest coverage was only 5% and it occurred when the prevalence of y_{ij} was high, the overall prevalence of x_{ij} was 0.4, ICC=0.3, and for the top quarter of the distribution of dispersion of prevalence rates of x_{ij}.

Coverage by 95% CIs estimated from RI logistic regression models was at the nominal level of 95%, and it varied very little (range: 94.6% to 95.7%) across different combinations of dispersion of cluster-specific rates of x_{ij}, prevalence rates of the outcome and the explanatory variable, and ICC levels (data not shown).

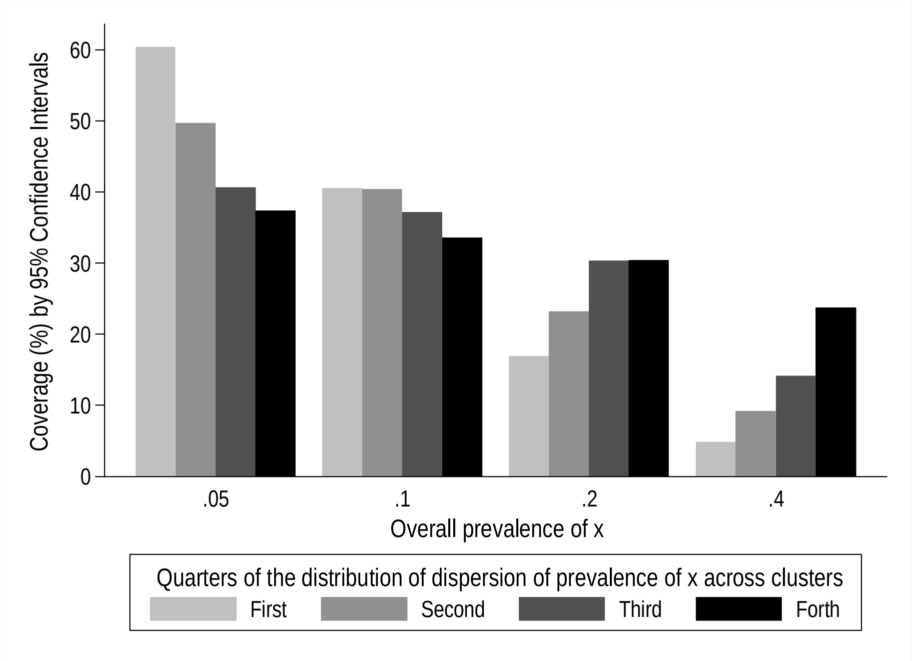


Figure 5.23. Coverage (%) by 95% confidence intervals of simulated effect OR=2 from the OL regression model for high prevalence of the outcome variable and for ICC=0.3, according to overall prevalence of explanatory variable x_{ij} (y-axis) and quarters of the distribution of dispersion of prevalence of x_{ij} across clusters (shown in different shades as per legend).

Type I error

Table 5.15 presents type I error from the OL model by quarters of the distribution of dispersion of prevalence of x_{ij} across clusters, overall prevalence rates of x_{ij} and y_{ij}, and by ICC. Type I error with the OL model when ICC was low (0.001 or 0.003) was very close to the nominal value of 5%; the average value for type I error in all combinations of the parameters examined was 5.1% (range: 3.8% to 7.1%). Increasing ICC (>0.003) increased type I error. The rate of increase of type I error with increasing ICC was higher when the prevalence of y_{ij} was higher, or the overall prevalence of x_{ij} was lower, or the dispersion of prevalence of x_{ij} across clusters was higher (top quarter compared to bottom quarter of the distribution). These associations are also illustrated in Figure 5.24, in which type I error is shown for different values of the parameters examined with ICC held at the highest value of 0.3.

Table 5.15. Proportion (%) of datasets for which the null hypothesis was rejected under the OL model according to quarters of the distribution of dispersion of prevalence of x_{ij} across clusters, ICC, and population prevalence of the outcome variable y_{ij}

	ICC	**Quarters of the distribution of dispersion of x_{ij} across clusters**							
		First	**Second**	**Third**	**Forth**	**First**	**Second**	**Third**	**Forth**
		Overall prevalence of x_{ij} = 0.05				**Overall prevalence of x_{ij} = 0.1**			
Population prevalence of y_{ij}=10%	0.001	4.6	5.0	5.0	4.7	4.2	4.6	4.5	5.1
	0.003	5.1	4.8	5.3	3.8	4.6	4.5	5.2	5.1
	0.01	5.9	4.4	5.1	3.1	5.1	5.0	5.5	4.0
	0.03	4.2	6.2	7.4	6.9	4.9	5.1	6.2	8.4
	0.1	5.5	9.9	15.3	15.9	5.1	9.4	12.5	16.6
	0.3	9.1	23.2	35.6	38.0	7.3	19.8	31.5	37.6
		Overall prevalence of x_{ij} = 0.2				**Overall prevalence of x_{ij} = 0.4**			
	0.001	4.9	5.5	5.5	5.1	4.2	5.9	5.0	4.9
	0.003	4.7	5.7	5.5	5.0	4.7	4.6	5.1	5.0
	0.01	4.9	4.4	5.2	5.3	4.1	5.2	5.3	4.6
	0.03	5.9	5.1	6.0	7.0	4.3	5.8	5.7	6.8
	0.1	4.8	6.3	11.6	14.6	5.1	6.5	9.0	13.5
	0.3	5.9	14.0	24.7	35.8	4.9	10.7	19.3	33.1
		Overall prevalence of x_{ij} = 0.05				**Overall prevalence of x_{ij} = 0.1**			
Population prevalence of y_{ij}=30%	0.001	5.4	5.0	4.8	4.2	5.7	5.6	4.6	4.1
	0.003	5.0	5.2	5.1	7.1	5.1	5.6	4.3	5.3
	0.01	5.4	5.9	5.8	5.2	4.0	5.6	5.1	7.5
	0.03	5.2	7.6	9.1	13.8	6.7	7.4	9.8	13.3
	0.1	7.3	14.7	20.0	25.6	5.8	11.9	18.1	24.0
	0.3	12.6	28.8	39.0	39.7	7.3	24.7	37.1	44.5
		Overall prevalence of x_{ij} = 0.2				**Overall prevalence of x_{ij} = 0.4**			
	0.001	6.2	6.0	5.5	6.3	4.6	4.8	5.4	5.3
	0.003	5.1	5.9	5.5	4.8	5.1	6.2	5.1	5.1
	0.01	4.9	5.0	6.0	5.9	6.2	5.7	4.3	5.4
	0.03	4.4	6.0	9.2	10.2	3.7	5.9	7.3	9.6
	0.1	4.9	8.5	15.5	21.6	3.4	7.5	13.2	19.5
	0.3	5.1	17.7	32.0	42.2	4.1	13.4	24.7	38.4

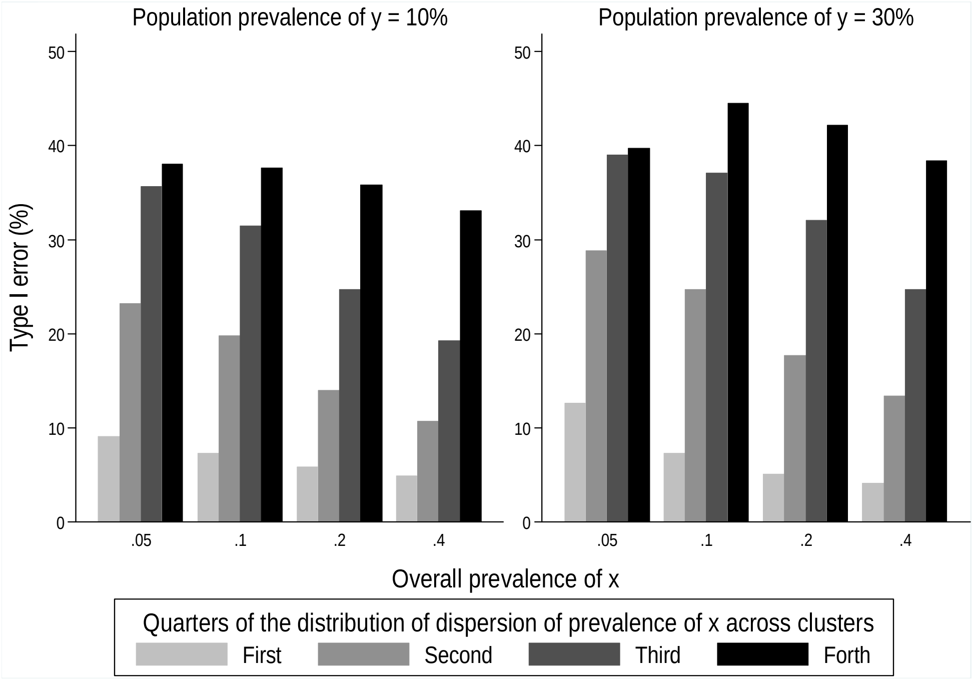


Figure 5.24. Proportion (%) of datasets for which the null hypothesis was rejected when ICC was 0.3 according to overall prevalence of explanatory variable x_{ij}, quarters of the dispersion of prevalence of x_{ij} across clusters and population prevalence rate of the outcome variable y.

The highest type I error was 44.5%, and it occurred for the top quarter of the distribution of dispersion of prevalence rates x_{ij} across clusters, overall prevalence of x_{ij} = 0.2, and for high prevalence of y_{ij}. Type I error was at the 5% level when overall prevalence of x_{ij} was high, prevalence of the outcome variable was low, and the dispersion of cluster-specific prevalence rates was small (bottom quarter of the distribution) (Figure 5.24). For high prevalence of y_{ij} and x_{ij}, and small dispersion of cluster-specific prevalence rates, type I error was below 5% when ICC was >0.01.

Type I error from the RI logistic regression model was on average 4.9% for the different combinations of prevalence rates of the two main variables, ICC levels, and dispersion of the prevalence rates of x_{ij} across clusters. The range of values was 3.1 to 7.4. No patterns of higher/lower values of type I error within that range (3.1 to 7.4) were observed in relation to the other parameters (data not shown).

5.5.3 Analysis of CUPID data

As in section 5.3.3, to explore whether observations on the consequences of ignoring clustering that were made using simulated data were supported by findings from real data, I used the binary variables from the CUPID study and I dichotomised those that were ordinal (by collapsing all but the null category together). This resulted in 116 binary variables (also including an ordinal variable of age band that was not used in section 5.3.3). Each of these variables served as an outcome and as an explanatory variable in univariate analysis models. For each combination of these binary variables (N=13,046 combinations after excluding combinations for which the outcome variable did not vary by levels of the explanatory variable, or vice versa), both an OL and a RI model were fitted. To ensure comparability of results with those from simulation studies, when for a given association the estimate of log odds ratio was positive under one of the two methods and negative under the other (N=1,080), estimates from that association were excluded from analyses presented here. From the remaining resulting estimates (N=11,966), I calculated the differences in log odds ratios and the ratios of the SEs. As in section 5.3.3, when log odds ratios from both analytical approaches were <0, the difference between them was replaced by the difference of their absolute values. That way, any positive/negative difference between estimates would indicate that the log odds ratio from the RI model was further/closer to the null value than that from the OL model.

The values of ICC estimated from the RI logistic models had a median value of 0.16 and varied from 0 to 0.72. The values of the estimated ICC are plotted in the histogram of Figure 5.25.

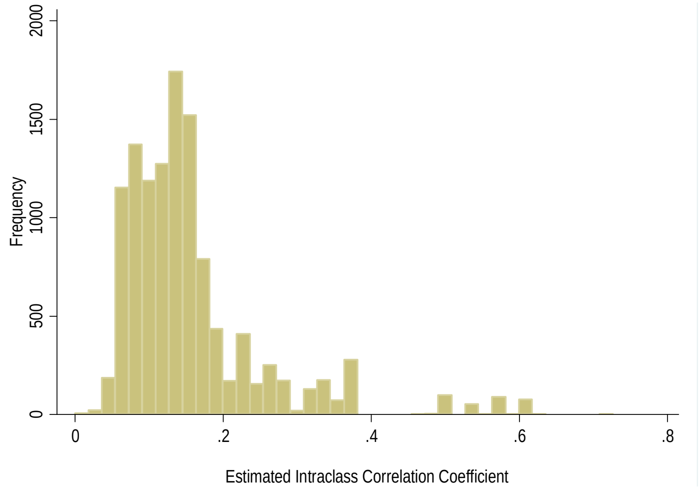


Figure 5.25. Distribution of estimated ICCs using data from the CUPID study

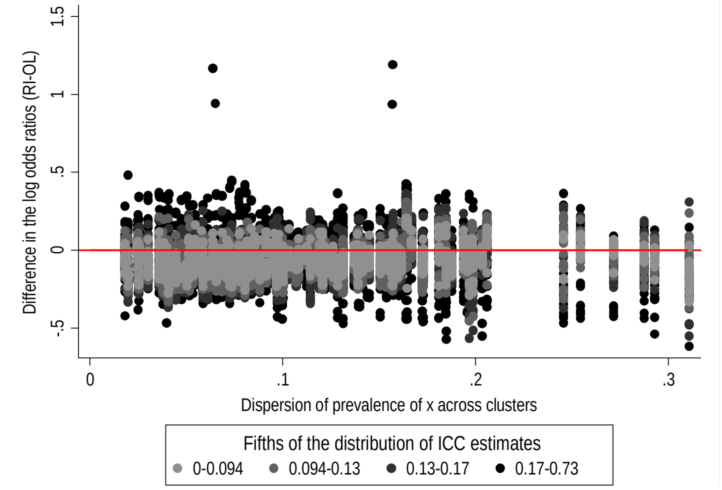


Figure 5.26. Differences in the log odds ratios ($\beta_1^{RI} - \beta_1^{OL}$) plotted against dispersion of the prevalence of x across clusters, for different levels of intraclass correlation (shades of grey as indicated in the legend)

Figure 5.26 shows the differences in log odds ratios estimated from the two models across increasing values of the dispersion of prevalence of x_{ij} across clusters, with darker shades of grey depicting higher estimated ICCs. As also seen in the analysis of simulated data, increasing differences in point estimates were observed for increasing ICC levels, and for increasing dispersion of the prevalence of the explanatory variables.

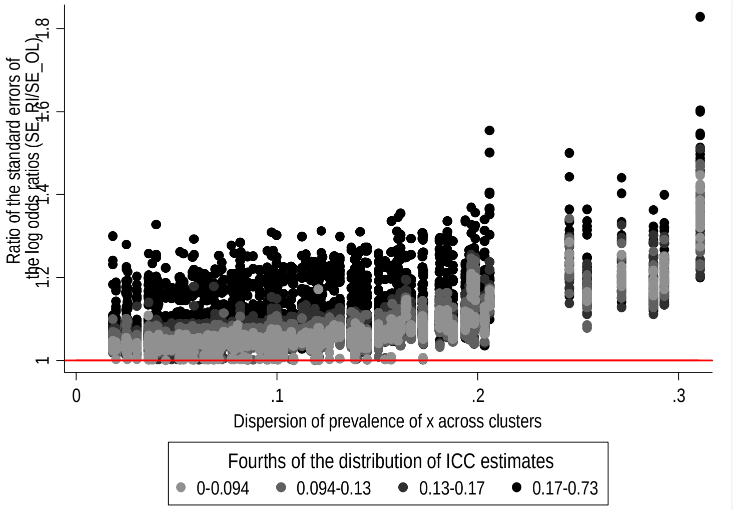


Figure 5.27. Ratios of standard errors of the log odds ratios estimated from RI and OL models ($\mathrm{SE}_{\beta_1^{RI}}/\mathrm{SE}_{\beta_1^{OL}}$) plotted against dispersion of explanatory variable x across clusters, for different levels of intraclass correlation (shades of grey as indicated in the legend)

Figure 5.27 shows the ratios of the SEs of the log odds ratios estimated from the two methods. The observation agrees with that from the analysis of simulated data (Figure 5.21); all ratios were >1 ($\mathrm{SE}_{\beta_1^{RI}} > \mathrm{SE}_{\beta_1^{\mathrm{OL}}}$), and increasing ICC, as well as increasing dispersion of x_{ij} across clusters increased the ratios of SEs to values further from 1.

5.6 Summary of results

5.6.1 Binary outcome - continuous predictor – 2 clusters

Adjusted point estimates were always lower than the unadjusted ones (i.e. closer to the null value). Difference between adjusted and unadjusted point estimates was greater when distance between the cluster-specific intercepts was bigger and when prevalence of the outcome variable was higher.

In most cases, SEs of the adjusted point estimates were larger than SEs of the unadjusted ones. The ratio of SEs of the adjusted to unadjusted ORs increased as the distance between the cluster-specific intercepts and the prevalence of the outcome variable increased.

Coverage by 95% CIs was 95% from the cluster-adjusted model and lower than 95% from the cluster-unadjusted model. Coverage decreased considerably as distance between the cluster-specific intercepts increased, and it was somewhat lower for higher prevalence of the outcome variable.

Type I error in the cluster-adjusted model was 5%, and higher than that in the cluster-unadjusted model. Type I error in the cluster-unadjusted model increased with increase in the distance between the cluster-specific intercepts, but was not affected by the prevalence of the outcome variable.

These results are further summarised in Table 5.16.

Table 5.16. Summary of simulation results when the outcome variable was binary, the explanatory variable was continuous, and observations were grouped in 2 clusters

$\beta_1^{Cluster-adjusted} < \beta_1^{Cluster-unadjusted}$	
$\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unadjusted}$	↑ when k ↑
	↑ when prevalence of y ↑
	↑ when $\lvert\bar{x}_{cluster\,1} - \bar{x}_{cluster\,2}\rvert$ approximately < 2SDs of x
	↓ when $\lvert\bar{x}_{cluster\,1} - \bar{x}_{cluster\,2}\rvert$ approximately > 2SDs of x

$SE^{cluster-adjusted} > SE^{cluster-unadjusted}$ $SE^{cluster-adjusted}/SE^{cluster-unadjusted}$ ↑ when k ↓ ↑ when prevalence of y ↑ ↑ when $\lvert\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\rvert$ ↑
Coverage by 95% CIs from the cluster-adjusted model: on average 95% Coverage by 95% CIs from the cluster-unadjusted model was < 95% and it ↓ when k ↑ ↓ when prevalence of y ↑
Type I error from the cluster-adjusted model: close to 5% Type I error from the cluster-unadjusted model: >5% and it ↑ when k ↑ ↓ when prevalence of y ↑

5.6.2 Binary outcome - continuous predictor – multiple clusters

ORs from OL were very similar to those from RI for ICC≤0.03. For ICC>0.03, increasing ICC increased differences between the ORs estimated from the two models, with ORs from OL being lower than those from RI (contrary to what would have been expected from the exploration of this research question in the case of two clusters). Differences were somewhat smaller for increasing dispersion of cluster-mean values of the explanatory variable. Also, the prevalence of the outcome variable did not seem to influence differences between estimates from the two approaches.

All ratios of SEs were higher than 1 indicating that SEs of the log-OR from the OL model were lower than those from the RI model. For small relative dispersion of the explanatory variable, ratios were close to 1 (SEs from the two models were close), and increasing relative dispersion increased the ratio of SEs up to a specific value of relative dispersion, after which the ratio of SEs did not increase further. No association was observed between the ratio of SEs and prevalence of the outcome variable.

Coverage of 95% CIs was at the nominal level of 95% for the RI models independent of ICC values, relative dispersion of the explanatory variable, and prevalence of the outcome variable. Coverage was also 95% for the OL model when the ICC was low (≤0.01). Coverage for the OL model was lower for higher ICC values, and it varied with increasing values of relative dispersion of the explanatory variable.

Type I error from the RI model was 5% independent of ICC values, relative dispersion of the explanatory variable, and prevalence of the outcome variable. It was also approximately 5% from the

OL model when the ICC was low (≤0.01), but increased steeply with increasing ICC values and increasing relative dispersion of the explanatory variable, and was somewhat higher for higher prevalence of the outcome variable.

Results are further summarised in Table 5.17.

Table 5.17. Summary of simulation results when the outcome variable was binary, the explanatory variable was continuous, and observations were grouped in 100 clusters

$\beta_1^{RI} \approx \beta_1^{OL}$ when ICC≤0.03 $\beta_1^{RI} - \beta_1^{OL}$ ↑ (with $\beta_1^{RI} > \beta_1^{OL}$) when ICC ↑ (above 0.03)
$\mathrm{SE}_{\beta_1^{RI}} \approx \mathrm{SE}_{\beta_1^{\mathrm{OL}}}$ for small ICC values small relative dispersion of x_{ij} $\mathrm{SE}_{\beta_1^{RI}}/\mathrm{SE}_{\beta_1^{\mathrm{OL}}}$ ↑ when relative dispersion of x_{ij} ↑ ↑ when ICC ↑
For the RI model, coverage by 95% CIs: on average 95% For the OL model, coverage by 95% CIs: on average 95% when ICC ≤0.01 ↓ when ICC ↑ (above 0.01) ↓ when prevalence of y_{ij} ↑
For the RI model, type I error: on average 5% For the OL model, type I error: on average 5% when ICC ≤0.01 ↑ when ICC ↑ (above 0.01) ↑ when relative dispersion of x_{ij} ↑ ↑ when prevalence of y_{ij} ↑

5.6.3 Binary outcome - binary predictor – 2 clusters

Absolute differences in point estimates increased from the null value of 0 when the difference between the cluster-specific prevalence rates of the explanatory variable, and the distance between cluster-specific intercepts increased. Prevalence of the outcome variable did not affect differences in ORs estimated from the two models.

Ratios of SEs from the two methods were approximately equal to 1 when the prevalence of the explanatory variable was the same in the two clusters. As difference in the prevalence of the explanatory variable between the two clusters increased, the ratio of the SEs increased giving a U-shaped pattern. Prevalence of the outcome variable, and distance between cluster-specific intercepts did not affect ratios of SEs.

Coverage by 95% CIs was at the nominal level of 95% when the model adjusted for clustering. However, it was poor for the cluster-unadjusted model; it was poorer for bigger differences in the cluster-specific prevalence rates of the explanatory variable, for larger values of k, and for higher overall prevalence of the outcome variable.

Type I error was approximately 5% under the cluster-adjusted model for any prevalence of the explanatory or the outcome variable, and value of k. It was also 5% for the cluster-unadjusted model, when prevalence of the explanatory variable was similar in the two clusters, and for small distance between cluster-specific intercepts. Type I error for the cluster-unadjusted model increased when difference between the cluster-specific prevalence rates of the explanatory variable, k, or prevalence of the outcome increased.

Results are further summarised in Table 5.18.

Table 5.18. Summary of simulation results when both the outcome and the explanatory variables were binary, and observations were grouped in 2 clusters

$\beta_1^{Cluster-adjusted}$:	very close to the true value
$\|\beta_1^{Cluster-adjusted} - \beta_1^{Cluster-unjusted}\|$	↑ when $\|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\|$ ↑ ↑ when k ↑ not associated with prevalence of y
$SE^{cluster-adjusted} \approx SE^{cluster-unadjusted}$	when prevalence of $x_{cluster\ 2}$ > prevalence of $x_{cluster\ 1}$
$SE^{cluster-adjusted}/SE^{cluster-unadjusted}$	↑ when $\|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\|$ ↑ not associated with k not associated with prevalence of y
For the RI model, coverage by 95% CIs:	on average 95%
For the OL model, coverage by 95% CIs:	↓ when $\|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}\|$ ↑ ↓ when k ↑ ↓ when prevalence of y ↑
For the RI model, type I error:	on average 5%

For the OL model, type I error: ↑ when $|\bar{x}_{cluster\ 1} - \bar{x}_{cluster\ 2}|$ ↑
↑ when k ↑
↑ when prevalence of y ↑

5.6.4 Binary outcome - binary predictor – multiple clusters

Differences in the log-odds ratios were small for small ICC values (≤0.03) whatever the prevalence of the outcome and explanatory variable. Average differences were higher for ICC larger than 0.03. They were not associated with dispersion of the prevalence of the explanatory variable, but were somewhat greater for higher prevalence of the outcome variable.

SEs of β_1^{OL} were lower than those of the β_1^{RI}. The ratio of SEs was always >1, and it increased with increasing ICC and prevalence of the outcome variable. The rate of increase of the ratios of SEs was higher for higher dispersion of the cluster-specific prevalence of the explanatory variable. Overall prevalence of the explanatory variable was also associated with rate of increase of ratios of SEs; it increased more rapidly when the overall prevalence of x_{ij} was lower than when it was higher.

Coverage by 95% CIs was 95% for the RI model and varied little for different combinations of the parameters explored. For the OL model, coverage by 95% CIs was at the nominal level of 95% for any prevalence of the outcome and explanatory variables and dispersion of the prevalence of x_{ij} when the ICC was ≤0.01. Coverage decreased with increase in ICC and increase in the prevalence of the outcome variable, but it decreased less when the overall prevalence of x_{ij} was lower, than when it was higher.

Type I error was 5% for the RI model for any combination of the parameters examined. That was also seen when the OL models were fitted, but only when ICC was lower than 0.01. For the OL model, type I error increased with increasing ICC, and increasing prevalence of the outcome variable. The rate of increase of type I error with increasing ICC was higher when the prevalence of y_{ij} was higher, or the overall prevalence of x_{ij} was lower, or the dispersion of prevalence of x_{ij} across clusters was higher.

Results are further summarised in Table 5.19.

Table 5.19. Summary of simulation results when both the outcome and the explanatory variables were binary, and observations were grouped in 100 clusters

$\beta_1^{RI} \approx \beta_1^{OL}$ when ICC≤0.03 $\beta_1^{RI} - \beta_1^{OL}$ ↑ (with $\beta_1^{RI} > \beta_1^{OL}$) when ICC ↑ (above 0.03) ↑ when prevalence of y_{ij} ↑ − when dispersion of prevalence of x_{ij} across clusters ↑
$\mathrm{SE}_{\beta_1^{RI}} > \mathrm{SE}_{\beta_1^{\mathrm{OL}}}$ $\mathrm{SE}_{\beta_1^{RI}}/\mathrm{SE}_{\beta_1^{\mathrm{OL}}}$ ↑ when ICC ↑ when dispersion of prevalence of x_{ij} across clusters ↑ Rate of ↑ of $\mathrm{SE}_{\beta_1^{RI}}/\mathrm{SE}_{\beta_1^{\mathrm{OL}}}$ was higher when overall prevalence of x_{ij} ↓
For the RI model, coverage by 95% CIs: on average 95% For the OL model, coverage by 95% CIs: on average 95% when ICC ≤0.01 ↓ when ICC ↑ (above 0.01) ↓ when prevalence of y_{ij} ↑ ↓ when overall prevalence of x_{ij} was higher
For the RI model, type I error: on average 5% For the OL model, type I error: on average 5% when ICC ≤0.01 ↑ when ICC ↑ (above 0.01) ↑ when prevalence of y_{ij} ↑ Rate of ↑ of type I error with ICC ↑ was higher when prevalence of y_{ij} ↑ when overall prevalence of x_{ij} ↓ when dispersion of prevalence of x_{ij} across clusters ↑

5.7 Discussion

In this chapter, the research question about consequences of ignoring clustering in statistical inference was extended from linear regression (presented in Chapter 4) to logistic regression modelling. The methods used closely mirrored those of the previous chapter on linear regression. However, statistical inference with and without adjustment for clustering of observations was made on the association between a binary outcome and a single explanatory variable x_{ij}, of which two different types were considered –continuous and binary. Also, implications of ignoring clustering were explored in relation to bias in the point estimates, and relative precision, coverage by 95% CIs, and type I error rates. Bias in point estimates was assessed by considering the difference in the estimated log odds ratios, and that in precision was assessed by calculating the ratios of the SEs of the log odds ratios estimated from the two models. As the thesis focuses on data that are hierarchically structured, the analysis for two clusters served simply as an introduction to the main results on multiple clusters, and will not be discussed here.

The limitations that need to be considered with regard to the findings are the same as those described in the discussion of consequences of ignoring clustering in linear regression (section 4.8). They are summarised here in brief. The sample size used in the simulation of data was large, as was the number of clusters and the number of observations per cluster. The decision to use such large samples was made to minimise random sampling variation, so that bias in point estimates and related precision could be characterised better. Moreover, the number of clusters, the number of observations per cluster, and the true effect estimate were fixed for all simulated data, as the focus of this thesis was not to explore how the structure of clustered data, or the strength of the association of the outcome with the explanatory variable might influence statistical inference when clustering is ignored. Also, the process of simulating data followed the model specification and assumptions of a RI logistic regression model, leading the RI models to function as intended. Lastly, the relative dispersion of the continuous x_{ij} variable varied from small to higher values, by increasing the between-cluster dispersion of the cluster-specific mean values of x_{ij} alone, keeping the within-cluster dispersion of x_{ij} constant. However, I have no reason to expect that another approach to producing increasing values of relative dispersion of x_{ij} would result in different conclusions.

Effects and related precision in logistic regression modelling are estimated using maximum likelihood in which an iteration process is followed. It therefore is challenging to investigate algebraically the problem of ignoring clustering and to derive formulae for bias in regression coefficients and their precision. There is, however, no reason to expect that the algorithm used for simulating data in this chapter would not function as well as that described in the previous chapter.

Simulation results showed that when the true log odds ratio was greater than zero, the RI logistic regression model produced log odds ratios that were on average higher than those derived from the OL model. Additionally, SEs from the RI model were larger than those from the naïve logistic regression model. Unlike for linear regression, in the case of the naive logistic regression model, the direction of bias in the point estimates of effect and corresponding SEs was independent of whether the x_{ij} was continuous or binary, but increased with increasing clustering in the outcome after adjustment for the x_{ij}. Coverage by 95% CIs was very close to the nominal levels when estimates were derived from the RI logistic model, and also when derived from the naïve model if clustering was low (ICC<0.03). However, for higher ICC values (≥0.03), it fell considerably below 95% when the OL model was fitted, especially with increasing prevalence of the outcome variable. Type I error rates were very close to 5% when the RI model was fitted, as one might expect, but also with the OL model when clustering was low (ICC<0.01). For higher values of ICC, Type I error rates increased with increasing prevalence of the outcome variable, and with increasing relative dispersion of a continuous x_{ij}, or with increasing dispersion of the cluster-specific prevalence rates of a binary x_{ij}.

Conclusions from the analysis presented in this chapter are very much in agreement with results from the published literature on the consequences of ignoring clustering in logistic regression. Specifically, bias in the point estimates derived from a naïve logistic regression model has been reported from studies of simulated data (108, 112, 116), which showed that when the analytic model fails to account for clustering, the resulting effect estimates can be underestimated by as little as approximately 0% up to as much as -26%, according to the structure of the data (overall number of observations, number of clusters, and number of observations per cluster), true effect size, and levels of clustering. Such observations from simulated studies have been supported by studies of real data which compared results from the two analytical models (109-111, 113-115). These studies have found log odds ratios derived from naïve logistic models that differed from those from RI logistic models by an average of -6% (range: -115% to 28%), with the majority of differences being negative. In this chapter, given that estimates from the RI model were very close to the true value (mean OR=2, independent of size of clustering effect (ICC)), differences in the estimates of effect can be viewed as bias when clustering was not accounted for in the regression model. Results showed that bias increased with increasing levels of clustering. For the highest level of clustering explored (ICC=0.30), point estimates from the OL model differed from those from the RI logistic by an average of -17%, when the x_{ij} was continuous, and for any prevalence of the outcome variable. In the case of a binary x_{ij}, the estimates of effect from the OL model were also about -17% different from those derived from the RI model, but in this case bias increased with increasing prevalence of the outcome variable (bias $\sim -21\%$ for prevalence of y_{ij} of 30%). This level of bias in the log odds ratios is considerable, and exceeds the cut point of 10% that has previously been taken to indicate a meaningful difference in a parameter estimate from the true value (137).

The published literature is also quite coherent regarding the effects of ignoring clustering on the precision of point estimates. Studies of both simulated and real data (108-113, 116-118) have shown that SEs are underestimated by the OL model, sometimes importantly (111). The size of bias in the estimated SEs, as in the case of bias in the point estimates, depended mostly on the level of clustering, and the magnitude of the effect of the x_{ij} on the outcome variable. The association between effect size and bias in SEs could not be assessed in this thesis, as the true effect estimate was fixed to the same value for all simulated data. However, like previous studies, the analysis in this chapter supports the importance of clustering of the outcome variable for the extent of bias in SEs when clustering is not accounted for. SEs from the RI logistic model were higher than those from the OL model in all the cases examined. The difference in precision of point estimates from the two models was larger as clustering of the outcome variable increased. However, even for the lowest value of clustering (ICC=0.001), SEs from the RI logistic model were on average 2% higher than those from the naïve logistic model.

In addition, in this chapter, I explored separately cases of continuous and binary x_{ij}, and for the latter I considered different overall prevalence rates. I found that when x_{ij} was binary, the ratios of the SEs ($SE_{\beta}^{RI}/SE_{\beta}^{OL}$) were higher when the overall prevalence of x_{ij} was lower (Figure 5.22), especially for higher values of dispersion of the cluster-specific prevalence rates of x_{ij} across clusters. These results contradict findings reported by Abo-Zaid et al (108) who showed that the ratio of standard errors from the RI model to those from the OL model was greater for x_{ij} of higher prevalence. However, this difference in findings could be a consequence of different conditions and assumptions under which simulated data were generated (small number of clusters, larger true effect estimate).

None of the studies mentioned above explored the effect of clustering in statistical inference for varying levels of clustering in x_{ij}. Results from this chapter show that bias in effect estimates did not depend on the relative between- to within-cluster dispersion of a continuous x_{ij}, nor on increasing dispersion between clusters in the prevalence of a binary x_{ij}. However, in the case of a binary x_{ij}, the variation around the average value of differences in point estimates increased with increasing dispersion of prevalence of x_{ij}. The ratio of SEs of effects estimated from the RI logistic model to those of effects estimated from the naïve logistic regression model increased when relative dispersion of the continuous x_{ij} increased up to a cut-point value, after which further increase did not influence relative precision of effects. When x_{ij} was binary, the ratio of SEs increased with increasing dispersion of the cluster-specific prevalence rates of x_{ij}. Moreover, I showed that bias in the precision of effect estimates was independent of the population prevalence of the outcome variable when the explanatory variable under investigation, x_{ij}, was continuous, but when x_{ij} was binary, it was somewhat greater for higher prevalence of the outcome than for lower prevalence.

I additionally analysed real data from the CUPID study to explore whether observations made from analysis of the simulated data regarding point estimates and related precision could be replicated. Almost all variables in the CUPID study were binary or categorical, only two being continuous and with an approximately normal distribution. To enable a direct comparison of results, all binary categorical variables were dichotomised, giving 115 binary variables with prevalence rates ranging from 2% to 98%. For the purpose of this comparison, all binary variables were used both as outcomes and explanatory variables in the univariate logistic regression models fitted, without considering what they represented. Estimated ICC values across the different RI logistic models ranged from 0 to 0.7. When the two continuous variables (age, and hours worked per week) were used as explanatory variables, log odds ratios from the RI logistic model were higher than those from the OL model in most of the cases, while SEs from the RI model were always further from the null than SEs from the OL model. As observed in the simulated data, ratios of the SEs were higher for higher ICC values. The pattern of differences in log odds ratios by increasing levels of clustering was not observed when participant's hours worked/week was used as the explanatory variable. The inconsistency of this observation between analysis of simulated data and that of real data from the CUPID study when hours worked/ week was used as an explanatory variable was due to the log odds ratios being very close to zero; their mean value was 0.004 (i.e. OR=1.0) and they ranged from 0 to 0.04, while the corresponding values for age were higher. Indeed, differences in log odds ratios when the two models were fitted in simulated data that assumed a true value of zero, were very close to zero and they did not vary by ICC level (data not shown).

Observations from the analysis of real data agreed with those from simulated data also when x_{ij} was binary; point estimates from the two models were on average the same (Figure 5.26), while SEs of the point estimates when derived from the RI model were higher than those derived from the OL model, with ratios of SEs between the two models being higher when ICC or dispersion of the prevalence of x_{ij} across clusters increased (Figure 5.27).

Coverage by 95% confidence intervals has been discussed by previous researchers, who have shown that when clustering is not accounted for in a regression model, coverage is lower than 95% (108, 116). That is more apparent in the case of a binary exploratory variable than a continuous one (108). The lower rates of coverage reported by Abo-Zaid et al (108) were 46% when x_{ij} was binary with a high prevalence and 84% when it was continuous, and they occurred when both clustering and the true effect value were high. Coverage rates by 95% confidence intervals under the RI logistic model have been shown to be 95% (108, 116), as they were in the simulation studies presented in this chapter. When the OL model was fitted, I showed that coverage by 95% confidence intervals was very close to 95% when clustering was low (ICC≤0.01) in the cases both of a continuous and a binary x_{ij}. However, as clustering or prevalence of the outcome variable increased, an undercoverage problem was observed for both types of x_{ij}. In the case of a binary x_{ij}, coverage also fell to values considerably lower than 95%, when the overall prevalence of x_{ij} or the dispersion of cluster-specific prevalence of x_{ij} increased. A previous study of simulated data (120) that explored consequences of ignoring clustering in relation to levels of clustering of a binary explanatory variable (ρ_x) drew similar conclusions, showing that as ρ_x increased, coverage decreased to values as low as 45%. An exception to this observation was the case of high clustering (ICC=0.3) and high prevalence of x_{ij} (0.4), when coverage decreased for lower ρ_x (i.e. lower dispersion of x_{ij} across clusters).

As for coverage rates, type I error rates were close to the nominal value of 5%, both for the RI logistic regression model, and for the OL regression model with low clustering of the outcome variable, whether x_{ij} was continuous or binary. As ICC increased further from 0.01, error rates for the OL model increased. They were also somewhat higher for higher prevalence of the outcome variable, lower overall prevalence of x_{ij}, and they increased with increasing ρ_x, expressed either as relative dispersion of a continuous x_{ij}, or as dispersion of the prevalence of a binary x_{ij} across clusters. Two studies have reported type I error rates when a naïve logistic regression model was fitted to clustered data (112, 138). However, in both cases, the rates reported were related to the effect of a cluster-level variable rather than that of an individual-level variable, of the sort considered in this chapter. Consequently, comparison of observations on type I error rates from my results and published studies cannot be made.

In summary, my findings support the use of RI logistic regression models when data are clustered (and conform to the assumptions of the RI model) and one wishes to estimate ORs. Failure to account for clustering can result in important underestimation, both of effects and their precision, a bias which is most apparent when ICC values exceed 0.03. Prevalence of the outcome variable does not influence bias in the case of a continuous explanatory variable. In contrast, when the explanatory variable is binary, bias is somewhat higher when prevalence of the outcome variable is higher, but it is not affected by overall prevalence of the explanatory variable. These findings are very much in agreement with those from the literature. However, unlike previous researchers, I also explored consequences of ignoring clustering for increasing relative between- to within-dispersion of a continuous explanatory variable, and for increasing dispersion of the cluster-specific prevalence rates of a binary explanatory variable, essentially representing increasing clustering in the explanatory variable. With regard to the latter, I found that ρ_x does not affect level of bias in the point estimates, but it increases bias in SEs estimated by a naïve logistic regression model. Other than estimates of effect and SEs, I also showed that higher levels of ICC lead to severe undercoverage problems and inflation of Type I error rates when clustering is not taken into account. Such errors were even more likely to occur when the population prevalence of the outcome variable was high, and in the case of a binary explanatory variable, when its prevalence was high (~40%). It is in those circumstances that statistical inference is likely to be importantly influenced by failing to consider clustering effects when analysing data. As with the case of a continuous outcome, in all circumstances in which the ICC is small, clustering is minimal, and there is little increase in Type 1 error when a naïve logistic regression model is fitted instead of a RI logistic regression model.

Chapter 6. Comparison of naïve regression models with dummy variables for the clusters to RI regression models

6.1 Introduction

The consequences of ignoring clustering were examined in detail in chapters 4 and 5 for linear and logistic regression, respectively. In that exploration, the regression models that ignored the effects of clustering were OLS for continuous y_{ij} and OL for binary y_{ij}, as described in equations 4.1 and 5.1. However, sometimes researchers who appreciate the potentially important effects of different centres/clusters from which observations are drawn, instead apply the same naïve regression models, but with incorporation of dummy variables for the centres/clusters. In that way, a fixed effect is fitted for each cluster, relative to a baseline cluster. The cluster is thus treated as representing an observed quantity, the effect of which needs to be fixed in the same way that a confounder would need to be accounted for in the regression model. That, however, is still different from the principle of the random intercept multilevel model in which the interest is in the association between the outcome and the explanatory variables when accounting for a random distribution of intercepts across clusters. In the second case, conclusions are drawn about the population from which the observed clusters were drawn, rather than the particular clusters themselves.

This approach has been described earlier in the thesis (section 2.3). I here extend the research question on consequences of ignoring clustering by comparing the estimates from a naïve regression model with a dummy variable (DV) for each cluster with those from the RI model.

6.2 Methods

The methods of this investigation were the same as those described in sections 4.3, 4.5, 5.3, and 5.5. However, only a subset of the scenarios explored in the previous chapters are considered here. These are described below

- Continuous outcome and continuous explanatory variable with relative dispersion of the explanatory variable ranging from 0-1 and 5-5.5
- Continuous outcome and binary explanatory variable with dispersion of prevalence of the explanatory variable ranging from 0 to 0.6 and overall prevalence of the explanatory variable 0.1 and 0.4

- Binary outcome of prevalence 0.1 and 0.3 and continuous explanatory variable with relative dispersion of the explanatory variable ranging from 0-1 and 5-5.5
- Binary outcome of prevalence 0.1 and 0.3 and binary explanatory variable with dispersion of prevalence of the explanatory variable ranging from 0 to 0.6 and overall prevalence of the explanatory variable 0.1 and 0.4

6.3 Results

Comparison of point estimates

Differences in regression coefficients estimated from RI and DV linear regression models are plotted against the relative dispersion of a continuous and the dispersion of prevalence rates of a binary explanatory variable x_{ij} in figures 6.1 and 6.3, respectively. Also, the difference in the estimated log odds ratios from the two modelling approaches (when the outcome was binary and logistic regression models were fitted) are plotted in figures 6.5 (for continuous x_{ij}) and 6.7 (for binary x_{ij}). For ease of interpretation of findings on differences between effect estimates, the dispersions of point estimates derived from the two analytical models (RI and DV) are illustrated in figures 6.2, 6.4, 6.6, and 6.8.

With the simulated data that I generated, the two modelling methods (RI and DV) produced effect estimates that were very close to each other, the largest differences being only about 4% of the true value. They were even lower when x_{ij} was binary with a high overall prevalence (figures 6.3 and 6.7).

In the case of linear regression, the average difference of the effect estimates was very close to the null value of zero irrespective of ICC values, and relative dispersion of mean values of the continuous x_{ij} (Figure 6.1) or dispersion of prevalence rates of the binary x_{ij} across clusters (Figure 6.3). However, in logistic regression (figures 6.5 and 6.7), a small bias was observed, with average values of differences in log odds ratios being <0 (mean value = -0.01 for all ICC values). That was due to small overestimation of β_1^{DV} (mean value 0.701 corresponding to an OR≈2.02) rather than bias in the estimation of β_1^{RI} (mean value 0.694 corresponding to OR≈2.00).

In all cases, increasing clustering of the explanatory variable (expressed either as increasing relative dispersion of a continuous x_{ij} or increasing dispersion of the prevalence rates of a binary x_{ij} across clusters) increased the dispersion of differences of effect estimates symmetrically around the mean value of differences (0 in linear and -0.01 in logistic regression). Also, differences in effect estimates were more narrowly spread around the mean value of differences when the ICC was high (ICC=0.3) than when it was low (ICC=0.003). These two patterns were due to larger dispersion of the effects derived from the DV approach ($\beta_1^{DV}{}'s$) compared to those from the RI approach ($\beta_1^{RI}{}'s$) when the ICC was low (ICC=0.003) and clustering in the explanatory variable was high (figures 6.2, 6.4, 6.6, and 6.8). Increasing ICC made dispersion of the effect estimates from the two approaches more comparable, irrespective of the clustering level of x_{ij}. That pattern was more pronounced when the explanatory variable of interest was continuous than when it was binary (difference in dispersion of estimates from the RI and the DV models in figures 6.2 and 6.6 as compared with figures 6.4 and 6.8).

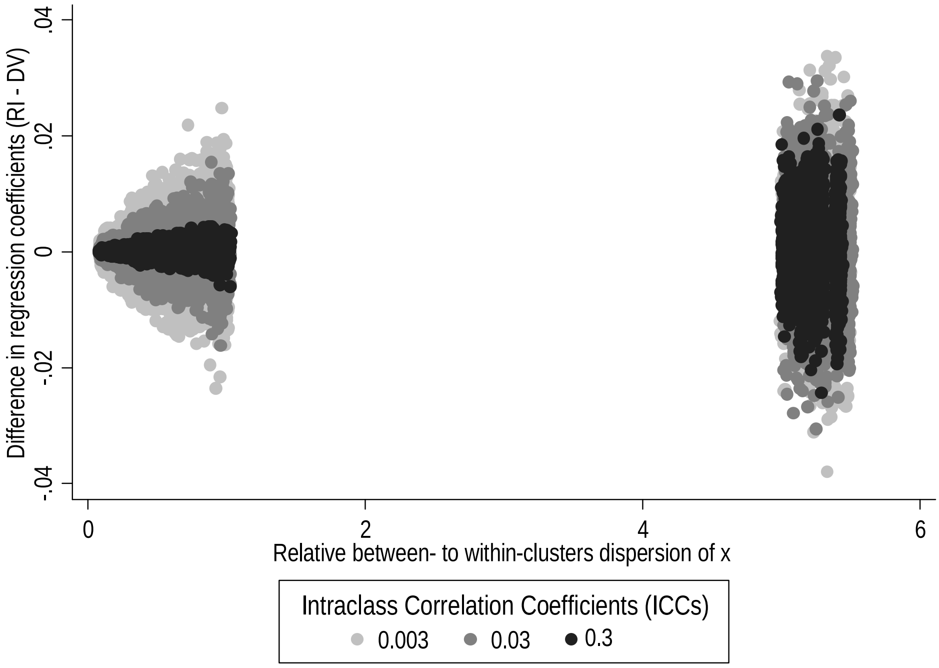


Figure 6.1. Difference between regression coefficients estimated from RI and DV linear regression models ($\beta_1^{RI} - \beta_1^{DV}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend) and for $SD_{e_{ij}} = 1$

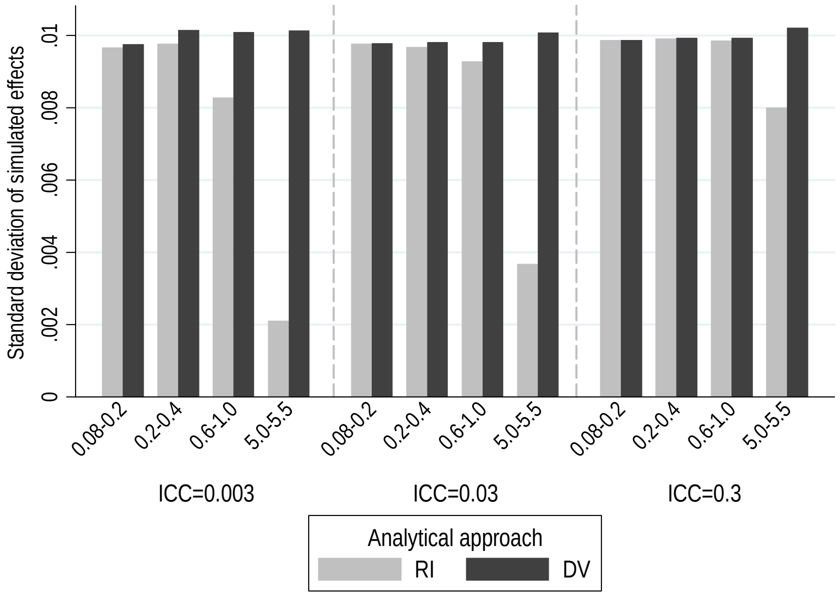


Figure 6.2. Standard deviation and range of estimated regression coefficients from the RI and the DV linear regression models for categories of relative between- to within-cluster dispersion of the explanatory variable x_{ij} (as indicated in the x-axis) and ICC values

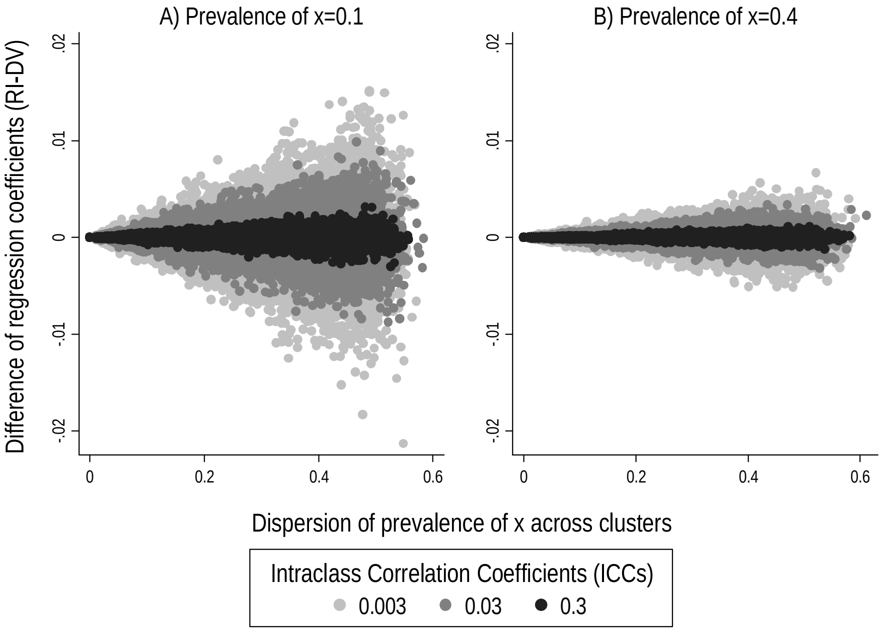


Figure 6.3. Difference between regression coefficients estimated from RI and DV linear regression models ($\beta_1^{RI} - \beta_1^{DV}$) plotted against dispersion of prevalence of the explanatory variable x_{ij} across clusters. Figure A: Overall prevalence of x_{ij} 0.1. Figure B: Overall prevalence of x_{ij} 0.4

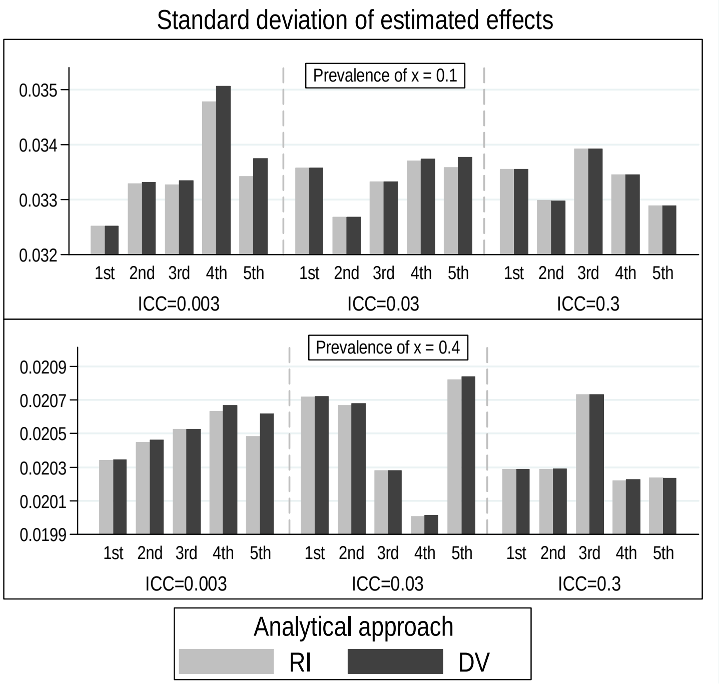


Figure 6.4. Standard deviation of estimated regression coefficients from the RI and the DV linear regression models for fifths of the distribution of dispersion of prevalence of the explanatory variable x_{ij} across clusters, overall dispersion of x_{ij}, and ICC values

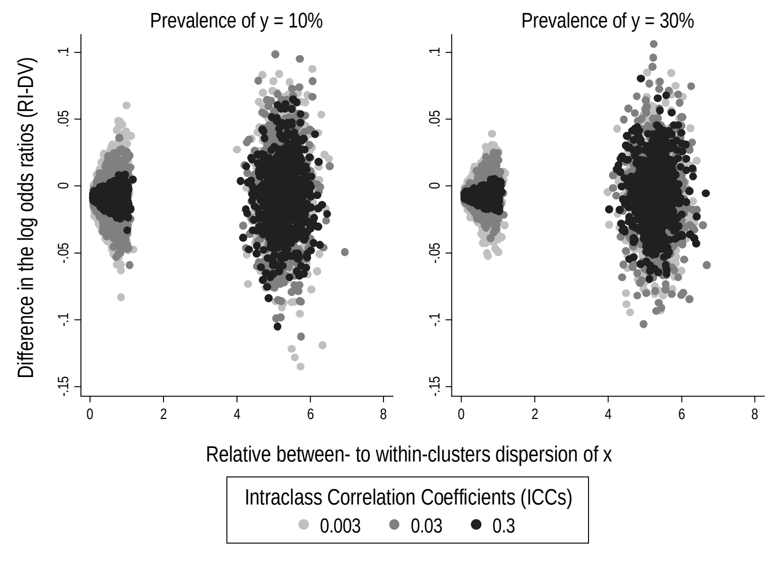


Figure 6.5. Difference between log odds ratios estimated from RI and DV logistic regression models ($\beta_1^{RI} - \beta_1^{DV}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij}, and for different levels of intraclass correlation (shades of grey as indicated in the legend)

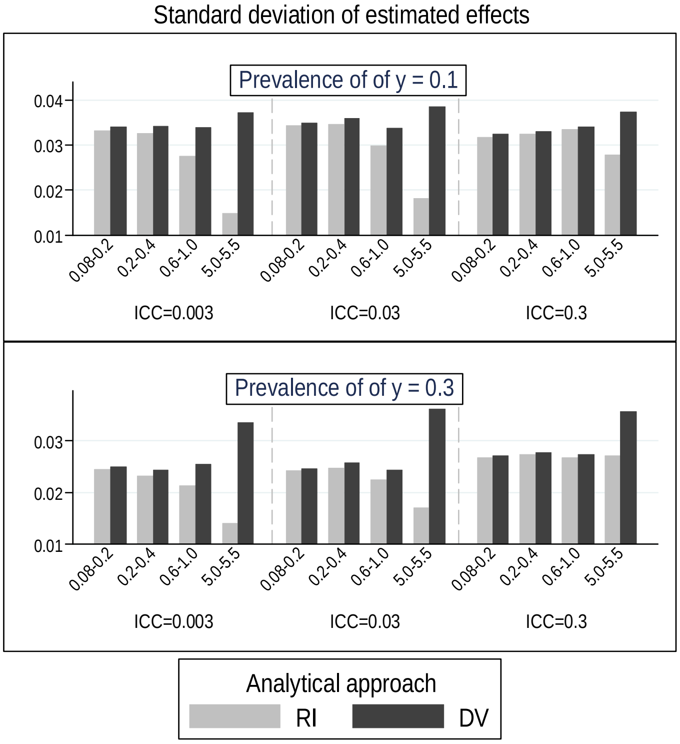


Figure 6.6. Standard deviation and range of estimated log odds ratios from the RI and the DV logistic regression models for categories of relative between- to within-cluster dispersion of the explanatory variable x_{ij} (as indicated in the x-axis) and ICC values

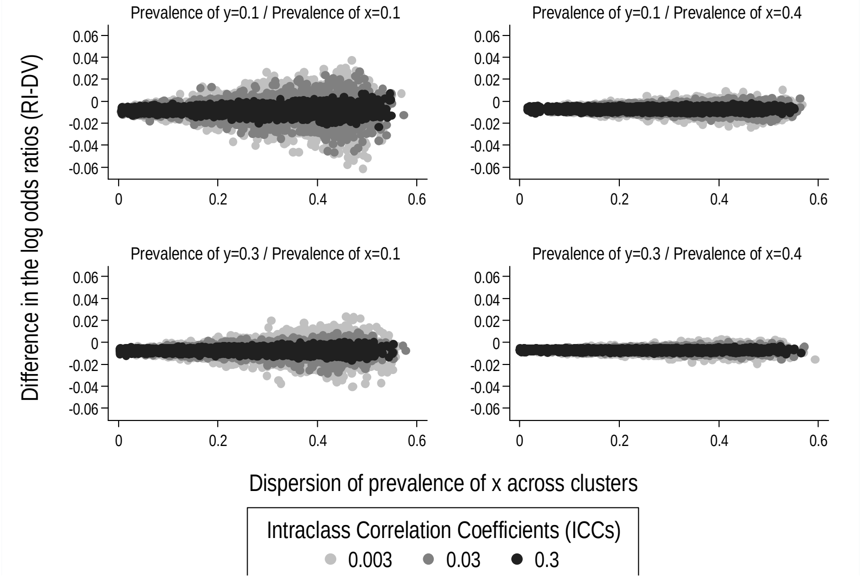


Figure 6.7. Difference between log odds ratios estimated from RI and DV logistic regression models ($\beta_1^{RI} - \beta_1^{DV}$) plotted against dispersion of prevalence of the explanatory variable x_{ij} across clusters

and for different ICC values, and overall prevalence of the outcome and the explanatory variable as indicated in each sub-plot

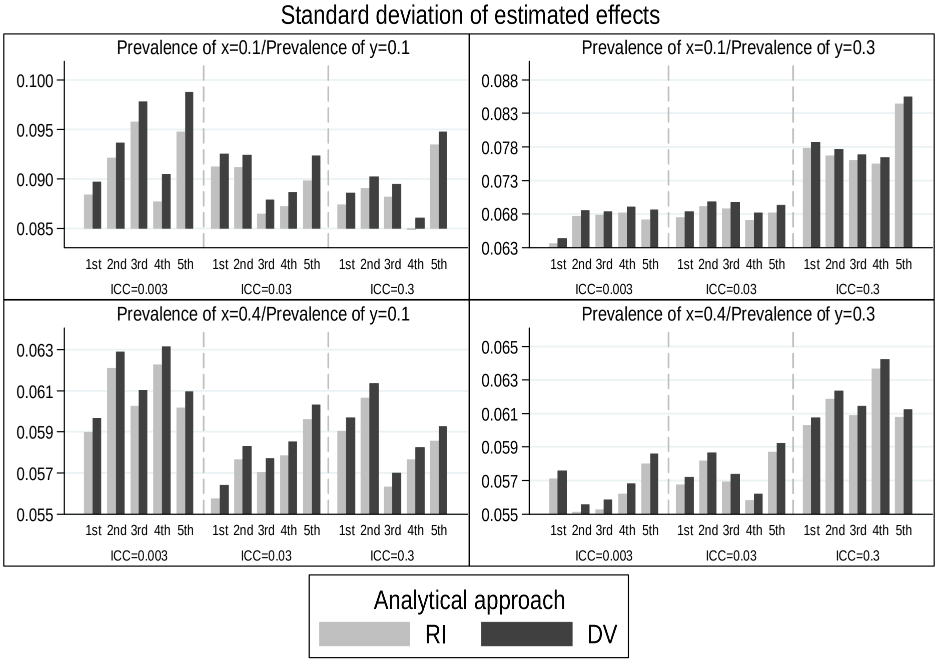


Figure 6.8. Standard deviation and range of log odds ratios estimated from the RI and the DV logistic regression models for categories of dispersion of prevalence of the explanatory variable x_{ij} across clusters, overall prevalence of x_{ij}, overall prevalence of y_{ij}, and ICC values

Comparison of standard errors of point estimates

The ratios of SEs of the point estimates from the two analytical methods ($SE_{\beta_1^{RI}}/SE_{\beta_1^{DV}}$) were plotted against the relative between- to within-cluster dispersion of the continuous explanatory x_{ij} in figures 6.9 and 6.12 and against the dispersion of prevalence rates across clusters of the binary x_{ij} in Figure 6.11 and 6.14. The pattern of ratios of SEs in relation to clustering of x_{ij} was very similar across the four combinations of distributions of y_{ij} (continuous and binary) and x_{ij} (continuous and binary) explored. For very low clustering of x_{ij} (small relative dispersion of the continuous x_{ij} or small dispersion of the prevalence rates across clusters of the binary x_{ij}), SEs from the two methods were very similar (ratios of SEs very close to 1). As clustering in x_{ij} increased, ratios of SEs increasingly fell below the null value of 1, indicating that SEs estimated from the DV approach were higher than those from the RI regression models. In addition, as ICC values decreased, the ratios of SEs were further from the null value of 1. Larger deviations from 1 in the ratios of SEs of estimated effects were seen when x_{ij} was continuous, while when x_{ij} was binary, ratios of SEs ranged only from 0.97 to 1.01. The lowest values of ratios of SEs (i.e. largest deviation of the ratio of SEs from 1) were seen when x_{ij} was continuous and for the highest level of clustering of x_{ij} and the lowest value of ICC.

In the circumstances in which the deviations of the ratios of SEs from the null value were greatest (i.e. when x_{ij} was continuous), the patterns of ratios of SEs were explored further by plotting the mean values of SEs of the point estimates from the two regression approaches (Figure 6.10 for continuous y_{ij}, and Figure 6.13 for binary y_{ij}). With a continuous x_{ij}, and either a continuous or a binary y_{ij}, the average values of SEs of effect estimates derived from the DV model were generally very similar across increasing categories of relative dispersion of x_{ij}. An exception, however, was the case of the highest category of relative dispersion of x_{ij} and high prevalence of y_{ij} (lower part of Figure 6.13) for which SEs from the DV model were on average higher than those for lower categories of relative dispersion of x_{ij}. In contrast, SEs from the RI model were on average higher for lower relative dispersion of x_{ij} and decreased as relative dispersion of x_{ij} increased. This decreasing trend was more prominent for lower ICC values.

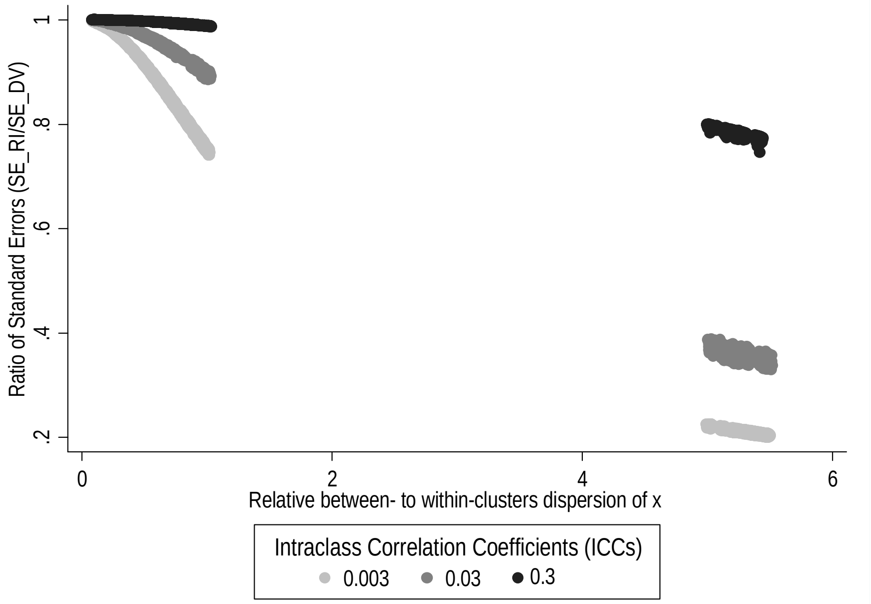


Figure 6.9. Ratios of standard errors estimated from RI and OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{DV}}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij} and ICC values

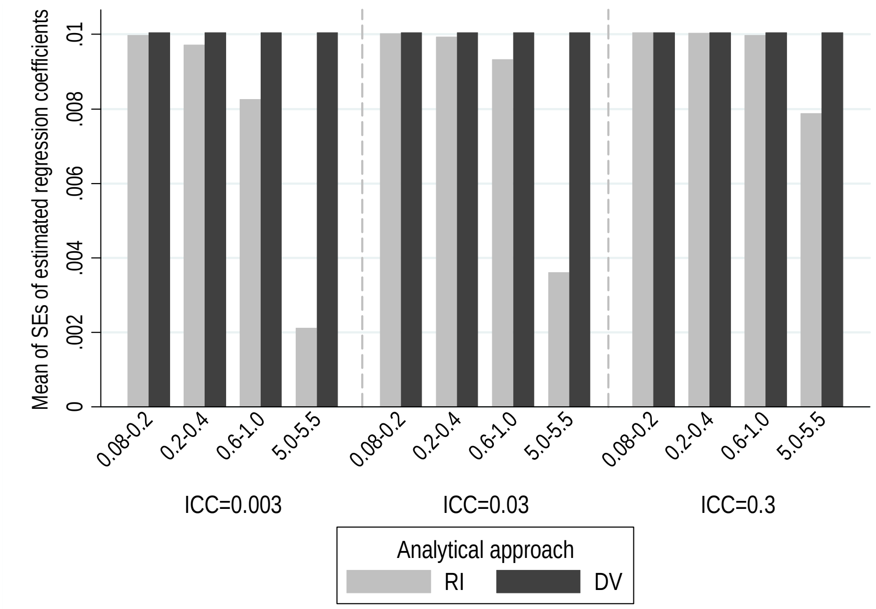


Figure 6.10. Mean values of SEs of estimated regression coefficients from the RI and the DV models for categories of relative between- to within-cluster dispersion of the explanatory variable x_{ij} (as indicated in the x-axis) and ICC values

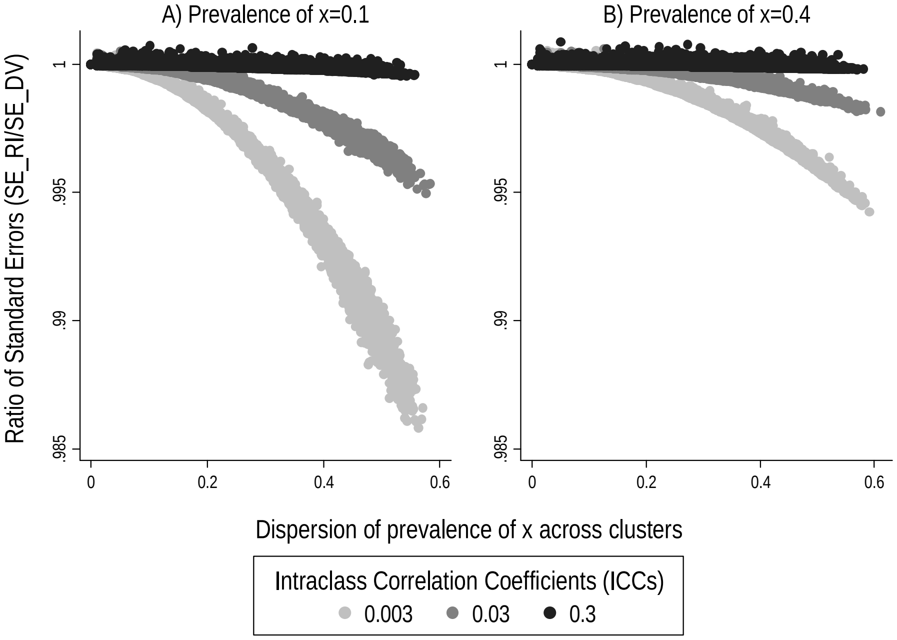


Figure 6.11. Ratios of standard errors of regression coefficients estimated from RI and DV linear regression models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{DV}}$) plotted against dispersion of prevalence of the explanatory variable x_{ij} across clusters. Figure A: Overall prevalence of x_{ij} 0.1. Figure B: Overall prevalence of x_{ij} 0.4

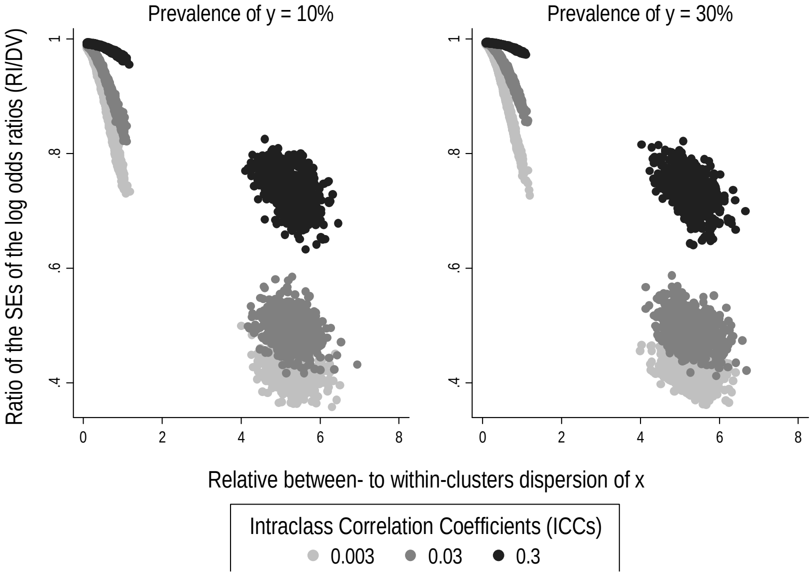


Figure 6.12. Ratios of standard errors of log odds ratios from RI and DV logistic regression models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{DV}}$) plotted against relative between- to within-cluster dispersion of explanatory variable x_{ij} and ICC values

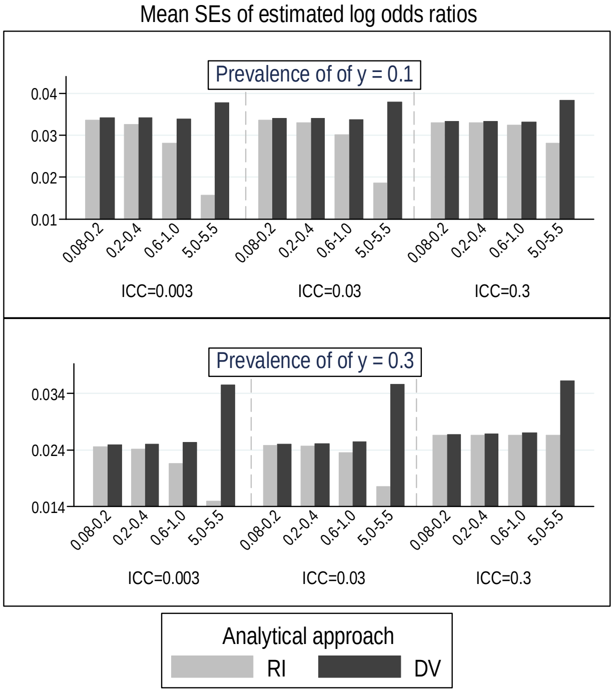


Figure 6.13. Mean values of SEs of estimated log odds ratios from the RI and the DV logistic regression models for categories of relative between- to within-cluster dispersion of the explanatory variable x_{ij} (as indicated in the x-axis) and ICC values

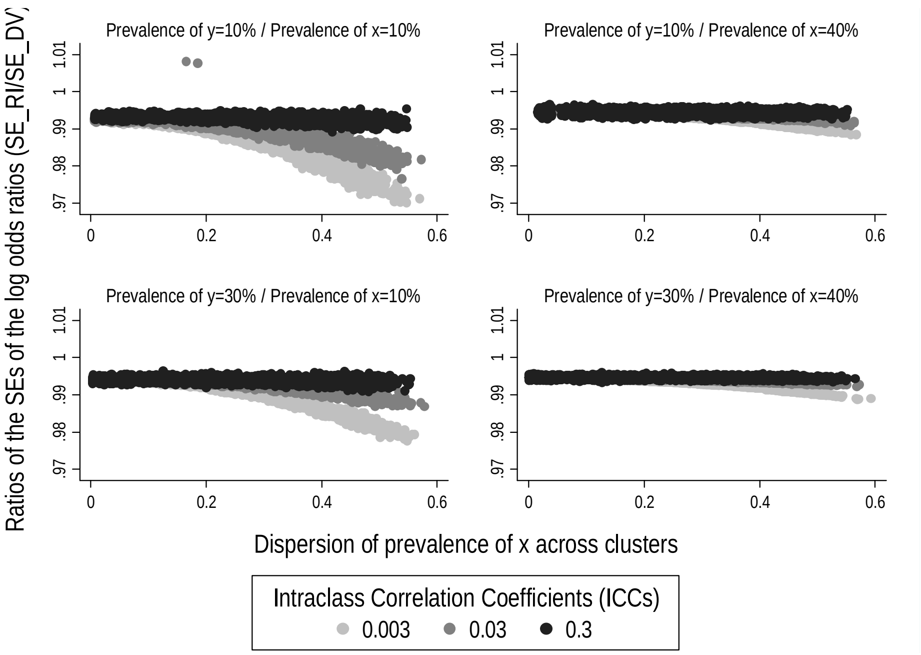


Figure 6.14. Ratios of standard errors of log odds ratios estimated from RI and DV logistic regression models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{DV}}$) plotted against dispersion of prevalence of the explanatory variable x_{ij} across clusters and for different ICC values, and overall prevalence of the outcome and the explanatory variable as indicated in each sub-plot

Comparison of rates of coverage and type I error

The rates of coverage by 95% confidence intervals from the RI and the DV regression models, and of type I error, are summarised according to ICC values and levels of clustering in the explanatory variable (expressed either as increasing relative dispersion of a continuous x_{ij} or increasing dispersion of the prevalence rates of a binary x_{ij} across clusters) in tables 6.1 to 6.4. Type I error rates from both analytical models were very close to the nominal value of 5% across all scenarios of distribution of y_{ij} (binary or continuous) and x_{ij} (binary or continuous), ICC values, and clustering of x_{ij}. They varied from 3% to 7.3% when a RI model was fitted and from 3% to 6.8% when a DV model was fitted. Similarly, coverage rates across the different scenarios were very close to 95%; those from the RI model varied from 93% to 97.1% and those from the DV model varied from 92.4% to 96%. The tight distributions of both types of rate around the nominal values were to be expected for the DV approach given the great similarity of the effect estimates and associated SEs derived from the DV to those from the RI model, as described above.

Table 6.1. Coverage (%) by 95% confidence intervals and Type I error rates for RI and DV linear regression models according to categories of relative between- to within-cluster dispersion of the explanatory variable x_{ij} and ICC values

	Relative between- to within-cluster dispersion of x_{ij}	ICC=0.003		ICC=0.03		ICC=0.3	
		RI	**DV**	**RI**	**DV**	**RI**	**DV**
Coverage (%) by 95% CIs	0.08-0.2	95.7	95.9	95.2	95.6	95.5	95.6
	0.2-0.4	95.1	95.1	95.4	95.3	95.2	95.1
	0.6-1.0	94.8	95.0	95.5	96.0	94.7	94.9
	5.0-5.5	94.3	95.0	94.5	95.0	94.6	94.7
Type I error rates (%)	0.08-0.2	4.9	4.9	5.5	5.6	5.7	5.6
	0.2-0.4	4.9	5.3	5.0	5.5	5.4	5.4
	0.6-1.0	4.8	4.9	4.8	5.1	5.1	5.1
	5.0-5.5	3.8	5.2	4.4	5.0	4.3	4.9

Table 6.2. Coverage (%) by 95% confidence intervals and Type I error rates for RI and DV models according to categories of dispersion of prevalence of the explanatory variable x_{ij} across clusters, overall dispersion of x_{ij}, and ICC values

	Overall prevalence of x_{ij}	Dispersion of prevalence of x_{ij} across clusters	ICC=0.003		ICC=0.03		ICC=0.3	
			RI	**DV**	**RI**	**DV**	**RI**	**DV**
Coverage (%) by 95% CIs	**0.1**	0-0.2	95.1	95.1	95.1	95.1	94.7	94.8
		0.2-0.4	94.4	94.1	94.4	94.4	94.9	94.9
		0.4-0.6	95.0	94.8	95.4	95.2	95.1	95.2
	0.4	0-0.2	95.2	95.1	94.8	94.9	95.4	95.3
		0.2-0.4	94.9	94.9	95.5	95.6	94.8	94.8
		0.4-0.6	95.4	95.1	93.8	93.8	95.8	95.7
Type I error rates (%)	**0.1**	0-0.2	4.9	4.9	4.9	4.9	5.3	5.2
		0.2-0.4	5.6	5.9	5.6	5.6	5.1	5.1
		0.4-0.6	5.0	5.2	4.6	4.8	4.9	4.8
	0.4	0-0.2	4.8	4.9	5.2	5.1	4.6	4.7
		0.2-0.4	5.1	5.1	4.5	4.4	5.2	5.2
		0.4-0.6	4.6	4.9	6.2	6.2	4.2	4.3

Table 6.3. Coverage (%) by 95% confidence intervals and Type I error rates for RI and DV logistic regression models according to categories of relative between- to within-cluster dispersion of the explanatory variable x_{ij} and ICC values

	Population prevalence of y_{ij}	**Relative between- to within-cluster dispersion of x_{ij}**	**ICC=0.003**		**ICC=0.03**		**ICC=0.3**	
			RI	**DV**	**RI**	**DV**	**RI**	**DV**
Coverage (%) by 95% CIs	**0.1**	0.08-0.2	94.9	94.9	93.9	94.4	96.3	94.1
		0.2-0.4	94.9	93.6	93.7	92.4	94.8	93.6
		0.6-1.0	94.9	93.5	94.6	95.4	95.0	94.1
		5.0-5.5	97.0	96.0	94.8	94.8	95.4	95.1
	0.3	0.08-0.2	94.7	93.7	95.7	94.8	95.1	94.6
		0.2-0.4	95.4	93.0	94.7	93.4	95.1	93.7
		0.6-1.0	95.0	93.7	96.4	94.9	95.1	93.3
		5.0-5.5	97.1	95.4	95.7	94.4	96.0	94.6
Type I error rates (%)	**0.1**	0.08-0.2	5.3	5.1	4.8	5.0	4.8	4.8
		0.2-0.4	4.6	4.4	4.3	5.2	4.9	4.9
		0.6-1.0	7.3	6.0	5.5	5.5	4.5	5.1
		5.0-5.5	5.2	4.6	4.9	5.7	4.6	3.0
	0.3	0.08-0.2	5.9	6.1	3.6	4.2	3.8	4.2
		0.2-0.4	4.5	3.2	4.5	4.9	5.0	4.5
		0.6-1.0	4.6	4.3	4.2	4.7	5.0	5.1
		5.0-5.5	3.0	4.4	4.6	6.3	5.9	6.3

Table 6.4. Coverage (%) by 95% confidence intervals and Type I error rates for RI and DV logistic regression models according to categories of dispersion of prevalence of the explanatory variable x_{ij} across clusters, overall dispersion of x_{ij}, overall prevalence of y_{ij}, and ICC values

	Overall prevale-nce of y_{ij}	**Overall prevale-nce of x_{ij}**	**Dispersion of prevalence of x_{ij} across clusters**	**ICC=0.003**		**ICC=0.03**		**ICC=0.3**	
				RI	**DV**	**RI**	**DV**	**RI**	**DV**
Coverage (%) by 95% CIs	**0.1**	**0.1**	0-0.2	95.0	94.2	94.6	94.6	95.4	95.1
			0.2-0.4	94.6	93.6	95.6	95.4	96.5	96.0
			0.4-0.6	93.4	93.1	95.1	94.3	95.3	95.0
		0.4	0-0.2	93.9	93.8	95.4	95.30	94.2	94.6
			0.2-0.4	93.9	94.0	95.5	94.9	94.2	94.4
			0.4-0.6	95.5	95.8	95.3	94.4	94.4	94.4
	0.3	**0.1**	0-0.2	94.8	94.7	95.0	94.6	94.9	95.0
			0.2-0.4	94.9	94.5	95.1	94.2	95.6	95.0

	Overall prevale-nce of y_{ij}	Overall prevale-nce of x_{ij}	Dispersion of prevalence of x_{ij} across clusters	ICC=0.003		ICC=0.03		ICC=0.3	
				RI	**DV**	**RI**	**DV**	**RI**	**DV**
			0.4-0.6	94.7	94.9	94.9	94.4	93.0	92.7
		0.4	0-0.2	94.7	94.2	95.0	94.6	95.1	94.5
			0.2-0.4	96.0	94.4	95.5	95.1	94.4	93.7
			0.4-0.6	94.4	94.8	94.1	93.2	95.4	94.4
Type I error rates (%)	**0.1**	**0.1**	0-0.2	5.1	5.3	4.7	4.9	4.7	4.9
			0.2-0.4	5.1	5.6	5.7	5.7	3.8	4.1
			0.4-0.6	4.5	4.8	5.9	6.4	7.0	6.8
		0.4	0-0.2	5.2	5.3	4.7	4.9	5.4	5.5
			0.2-0.4	4.9	5.1	5.3	5.4	4.7	4.7
			0.4-0.6	4.2	3.8	3.9	4.1	5.3	5.1
	0.3	**0.1**	0-0.2	5.7	5.9	4.4	4.6	4.5	4.6
			0.2-0.4	4.8	4.8	5.6	5.7	6.0	6.0
			0.4-0.6	3.7	3.9	4.1	4.5	5.1	5.1
		0.4	0-0.2	5.4	5.4	5.4	5.6	4.1	4.1
			0.2-0.4	5.3	5.3	5.9	6.0	4.8	5.0
			0.4-0.6	5.6	5.6	4.2	3.9	4.0	4.0

6.4 Summary of findings

In this chapter, a different analytical approach to clustering was explored, namely OL regression modelling with the incorporation of dummy variables (DVs) for clusters in the prediction part of the model. As in the previous two chapters, the estimated effects and the corresponding precision derived from the DV approach were compared with those from the RI model for several scenarios of distribution and level of clustering of the outcome and explanatory variables. In all of the cases explored, point estimates from the two methods (DV and RI) were very close to each other while small differences were seen in the corresponding SEs from the two approaches mainly when the explanatory variable x_{ij} was continuous, with large relative dispersion of cluster-specific mean values of x_{ij}, and small ICC values. Despite these small differences observed in SEs, statistical inference as assessed by error and coverage rates was not affected by fitting fixed effects for the clusters instead of incorporating random effects in the model; both types of rate were very close to the nominal values. These results provide some reassurance that when a naïve DV model is fitted instead of a RI model when data are clustered, conclusions drawn from the analyses are unlikely to be as misleading as from a naïve model without the use of dummy variables for clusters. It should, however, be noted that such a model can be applied only in situations where the association under investigation is between the outcome and an individual-level explanatory variable. When the explanatory variable of interest does not vary within clusters, its effect cannot be estimated as it cannot be separated from those of the cluster dummy variables. Additionally, the DV model is more suitable when the number of clusters is small, or when the size of clusters is large. Otherwise (when the number of clusters is large) many degrees of freedom are used resulting in loss of statistical power. A more detailed description of the problems associated with fitting a DV model in the presence of clustering has also been given above in section 2.3.

Chapter 7. Comparison of meta-analytical to pooled analysis estimates

7.1 Introduction

Chapter 4 and Chapter 5 described investigations into the consequences of ignoring clustering in regression analysis, and identified circumstances in which conclusions drawn from statistical analysis of hierarchical data are likely to be misleading when clustering is ignored. One situation in which clustered data are commonly encountered is meta-analysis, and in this chapter I exploit the CUPID dataset to compare various statistical methods for estimating odds ratios through meta-analysis.

Meta-analysis is a statistical technique that is used to combine evidence from independently conducted studies that provide data relevant to the same specified research question. Usually, the objective is to generate an overall estimate of the effect of an exposure on an outcome of interest, calculated as a weighted average of separate effect estimates reported in different studies. Meta-analysis is relatively quick and inexpensive as it uses information available from published results. Also, when it forms part of a well-conducted systematic review, it is considered to provide the strongest evidence concerning an association, as it incorporates and synthesises evidence from multiple studies (139). For these reasons, the use of meta-analysis has increased over the years, with 302 meta-analyses found in PubMed in 1990 and 1,301 in 2000, increasing to 2,885 in 2005 and to 3,095 in 2008 (140).

Even though conclusions drawn from meta-analyses are widely accepted, there has been criticism of the statistical technique insofar as it combines results from studies that can be very different in the methods used, population samples chosen, and effect sizes reported (heterogeneity of results) (141). Another potential problem in meta-analysis of published results is publication bias, which can arise when there is selective presentation of research findings in the published literature (142). This is a concern because results that are selected for publication tend to be biased in favour of positive or more interesting findings. Publication bias can thus be regarded as a form of selection bias.

As an alternative, when individual-level data are available from multiple studies that explore the same research question, evidence can be synthesised by conducting an analysis that uses individual information from the entire dataset. In the context of clinical trials, this method is sometimes called meta-analysis of individual patient data (143), but in epidemiological studies it is commonly referred to as pooled analysis of individual data (73).

Pooled analysis has important advantages over meta-analysis of published results (144). Firstly, it allows researchers to obtain information about participants that is not available in published reports, and to use that information in statistical analysis. As individual information is available, subgroup analyses can be conducted, such as by gender, which may not be possible otherwise. Also with this statistical approach, publication bias, which is a major problem in meta-analysis of published results, may be reduced if data from unpublished studies can be included (73). Effect modification can also be examined when individual level data are available. Finally, to obtain patient data from different studies, several research teams must cooperate, effectively enhancing collaboration between study centres.

Due to its important advantages over meta-analysis of published results, pooled analysis has been described as the gold standard for synthesising evidence from different studies (73, 145-147). Since the method was initially proposed, the number of systematic reviews that have conducted pooled analysis of individual data has increased, similar to the increase seen in meta-analyses. As reported by Simmonds et al (148), the method was identified in only 6 publications in the period 1991-1992, and the number increased to 16 during 1993-1995, 33 during 1996-1998, and 43 during 1999-2001. Riley et al have also reviewed the articles that have applied this method and have reported a total of 383 published articles up to March 2009, with an average number of 49 articles per year between 2005 and 2009 (144). They suggested that this increase was due to an increased awareness of the benefits of pooled analysis as compared to the meta-analysis of published results.

Despite the strong advantages of pooled analysis of individual data, the method also has several drawbacks. It is resource-intensive and requires a variety of skills (i.e. scientific, computing, data management) (143). Moreover, it is time-consuming because it involves negotiating collaborations, and a long period normally elapses from the first communication with the authors until final acquisition of the data, data management and eventually data analysis. In addition, if authors are not contactable and thus do not provide individual data, only a selection of all studies will eventually be included in the pooled analysis, with the possibility that those that are available are systematically unrepresentative, leading to selection bias (149).

As both statistical approaches for synthesising evidence are prone to selection bias, and given the time and resource requirements of pooled analysis of individual data, meta-analysis will continue to find applications. Therefore, it is important to explore how well estimates from the two methods compare, and what determines any differences in results when they occur. This might enable confident use of meta-analysis of results when derived estimates were likely to be very close to those that one would get from pooled analysis, with limitation of pooled analysis of individual data to circumstances in which this did not apply.

Several studies have compared results from pooled analysis of individual data and from meta-analysis of published results. They have found that meta-analysis estimates were very close to those derived from pooled analysis. However, most of them compared results from application of the two methods to data that were not derived from identical populations (i.e. different studies were included for the two methods) (150-153), or analyses based on only a small number of studies (154, 155). No study so far has compared estimates from the two methods using multiple studies in which methods of data collection and ascertainment of exposure and disease were standardised.

The CUPID study is a multicentre investigation with participants recruited in 18 countries from up to 4 occupational groups. As described in section 0, the participants in the study are uniquely grouped into 47 occupational groups (Table 1.1) that are defined by type of work and country. These 47 groups can be regarded as providing 47 distinct and independently conducted studies for which the methods of data collection and ascertainment of exposure and health outcomes were standardised. Thus, the CUPID study provides a good opportunity to explore agreement between the two statistical approaches for synthesising evidence in a setting of 47 similarly conducted studies. Also, identifying situations in which the two methods may disagree, even when combining effect estimates derived from studies of standardised methodology, would be valuable information for the choice of analytical approach in future studies of secondary research.

This chapter compares estimates of effect and related precision derived from several methods of meta-analysis and pooled analysis of individual data, using data from the CUPID study. To assess possible influences on differences in the estimation of effect from the two methods, various combinations of outcome and binary predictor variables were examined.

7.2 Methods

The outcome variables considered were the one-month prevalence of disabling low back, wrist/hand, and elbow pain, hereafter described as ‘back pain’, ‘hand pain’ and ‘elbow pain’, chosen to cover a range of prevalence rates (6% to 22%). Pain was defined as disabling if it had made it difficult or impossible for the participant to carry out any of a specified list of everyday activities (36).

The explanatory variables examined were chosen because they had been shown previously to be significantly associated with the pain outcomes of interest. They were relevant occupational activities (in an average working day), reported distress from common somatic symptoms (somatising tendency), age, and sex (26, 156). The occupational activities examined were specific to the pain outcome. Lifting weights of 25kg or more by hand, use of a keyboard for >4 h in total and/or other tasks involving repeated movements of the wrists or fingers for >4 h in total, and repeated elbow bending for one hour or more were examined as risk factors for back, wrist/hand, and elbow pain, respectively. Somatising tendency was classified according to the number of somatic symptoms from a total of 5 that had been at least moderately distressing in the past week (categories: 0, 1, and 2+). Age was categorised in 10 year age bands (20-29, 30-39, 40-49, and 50-59).

Unadjusted and mutually adjusted effects of exposure variables on the pain outcome variables were estimated from models fitted by following the statistical approaches described below.

Meta-analysis

As described in the introduction, meta-analysis combines summary results from different studies to create an overall pooled estimate. Here, to meta-analyse summary estimates from the 47 occupational groups, standard logistic regression models, as described in the equation

$$logit(p_{ij}) = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \sum_{k=1}^{n} \beta_{kj} x_{kij}\, , for\ each\ j = 1, \ldots, 47 \tag{7.1}$$

$$y_{ij} \sim Bernoulli(p_{ij}), i{:}\ individual, j{:}\ occupational\ group$$

were initially fitted to explore associations between k $(k = 1, \ldots, n)$ exposure variables (x_{kij}) and the outcome (y_{ij}) variable separately in each occupational group. In this model, for a given occupational group j, β_{0j} is the log odds of the outcome for individuals with exposures x_{kij} $(k = 1, \ldots, n)$ equal to zero (unexposed individuals), also referred to as baseline risk, and $\beta_{kj}'s$ are the log odds ratios comparing the odds of the outcome between exposed $(x_{kij} = 1)$ and unexposed $(x_{kij} = 0)$ individuals. Also p_{ij} is the probability of the occurrence of the outcome conditional on the exposure $(p_{ij} = \Pr(y_{ij}|x_{kij}))$. From each group-specific logistic regression model, estimates of β_{kj} were derived and saved and then combined in meta-analysis models. As for some occupational groups the log odds ratio β_{kj} could not be estimated (due either to there being zero events for the outcome or exposure variable, or to problems with estimation), the number of occupational groups finally included in the meta-analysis was recorded. Estimates of baseline risk β_{0j} for the 47 occupational groups were also saved.

The (up to 47) log odds ratio estimates ($\hat{\beta}_{kj}$) were used to test for heterogeneity using the *Q*-statistic (157), which follows a χ^2_{j-1} distribution. The derived p-value from this test was taken as a measure of the level of statistical heterogeneity with lower values indicating more heterogeneous effect estimates.

Fixed effects model (MA_FE)

The fixed effects meta-analysis model assumes a common effect estimate across the different studies included in the meta-analysis, with a small variation that is due to sampling error alone. Given homogeneity of effect estimates, several methods have been suggested for pooling log odds ratios, including the Mantel-Haenszel method (158), the inverse variance method (159), and Peto's method (160). The inverse variance method was chosen here as it is widely used and can be directly compared with the random effects method for heterogeneous effect estimates.

Following the inverse variance method, also known as Woolf's method, the group-specific estimates $\hat{\beta}_{kj}$ were weighted according to the inverse of their variance $\left(w_j = 1/var\left(\hat{\beta}_{kj}\right)\right)$. The pooled estimate was then calculated as the weighted average of all separate estimates $\left(\hat{\beta}_k = \sum_{j=1}^{47} w_j \hat{\beta}_{kj} / \sum_{j=1}^{47} w_j\right)$.

Random effects model (MA_RE)

The random-effects meta-analysis model was suggested by DerSimonian and Laird (161) as an alternative method when variation of the effect estimates across studies included in the meta-analysis cannot be explained by sampling error alone, and systematic differences between estimates are assumed to be present. With this method, the effects estimated from logistic models fitted separately in each occupational group were weighted by $w_j = 1/\left(var\left(\hat{\beta}_{kj}\right) + \sigma^2\right)$, with $var\left(\hat{\beta}_{kj}\right)$ being the variance of the effect estimate for the occupational group j, and σ^2 the variance between occupational groups. The pooled effect was then calculated as in the MA_FE model.

Pooled analysis of individual data

Logistic regression – no group adjustment (LR_nga)

In this approach the clustering of observations within occupational groups is ignored and data are analysed together as if they come from a single group. The standard logistic regression model for a binary outcome y_{ij} and k exposure variables x_{kij} ($k = 1, \dots, n$), is similar to that described in 7.1:

$$logit(p_{ij}) = \beta_0 + \sum_{k=1}^{n} \beta_k x_{kij}, \quad \forall j \tag{7.2}$$

As in equation 7.1, β_0 is the log odds of the outcome for individuals with x_{kij} equal to zero for all k's (unexposed individuals), and β_k is the log odds ratio comparing the odds of the outcome between exposed ($x_{kij} = 1$) and unexposed individuals ($x_{kij} = 0$) for a given exposure k, keeping the rest of the covariates constant.

Logistic regression – dummy variable group adjustment (LR_dvga)

With this approach the model described in equation 7.2 is expanded by including dummy variables for the occupational groups to which individuals belong. With these covariates in the model, the main effect of the exposure variables x_{kij}, $k = 1, \dots, n$, is corrected for the effect that each occupational group has on the outcome variable. As there were 47 occupational groups, the model incorporated 46 dummy variables and was specified as:

$$logit(p_{ij}) = \beta_0 + \sum_{k=1}^{n} \beta_k x_{kij} + \sum_{m=2}^{47} \beta_m group_{im}$$

where β_m $(m = 2, \dots, 47)$ was the log odds of the outcome of the m-th occupational group as compared to the first group.

Pooled analysis random intercept model (PA_RI)

This model has been described in Chapter 5. In brief, the occupational group effects are now incorporated in the model as group-specific intercepts that follow a normal distribution around a mean value of β_0 with variance σ_u^2. The random intercept model can be described by the following equation:

$$logit(p_{ij}) = \beta_{0j} + \sum_{k=1}^{n} \beta_k x_{kij}$$

with index i denoting the individual, j the occupational group to which the individual belongs, and k the exposure. Here, the intercept of the model, β_{0j}, represents the log odds of the outcome when x_{kij} for all k's is zero, and is specific for each occupational group. In other words, based on this analytical approach each occupational group is allowed to have a different baseline risk (log odds of the outcome event for the unexposed) distributed normally around an overall log odds β_0. This model also assumes that the log odds of the outcome event among the exposed as compared to the log odds among the unexposed is constant across occupational groups and equal to β_k for a given exposure k, while the rest remain constant.

Pooled analysis random coefficient model (PA_RC)

This model is an expansion of the multilevel random intercept model such that the ratio of log odds between exposed and unexposed individuals ($\beta_k, k = 1, \dots, n$) is allowed to vary across occupational groups. This model has also been described in section 2.3 as the RE model. The model specification is very similar to that of the random-intercept model, and is described by the equation:

$$logit(p_{ij}) = \beta_{0j} + \sum_{k=1}^{n} \beta_{kj} x_{kij}$$

Here, β_{kj} is the log odds ratio that is specific to each occupational group, for a given exposure k. In this model, β_{0j} and β_{kj} are distributed normally around an overall β_0 and β_k, respectively. Based on the model, any assumption about common baseline risk and risk ratio across the different groups is relaxed as they both vary.

Statistical analysis

Effects of the explanatory variables were estimated from the models described above, initially in univariate analysis and then mutually adjusted. Parameter estimates $\left(\widehat{OR} = \exp(\hat{\beta}_k)\right)$ from the fitted statistical models and the corresponding 95% confidence intervals were saved, and were compared graphically. In addition to $\widehat{OR}$s, SEs of $\hat{\beta}_k$ were saved and compared between different models. For each two analytical methods, the measures of comparison for $\widehat{ORs}$ and $se(\hat{\beta}_k)$ were defined as

$$Ratio\ of\ OR\ (ROR) = \frac{OR\ from\ first\ method}{OR\ from\ second\ method}$$

$$Ratio\ of\ SEs\ (RSE) = \frac{SE(\hat{\beta}_k)\ from\ first\ method}{SE(\hat{\beta}_k)\ from\ second\ method}$$

7.3 Results

Description of the outcome and exposure variables across the 47 occupational groups

The overall prevalence of the outcome and exposure variables, and the distribution of prevalence rates across the 47 occupational groups, are described in Table 7.1 and illustrated graphically in Figure 7.1. Among the three pain outcomes, back pain was the most prevalent overall (22.5%), and elbow pain the least prevalent (6.2%). Also, elbow pain was the outcome for which prevalence across the 47 occupational groups varied the least. Overall prevalence rates were high for most of the exposure variables examined (21% to 74%), but were somewhat lower for reporting distress from two or more common somatic symptoms (18%) and being 50-59 years old (16%). The prevalence of relevant occupational activities, young age (20-29), and the proportion of females varied considerably, while that of reporting distress from one common somatic symptom varied less across occupational groups (range for one somatic symptom: 8-31%).

Ratios of the effect estimates and of the associated precision from the univariate models are presented in Table 7.2, while those from the models in which the effects of the explanatory variables were mutually adjusted are presented in Table 7.3. For ease of description and interpretation, only ratios higher than 1.1 corresponding to differences greater than 10% or an equivalent of 0.91, are shown.

Graphical illustration of the comparison of estimates of effects and related precision from the different models is given in Appendix 1.

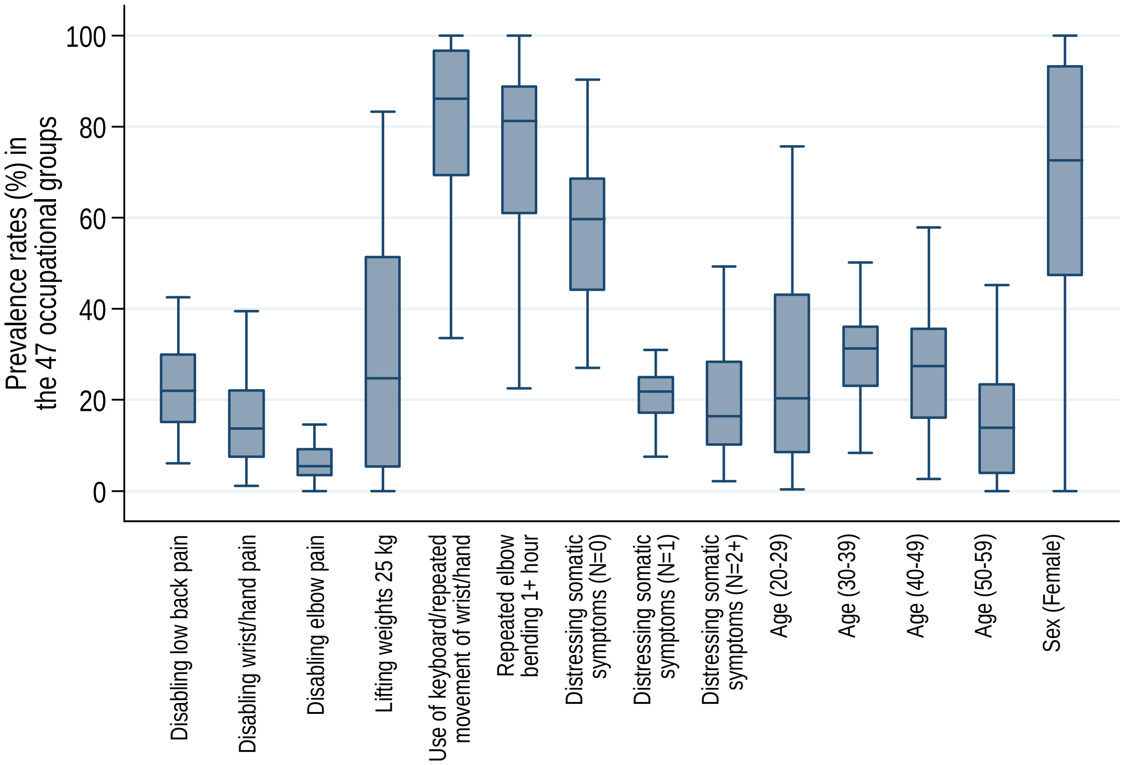


Figure 7.1. Box and whisker plot showing the prevalence (%) of the outcome and the exposure variables across the 47 occupational groups. For each outcome/exposure variable presented here, the line that cuts through each box represents the median prevalence rate, the top and bottom of the box

represent the 75th centile (upper quartile) and 25th centile (lower quartile) of the distribution of the prevalence rates, and the lines above and below the box are the two whiskers (indicating the largest and the smallest extremes of the distribution excluding exceptional outliers)

Table 7.1. Prevalence rates (%) of the outcome and the explanatory variables across the 47 occupational groups

	Mean of prevalence rates	**Median of prevalence rates**	**Range of prevalence rates**	**Interquartile range (IQR) of prevalence rates**
Outcomes				
Disabling low back pain	22.5	22.0	6.1-42.6	15.1-30.0
Disabling wrist/hand pain	15.8	13.7	1.1-39.5	7.5-22.1
Disabling elbow pain	6.2	5.5	0.0-14.6	3.5-9.2
Explanatory variables				
Occupational activities				
Lifting weights 25 kg or more	29.7	24.8	0.0-83.3	5.4-51.3
Use of keyboard or other repeated movement of wrist/hand	80.4	86.1	33.6-100.0	69.3-96.7
Repeated elbow bending 1+ hour	73.0	81.3	22.6-100.0	60.9-88.8
Number of distressing somatic symptoms				
0	58.5	59.7	27.1-90.3	44.1-68.6
1	21.3	21.8	7.5-31.0	17.2-25.0
2+	19.3	16.4	2.1-49.3	10.1-28.4
Age (years)				
20-29	26.4	20.3	0.4-75.7	8.5-43.1
30-39	30.7	31.3	8.4-67.0	23.0-36.1
40-49	27.4	27.4	2.6-57.9	16.1-35.6
50-59	15.5	13.9	0.0-45.2	4.0-23.5
Sex (Female)	66.3	72.6	0.0-100.0	47.4-93.2

Comparison of estimates from different methods of meta-analysis

In both unadjusted and adjusted analyses, effect sizes estimated from the MA_FE models were very similar to those estimated from the MA_RE models (RORs ranged from 0.97 to 1.03). However, the estimated SEs differed between the two models with the RSEs diverging from the value of 1 as heterogeneity increased. Specifically, RSEs higher than 1.1 or lower than 0.91 were observed when p-values of the *Q*-statistic were approximately equal to 0.2 or less, with higher SEs estimated from MA_RE than from MA_FE (all RSEs<1). The lowest ratios of SEs estimated from the two meta-analytical models was for the effect of the highest age band on wrist/hand pain, for which the occupational group-specific effects were highly heterogeneous (p=0.002 and p=0.001, respectively for the unadjusted and adjusted analyses).

Comparison of estimates from different methods for pooled analysis of individual data

Pooled analysis random intercept (PA_RI) versus Pooled analysis random coefficient (PA_RC) model

Effect estimates derived from PA_RI and PA_RC were very similar for all associations examined. RORs ranged from 0.99 to 1.09 with RORs further from 1 seen where group specific effects were more heterogeneous. Relative differences in SEs from the two models were greater than 10% (RSEs larger than 1.1 or lower than 0.91) mostly where group-specific effect estimates were heterogeneous at a 20% significance level. Exceptions were differences in SEs greater than 10% observed in the unadjusted associations between age (40-40 v 20-29) and wrist/hand pain (RSE=0.880, heterogeneity p-value=0.428), and occupational activity and wrist/hand pain (RSE=0.896 heterogeneity p-value=0.407); and in the adjusted association between somatising tendency (2+ v 0) and wrist/hand pain (RSE=0.861, heterogeneity p-value=0.420). In all cases, SEs from the PA_RC models were higher than those from the corresponding PA_RI models (RSEs<1).

Logistic regression – dummy variable group adjustment (LR_dvga) versus Pooled analysis random coefficient (PA_RC) model

Effect estimates and associated SEs from the PA_RI model were generally very similar to those from the PA_RC model. Thus, comparisons between estimates from any given model and those from the PA_RI model closely mirrored comparisons between the same estimates and those from the PA_RC model. For simplicity, in what follows, I will only describe comparisons with estimates from the PA_RC model.

Effect sizes from LR_dvga and PA_RC were also similar for all associations. All ORs from LR_dvga were less than 10% different from those derived from the PA_RC model (RORs ranged from 0.92 to 1.09). Estimates of SEs from the two methods differed by more than 10% for 10 of the 42 associations examined (7 unadjusted and 3 adjusted), with SEs from LR_dvga models being lower than those from PA_RC models (RSEs<1). For almost all of the associations with differing estimates of SE, the group-specific effect estimates were heterogeneous at a 20% level (p-value for heterogeneity <0.2), and differences were greater when p-values for the heterogeneity of group effects were lower. Differences in estimated SEs greater than 10% for homogenous group specific estimates (heterogeneity p-value >0.2) were observed for the three associations reported in the comparison of SEs from the PA_RI and PA_RC models described above (i.e. unadjusted associations of age (40-49 v 20-29) and occupational activity with wrist/hand pain, and adjusted association of somatising tendency (2+ v 0) with wrist/hand pain).

Logistic regression – no group adjustment (LR_nga) versus Pooled analysis random coefficient (PA_RC) model

Estimates of effect sizes and SEs from LR_nga models differed greatly from those from PA_RC models. All RORs were higher (i.e. further from the null), and all RSEs were lower than 1, indicating that point estimates appeared higher and more precise when no adjustment was made for the clustering of observations within occupational groups. Larger differences in the point estimates were observed for the (unadjusted and adjusted) association between sex and wrist/hand pain, and the largest RSE was seen for the association between age (40-49 v 20-29) and wrist/hand pain. For the latter association, the SE from the PA_RC model was almost twice as high as that from the LR_nga model.

<u>Comparison of estimates from meta-analysis and pooled analysis of individual data</u>

Meta-analysis Random-effects (MA_RE) versus Pooled analysis random coefficient (PA_RC) model

Relative differences between MA_RE and PA_RC models of 10% or more were seen for most of the (unadjusted and adjusted) associations examined, mainly in the SEs of ln(ORs) ($se(\hat{\beta}_1)$) derived from the two models. Most RSEs were greater than 1 (few RSEs not presented in the tables were <1 as they were >0.91), meaning that the estimated precision of the effect from the MA_RE model was always lower than that from the PA_RC model. The differences in the SEs generally increased as the number of occupational groups that did not contribute to the meta-analysis increased (Spearman correlation between number of studies excluded from the estimation of ORs and RSEs = 0.56). Exceptions were the unadjusted effect of age (40-49 v 20-29) on back pain, and the mutually adjusted effects of somatisation (2+ v 0) and age (40-49 v 20-29) on back pain, in which the relative difference in estimated SEs from the two methods was more than 20% when the number of occupational groups that did not contribute to the meta-analysis was five or fewer. However, for these associations (especially the adjusted ones) heterogeneity of the group-specific effects was high, resulting in increased imprecision from the MA_RE model. For the associations for which group specific estimates were homogeneous across occupational groups, and for which most of the occupational groups contributed to the estimation of the pooled effect, RSEs were very close to 1 (range: 0.95 – 1.14).

Most RORs were less than 1, indicating lower estimates (i.e. closer to the null) from the MA_RE model. As for the RSEs, RORs were also further from 1 as the number of occupational groups excluded from the meta-analysis increased (Spearman correlation = 0.71). The largest differences in estimated ORs from the two methods were seen when the outcome was elbow pain. Exceptions were the (adjusted and unadjusted) associations of age (50-59 v 20-29) with wrist/hand pain and that of the same predictor with back pain. However, unlike RSEs, RORs were not influenced by heterogeneity of group specific effects.

To explore further the differences in estimates derived from MA_RE and PA_RC, PA_RC models were refitted after exclusion of the occupational groups that were not included in the meta-analysis owing to zero events or problems with estimation (last column in tables 2 and 3). With regard to estimated precision, results were very similar to those without exclusion of the occupational groups, with discrepancies between the meta-analysis and the pooled analysis remaining. However, several differences in point estimates observed between MA_RE and PA_RC reduced in size (ROR close to 1) when the occupational groups excluded from the meta-analysis were also excluded from the pooled analysis of individual data.

Table 7.2. Comparisons between unadjusted estimates of associations with disabling pain from different statistical models. Only RORs and RSEs >1.1 or <0.91 are shown

	Explanatory variable (Unadjusted)	N groups excluded from MA	Hetero-geneity χ^2 value	Hetero-geneity P value	SD of Intercepts	MA_FE vs MA_RE		PA_RI vs PA_RC		LR_dvga vs PA_RC		LR_nga vs PA_RC		MA_RE vs PA_RC		MA_RE vs PA_RC [¥]	
						ROR	RSE	ROR	RSE	ROR	RSE	ROR	RSE	ROR	RSE	ROR	RSE
Disabling Low Back Pain	Occupational activity	5	48.46	0.197	0.65		0.886						0.749		1.107		1.108
	Somatising (1 v 0)	0	44.3	0.544	0.58												
	Somatising (2 v 0)	2	61.69	0.04			0.813		0.859		0.864	1.151	0.801		1.110		1.113
	Age 30-39 v 20-29	5	40.78	0.48	0.69								0.891				
	Age 40-49 v 20-29	5	50.31	0.151			0.866						0.904		1.272		1.249
	Age50-59 v 20-29	13	38.81	0.224			0.892						0.855		1.311		1.192
	Female	10	53.06	0.033	1.24		0.782		0.883		0.905	1.183	0.658		1.191		1.175
Disabling Wrist/Hand Pain	Occupational activity	16	31.17	0.407	0.90				0.896		0.900	1.240	0.809	0.901			
	Somatising (1 v 0)	3	36.55	0.746	0.84							1.156					
	Somatising (2 v 0)	3	50.61	0.198			0.900		0.866		0.868	1.310	0.803				
	Age 30-39 v 20-29	9	27.28	0.879	0.92												
	Age 40-49 v 20-29	10	36.87	0.428					0.88		0.886		0.780				
	Age50-59 v 20-29	16	56.37	0.002			0.696		0.668		0.673		0.585	1.158	1.230		1.179
	Female	11	42.14	0.189	1.24		0.879					1.329	0.743				
Disabling Elbow Pain	Occupational activity	14	24.15	0.839	0.83							1.217	0.852	0.733		0.860	
	Somatising (1 v 0)	6	35.12	0.689	0.73												
	Somatising (2 v 0)	7	28.88	0.882								1.166	0.891				
	Age 30-39 v 20-29	24	11.13	0.973	0.88									0.838	1.203		1.101
	Age 40-49 v 20-29	23	11.55	0.977											1.179		
	Age50-59 v 20-29	27	11.73	0.897					0.905				0.831	0.884	1.299		
	Female	20	33.69	0.143	1.38		0.844		0.855		0.900		0.671		1.165		

[¥] Analysis for the PA_RC restricted to the occupational groups that were included in MA_RE model

ROR: Ratio of ORs; **RSE**: Ratio of SEs of ln(OR); **MA_FE**: meta-analysis fixed effect model; **MA_RE**: meta-analysis random effect model; **LR_nga**: logistic regression model without group adjustment; **LR_dvga**: logistic regression model with dummy variable group adjustment; **PA_RI**: pooled analysis random-intercept model; **PA_RC**: pooled analysis random-coefficient model;

Table 7.3. Comparisons between mutually adjusted estimates of associations with disabling pain from different statistical models. Only RORs and RSEs >1.1 or <0.91 are shown

	Predictor (Mutually adjusted)	Number of groups excluded from MA	Hetero-geneity χ^2 value	Hetero-geneity P value	MA_FE vs MA_RE		PA_RI vs PA_RC		LR_dvga vs PA_RC		LR_nga vs PA_RC		MA_RE vs PA_RC		MA_RE vs PA_RC ¥	
					ROR	*RSE*	*ROR*	*RSE*	*ROR*	*RSE*	*ROR*	*RSE*	*ROR*	*RSE*	*ROR*	*RSE*
Disabling Low Back Pain	Occupational activity	6	42.4	0.368								0.774		1.101		
	Somatising (1 v 0)	1	42.52	0.577												
	Somatising (2 v 0)	3	65.11	0.016		0.778		0.865		0.871	1.125	0.819		1.205		1.197
	Age 30-39 v 20-29	4	45.26	0.338										1.144		1.135
	Age 40-49 v 20-29	5	56.7	0.052		0.805								1.410		1.383
	Age50-59 v 20-29	13	38.73	0.227		0.893							0.900	1.420	0.894	1.314
	Female	11	50.89	0.04		0.787						0.716		1.363		1.291
Disabling Wrist/Hand Pain	Occupational activity	16	31.82	0.376							1.136			1.121		1.103
	Somatising (1 v 0)	3	39.25	0.635												
	Somatising (2 v 0)	3	44.21	0.42				0.861		0.864	1.18	0.812				
	Age 30-39 v 20-29	9	30.97	0.747										1.119		
	Age 40-49 v 20-29	10	41.99	0.227		0.901		0.905				0.818		1.182		1.174
	Age50-59 v 20-29	16	59.83	0.001		0.675		0.663		0.672	1.117	0.594	1.116	1.291		1.244
	Female	11	38.43	0.317							1.216	0.812		1.143		
Disabling Elbow Pain	Occupational activity	20	21.12	0.736							1.108		0.812	1.216	0.895	
	Somatising (1 v 0)	12	42.27	0.156		0.868						0.908	1.101	1.232		1.155
	Somatising (2 v 0)	14	25.76	0.774							1.114	0.896		1.112		
	Age 30-39 v 20-29	24	13.47	0.919									0.857	1.27		1.131
	Age 40-49 v 20-29	24	15.32	0.848									0.900	1.256		1.136
	Age50-59 v 20-29	28	15.47	0.63								0.856	0.768	1.415	0.883	1.175
	Female	26	25.19	0.194		0.856						0.742		1.456		1.230

¥ Analysis for the PA_RC restricted to the occupational groups that were included in MA_RE model

ROR: Ratio of ORs; **RSE**: Ratio of SEs of ln(OR); **MA_FE**: meta-analysis fixed effect model; **MA_RE**: meta-analysis random effect model; **LR_nga**: logistic regression model without group adjustment; **LR_dvga**: logistic regression model with dummy variable group adjustment; **PA_RI**: pooled analysis random-intercept model; **PA_RC**: pooled analysis random-coefficient model;

7.4 Discussion

Combining evidence from different studies is important in gauging the likelihood and strength of associations between exposures and health outcomes. Thus, it is essential to assess the validity of estimates derived from the analytical approaches that are widely used. The focus of this chapter was to compare meta-analysis of results with pooled analysis of individual data. I found that in most cases, the two analytical approaches produced point estimates that were very similar, the differences being of little practical importance (most RORs ranged between 0.9 and 1.1, indicating differences <10%). However, when either the outcome (i.e. elbow pain) or the exposure under investigation (i.e. highest age band) was rare, estimates derived from meta-analysis of study-specific results differed from those derived from pooled analysis of individual data. In most such cases, effect estimates from meta-analysis were lower (i.e. closer to the null) than those from pooled analyses of individual data that took into account within-study clustering of observations. In contrast to point estimates, precision from the two approaches differed more. Most standard errors estimated from meta-analysis differed from those estimated from pooled analysis by more than 10%. Those derived from meta-analysis of results were greater than those derived from pooled analysis. Also, comparison of different statistical approaches to pooled analysis of individual data showed that accounting for clustering of observations within studies resulted in different estimates of effect and related precision than when no adjustment for clustering was made. All standard errors of effect estimates from pooled analysis of individual data that adjusted for clustering were larger than from those that did not, while most of the effect estimates were lower.

The comparison of meta-analytical estimates to those derived from pooled analysis of individual data was based on real data taken from the CUPID study rather than simulated data. Using real data, assessment of the appropriateness of a statistical model for the estimation of a parameter cannot be conclusive as the true underlying value of the parameter remains unknown. Also many statistical properties are based on large sample approximations, but large samples are not always possible in real studies. Moreover, conclusions from analysis of real data are restricted to the scenarios that arise in the specific dataset, while in simulations a wide range of conditions can be examined. However, the CUPID study is a rich data source in which outcomes and risk factors of varying prevalence rates could be analysed covering a range of different data scenarios.

Although there is a wide range of meta-analytical models one could fit when the outcome of interest is binary, in this analysis only the inverse variance and the DerSimonian and Laird methods were used. However, the inverse variance method is the method most commonly applied in published studies (162), while the DerSimonian and Laird method is an expansion of the inverse variance method that accounts for significant heterogeneity of the effect estimates across studies.

Unlike Chapters 4 and 5 that only considered RI models to account for clustering, the analysis presented here also included the RC model. This model was also fitted as it is more comprehensive than the RI model in accounting for potential heterogeneity of the effect estimates across studies. In addition, in such a case of high heterogeneity of effect estimates, the RC model would offer a more appropriate comparator with the MA_RE model.

The merits of using primary data from all studies in a pooled analysis have been well described in previous reports (144, 163, 164). In brief, it allows for unpublished data to be included when synthesising information, while enabling researchers to apply more sophisticated modelling methods and to account statistically for individual-level effects and test for possible interactions. However, it is time-consuming and costly as a collaboration between multiple research teams must be established. Therefore, meta-analysis of published data is still potentially of great value, and it is important to check its validity.

Some authors have argued that aggregating findings from clinical trials presents fewer problems than meta-analysis of observational studies, as they are more uniform in design and less prone to bias. Also, unlike in observational studies, in clinical trials, the prevalence of relevant exposures (treatments) tends to be similar across studies, and generally not far from 50%. Previous studies have usually shown good agreement between conclusions drawn from meta-analysis of published results and from pooled analysis of individual data from clinical trials (165, 166). In one investigation, the results differed considerably but the studies analysed by the two approaches were not the same (145).

Combining evidence from observational studies is challenging due to the variety of the study designs employed, with differing methods of data collection and ascertainment of disease and exposure. However, observational studies are important as trials may be unnecessary, inadequate, inappropriate, or impossible (167). Several researchers have compared results from the two analytical approaches for combining evidence from observational studies, but in most cases, the two methods were not applied to exactly the same studies (150-153), which complicates interpretation. Gordon et al (154) combined information from 12 case-control studies of the association between sinonasal cancer and occupational exposure to wood dust, using meta-analysis of results and pooled analysis of individual data, and showed that the two methods gave fairly consistent results. However, the confounders for which the main association was adjusted differed between the studies included in the meta-analysis, which could be considered as an additional source of error in the estimation of the overall effect. A later investigation by Tobias et al (155) used a subset of the studies presented in Gordon et al (154) to explore the same research question, using the same adjustments in the study-specific analyses before combining the results in the meta-analysis model. This study also compared results from different approaches to pooled analysis of primary data (fixed- and random-effects models). The odds ratio for sinonasal cancer from wood dust exposures was estimated as 2.4 from the meta-analysis model (random-effects model due to significant heterogeneity of effect estimates) and 2.9 from the pooled analysis of individual data (same estimate from both fixed-effects and random-effects). The authors of the study commented that estimates from pooled analyses are more precise than estimates derived from meta-analysis, but no evidence of this was given.

In contrast to earlier investigations, the analysis presented in this chapter combined data from a large number of occupational groups (N=47) (regarded as separate studies) that all came from a single multicentre study. This offered the advantage that procedures for data collection and definitions of variables were standardised, creating an ideal situation in which the question about comparability of estimates derived from the two statistical approaches could be explored. That is an advantage over what usually happens in practice where estimates from studies of varying designs and methods are combined, as any differences observed in the CUPID data between estimates and related precision can only be attributed to the differences in the underlying statistical models fitted rather than methodological heterogeneity between studies. As in previous studies, I found that in most cases, estimates from the two methods were fairly similar. The great majority of relative differences in point estimates were <10%. However, differences were sometimes greater when the prevalence of either the outcome or the exposure was very low (or very high). In this circumstance, zero cells occurred in the 2x2 tables of exposure versus outcome for some of the occupational groups (sub-studies), and therefore the effect estimates for those studies were not included in the meta-analysis, although the studies did contribute to the pooled analysis. Unlike point estimates, differences >10% in standard errors occurred in most of the (unadjusted and adjusted) associations explored. Ratios of standard errors from meta-analysis to those from pooled analysis of individual data ranged from 0.95 to 1.46, with approximately half of the ratios being >1.15. Most of the ratios of standard errors (86%) were >1 indicating lower precision of the summary estimate from the meta-analytical approach. As for the point estimates, estimated standard errors differed more in the case of rare outcomes and when the number of studies (i.e. occupational groups) excluded from the meta-analysis was larger.

The problem of sparse data for the estimation of an effect and its corresponding precision, has been identified and described before (168, 169). A number of different corrections have been suggested to account for the exclusion of a considerable number of studies from meta-analysis due to zero events. These correction methods have been applied to both simulated and real data and the relevant merits and problems have been discussed. The work presented in this chapter did not attempt to explore how estimates derived from meta-analysis with and without correction compare to each other, or how estimates from pooled analysis of individual data compare to those from meta-analysis after applying zero-cells corrections.

To explore reasons for the differences in estimates from the two methods, I restricted the pooled analysis of individual data to those studies that were included in the meta-analysis. Most of the differences in the point estimates were resolved. However, with regard to estimated standard errors, the findings from this sensitivity analysis were very much the same, differences remaining between the two methodological approaches. The direction of the differences in estimated precision was always the same, with standard errors of the meta-analysis effect estimates being higher than those for the pooled analysis estimates. The explanation for discrepancies when the outcome was rare therefore remains unclear. Future work in this area should explore the question further by simulation studies, but will not be taken further here.

This chapter also compared results from different methods of pooled analysis. The methods investigated were logistic regression a) without adjustment for group/study participation, b) with dummy variables to adjust for group/study participation, c) with random intercepts, and d) with random coefficients. As individual observations were clustered within occupational groups/studies, a comparison of the first method with the rest was essentially a comparison between methods that accounted for clustering and those that did not. Such comparison has been described in detail in Chapter 5. In the analysis presented in this chapter, I found that when clustering within occupational groups was not taken into account in the statistical model, effect estimates were higher (i.e. further from the null) than those derived from models that controlled for clustering, and that the corresponding standard errors were downwardly biased. The direction of bias in standard errors was the same as that found in analysis of simulated data described in the previous chapter. Unlike bias in precision, however, bias in point estimates was the opposite to that which was found in the more extended exploration of consequences of ignoring clustering using simulations. As shown in sections 5.3.2 and 5.5.2, point estimates tended to be underestimated (i.e. closer to the null value of one) when the statistical approach failed to account for clustering. However, differences in effect estimates derived from a cluster-adjusted (RI) and a cluster-unadjusted (OL) model varied quite considerably around the average value of bias (Figure 5.20), with many differences indicating a higher value for the point estimate from the naïve logistic model than from the cluster-adjusted model (negative values in Figure 5.20). Given the wide variation of differences in point estimates from the two analytical approaches and the fact that observations from the simulated data agree with numerous previous studies that have reported a bias towards the null when clustering is ignored, the discrepant observation made in this chapter may be attributable to chance.

In a similar context to this chapter, Abo-Zaid et al (108) applied a one-step and a two-step pooled analytical approach to the analysis of data from different studies (real and simulated data). The former approach was a simple pooling of data that ignored information on participants' study participation, while the latter was a method that accounted for study. They showed that when clustering was ignored, there was bias of effect estimates towards the null and low coverage by the 95% confidence intervals. Bravata & Olkin (170) have also suggested that simple pooling of data should be avoided as it is likely to yield misleading results and conclusions.

The three methods of pooled analysis that accounted for clustering of data in occupational groups yielded very similar results. Similarity of results from random-intercept logistic regression models and simple logistic regression models with dummy variables for the occupational groups (clusters) is not surprising as the latter is a method of analysis one could use in the absence of effect modification, when the number of clusters is small or when estimation of cluster-level effects was of interest. Also, results from the multilevel logistic random-intercept and random-coefficient models were very similar. That was to be expected where there was no heterogeneity of effect estimates across occupational groups, as was the case for most of the associations examined. In the small number of cases that heterogeneity was significant at a 20% significance level, differences in estimates from random-intercept and random-coefficient models increased with increasing level of heterogeneity. The maximum relative difference (SE from PA_RI < SE from PA_RC by 34%) between standard errors of effect estimates from the two approaches to pooled analysis was observed in the adjusted association between age (50-59 v 20-29) and wrist/hand pain for which effects across studies were markedly heterogeneous (p-value=0.001).

Results from work presented in this chapter give some reassurance that when synthesising evidence from different studies, point estimates from meta-analysis are often close to those obtained from pooled analysis of individual data that takes into account nesting of observations within studies. However, precision of ORs was lower (higher standard errors) when estimated from meta-analysis, and higher when estimated from pooled analysis of individual data. Discrepancies between the two methods arise when the prevalence of the outcome is very low or very high. Reasons for inconsistencies remain unknown. These could be explored further using simulation studies, but is outwith the scope of this thesis.

Chapter 8. Estimating relative risks from clustered data

8.1 Introduction

Binary outcome variables are commonly studied in epidemiological research. To estimate the effect of an exposure on a binary outcome, a logistic regression model is often used, from which the estimated effect is expressed as an odds ratio (OR). Previous chapters on the consequences of ignoring clustering and comparison of meta-analytical to pooled analysis approaches were based on estimation of ORs. The use of ORs as a measure of effect is particularly suitable for case-control studies, although they can also be applied in other study designs such as cohort or cross-sectional studies. However, their use in study designs other than case-control studies has been criticised with regard to their usefulness, appropriateness and interpretability (171). Also ORs have the disadvantage of non-collapsibility, which occurs when the estimated OR for a covariate of interest is the same across different levels of a second covariate (which is not a confounder), but different from the overall OR (172).

For these reasons, a measure of relative prevalence of the outcome in the exposed and unexposed groups, such as the relative risk (RR), is frequently preferred in epidemiological studies. Also, as the RR is expressed as proportion of the outcome in those exposed compared to those unexposed to a given risk factor, it is conceptually the obvious choice as a measure of association. However, it too has limitations. For example, RRs are subject to a ‘ceiling effect’, in that they are constrained by the maximum possible prevalence of 100%. This matters most when the outcome is common among unexposed persons. For example, if that prevalence is 50%, then RRs cannot exceed 2. Another limitation of RRs is the lack of symmetry for the association of non-exposure with absence of the outcome. For example, if the outcome has a prevalence of 70% in unexposed persons and 90% in those who are exposed, the relative risk of the outcome from exposure is 90/70 = 1.29, whereas the relative risk of not having the outcome in those who are unexposed is 30/10 = 3.0.

ORs have been shown to approximate RRs when the outcome variable of interest has a low incidence or prevalence rate, but when the outcome is more frequent, they are further from the null value of one. Therefore, especially in cohort and cross-sectional studies where the prevalence of an outcome variable is higher than 10%, RRs have been recommended as preferable, with direct estimation by log-binomial regression the analytical method of choice (173). However, the log-binomial model often fails to converge in the iterative process of parameter estimation. Several researchers have highlighted the problem, and have suggested various alternative analytical methods for the estimation of RRs. These include the derivation of RRs from ORs estimated by logistic regression, alternative regression models for direct estimation of RRs, or use of similar measures that can be interpreted as RRs. The main suggested alternative regression models are the Poisson model with robust SEs (also referred to as modified Poisson), Cox, and complementary log-log (CLL) models (74).

Previous studies have focused on comparisons between these models in the absence of clustering. However, the challenges in estimation of RRs apply also to data that are clustered, and comparison of methods for estimating RRs from hierarchically structured data has received little attention. Only three studies have addressed this problem. In the first (174), the authors used Poisson regression with robust SEs, and they also derived RRs indirectly from logistic regression models. The second study (77) entailed a large simulation exercise, comparing estimates from log-binomial regression to those from Poisson regression with robust SEs. The third study (75) also focused on the performance of the Poisson model. A later publication by Janani et al (175) discussed the methods that are available to researchers for estimation of RRs in a setting of clustered data, but did not present numerical results of comparisons.

The aim of this chapter is to review the main methods suggested in literature as ways to estimate RRs in the context of hierarchical data, and to apply those methods to data from the CUPID study, comparing the estimates of RR and their precision.

8.2 Methods

As in Chapter 6, the outcome variables considered were disabling low back, wrist/hand, and elbow pain in the past month, as reported at baseline. These were chosen to give outcome prevalence rates ranging from 6% to 22%. The explanatory variables considered were sex, age (in four 10-year age bands), occupational activity, and somatising tendency. The occupational activities examined were specific to the outcomes; lifting weights of 25kg or more for low back pain; use of a keyboard for >4 hours and/or other tasks involving repeated movements for pain in the wrist or hand; and repeated elbow bending for an hour or more for elbow pain. For each of the three outcome variables, relevant risk factors were first analysed univariately and then mutually adjusted, using the analytical approaches described below.

Log-binomial random-intercept (RI) model

The log-binomial model is a generalized linear model for binary outcomes with a *log* link function. For a binary outcome variable y_i and k explanatory variables $x_{i1}, x_{i2}, \dots, x_{ik}$, in the absence of clustering of observations of the outcome variable, the model can be written as

$$log\,P\,(y_i = 1|x_{1i}, x_{2i}, \dots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \qquad 8.1$$

where $\beta_1, \beta_2, \dots, \beta_k$ are the parameters to be estimated in the regression model. If x_{i1} is the binary exposure of interest, then $\exp(\beta_1)$ expresses the risk of $y_i = 1$ when the individual is exposed relative to being unexposed after adjustment for the effect of the other variables in the model $(x_{i2}, \dots, x_{ik})$. In the presence of clustering, to relax the assumption of independence of observations, a RI u_j, also described as a cluster-level error term, can be added to the prediction part of the log-binomial model in equation 8.1

$$log\, P\left(y_{ij} = 1 \middle| x_{1i}, x_{2i}, \dots, x_{ki}, u_j\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + u_j \qquad 8.2$$

with $u_j | x_{1i}, x_{2i}, \dots, x_{ki} \sim N\left(0, SD_{u_j}^2\right)$. The model described in equation 8.2 is the RI log-binomial model.

From equation 8.2, $P\left(y_{ij} = 1 \middle| x_{ij1}, x_{ij2}, \dots, x_{ijk}, u_j\right) = \exp\{x_{ij}^T \beta + u_j\}$. As the left hand side of this equation is a probability, it follows that $0 \leq \exp\{x_{ij}^T \beta + u_j\} \leq 1$, or that $x_{ij}^T \beta + u_j \leq 0$. The parameters estimates $\beta_1, \beta_2, \dots, \beta_k$ are obtained using the method of maximum likelihood (ML), which can sometimes occur on the boundary of the parameter space ($x_{ij}^T \beta + u_j \leq 0$), resulting in failure of the model to converge. The same may also occur when the default starting values are not appropriate, in which case, the problem of convergence may be solved by choosing different starting values.

Modified random-intercept (RI) logistic regression model

The RI logistic regression model has been described in previous chapters. In brief, the model uses the *logit* link function to describe the association between a binary outcome variable y_{ij} and a set of exposure variables, after adjustment for a cluster level error term u_j, as follows

$$\begin{aligned} logit\, P\left(y_{ij} = 1 \middle| x_{1i}, x_{2i}, \dots, x_{ki}, u_j\right) = logit(p_{ij}) = ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) \\ = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + u_j \end{aligned} \qquad 8.3$$

In this model, the exponential of a parameter β_k is the odds ratio describing the odds of $y_{ij} = 1$ in the exposed group of x_{kij} relative to the unexposed group after adjustment for the effect of the other variables in the model.

After fitting a RI logistic model, the estimated adjusted odds ratio (OR_{adj}) can be used to obtain an adjusted RR using the formula suggested by Zhang and Yu (176)

$$RR = \frac{OR_{adj}}{(1 - p_0) + \left(p_0 \times OR_{adj}\right)} \qquad 8.4$$

where p_0 is the prevalence of the outcome in the unexposed group. This formula was derived from simple algebraic manipulations presented in Appendix 2.

Because of bias when RRs are estimated in the context of complex patterns of covariates, Dwinvedi et al (177) suggested a modification in the Zhang and Yu formula in which p_0 is replaced by an adjusted predicted proportion of outcome in the unexposed group, $\widehat{p_0}$. This is expressed as

$$\widehat{p_0} = \frac{\exp(\widehat{\beta_0} + \widehat{\beta_2} + \cdots + \widehat{\beta_k})}{1 + \exp(\widehat{\beta_0} + \widehat{\beta_2} + \cdots + \widehat{\beta_k})} \qquad 8.5$$

where the parameter estimates $\widehat{\beta_0}$ to $\widehat{\beta_k}$ are derived from the random-intercept logistic model (equation 8.3).

Random-intercept (RI) Poisson model

Binomial data can be viewed as following a Poisson distribution with the probability of a value of 2 or more being very low, resulting in values of 0 or 1 for all individuals in a sample. As the Poisson model also uses a *log* link function, like the log-binomial model, its specification is the same as that described in equation 8.2. Thus exponentiation of the estimated parameters in a Poisson model approximates the RRs that one would get from a log-binomial model. The difference from the log-binomial model is that the errors in the Poisson model are assumed to follow a Poisson distribution.

It is well reported that the Poisson model, when used to approximate the binomial model, overestimates the SEs of RRs. The use of robust Sandwich estimation has been proposed as a method to correct this bias in SEs. The method has been applied by a number of researchers, and in the context of clustered data it has also been described as modified Poisson regression (76, 77).

Shared frailty model

Survival analysis techniques, particularly the Cox model, have been used to estimate hazard ratios (HRs). When survival time is constant across follow-up, the HRs are equivalent to RRs, and the Cox model can be said to equate to the Poisson model described above (Appendix 3). The Cox model assumes homogeneity of the population, meaning that all individuals in the study sample have the same risk, after adjustment for known risk factors. Often such an assumption is violated because of unmeasured covariates that are associated with the outcome of interest, resulting in clustering of observations.

When data are clustered, individuals belonging to the same cluster are likely to share the same excess risk, typically called the frailty value (178). Consequently, individuals are dependent within cluster, and independent conditional on the frailty. Shared frailty models thus provide an extension of the survival analysis models in the sense that they account for variation in the background frailty across clusters.

The hazard function in the shared frailty model is described by the following equation

$$\begin{aligned} h_{ij}(t) &= h_0(t)\exp\{\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + u_j\} \\ &= h_0(t) w_j \exp\{\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij}\} \end{aligned} \qquad 8.6$$

where w_j is the multiplicative effect for the shared frailty of individuals within the same cluster j.

Complementary log-log (CLL) model with robust SEs

The complementary log-log (CLL) model is an alternative to the logistic model, which uses the link function $\log(-\log(1-p_i))$. When observations are independent, the model is specified as

$$cloglog(p_i) = \log(-\log(1-p_i)) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \qquad 8.7$$

When observations are clustered, robust standard errors can be used to correct the standard errors derived from the above equation. Robust standard errors relax the assumption of independence of observations required by the CLL estimator to an assumption of independence only between clusters.

Random-intercept (RI) complementary log-log method

An alternative way of accommodating the dependence of observations when using the model described in equation 8.7, is to include a RI u_j as described in several models above. The model specification then becomes

$$cloglog(p_i) = \log\left(-\log\left(1-p_{ij}\right)\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + u_j \qquad 8.8$$

The exponentials of the effect estimates can be expressed as RRs.

Statistical analysis

Unadjusted and adjusted estimates of effect and their associated precision were derived for risk factors from the statistical models described above.

The log-binomial model has been described by various authors as the preferred method by which to estimate adjusted RRs directly, both when observations are independent (74, 179, 180) and clustered (77). For that reason, estimates obtained from the log-binomial model were used as the reference against which those derived from the other models were compared. For each of the models (i.e. the methods described by Zang & Yu (Z&Y), modified Zang & Yu (mod Z&Y), RI Poisson without and with robust SES, shared frailty, CLL with robust SEs, and RI CLL) comparisons of point estimates were summarised as the ratio of the RR from the method under consideration to that from the RI log-binomial model :

$$\frac{\text{RR from method}}{\text{RR from log} - \text{binomial method}}$$

To compare the precisions of point estimates obtained by different methods, SEs of the ln(RRs) derived from the method under consideration were compared to those from the log-binomial model:

$$\frac{\text{SE of}\ln(\text{RR})\text{ from method}}{\text{SE of}\ln(\text{RR})\text{ from log} - \text{binomial method}}$$

To assess the level of confounding in the mutually adjusted model, for each combination of outcome and explanatory variable, I calculated the proportional change in the estimate of RR following adjustment, using the formula

$$confounding = \frac{RR_{adjusted} - RR_{unadjusted}}{RR_{adjusted}} \times 100$$

When the resulting value was negative, the adjusted estimate was smaller than the unadjusted one.

8.3 Results

The RI log-binomial model converged for all but one of the models fitted (including both univariate and mutually adjusted models). Thus, for almost all analyses, the point estimates from the other methods examined, along with their precision, were compared with those from the RI log-binomial model. The only exception was the mutually adjusted model with low back pain as an outcome. For that model, estimates were compared with those derived from the RI Poisson model with robust SEs, as this has been shown to be a method that produces estimates adequately close to those from the log-binomial model in situations of both unclustered (74, 76) and clustered data (77).

Comparison of point estimates

As expected, the point estimates were the same for the RI Poisson models with and without robust SEs. Therefore, in the description that follows, no distinctions will be made between the two RI Poisson models, with regard to point estimates.

The point estimates derived from the RI Poisson, shared frailty, and RI CLL models were very similar to those derived from the RI log-binomial model in both univariate and multivariate analyses. In univariate analyses (sub-figures in left-hand columns of Figure 8.1 to Figure 8.3, and

Table 8.1), the effects estimated from the three models (RI Poisson, shared frailty, and RI CLL) were higher than those from the corresponding RI log-binomial models by an average of 1%, the differences ranging from 0% to 6%. In adjusted analyses (sub-figures in right-hand columns of Figure 8.1 to Figure 8.3, and Table 8.1), the differences in the point estimates from these three models had an average value of 1% and ranged from 0% to 10%, most (60/63) being between 0% and 5%. Overall differences (for RRs from these three models and those from the RI log-binomial model) between crude estimates and those that were mutually adjusted were statistically similar (p=0.3). For both univariate and multivariate analyses, the majority of differences for point estimates from the RI Poisson, shared frailty, and RI CLL models when compared with those from the RI log-binomial models (95/126 estimates) were <2%. The largest differences in point estimates were found for the adjusted association between older age (50-59 v 20-29) and wrist/hand pain; all three models (RI Poisson, shared frailty, and RI CLL) resulted in estimates for this association that differed from those from RI log-binomial by an average of 10% (Appendix 4 Table 5).

From the models that directly estimate RRs, the largest deviations from the point estimates derived from the RI log-binomial model were found for the CLL model with robust SEs. Point estimates from this model were up to 33% higher (average 7%) compared to those from the RI log-binomial model (Table 8.1). Estimates differed from those from the RI log-binomial model irrespective of adjustment for other explanatory variables, but they were higher when the outcome was pain in the wrist/hand; mean ratio of RR 1.16 compared to 1.04 for the low back and elbow pain outcomes. Differences were also <5% for most explanatory variables, but those for sex and occupational activity were ≈ 12% and those for somatising tendency (2+ v 0 somatic symptoms) were on average 14%.

Estimates of RR produced by the indirect method suggested by Z&Y, were very similar to those from the RI log-binomial model. Most ratios of RRs (39/42) were between 0.92 and 1.08, indicating differences ≤8%. Exceptions were the adjusted effects of older age (40-49, and 50-59 v 20-29) and somatising tendency (2+ v 0) on wrist/hand pain, for which RRs were 11.5%, 24.3%, and 17.0%, higher than the corresponding ones derived from the RI log-binomial model (Appendix 4 Table 5). Unlike the Z&Y model, the mod Z&Y model performed poorly. The estimates derived from the mod Z&Y method were on average 20% different from those derived from the RI log-binomial model. Differences varied markedly, with RRs from the mod Z&Y model being almost 3-fold lower than those derived from the RI log-binomial model (Table 1).

Effect estimates from the methods described above differed (to a greater or lesser extent) from those produced by the RI log-binomial model (or the RI Poisson model in the case of the adjusted model for low back pain), but the direction of the differences was generally inconsistent. An exception was the mod Z&Y model, estimates from which were mostly lower than those from the reference models.

Summary statistics for differences from RRs derived using the RI log-binomial model, according to outcome variable and predictor variable, showed no clear pattern (Table 8.1).

Comparison of precision of point estimates

Ratios of SEs of the ln(RRs) from the several methods to those from the RI log-binomial model are summarised by outcome and predictor in Table 8.2.

From the models that directly estimate RRs, the RI CLL model gave the closest average precision to that from the RI log-binomial model; ratios of SEs had an average value of 0.97. However, the model performed better in the unadjusted analyses than in the adjusted ones. Ratios of SEs in the unadjusted models ranged only from 0.92 to 1.05, while they ranged from 0.70 to 1.04 in the adjusted ones (Table 8.2). In the adjusted models, ratios of SEs lower than 0.9 (indicating differences in SEs > 10%) were found when the outcome was pain in the lower back (Appendix 4, Table 4). That observation could simply indicate differences between the RI CLL and the RI Poisson with robust SEs model, as for the multivariable model of low back pain as an outcome, the RI log-binomial model did not converge and estimates from the RI Poisson with robust SEs model were used for comparison instead.

The RI Poisson without robust SEs, shared frailty, and Poisson with robust SEs models, also showed precision of point estimates close to that of the comparator, with average ratios of SEs 1.08, 1.08, and 1.11, respectively. However, from these models, the Poisson without robust SEs and shared frailty models had the narrowest range of relative precision across all adjusted and unadjusted analyses (0.87-1.21 in the two models). Differences between SEs from these models and those from the RI log-binomial model were higher than 10% (ratio of SEs <0.9 or >1.1) mainly for the adjusted associations with wrist/hand pain as an outcome.

On average the RI Poisson model with robust SEs produced SEs adequately close to those from the RI log-binomial model (mean ratio of SEs = 1.11). However, ratios of SEs varied widely from 0.87 to 1.94. The biggest differences between SEs derived from the RI Poisson with robust SEs and those from RI log-binomial were seen in associations with wrist/hand pain. The largest deviation of the ratio of SEs from the null value of 1 was for the unadjusted and adjusted associations between the highest age band (50-59 v 20-29) and wrist/hand pain, with SEs of the point estimates from the Poisson model with robust SEs being almost twice as high as those derived from the RI log-binomial model. The RI Poisson with robust SEs also underestimated precision of the unadjusted effect estimates of somatising tendency with back pain.

The CLL with robust SEs model produced SEs of point estimates that were considerably higher than those derived from the RI log-binomial model. The average value of ratios of SEs was 1.46 and it ranged from 0.88 to 2.83. Ratios were closer to 1 in the adjusted analyses than the unadjusted (average ratio of SEs 1.29 and 1.64, respectively). The largest deviations of the ratios of SEs from 1 were for the unadjusted associations with wrist/hand pain as an outcome (mean ratio of SEs: 2.08, range: 1.12-2.83).

Considering average values of relative precision of point estimates when RRs where indirectly derived from RI logistic model, I found that the Z&Y method produced SEs very close to those from the RI log-binomial model (average ratio of SEs: 0.97, range: 0.64-1.19). Agreement in SEs from the two models was better for the adjusted analyses than the unadjusted (ratios of SEs: 1.00 and 0.93, respectively). However, SEs derived from the Z&Y method for the association of somatising tendency with back and wrist/hand pain were considerably lower than those from the RI log-binomial model (ratios of SEs < 0.9).

Unlike the method proposed by Z&Y, the modified Z&Y model severely underestimated SEs when compared to the RI log-binomial model. Most ratios of SEs were < 1. The underestimation of SEs was more prominent in the adjusted analyses with ratios of SEs ranging from 0.06 to 0.35. No pattern of underestimation of SEs was observed in relation to specific outcomes or predictors examined.

Level of confounding did not seem to affect ratios of SEs in the adjusted associations for most of the methods. However, ratios of SEs from the mod Z&Y method to those from the RI log-binomial model increased as confounding increased (Spearman p-value = 0.05).

Table 8.1. Summary statistics (mean (range)) for ratios of RRs derived from the method under consideration to those from the RI log-binomial models

			RI logistic Z&Y	**RI logistic modified Z&Y**	**RI Poisson**	**Shared frailty**	**CLL (robust SEs)**	**RI CLL**	**All**
UNADJUSTED	Outcomes	LBP	0.99 (0.95-1.03)	0.92 (0.77-1.03)	1.01 (1.00-1.02)	1.00 (1.00-1.02)	1.04 (0.99-1.14)	1.00 (0.98-1.02)	1.00 (0.77-1.14)
		WHP	1.03 (0.99-1.06)	0.98 (0.85-1.05)	1.02 (1.01-1.06)	1.02 (1.00-1.05)	1.16 (1.01-1.33)	1.02 (0.99-1.05)	1.03 (0.85-1.33)
		ELP	0.98 (0.92-1.01)	0.89 (0.81-1.02)	1.00 (1.00-1.01)	1.00 (0.99-1.01)	1.04 (0.93-1.22)	0.99 (0.97-1.01)	0.99 (0.81-1.22)
	Explanatory variable	Sex	1.03 (1.01-1.06)	1.03 (1.02-1.05)	1.00 (1.00-1.01)	1.00(0.99-1.00)	1.17 (1.02-1.33)	1.02 (1.01-1.03)	1.03 (0.99-1.33)
		Occupational activity	1.03 (1.01-1.05)	1.03 (1.01-1.04)	1.01 (1.00-1.01)	1.00 (0.99-1.01)	1.16 (0.99-1.27)	1.01 (1.00-1.02)	1.03 (0.99-1.27)
		Older age (30-39 v 20-29)	1.00 (0.99-1.02)	0.95 (0.89-1.00)	1.01 (1.00-1.01)	1.01 (1.00-1.01)	1.00 (0.97-1.01)	1.00 (0.99-1.01)	1.00 (0.89-1.02)
		Older age (40-49 v 20-29)	0.99 (0.94-1.03)	0.92 (0.82-1.00)	1.01 (1.00-1.03)	1.01 (1.00-1.03)	1.02 (0.98-1.06)	1.01 (0.98-1.03)	1.00 (0.82-1.06)
		Older age (50-59 v 20-29)	0.99 (0.92-1.06)	0.93 (0.81-1.04)	1.02 (1.00-1.06)	1.02 (1.00-1.05)	0.99 (0.93-1.05)	1.01 (0.97-1.05)	1.00 (0.81-1.06)
		Somatising tendency (1 v 0)	0.98 (0.97-0.99)	0.84 (0.79-0.87)	1.00 (1.00-1.01)	1.00 (0.99-1.01)	1.05 (1.01-1.10)	0.99 (0.98-0.99)	0.98 (0.79-1.10)
		Somatising tendency (2+ v 0)	0.97 (0.95-0.99)	0.82 (0.77-0.86)	1.01 (1.00-1.02)	1.01 (0.99-1.02)	1.17 (1.11-1.26)	1.02 (1.00-1.03)	1.00 (0.77-1.26)
	All		**1.00 (0.92-1.06)**	**0.93 (0.77-1.05)**	**1.01 (1.00-1.06)**	**1.01 (0.99-1.05)**	**1.08 (0.93-1.33)**	**1.01 (0.97-1.05)**	**1.01 (0.77-1.33)**
MUTUALLY ADJUSTED	Outcomes	LBP	1.04 (1.01-1.06)	0.72 (0.46-0.84)	1.00 (1.00-1.00)	1.00 (0.99-1.00)	1.03 (0.99-1.06)	0.99 (0.97-1.00)	0.97 (0.46-1.06)
		WHP	1.11 (1.04-1.24)	0.74 (0.51-0.97)	1.04(1.00-1.10)	1.03(1.00-1.10)	1.11 (1.02-1.21)	1.03 (0.98-1.09)	1.01 (0.51-1.24)
		ELP	1.03 (1.01-1.05)	0.56 (0.34-0.77)	1.01(1.01-1.02)	1.01 (1.00-1.02)	1.02 (0.96-1.11)	0.99 (0.96-1.01)	0.95 (0.34-1.11)
	Explanatory variable	Sex	1.05 (1.03-1.08)	0.78 (0.75-0.81)	1.01(1.00-1.02)	1.00 (0.99-1.01)	1.08 (0.97-1.21)	1.01 (1.00-1.03)	0.99 (0.75-1.21)
		Occupational activity	1.05 (1.03-1.08)	0.73 (0.67-0.84)	1.01(1.00-1.02)	1.01(1.00-1.02)	1.09 (1.01-1.15)	1.01 (1.00-1.03)	0.99 (0.67-1.15)
		Older age (30-39 v 20-29)	1.03 (1.01-1.04)	0.85 (0.72-0.97)	1.01 (1.00-1.03)	1.01 (1.00-1.02)	1.02 (1.00-1.04)	1.00 (0.99-1.02)	0.99 (0.72-1.04)
		Older age (40-49 v 20-29)	1.07 (1.04-1.12)	0.64 (0.39-0.82)	1.02 (1.00-1.04)	1.02 (1.00-1.04)	1.05 (1.01-1.09)	1.00 (0.98-1.04)	0.97 (0.39-1.12)
		Older age (50-59 v 20-29)	1.11 (1.03-1.24)	0.60 (0.34-0.74)	1.04 (1.00-1.10)	1.04 (1.00-1.10)	1.03 (0.96-1.12)	1.01 (0.96-1.09)	0.98 (0.34-1.24)
		Somatising tendency (1 v 0)	1.02 (1.01-1.04)	0.67 (0.63-0.71)	1.00 (1.00-1.01)	1.00 (1.00-1.00)	1.01 (0.99-1.02)	0.98 (0.97-0.99)	0.96 (0.63-1.04)
		Somatising tendency (2+ v 0)	1.09 (1.05-1.17)	0.46 (0.40-0.51)	1.02 (1.00-1.03)	1.01 (1.00-1.03)	1.11 (1.06-1.16)	1.01 (0.99-1.05)	0.96 (0.40-1.17)
	All		**1.06 (1.01-1.24)**	**0.67 (0.34-0.97)**	**1.02 (1.00-1.10)**	**1.01 (0.99-1.10)**	**1.05 (0.96-1.21)**	**1.01 (0.96-1.09)**	**0.98 (0.34-1.24)**
All (unadjusted and adjusted)			**1.03 (0.92-1.24)**	**0.80 (0.34-1.05)**	**1.01 (1.00-1.10)**	**1.01 (0.99-1.10)**	**1.07 (0.93-1.33)**	**1.01 (0.96-1.09)**	**0.99(0.34-1.33)**

Table 8.2. Summary statistics (mean (range)) for relative precision (SEs of the ln(RRs) derived from the method under consideration to those from the log-binomial model) of point estimates

			RI logistic Z&Y	**RI logistic modified Z&Y**	**RI Poisson (no robust SEs)**	**RI Poisson (robust SEs)**	**Shared frailty**	**CLL (robust SEs)**	**RI CLL**	**All**
UNADJUSTED	Outcomes	LBP	0.91 (0.74-1.04)	0.75 (0.44-1.05)	1.16 (1.14-1.21)	1.13 (0.95-1.35)	1.16 (1.14-1.21)	1.61 (1.14-2.35)	0.99 (0.92-1.05)	1.09 (0.44-2.35)
		WHP	0.96 (0.82-1.02)	0.84 (0.61-1.01)	1.11 (1.08-1.15)	1.21 (0.95-1.80)	1.11 (1.08-1.15)	2.08 (1.12-2.83)	1.01 (0.96-1.04)	1.17 (0.61-2.83)
		ELP	0.93 (0.83-1.00)	0.79 (0.67-1.01)	1.03 (1.02-1.05)	1.02 (0.87-1.13)	1.03 (1.02-1.05)	1.23 (1.03-1.67)	0.97 (0.93-1.02)	1.00 (0.67-1.67)
	Explanatory variable	Sex	1.01 (1.00-1.02)	1.00 (0.99-1.01)	1.09 (1.04-1.14)	1.12 (1.03-1.21)	1.09 (1.03-1.14)	2.16 (1.67-2.83)	1.03 (1.02-1.04)	1.19 (0.99-2.83)
		Occupational activity	1.01 (0.99-1.04)	1.01 (0.99-1.05)	1.09 (1.03-1.18)	1.12 (1.08-1.13)	1.09 (1.02-1.18)	2.08 (1.34-2.56)	1.03 (1.01-1.05)	1.18 (0.99-2.56)
		Older age (30-39 v 20-29)	0.96 (0.93-1.01)	0.75 (0.67-0.86)	1.09 (1.02-1.14)	1.00 (0.95-1.06)	1.09 (1.02-1.14)	1.37 (1.14-1.71)	0.99 (0.97-1.02)	1.03 (0.67-1.71)
		Older age (40-49 v 20-29)	0.90 (0.85-0.97)	0.77 (0.68-0.88)	1.10 (1.03-1.15)	1.13 (0.96-1.38)	1.10 (1.03-1.15)	1.60 (1.10-2.11)	0.98 (0.94-1.02)	1.07 (0.68-2.11)
		Older age (50-59 v 20-29)	0.89 (0.83-0.94)	0.79 (0.69-0.92)	1.12 (1.04-1.16)	1.30 (1.05-1.80)	1.12 (1.04-1.16)	1.69 (1.26-2.35)	0.98 (0.93-1.03)	1.11 (0.69-2.35)
		Somatising tendency (1 v 0)	0.92 (0.88-0.96)	0.59 (0.44-0.72)	1.09 (1.04-1.15)	1.06 (0.95-1.21)	1.09 (1.04-1.15)	1.10 (1.03-1.14)	0.95 (0.92-0.97)	0.98 (0.44-1.21)
		Somatising tendency (2+ v 0)	0.83 (0.74-0.91)	0.63 (0.49-0.74)	1.14 (1.05-1.21)	1.12 (0.87-1.35)	1.13 (1.05-1.21)	1.47 (1.04-1.86)	0.99 (0.98-1.00)	1.04 (0.49-1.86)
	All		**0.93 (0.74-1.04)**	**0.79 (0.44-1.05)**	**1.10 (1.02-1.21)**	**1.12 (0.87-1.80)**	**1.10 (1.02-1.21)**	**1.64 (1.03-2.83)**	**0.99 (0.92-1.05)**	**1.08 (0.44-2.83)**
MUTUALLY ADJUSTED	Outcomes	LBP	0.91 (0.64-1.08)	0.13 (0.06-0.17)	1.02 (0.87-1.15)	1.00 (1.00-1.00)	1.02 (0.87-1.15)	1.15 (0.88-1.43)	0.84 (0.70-0.96)	0.87 (0.06-1.43)
		WHP	1.11 (1.02-1.19)	0.25 (0.15-0.35)	1.14(1.09-1.21)	1.27 (0.96-1.94)	1.14 (1.09-1.21)	1.61 (0.92-2.02)	1.01 (0.95-1.04)	1.07 (0.15-2.02)
		ELP	0.99 (0.95-1.05)	0.14 (0.08-0.20)	1.04 (1.03-1.06)	1.04 (0.94-1.14)	1.04 (1.03-1.05)	1.10 (0.98-1.33)	0.96 (0.90-1.01)	0.90 (0.08-1.33)
	Explanatory variable	Sex	1.00 (0.87-1.08)	0.19 (0.14-0.25)	1.03 (0.94-1.11)	1.04 (1.00-1.09)	1.03 (0.93-1.10)	1.54 (1.33-1.87)	0.96 (0.83-1.03)	0.97 (0.14-1.87)
		Occupational activity	1.02 (1.01-1.03)	0.18 (0.16-0.21)	1.06 (1.04-1.09)	1.03 (0.99-1.10)	1.06 (1.03-1.09)	1.34 (1.09-1.75)	0.98 (0.90-1.02)	0.95 (0.16-1.75)
		Older age (30-39 v 20-29)	1.08 (1.00-1.17)	0.24 (0.17-0.35)	1.10 (1.03-1.15)	1.07 (1.00-1.19)	1.10 (1.03-1.15)	1.23 (1.07-1.48)	0.98 (0.95-1.02)	0.97 (0.17-1.48)
		Older age (40-49 v 20-29)	1.01 (0.91-1.18)	0.16 (0.09-0.28)	1.07 (1.02-1.16)	1.17 (1.00-1.48)	1.07 (1.02-1.16)	1.37 (1.03-1.82)	0.93 (0.84-1.03)	0.97 (0.09-1.82)
		Older age (50-59 v 20-29)	1.06 (0.95-1.19)	0.15 (0.08-0.23)	1.13 (1.05-1.21)	1.36 (1.00-1.94)	1.13 (1.05-1.21)	1.46 (1.14-2.02)	0.95 (0.90-1.04)	1.03 (0.08-2.02)
		Somatising tendency (1 v 0)	0.97 (0.83-1.08)	0.17 (0.11-0.24)	1.04 (0.95-1.11)	1.01 (0.96-1.08)	1.04 (0.95-1.11)	0.95 (0.88-1.05)	0.89 (0.76-0.96)	0.87 (0.11-1.11)
		Somatising tendency (2+ v 0)	0.89 (0.64-1.03)	0.10 (0.06-0.15)	1.03 (0.87-1.15)	1.03 (0.94-1.15)	1.02 (0.87-1.15)	1.11 (0.96-1.38)	0.89 (0.70-0.99)	0.87 (0.06-1.38)
	All		**1.00 (0.64-1.19)**	**0.17 (0.06-0.35)**	**1.07 (0.87-1.21)**	**1.10 (0.94-1.94)**	**1.06 (0.87-1.21)**	**1.29 (0.88-2.02)**	**0.94 (0.70-1.04)**	**0.95 (0.06-2.02)**
	All (unadjusted and adjusted)		**0.97 (0.64-1.19)**	**0.48 (0.06-1.05)**	**1.08 (0.87-1.21)**	**1.11 (0.87-1.94)**	**1.08 (0.87-1.21)**	**1.46 (0.88-2.83)**	**0.97 (0.70-1.05)**	**1.02 (0.06-2.83)**

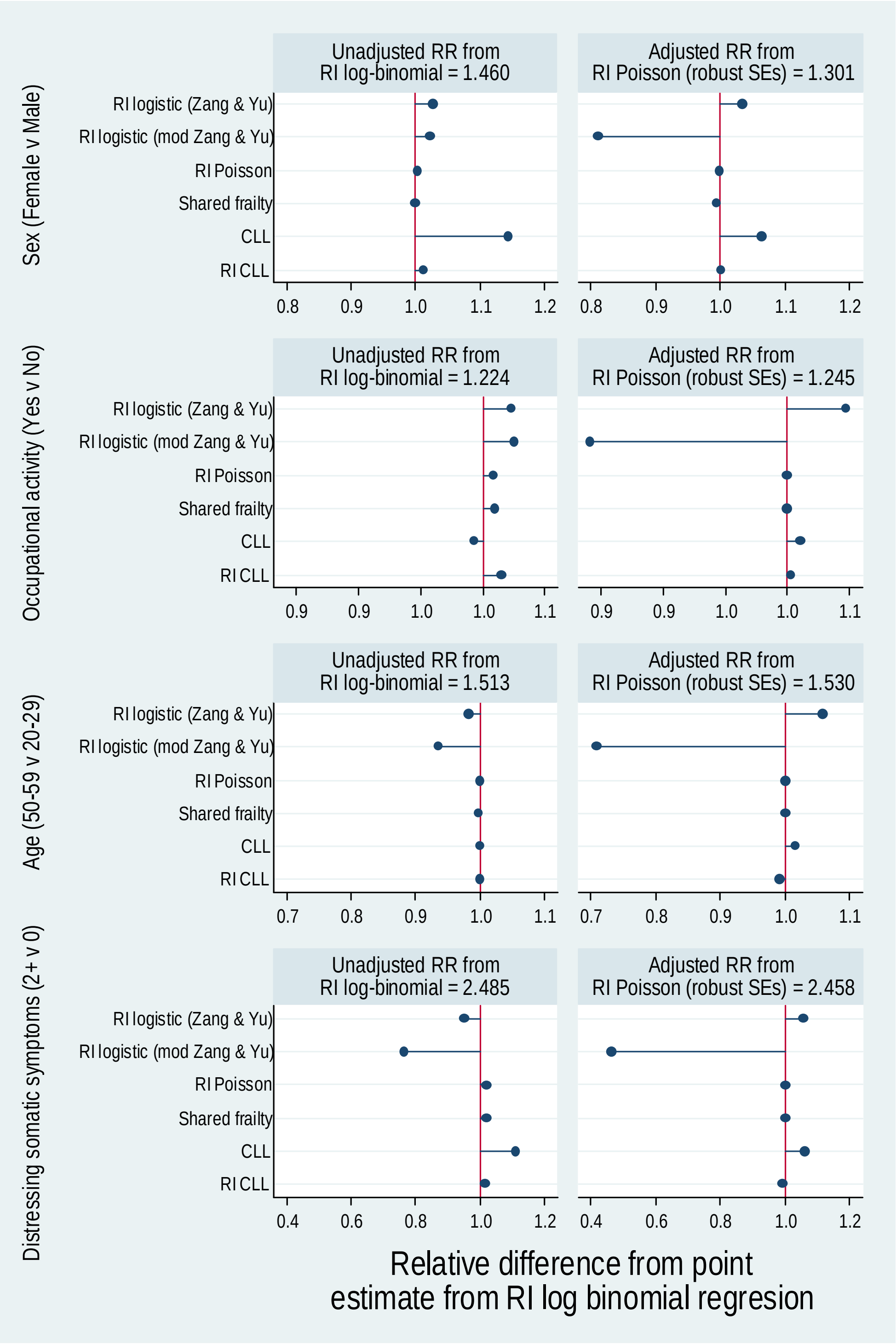


Figure 8.1. Relative difference from point estimates derived from RI log-binomial in the unadjusted models (left column) and from the RI Poisson (robust SEs) in the mutually adjusted models (right column) when outcome was low back pain

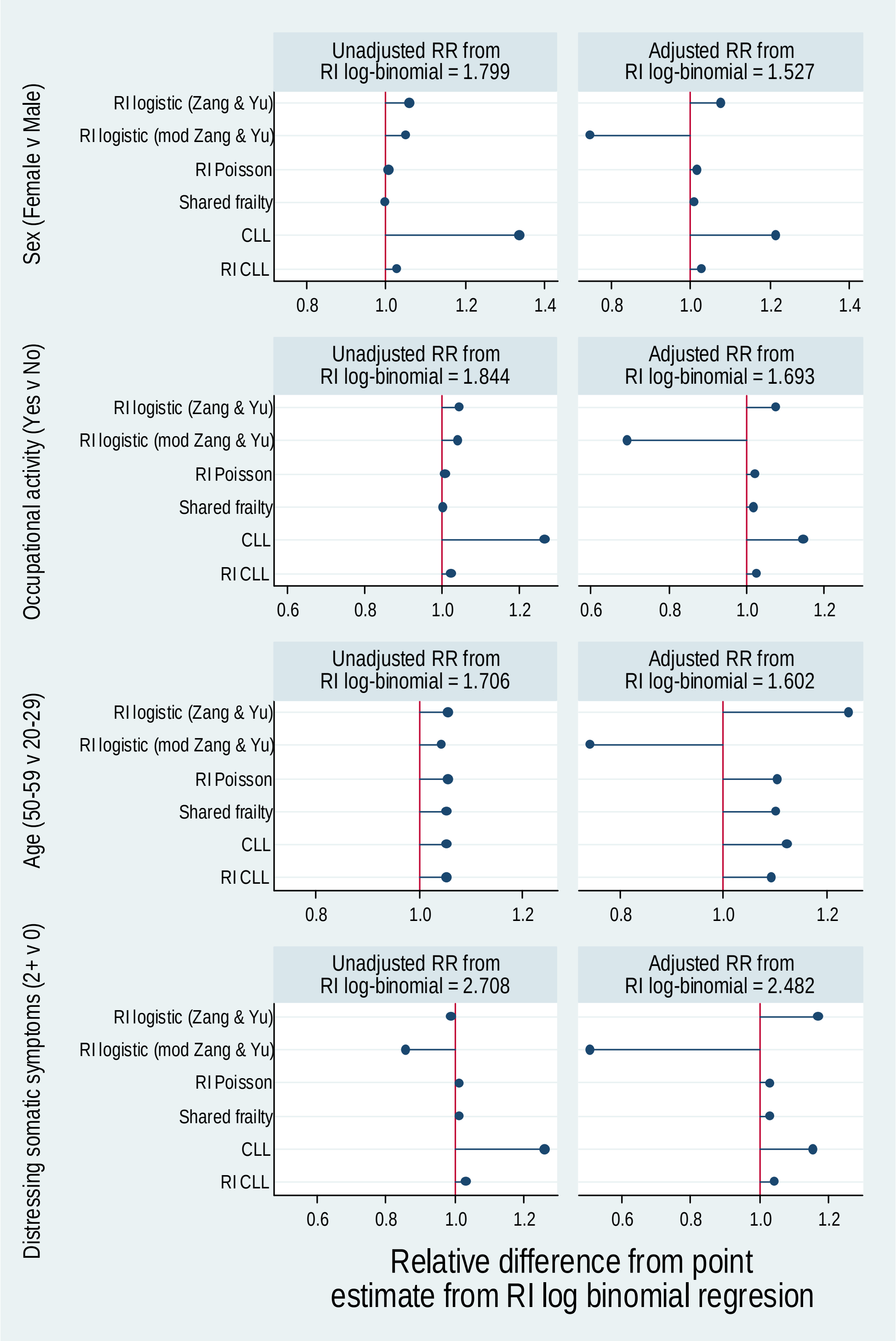


Figure 8.2. Relative difference from point estimates derived from RI log-binomial in the unadjusted models (left column) and mutually adjusted models (right column) when outcome was wrist/hand pain

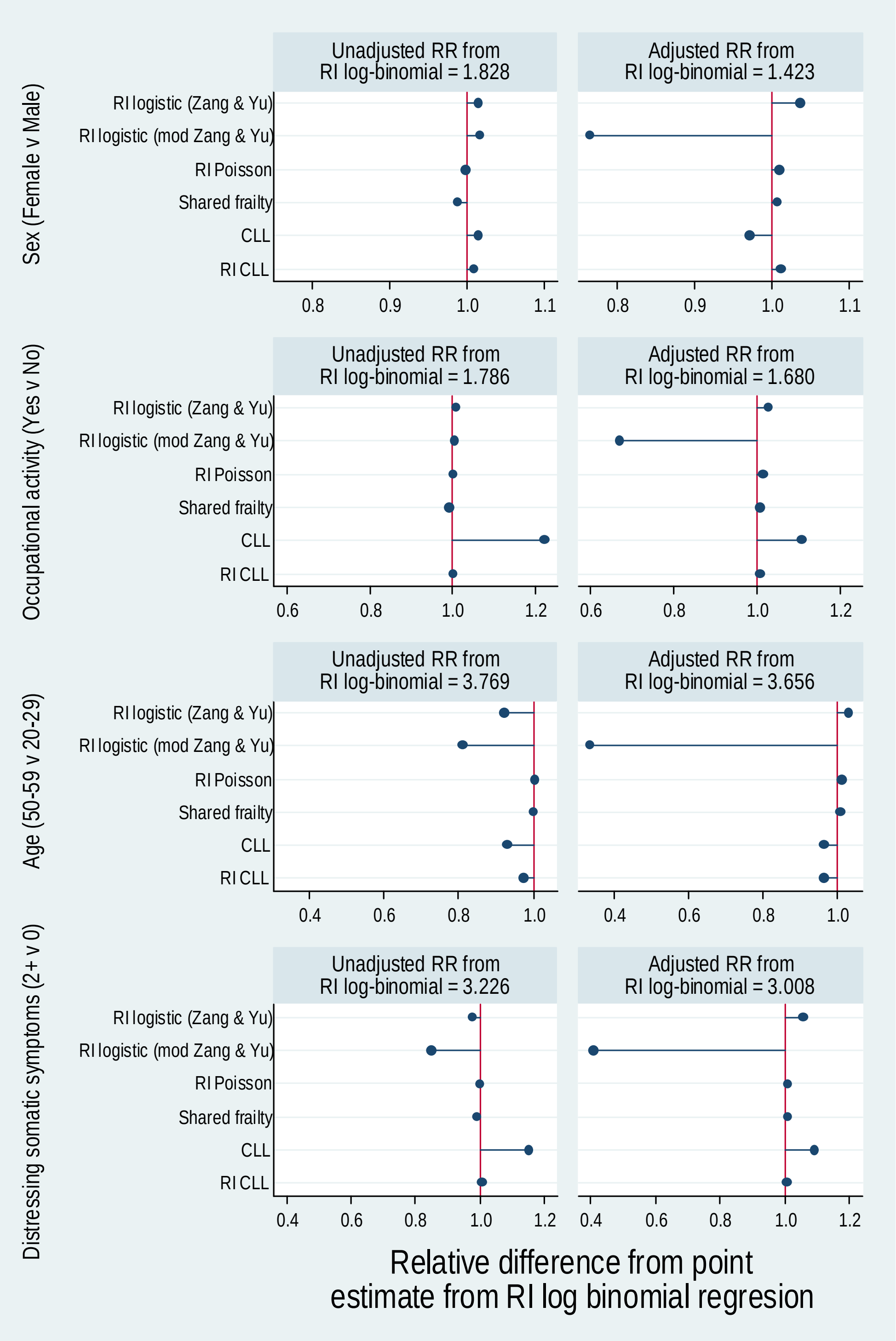


Figure 8.3. Relative difference from point estimates derived from RI log-binomial in the unadjusted models (left column) and mutually adjusted models (right column) when outcome was elbow pain

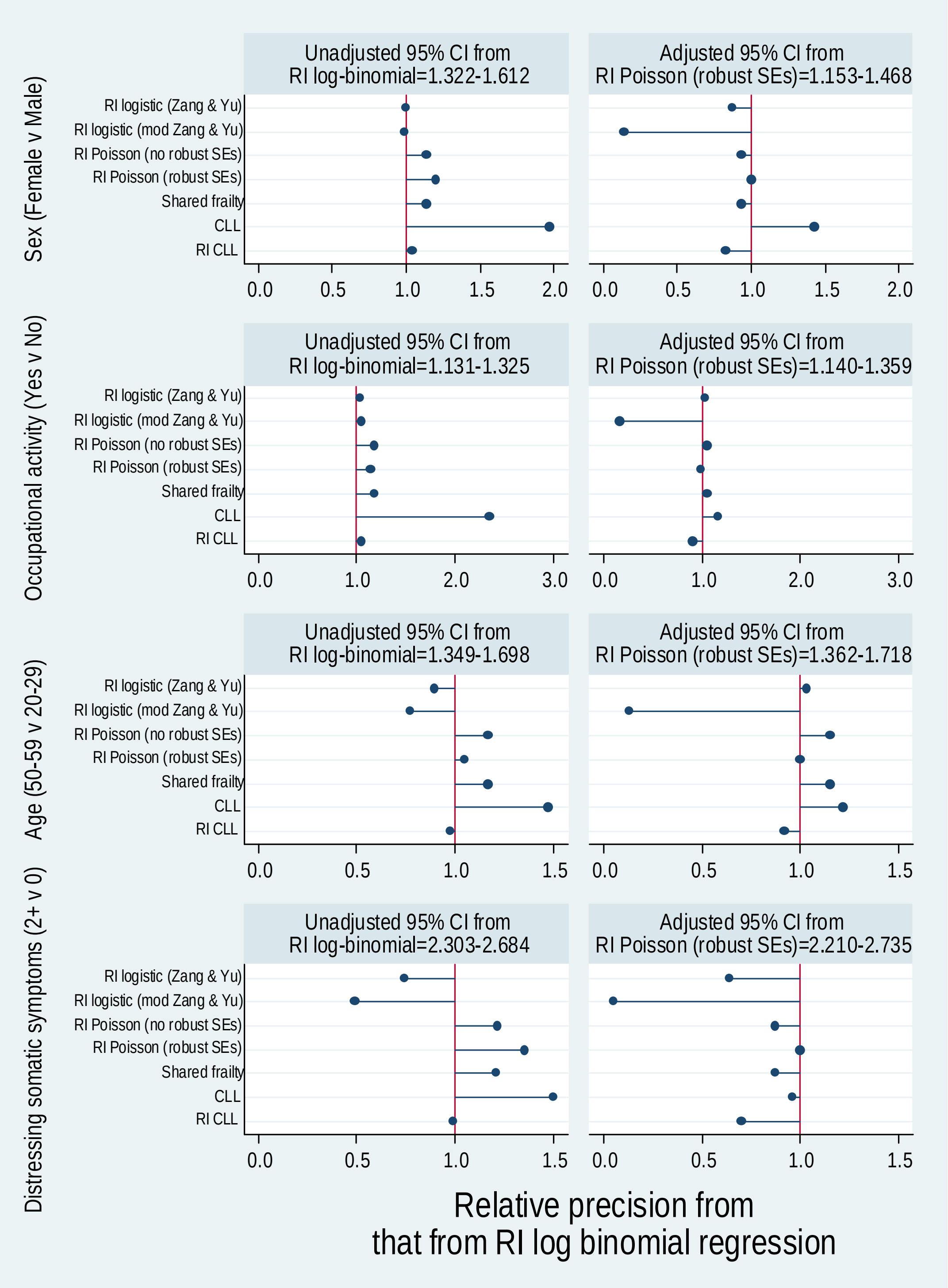


Figure 8.4. Relative precision from that derived from the RI log-binomial model in the unadjusted (left column) and from the RI Poisson (robust SEs) mutually adjusted (right column) models when outcome was low back pain

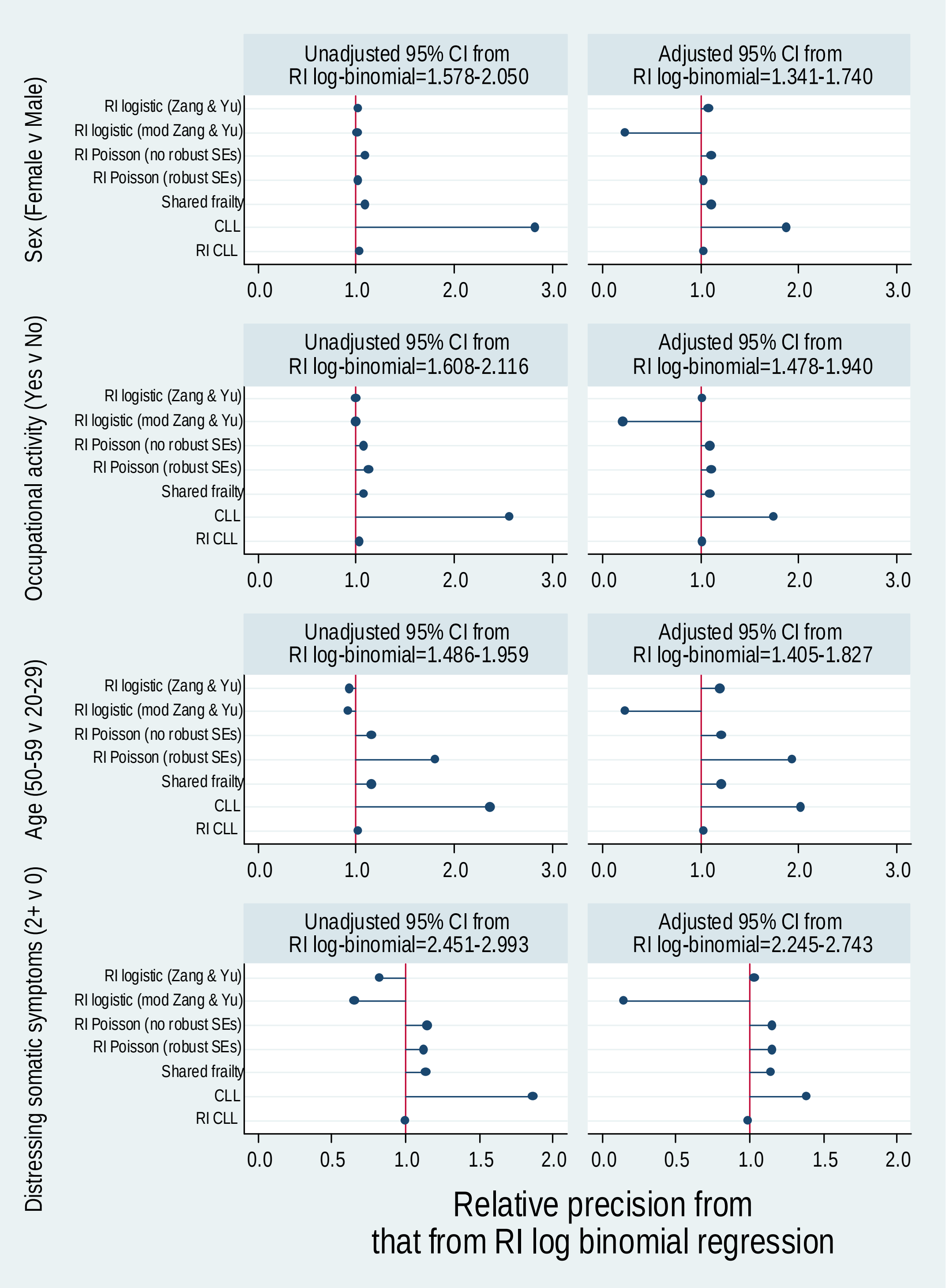


Figure 8.5. Relative precision from that derived from the RI log-binomial model in the unadjusted (left column) and mutually adjusted (right column) models when outcome was wrist/hand pain

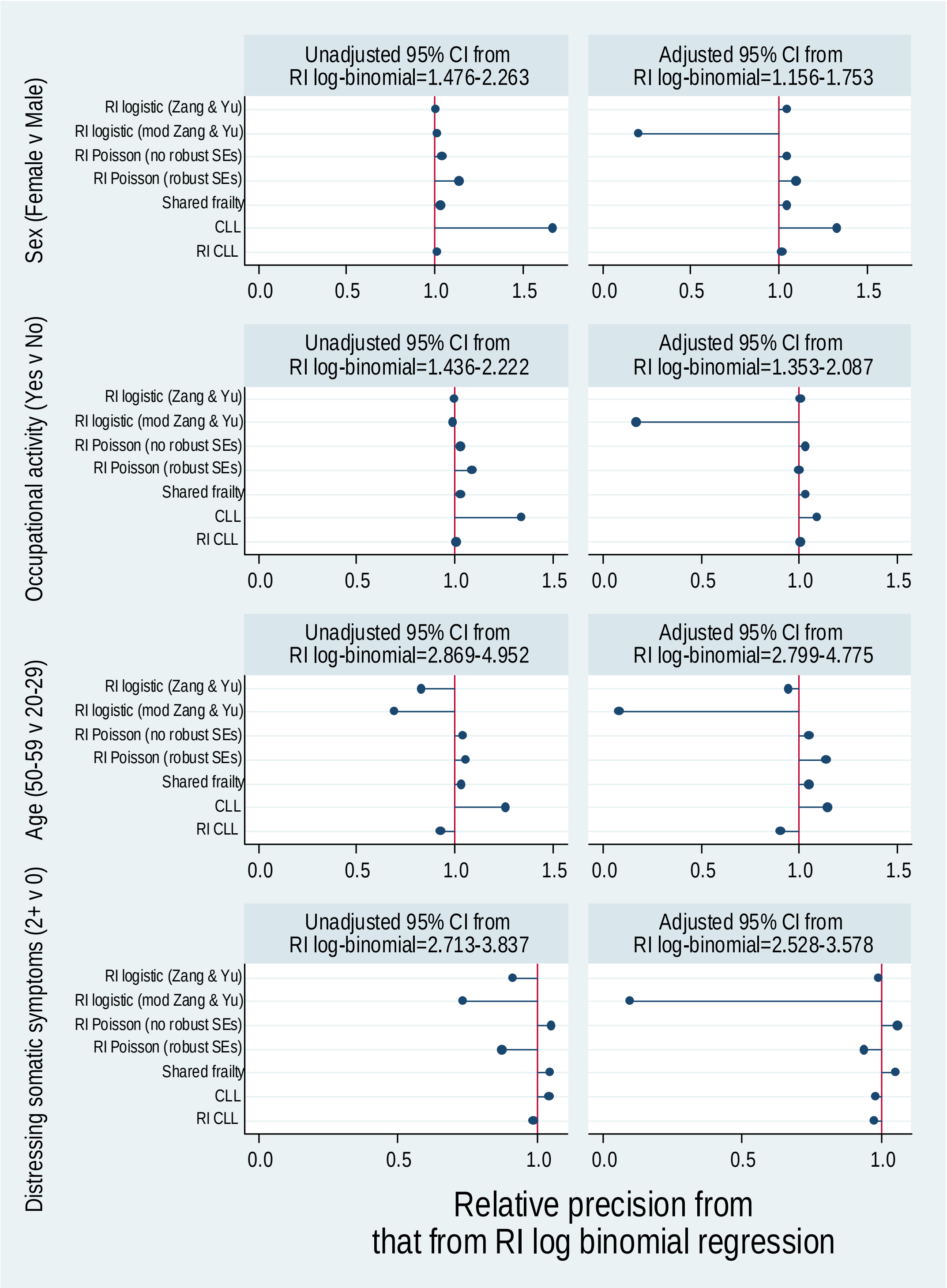


Figure 8.6 Relative precision from that derived from the RI log-binomial model in the unadjusted (left column) and mutually adjusted (right column) models when outcome was elbow pain

8.4 Discussion

The aim of this chapter was to explore different approaches to estimation of RRs when the outcome of interest is a binary variable and data are hierarchically structured. For that I used data from the CUPID study, covering a number of different scenarios for the prevalence of the outcome and explanatory variables. Point estimates from different analytical approaches, and their precision, were compared with those derived from RI log-binomial models. I found that the point estimates produced from the RI Poisson, the shared frailty, and the RI CLL model were very similar to those from the RI log-binomial model (maximum difference <10% and most differences <5%), whereas estimates from other approaches deviated further from those derived from the RI log-binomial model. The RRs estimated indirectly from ORs using the Z&Y method were adequately close to those from the RI log-binomial model. However, those from the modified Z&Y method performed poorly, with a tendency to underestimate RRs, a bias more apparent in the mutually adjusted models.

From the different analytical methods examined, the RI CLL model produced SEs that were the closest to those derived from the RI log-binomial model. SEs from the two models were only up to 11% different (considering only the analyses for which the RI log-binomial successfully converged). The shared frailty and the RI Poisson (with and without robust SEs) models indicated precision that was slightly lower but broadly comparable to that from the RI log-binomial model. Surprisingly, ratios of SEs from the shared frailty and the Poisson without robust SEs models to those from the RI log-binomial were more narrowly spread across the different (unadjusted and adjusted) analyses than the corresponding ratios of SEs for the Poisson with robust SEs models. In contrast, the other methods examined gave either much lower (CLL model with robust SEs) or much greater (Z&Y, and modified Z&Y models) precision for risk estimates.

The comparison of estimates and related precision derived from different analytical approaches was made using real data from the CUPID study. That constrained the range of scenarios in which the estimates could be compared (e.g. different effect sizes, and prevalence of outcome and explanatory variables). However, the CUPID study is a rich source of data, and it was nevertheless possible to examine a variety of scenarios; prevalence rates of outcome variables varied from 6% to 22%, and of explanatory variables from 16% to 74%, while effect sizes (estimated from the RI Poisson model) ranged from to 1.09 to 3.79 (the range of the effect estimates did not differ between the adjusted and unadjusted models).

Another limitation of using real rather than simulated data, was that true values for parameters were unknown, and for this reason, conclusions from the comparisons of analytical methods should be cautious. In the absence of known values for parameters, risk estimates and their precision were compared with those derived from the RI log-binomial model, which has widely been considered the preferred analytical approach for estimation of RRs.

Also limiting was the similarity of most unadjusted risk estimates to those that were adjusted for the other risk factors. Confounding ranged from -60% to 11%, with approximately half of the unadjusted RRs (72 out of 147) very close to the adjusted values (between -5% and +5% confounding). For that reason, conclusions from this investigation regarding adjusted RRs must be guarded. However, despite the weak confounding, the CUPID data could still be used to identify methods that performed poorly in comparison with the RI log-binomial model in the unadjusted models, and it seems unlikely that they would perform better in adjusted models.

The models that were fitted were the RI log-binomial, the RI logistic regression following the Z&Y method, a modification of the Z&Y method, the RI Poisson model (without and with robust SEs), the shared frailty model, and the CLL model (with robust SEs, and with RI). Among these models, the RI logistic, RI log-binomial, and the CLL regression models assume the same underlying binomial distribution of the outcome variable, but with different link functions, while the others assume different distributions.

The difference between estimates from the logistic models and the log-binomial has been discussed in numerous reports (180-182). Following these reports, Z&Y suggested deriving RRs from ORs, based on the formula presented in equation 8.5 which was derived from simple algebraic manipulations (Appendix 2). However, this method has been criticised as it uses a summary value for the proportion of the outcome in the unexposed group (p_0) ignoring the correlation structure of the covariate variables (173). This simplification can lead to bias in estimates of risk from both real and simulated data, as has been demonstrated in several publications (173, 177, 183). In response, the modified Z&Y method was proposed, which uses the estimated prevalence of the outcome in the group unexposed to the main exposure variable after adjustment for confounding effects (177). Even though those two methods have been referred to or used in several studies, neither has been tested in clustered data. Here, these methods were adjusted to incorporate a RI in the logistic model before transformation of the ORs into RRs. I found that estimates from the Z&Y method were very close to those from the RI log-binomial model (39/42 estimates of RR ≤10% different from RRs produced by a RI log-binomial model). In contrast, estimates from the modified Z&Y method were highly biased; only 7/42 estimates were <10% different from those produced by the RI log-binomial model. These results agree with findings from simulated data reported by Dwivedi et al (177), in which the modified Z&Y method gave estimates of RR that were slightly biased, whereas RRs estimated by the Z&Y method were adequately close to the true effects. The authors attributed the small bias in the Z&Y method to the absence of major confounding effects in their dataset. Likewise, the relatively good performance of the Z&Y method in the models fitted in this chapter using data from CUPID may reflect only small/moderate confounding effects. It is unclear, however, why in such circumstances of low confounding, precision of point estimates from the modified Z&Y method differs from that of the Z&Y method.

In comparison with the other methods examined in this chapter, the CLL model is used less frequently for the estimation of RRs. A search in Scopus identified only 15 epidemiological studies in which the method had been applied. It was first suggested by Martuzzi et al (184) who, based on the analysis of real data from a large survey, concluded that the CLL link function provides a useful alternative for the estimation of RRs. The model has been under-used as described by Nelder in 2001 (185), whose report was followed by methodological publications that applied the CLL model (186, 187) and compared it with other methods more commonly implemented (188). Based on simulation results reported by Penman et al (188), the CLL model performs well, with small bias in estimates of effect and satisfactory precision. That methodological work, however, has not been expanded to data that are grouped within clusters. Using data from the CUPID study, I showed that the RI CLL model produced risk estimates that were very similar to those from RI log-binomial models. On average, they were only 1% different from the RRs from the RI log-binomial model, and this difference was not influenced by adjustment for covariates. The RI CLL model also produced precision that was very close to that from the RI log-binomial model with ratios of SEs from the two methods being very close to one.

When clustering of observations was accounted for in the CLL model using robust SEs instead of RI to account for dependence of observations in the precision of the point estimates, both the estimated RRs and their precision differed from those obtained with the RI log-binomial model, making it the least satisfactory of the methods that were considered in this chapter for direct estimation of RRs when data are clustered. The poor performance of the method (overestimation of both RR and SEs of ln(RRs)) is not surprising given that previous research comparing methods that correct for clustering effects has favoured multilevel analytical approaches over traditional models that simply apply corrections to SEs (189).

The use of the Cox proportional hazards model for estimation of RRs, with an assumption that the period at risk is the same for all participants in a study, was introduced in 1994 (182), and compared with the log-binomial model a few years later (190). Because of overestimated standard errors, robust standard errors have been proposed (74). Unlike the CLL model, use of the Cox model with a constant observation time is quite widely reported in the literature. The extension of this model to analysis of data that are clustered, the shared frailty model, is commonly used in cases of recurrent events. However, as with the other methods that I considered, its performance when data are clustered has not been explored in methodological papers. When the performance of the shared frailty model was tested against the RI log-binomial model, it produced unbiased estimates of risk (average differences in point estimates ∼1%, and maximum difference <10%), with precision very close to that from the RI log-binomial model (average relative precision = 1.08). Also, the shared frailty model showed the narrowest dispersion of values of relative precision around the average value, indicating that agreement of SEs between this model and the RI log-binomial model was fairly consistent across associations of different outcomes and risk factors. Nonetheless, values of relative precision indicated differences in SEs up to 21%, but no pattern of prevalence rates of the risk factors and the outcomes with differences in SEs >10% was observed.

The Poisson model with robust SEs has been discussed by several researchers and has been proposed as an alternative approach to the estimation of RRs when the log-binomial fails to converge (74, 76, 191-193). However, most publications have focused on data that are independent (i.e. not clustered). Although a need for research on performance of the model in the context of clustered data was first identified more than a decade ago (76), interest in the topic has been quite limited. Three studies have explored the performance of the Poisson model with robust variance model using simulated data that were hierarchically structured (75, 77, 174), and they all concluded that the resulting effect estimates were adequately close to the real effect size. However, in two of these publications (75, 77), generalised estimating equation models with exchangeable correlation structure were fitted to account for the clustering, while the Poisson model in the third publication (174), although termed "random effects Poisson", was not clearly specified. As none of these studies used RI Poisson regression with robust SEs, comparisons with results shown in this chapter cannot be made. Nonetheless, my findings suggest that the RI Poisson model with robust SEs performs well with regard to estimation of effects, with RRs being very close to those derived from the RI log-binomial model (average ratio of RRs ≈ 1.01). Also, the average relative precision (ratio of SEs) indicated small average differences (1.11) in SEs from the two methods. Surprisingly, however, relative precision varied considerably across the (unadjusted and adjusted) analyses, with SEs from the Poisson model sometimes being markedly higher (i.e. low relative precision) than those from the log-binomial model. Disagreement in precision from the two methods was most notable in the associations of higher age bands (compared to the lower one) with wrist/hand pain. It remains unclear, however, why greater disagreement in precision was observed when a) the outcome variable (wrist/hand pain) had neither high (like low back pain) nor low (like elbow pain) prevalence; and its standardised dispersion of prevalence rates across clusters (SD/mean of prevalence rates) was comparable to other outcomes examined (i.e. elbow pain); and b) the risk factor presented similar characteristics in terms of prevalence rates and dispersion of prevalence rates across clusters to other risk factors explored.

As expected, the RI Poisson model without robust SEs produced estimates of effect and related precision that were very similar to those derived from the shared frailty model. As such, any comparison of the RI Poisson without robust SEs with the RI log-binomial model closely mirrored that for the shared frailty model discussed above. However, it was surprising that the two models (RI Poisson without robust SEs and shared frailty) produced SEs of point estimates that were closer to those from the RI log-binomial model than SEs from the RI Poisson with robust SEs were. These findings contradict reports (74, 75, 77) which describe that when the outcome is rare, the variance of point estimates produced by the Poisson model is very close to that from the log-binomial model, but it is overestimated otherwise, suggesting the use of robust variance estimation as a solution to this problem. The reasons for inconsistency of findings in this chapter with the published literature are unclear. Possible explanations might be the relative dispersion of the outcome or the risk factors under consideration, size of the effect estimate, level of clustering or a combination of the above.

In this chapter I reviewed the main methods that are available for the estimation of RRs when data are hierarchically clustered, and compared them using real data from a large multi-centre study. Estimates of relative risk from these methods and their relative precision were compared to those from the RI log-binomial model, which is widely accepted as the gold standard for estimation of RRs. The results presented support the use of RI CLL regression for estimating RRs and their variance when the RI log-binomial fails to converge. Alternative models may be used when the focus is more on point estimates of RRs than their precision and, specifically, the shared frailty and RI Poisson models appear to provide useful alternatives to the RI log-binomial model. Future work based on simulated data is needed to identify situations under which i) RI Poisson regression with robust SEs results in less satisfactory estimation of effects and related precision (as was seen in the analysis of CUPID data), and ii) other models are likely to produce unbiased RRs with appropriate 95% CIs.

Chapter 9. General conclusions and future work

In this thesis I have used simulated data and data from a large multicentre study of musculoskeletal symptoms to address three main research questions concerning statistical modelling of clustered data. In this final chapter, the key findings are summarised, implications are discussed, and directions for further work are suggested. This is done separately for the three questions explored in the thesis.

Consequences of ignoring clustering

The first aim was to explore implications of failing to account for clustering in statistical inference when data are hierarchically structured. The purpose was to identify the nature and extent of errors that might occur, and circumstances in which errors might be sufficiently large to be of practical importance. Specific assumptions were made about the association between the outcome and the explanatory variable, and also about the distribution of error terms and of average values of the outcome across different clusters. With these assumptions I then simulated data under various conditions, including different: types of outcome and explanatory variables (continuous or binary); levels of clustering; and relative between- to within- cluster dispersion of the explanatory variable. In analysis of those simulated data, I showed that when clustering was ignored and effects were estimated through a naïve regression model, they were on average unbiased when the outcome under investigation was continuous, but were underestimated (i.e. biased towards the null) when the outcome was binary. The precision of effect estimates was overestimated when the outcome of interest was binary, and also when both the outcome and the explanatory variable were continuous. However, in linear regression with a binary explanatory variable, the standard errors of effects estimated from the naïve model were falsely high (i.e. precision was underestimated). The magnitude of the bias, both in point estimates and in their precision, increased with greater clustering of the outcome variable (after adjustment for the explanatory variable). I also demonstrated that biases were influenced by the distribution of the explanatory variable across clusters, and the extent to which it was clustered. To supplement these findings of consequences of ignoring clustering in statistical inference I explored rates of Type I error and 95% confidence interval coverage. These were very close to the nominal values of 5% and 95%, respectively, when the RI regression model was fitted. However, they deviated from the nominal values when a naïve analytical approach was employed, verifying that differences between results from the naïve model and the hierarchical model reflected errors in the former and not the latter.

These results should be interpreted cautiously due to several limitations. The simulations did not cover the full range of possible scenarios which might be of interest, and this could be a topic for further research. For example, data presented here were simulated so that the assumptions of the random-intercept model were met (i.e. normality of cluster-level residuals, and independence of cluster-level and individual-level residuals). Through the algebraic modelling developed for the case of a continuous outcome, I showed that conclusions were not critically dependent on the normality of the cluster-level residuals. It was assumed though that the cluster-level and individual-level residuals were uncorrelated. Moreover, no similar algebraic modelling was possible in the case of a binary outcome, and conclusions about the importance of assumptions in the simulations cannot be necessarily be extrapolated to the RI logistic regression modelling. In practice, the distributions of real data might deviate from the assumptions that were made in the simulations, and it would be of interest to understand their importance, particularly in logistic regression modelling.

In addition, the numbers of clusters and of observations per cluster in the simulated datasets were large, aiming to minimise random sampling variation so that bias could be better characterised. However, in practice, either the number of clusters or the number of observations per cluster (or both) could be much smaller, and in these situations clustering effects might differ from those described here. Moreover, cluster sizes, and the magnitude of the true effect did not vary in my simulated data, whereas in most real studies, clusters vary in size, and effect sizes can be much larger or smaller than those that I considered.

Finally, I also assumed that the outcome depended on only one explanatory variable, while in practice several covariates may be relevant, and the association between them is often complex. Thus, future research could usefully focus on situations in which effects of one explanatory variable are modified by another. That might be best achieved by theoretical modelling of the problem, and then complementary simulation studies.

Meanwhile, despite the limitations, conclusions can be drawn about a number of circumstances in which simple regression analyses that ignore clustering are liable to be misleading. These occasions occur when researchers are principally interested in

- estimation of effect of an explanatory variable on a binary outcome through logistic regression and the ICC value of the outcome variable is >0.1
- precision of the effect of a continuous explanatory on an (either continuous or binary) outcome variable, when the variance of the explanatory variable within clusters is different (either higher or lower) from that between clusters, and the ICC of the outcome adjusted for the predictor is >0.01

- precision of the effect of a binary explanatory on a binary outcome variable, when the prevalence rates vary considerably across clusters, and the ICC of the outcome adjusted for the predictor is >0.03

Most importantly, rates of false positive results and of coverage intervals importantly further from the nominal values (for example <85% for coverage rates and >10% for type I error rates) when a naïve regression model is fitted occur when:

- ICC is ≥0.01 and both the outcome and the explanatory variables are continuous
- ICC is >0.01 for any binary outcome and continuous explanatory variable
- ICC is >0.1, the outcome is continuous and the explanatory variable is binary with a large dispersion of its cluster-specific prevalence rates, and the overall prevalence of the explanatory variable is low (~5%)
- ICC is ≥0.1, the outcome is binary and the explanatory variable is binary with a large dispersion of its cluster-specific prevalence rates

Further work based on simulated data was done to explore how estimates derived from naïve regression models using dummy variables (DV) for the clusters compared to those from RI models for different distributions and levels of clustering of the outcome and explanatory variables. I found that despite small differences in the precision of point estimates from the two models (DV and RI) (occurring mainly when the explanatory variable was continuous with a high relative dispersion of x across clusters), statistical inference was unlikely to be misleading when DV models were fitted instead of RI models. However, the number of scenarios explored was limited and further work is needed to explore comparisons of the two approaches in situations of varying cluster sizes and number of clusters, and also in situations in which cluster effects follow distributions other than normal.

Comparison of meta-analysis to pooled analysis

A second aim of this thesis was to explore alternative analytical approaches to the clustering of data that occurs in meta-analysis. When all individual observations were available, one would fit a multi-level model that allowed for study/cluster effects, in what is termed pooled analysis of individual data. To explore how estimates of effect from meta-analysis of summary results at cluster level compare with those from pooled analysis of individual data, I used real data from the CUPID study, a multicentre study in which participants are uniquely grouped into 47 groups/clusters. For the purpose of this investigation, the 47 clusters were regarded as 47 independent studies for which the methods of data collection and ascertainment of exposure and health outcomes were standardised. Estimates of the effects of specific exposures on musculoskeletal pain outcomes were summarised by odds ratios. These were derived separately from each of the groups/studies and a meta-analytical odds ratio was derived and compared with that derived from pooled analysis using multi-level modelling. Although several alternative analytical models were applied for each of the two approaches, the main comparison was between the random effects (or DerSimonian and Laird) model (a method of meta-analysis) and the random coefficients model (a method of pooled analysis). Comparison was based on associations between exposures and outcomes that covered a wide range of prevalence rates and sizes of effect.

Comparison of the two approaches showed that in most cases meta-analytical odds ratios were similar to those from pooled analyses. However, when either the outcome or exposure of interest was rare, odds ratios from the two analytical approaches differed, with those from meta-analysis tending to be biased towards the null value of one. The largest differences in point estimates were observed for the rarest outcome considered, with those from meta-analysis being lower than those from pooled analysis by up to 27%. Unlike the point estimates, the precision of the resulting odds ratios was different between the two models; it was lower when derived from meta-analysis than from pooled analysis of individual data. Again, the largest differences were seen mainly for the rarest outcome explored, with standard errors of point estimates from meta-analysis being up to 46% higher than those from pooled analysis.

Despite the advantage that the studies participating in the meta-analyses had used exactly the same methods of ascertainment of exposure and outcome and that comparisons were based on analyses of real data, showing that differences of the size observed can occur in practice, it should be noted that they all came from a single large study, using only a limited set of variables. Similar investigations could usefully be carried out using data from other studies to better characterise relationships. Meanwhile, results from my comparison of the two analytical approaches, provide some reassurance that when evidence from different studies is synthesised using meta-analysis, point estimates will often be close to those that would be obtained from multi-level modelling of pooled individual data. Comparison of the two approaches did, however, suggest a lower level of agreement when either the prevalence of the outcome or of the exposure of interest is very low or very high. The reasons for discrepancies remain unknown, and could be further explored using simulation studies.

Estimating relative risks from clustered data

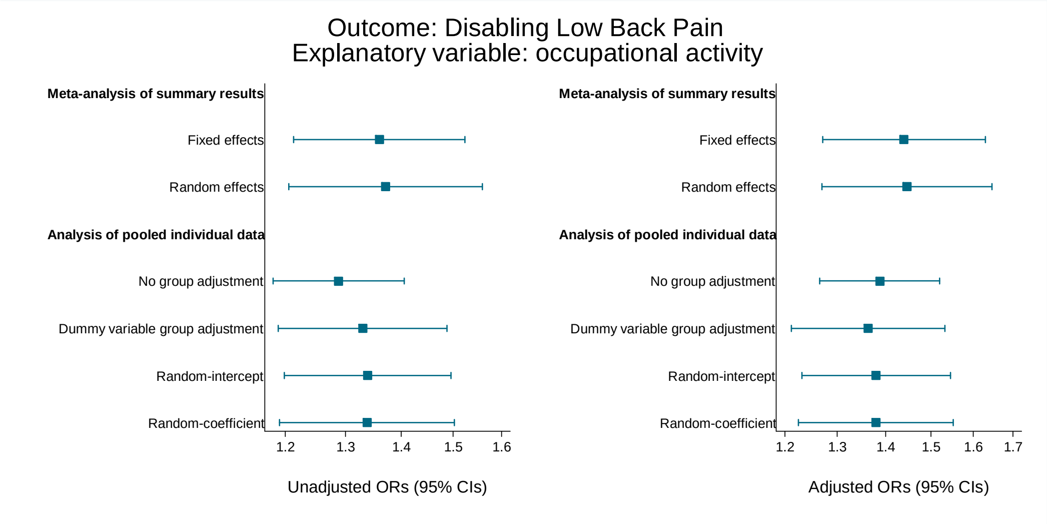
The third objective of this thesis was to compare methods for estimating relative risks from clustered data. Having identified analytical techniques that have been proposed for estimation of relative risks in a setting of independent observations, I described how these can be adjusted to account for clustering in the outcome variable. I then used real data from the CUPID study, to compare results from other methods with those from the random intercept log-binomial model (assumed to be the best approach if convergence can be achieved). Results suggested that the random intercept complementary log-log model produces estimates of effect and precision that are similar to those from the random intercept log-binomial model. Other models were found to give effect estimates adequately close to those from the log-binomial model, although with differing estimates of the related precision. However, in contrast to what might be expected from the literature on estimation of relative risks from unclustered data, the random-intercept Poisson model with robust variance yielded point estimates with SEs that were on average higher than those from the random intercept log-binomial model.

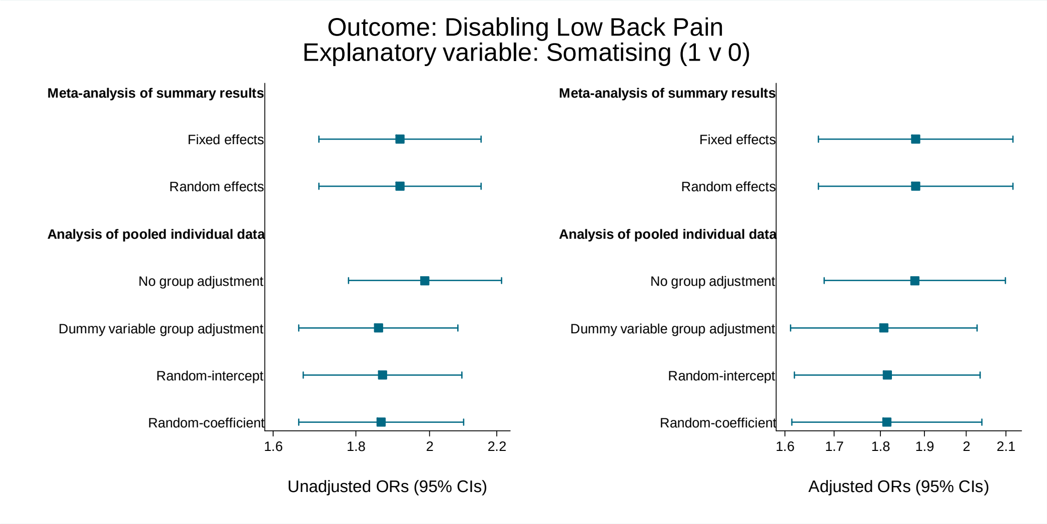
These findings support use of the random-intercept complementary log-log model to estimate relative risks from clustered data, where the random intercept log-binomial model cannot be employed. However, to date, its implementation has been rarely reported in the literature, and other alternative models could reasonably be used where the focus was more on point estimates of relative risks than their precision. That said, the analysis presented was based on only a single set of data, and does not allow firm recommendations regarding situations in which methods other than the random-intercept log-binomial model will prove satisfactory. To answer that question will require further research. Future work in this area could usefully be carried out using simulated data, and given the scope for error that I have demonstrated, should focus primarily on the analytical models that directly (i.e. not through the use of odds ratios from logistic regression) estimate relative risks. Varying conditions (for example prevalence rates of outcome and explanatory variables, relative dispersion of the explanatory variable across clusters, size of true effect, number of clusters etc.) one at a time in different sets of simulations would provide a thorough examination of the problem that would enable a more confident choice of analytical model for future estimation of relative risks from clustered data.

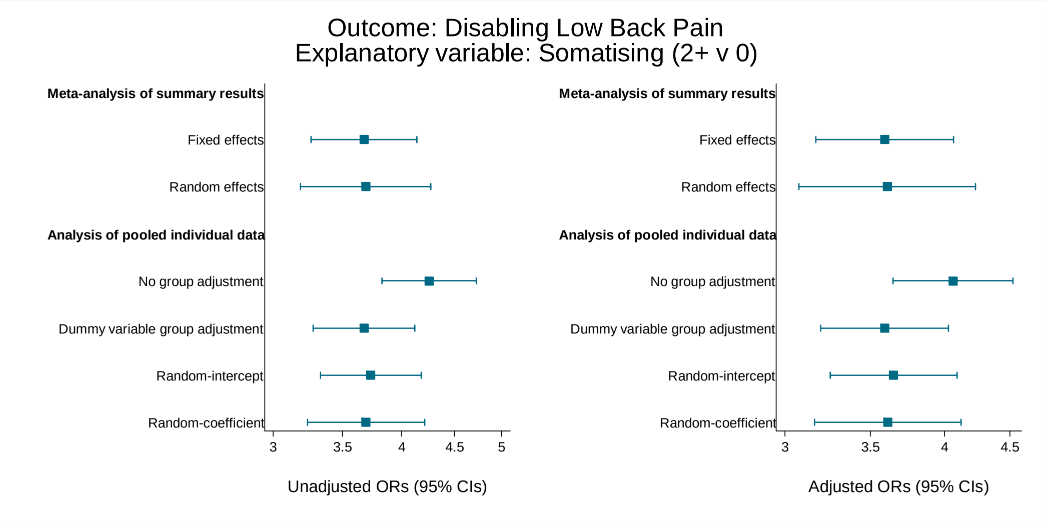
In all of the analyses presented in this thesis, I employed conditional methods to take clustering into account. However, marginal methods can alternatively be used. Indeed, they are often the choice of analytical approach when data are clustered, particularly when the interest is estimation of effects of variables defined at cluster level. Also, the structure of the datasets (both simulated and real data from the CUPID study) used in this thesis was such that the unit of analysis was the individual, with individuals clustered in groups, and the explanatory variables of interest were measured at the individual level. However, clustering also occurs when multiple measurements are taken on the same individual at several points in time (longitudinal data), with the interest being in effects of explanatory variables defined at the cluster level (78). To draw a more complete picture about the research questions addressed in this thesis, the work should be expanded to longitudinal data structures, considering effects of explanatory variables measured at the higher level of data hierarchy, and applying marginal methods.

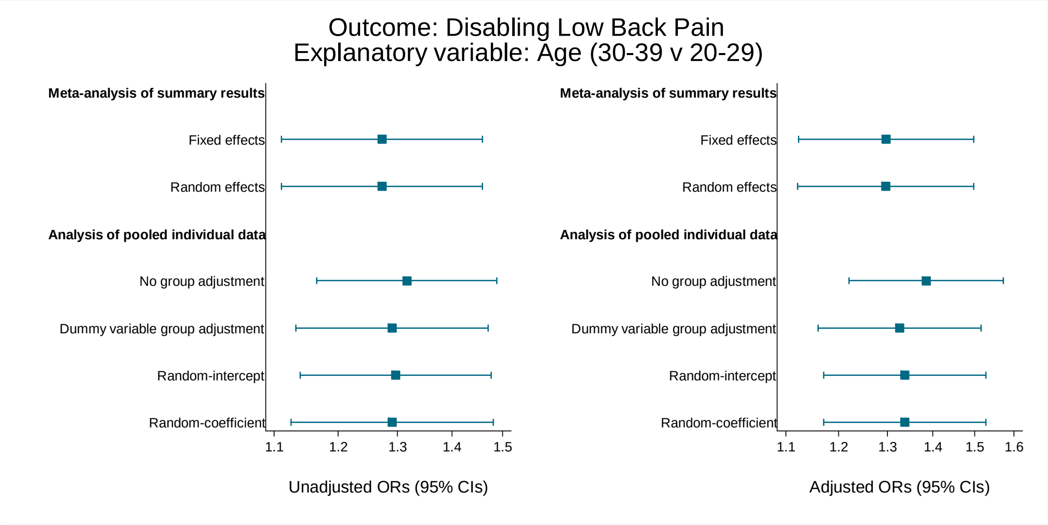
Appendices

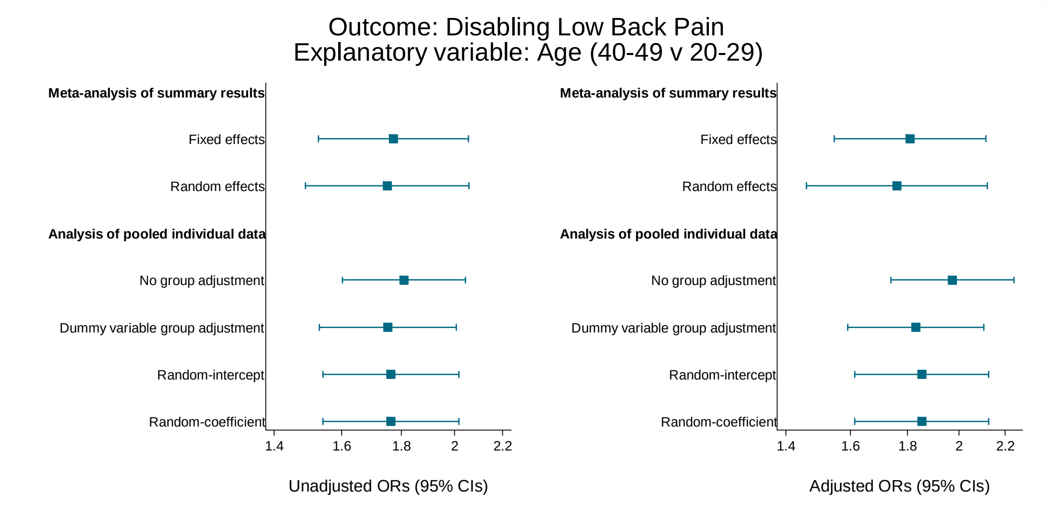
Appendix 1. Graphical illustration of the comparison of estimates of effects and related precision from the different models presented in Chapter 6

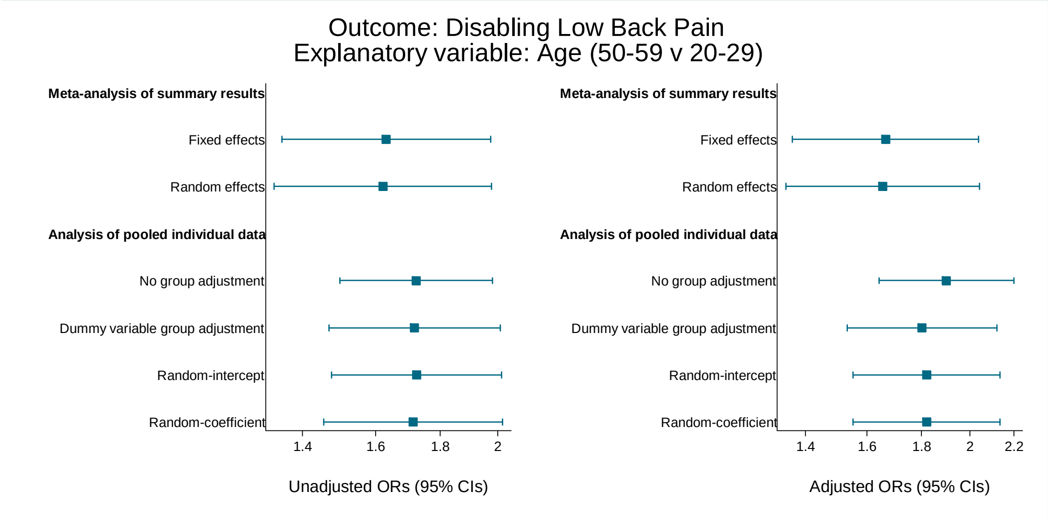


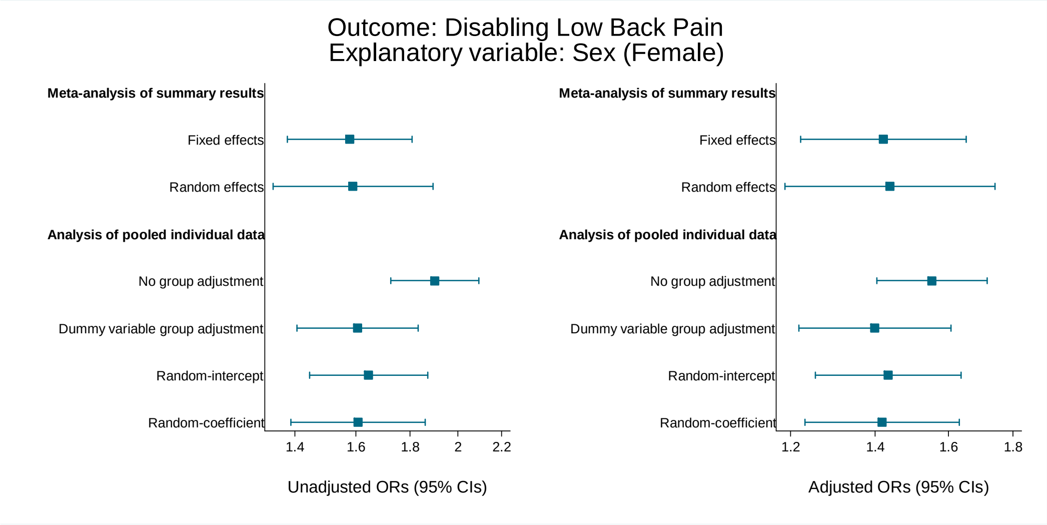


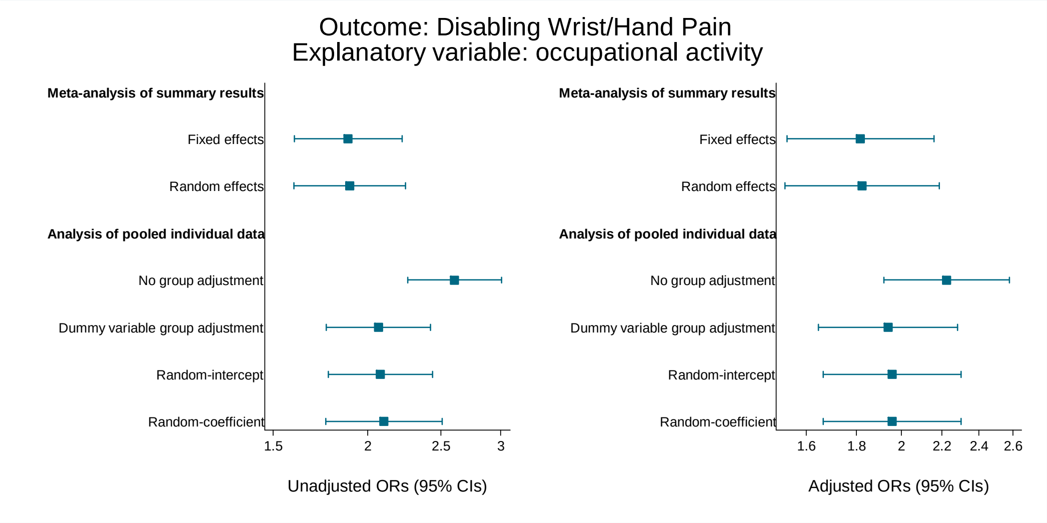


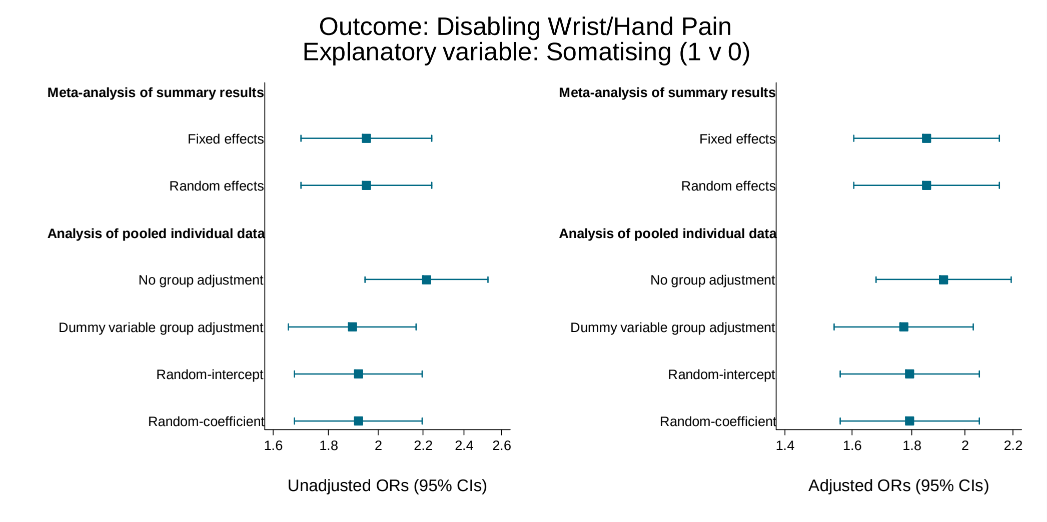

Outcome: Disabling Low Back Pain
Explanatory variable: Age (30-39 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.1
1.2
1.3
1.4
1.5
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.1
1.2
1.3
1.4
1.5
1.6
Adjusted ORs (95% CIs)

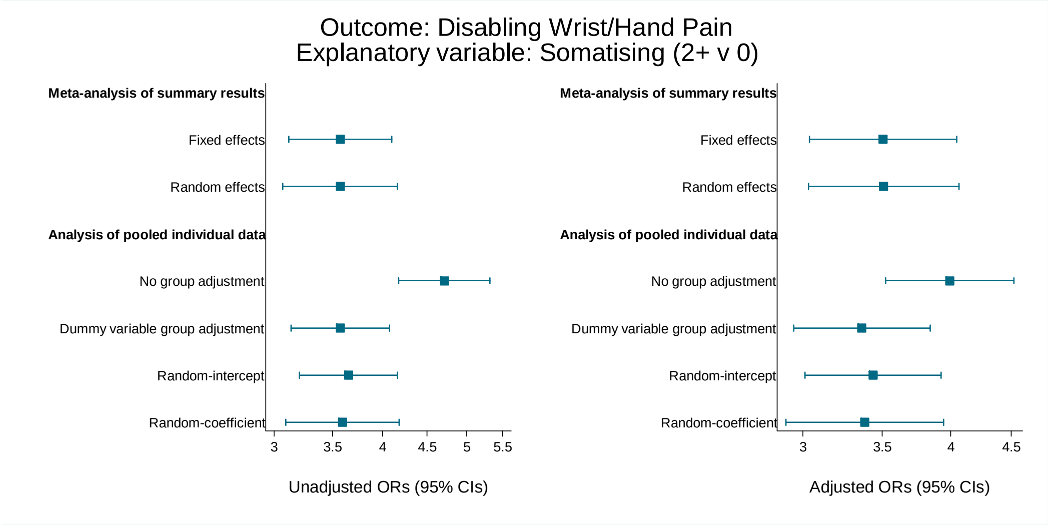

Outcome: Disabling Low Back Pain
Explanatory variable: Age (40-49 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.4
1.6
1.8
2
2.2
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.4
1.6
1.8
2
2.2
Adjusted ORs (95% CIs)


Outcome: Disabling Low Back Pain
Explanatory variable: Age (50-59 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.4
1.6
1.8
2
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.4
1.6
1.8
2
2.2
Adjusted ORs (95% CIs)

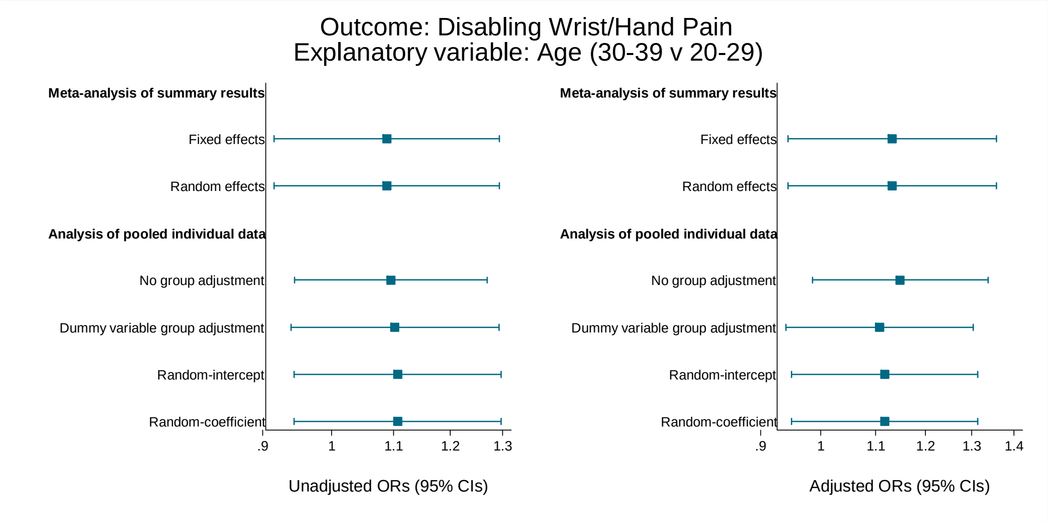

Outcome: Disabling Low Back Pain
Explanatory variable: Sex (Female)


Outcome: Disabling Wrist/Hand Pain
Explanatory variable: occupational activity

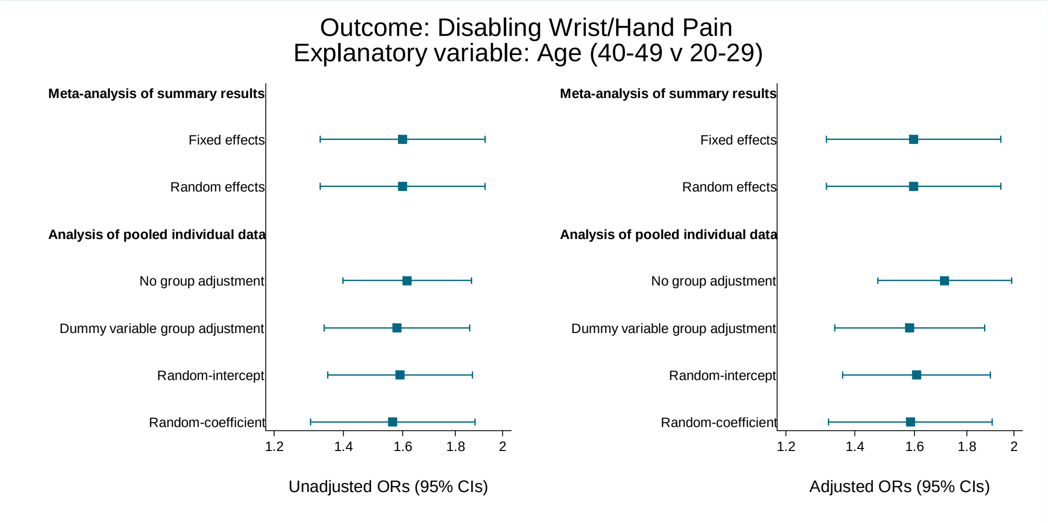

Outcome: Disabling Wrist/Hand Pain
Explanatory variable: Somatising (1 v 0)



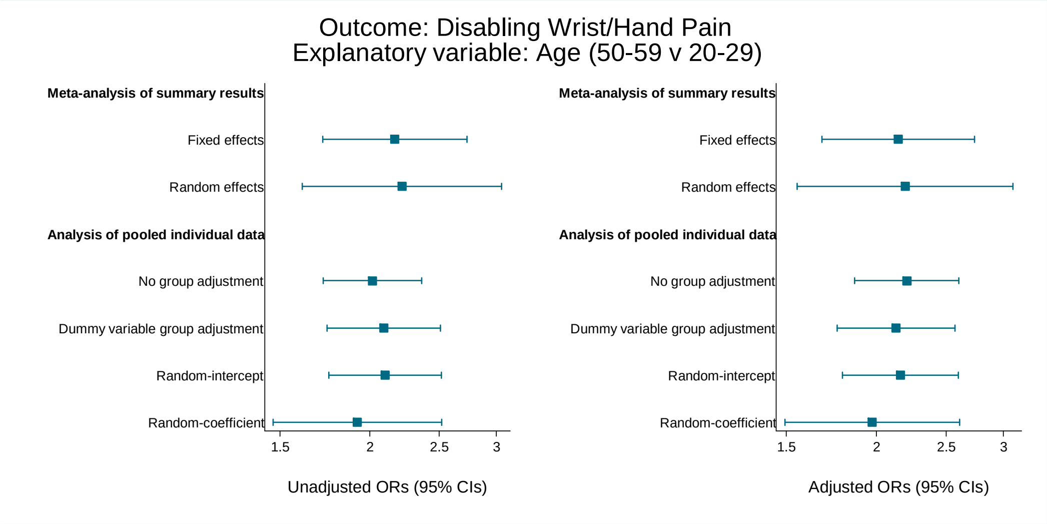
Outcome: Disabling Wrist/Hand Pain
Explanatory variable: Somatising (2+ v 0)

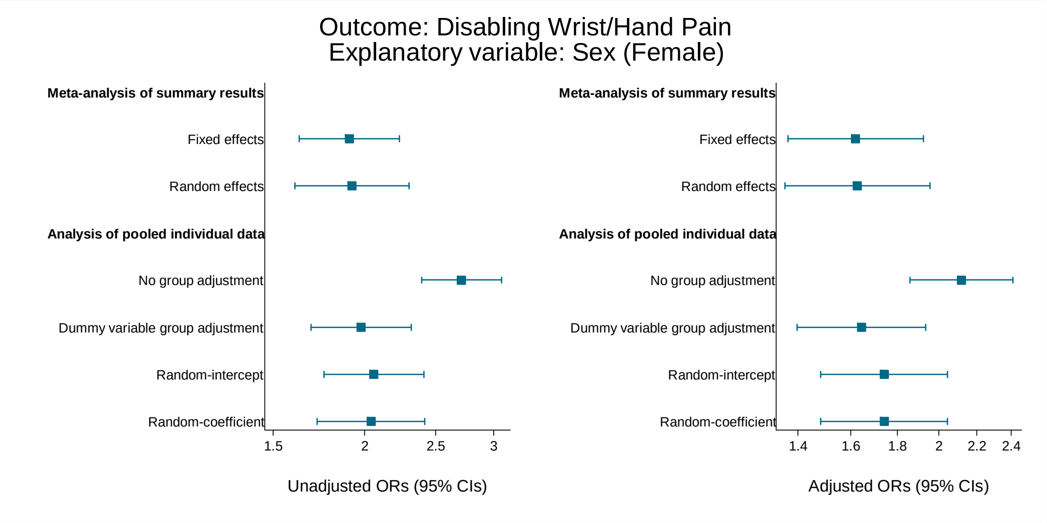


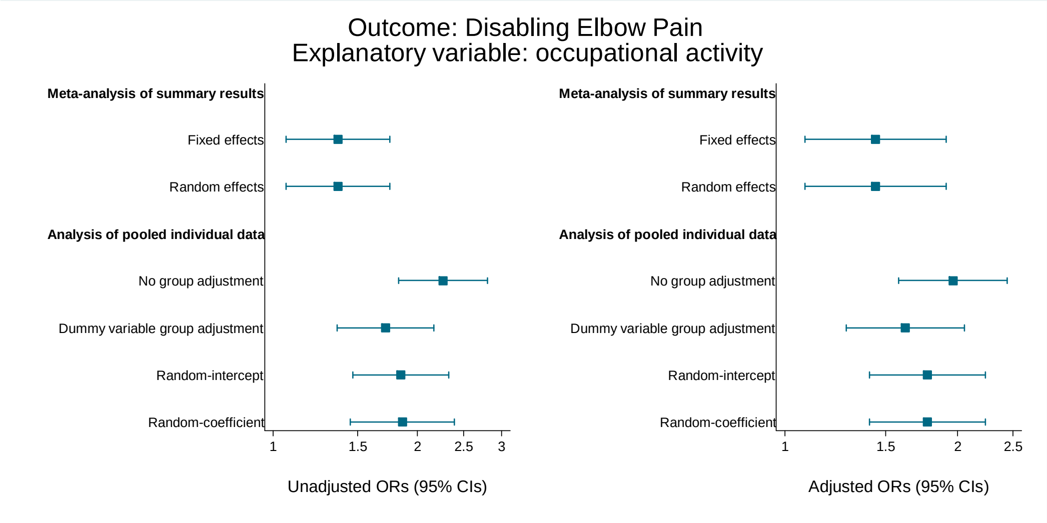
Outcome: Disabling Wrist/Hand Pain
Explanatory variable: Age (30-39 v 20-29)

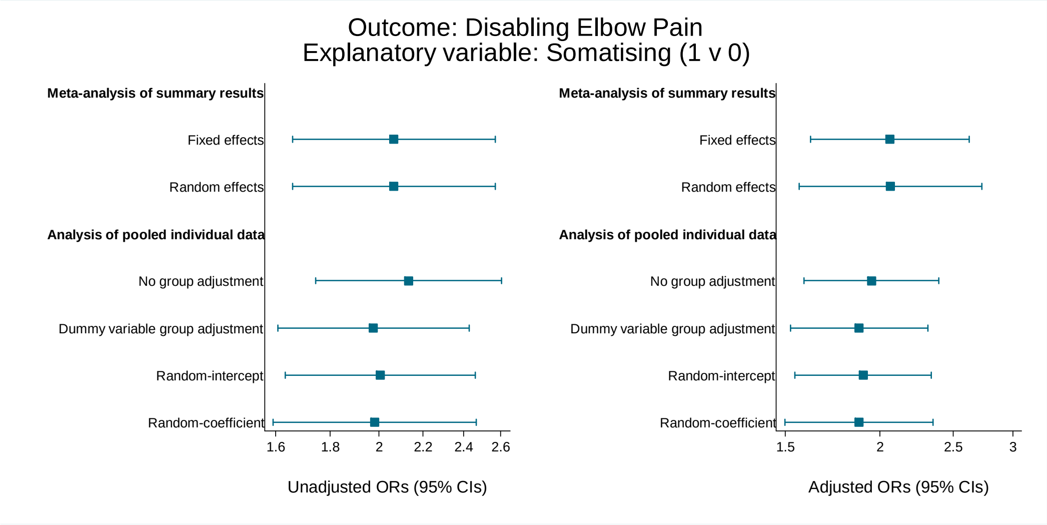


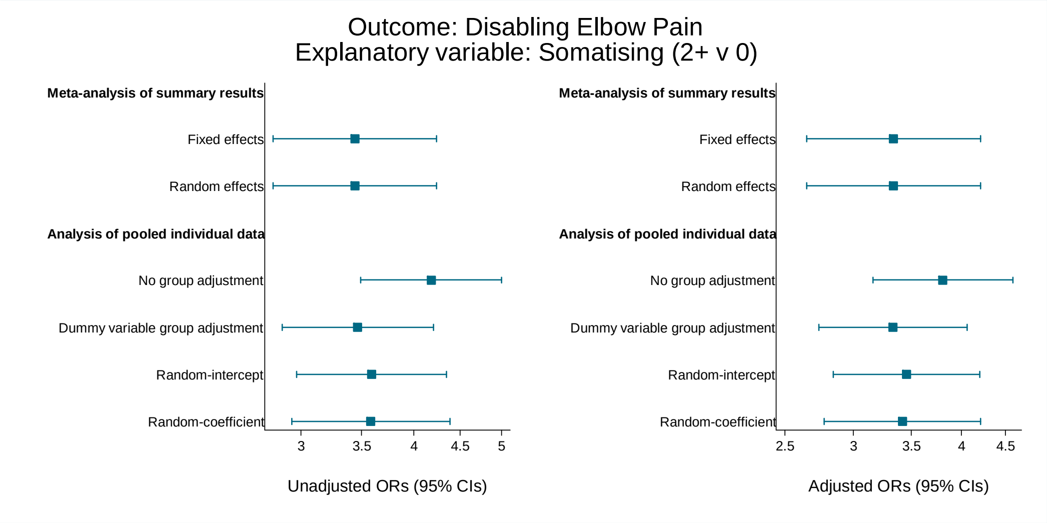
Outcome: Disabling Wrist/Hand Pain
Explanatory variable: Age (40-49 v 20-29)

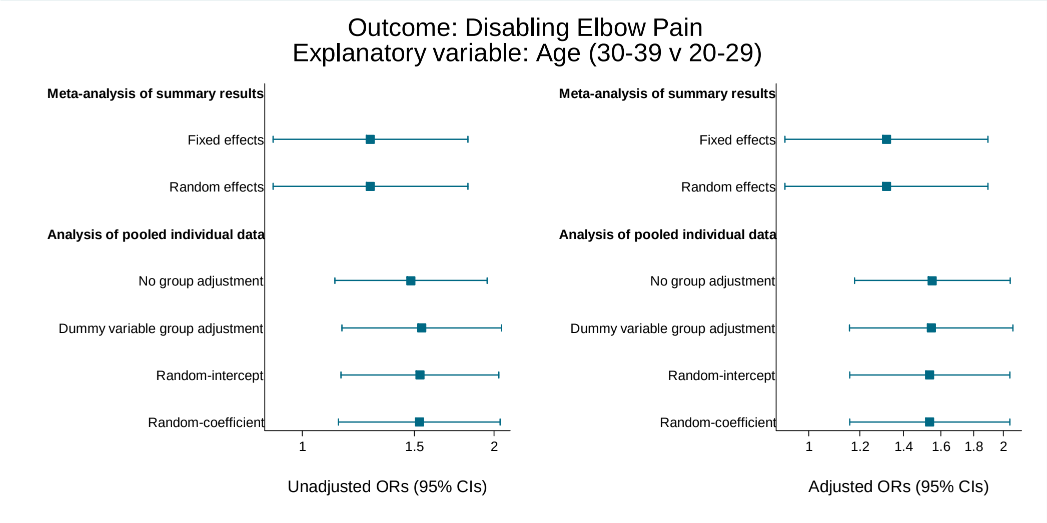

Outcome: Disabling Wrist/Hand Pain
Explanatory variable: Age (50-59 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.5
2
2.5
3
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.5
2
2.5
3
Adjusted ORs (95% CIs)

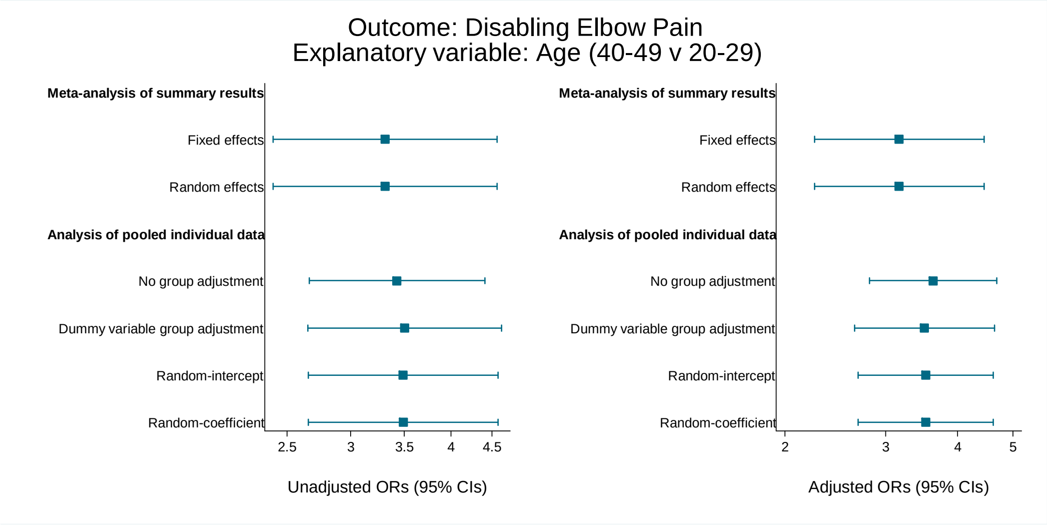

Outcome: Disabling Wrist/Hand Pain
Explanatory variable: Sex (Female)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.5
2
2.5
3
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.4
1.6
1.8
2
2.2
2.4
Adjusted ORs (95% CIs)

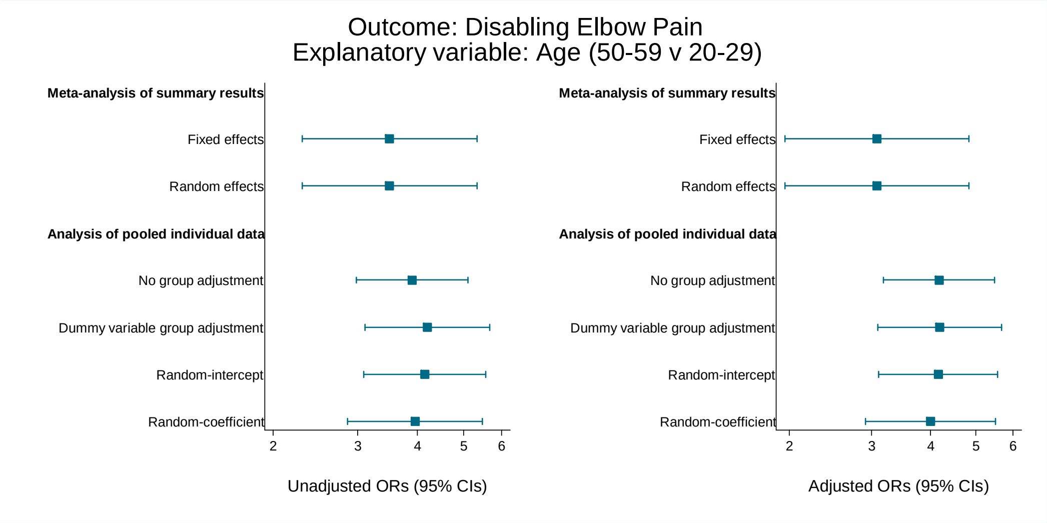

Outcome: Disabling Elbow Pain
Explanatory variable: occupational activity
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1
1.5
2
2.5
3
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1
1.5
2
2.5
Adjusted ORs (95% CIs)

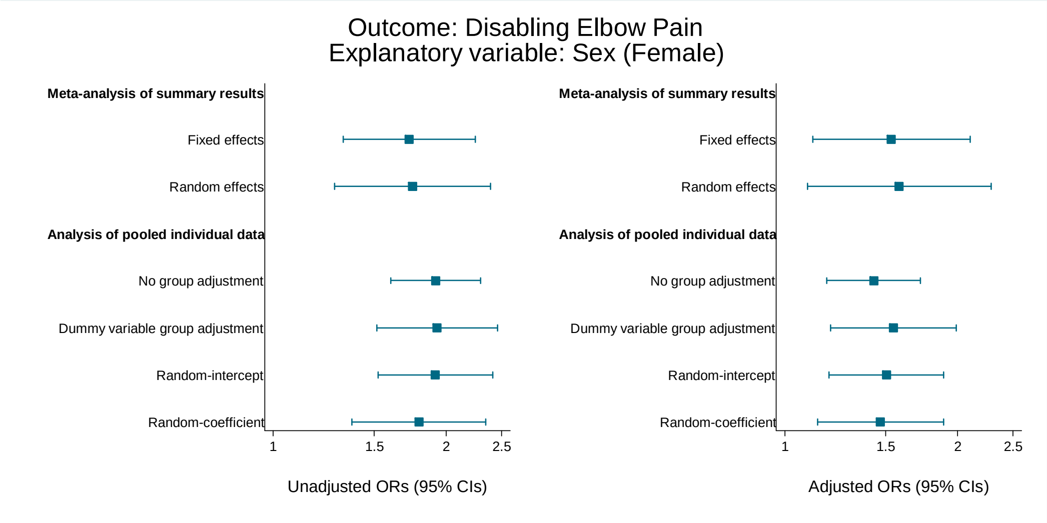

Outcome: Disabling Elbow Pain
Explanatory variable: Somatising (1 v 0)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.6
1.8
2
2.2
2.4
2.6
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1.5
2
2.5
3
Adjusted ORs (95% CIs)


Outcome: Disabling Elbow Pain
Explanatory variable: Somatising (2+ v 0)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
3
3.5
4
4.5
5
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
2.5
3
3.5
4
4.5
Adjusted ORs (95% CIs)


Outcome: Disabling Elbow Pain
Explanatory variable: Age (30-39 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1
1.5
2
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1
1.2
1.4
1.6
1.8
2
Adjusted ORs (95% CIs)


Outcome: Disabling Elbow Pain
Explanatory variable: Age (40-49 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
2.5
3
3.5
4
4.5
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
2
3
4
5
Adjusted ORs (95% CIs)


Outcome: Disabling Elbow Pain
Explanatory variable: Age (50-59 v 20-29)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
2
3
4
5
6
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
2
3
4
5
6
Adjusted ORs (95% CIs)


Outcome: Disabling Elbow Pain
Explanatory variable: Sex (Female)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1
1.5
2
2.5
Unadjusted ORs (95% CIs)
Meta-analysis of summary results
Fixed effects
Random effects
Analysis of pooled individual data
No group adjustment
Dummy variable group adjustment
Random-intercept
Random-coefficient
1
1.5
2
2.5
Adjusted ORs (95% CIs)

Appendix 2. Derivation of relative risks (RRs) from odds ratios (ORs)

In the equations shown below, p_1 and p_0 are the proportion of outcome in the exposed and the unexposed groups, respectively.

$$RR = \frac{p_1}{p_0}$$

$$OR = \frac{p_1/p_0}{(1-p_1)/(1-p_0)} = \frac{p_1 \times (1-p_0)}{p_0 \times (1-p_1)} = RR \times \frac{(1-p_0)}{(1-p_1)}$$

$$RR = OR \times \frac{(1-p_1)}{(1-p_0)} = OR \times \frac{(1 - RR \times p_0)}{(1-p_0)} = \frac{OR - OR \times RR \times p_0}{(1-p_0)}$$

$$RR - RR \times p_0 = OR - OR \times RR \times p_0$$

$$RR = \ OR - OR \times RR \times p_0 + RR \times p_0$$

$$RR = OR + RR \times p_0 \times (1 - OR)$$

$$RR - RR \times p_0 \times (1 - OR) = OR$$

$$RR = \frac{OR}{(1-p_0) + (OR \times p_0)}$$

Appendix 3. Cox regression model with constant time at risk

In survival analysis, the outcome of interest is the time T until an event. The probability density function (pdf), $f(t)$ describes the likelihood of observing T at time t. Then, the cumulative distribution function (cdf), $F(t)$, describes the probability of observing T before by time t ($\Pr(T \leq t)$). Then the cdf will be

$$F(t) = \int_0^t f(t)dt$$

The probability of surviving beyond time t is the described in the survival function, $S(t)$, that can be derived from the cdf, as follows

$$S(t) = 1 - F(t)$$

In survival analysis, the main interest is then to estimate the hazard function which is the ratio of the pdf to the survival function

$$h(t) = \frac{f(t)}{S(t)}$$

The cumulative hazard function is then obtained from integrating the hazard function over a time interval

$$H(t) = \int_0^t h(u)du = -\ln\big(1 - F(t)\big)$$

Thus,

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}$$

Assuming a time-constant baseline hazard, $h_0(t) = \lambda$, the survival function becomes

$$S(t) = \exp\{-\lambda t\}$$

and the pdf becomes

$$f(t) = \lambda\exp\{-\lambda t\}$$

which is the pdf of an exponential random variable with expectation λ^{-1}. This leads to a Cox model with hazard function

$$h_i(t) = \lambda\exp\{x_i'\beta\}$$

The parameters can be estimated using the likelihood method. The log-likelihood is given by

$$l = \sum_i \{d_i \log\big(h_i(t)\big) - t_i h_i(t)\}$$

where d_i is the event indicator, taking values 1 if the event has occurred and 0 otherwise.

With an additive constant, this is the same expression as the log-likelihood of the $d_i's$ seen as a realisation of the Poisson variable with mean $\mu_i = t_i h_i(t)$. As such, estimates can be obtained from the Poisson model

$$\log(\mu_i) = \log(t_i) + \beta_0 + x_i'\beta$$

where $\beta_0 = \log(\lambda)$.

Appendix 4. Relative risks and 95% confidence intervals for the associations between disabling pain outcomes and risk factors

Table 1. Relative risks and 95% CIs for the univariate associations between disabling LBP and risk factors

	Point Estimate		95% Confidence Intervals			
	RR	Ratio of RRs	lower	upper	SE(ln(RR))	Ratio of SEs
Outcome: LBP (22%) Explanatory variable: Sex (Female)						
RI log-binomial	1.460	--	1.322	1.612	0.051	--
Zang & Yu	1.497	1.025	1.353	1.651	0.051	1.003
Modified Zang & Yu	1.492	1.022	1.350	1.645	0.050	0.994
RI Poisson (no robust SEs)	1.465	1.004	1.308	1.640	0.058	1.139
RI Poisson (robust SEs)	1.465	1.004	1.299	1.651	0.061	1.206
Shared frailty model	1.459	0.999	1.303	1.633	0.058	1.136
CLL (robust SEs)	1.671	1.144	1.374	2.031	0.100	1.969
RI CLL	1.479	1.013	1.334	1.640	0.053	1.038
Outcome: LBP (22%) Explanatory variable: Age (30-39 v 20-29)						
RI log-binomial	1.228	--	1.110	1.358	0.051	--
Zang & Yu	1.216	0.990	1.105	1.334	0.048	0.932
Modified Zang & Yu	1.170	0.953	1.085	1.258	0.038	0.734
RI Poisson (no robust SEs)	1.227	0.999	1.094	1.376	0.058	1.135
RI Poisson (robust SEs)	1.227	0.999	1.115	1.351	0.049	0.952
Shared frailty model	1.227	0.999	1.095	1.376	0.058	1.135
CLL (robust SEs)	1.241	1.011	1.094	1.409	0.065	1.254
RI CLL	1.226	0.998	1.110	1.354	0.051	0.986
Outcome: LBP (22%) Explanatory variable: Age (40-49 v 20-29)						
RI log-binomial	1.528	--	1.381	1.691	0.052	--
Zang & Yu	1.503	0.983	1.372	1.639	0.045	0.875
Modified Zang & Yu	1.435	0.939	1.326	1.546	0.039	0.757
RI Poisson (no robust SEs)	1.532	1.002	1.363	1.722	0.060	1.152
RI Poisson (robust SEs)	1.532	1.002	1.379	1.702	0.054	1.040
Shared frailty model	1.531	1.002	1.362	1.721	0.060	1.152
CLL (robust SEs)	1.574	1.030	1.341	1.848	0.082	1.583
RI CLL	1.538	1.006	1.391	1.700	0.051	0.987
Outcome: LBP (22%) Explanatory variable: Age (50-59 v 20-29)						
RI log-binomial	1.513	--	1.349	1.698	0.059	--
Zang & Yu	1.484	0.981	1.336	1.641	0.052	0.894
Modified Zang & Yu	1.416	0.936	1.293	1.543	0.045	0.769
RI Poisson (no robust SEs)	1.513	1.000	1.324	1.730	0.068	1.164
RI Poisson (robust SEs)	1.513	1.000	1.341	1.707	0.062	1.049
Shared frailty model	1.511	0.998	1.322	1.727	0.068	1.163
CLL (robust SEs)	1.515	1.001	1.280	1.794	0.086	1.468
RI CLL	1.513	1.000	1.352	1.694	0.057	0.978
Outcome: LBP (22%) Explanatory variable: Occupational activity						
RI log-binomial	1.224	--	1.131	1.325	0.040	--
Zang & Yu	1.252	1.023	1.152	1.358	0.042	1.038
Modified Zang & Yu	1.256	1.026	1.154	1.364	0.043	1.054
RI Poisson (no robust SEs)	1.236	1.009	1.126	1.357	0.048	1.179
RI Poisson (robust SEs)	1.236	1.009	1.130	1.352	0.046	1.135
Shared frailty model	1.237	1.010	1.126	1.357	0.048	1.177
CLL (robust SEs)	1.217	0.994	1.010	1.466	0.095	2.349
RI CLL	1.243	1.015	1.144	1.350	0.042	1.048

Outcome: LBP (22%)						
Explanatory variable: Distressing somatic symptoms (1 v 0)						
RI log-binomial	1.624	--	1.490	1.769	0.044	--
Zang & Yu	1.579	0.973	1.463	1.700	0.038	0.877
Modified Zang & Yu	1.288	0.793	1.239	1.335	0.019	0.436
RI Poisson (no robust SEs)	1.633	1.006	1.480	1.802	0.050	1.147
RI Poisson (robust SEs)	1.633	1.006	1.472	1.811	0.053	1.208
Shared frailty model	1.632	1.005	1.479	1.800	0.050	1.147
CLL (robust SEs)	1.647	1.015	1.494	1.817	0.050	1.141
RI CLL	1.586	0.977	1.465	1.717	0.040	0.924
Outcome: LBP (22%)						
Explanatory variable: Distressing somatic symptoms (2+ v 0)						
RI log-binomial	2.485	--	2.303	2.681	0.039	--
Zang & Yu	2.365	0.952	2.232	2.498	0.029	0.740
Modified Zang & Yu	1.903	0.766	1.831	1.973	0.019	0.490
RI Poisson (no robust SEs)	2.531	1.019	2.307	2.776	0.047	1.214
RI Poisson (robust SEs)	2.531	1.019	2.283	2.805	0.052	1.351
Shared frailty model	2.528	1.017	2.305	2.771	0.047	1.209
CLL (robust SEs)	2.759	1.110	2.462	3.092	0.058	1.498
RI CLL	2.525	1.016	2.342	2.723	0.038	0.989

Table 2. Relative risks and 95% CIs for the univariate associations between disabling WHP and risk factors

	Point Estimate		**95% Confidence Intervals**			
	RR	**Ratio of RRs**	**lower**	**upper**	**SE(ln(RR))**	**Ratio of SEs**
Outcome: WHP (14%)						
Explanatory variable: Sex (Female)						
RI log-binomial	1.799	--	1.578	2.050	0.067	--
Zang & Yu	1.906	1.060	1.665	2.176	0.068	1.022
Modified Zang & Yu	1.892	1.052	1.656	2.154	0.067	1.006
RI Poisson (no robust SEs)	1.808	1.005	1.567	2.086	0.073	1.092
RI Poisson (robust SEs)	1.808	1.005	1.581	2.068	0.069	1.027
Shared frailty model	1.794	0.998	1.556	2.069	0.073	1.089
CLL (robust SEs)	2.401	1.335	1.658	3.476	0.189	2.827
RI CLL	1.850	1.029	1.615	2.119	0.069	1.038
Outcome: WHP (14%)						
Explanatory variable: Age (30-39 v 20-29)						
RI log-binomial	1.077	--	0.945	1.226	0.066	--
Zang & Yu	1.096	1.018	0.959	1.248	0.067	1.011
Modified Zang & Yu	1.081	1.004	0.964	1.206	0.057	0.858
RI Poisson (no robust SEs)	1.093	1.015	0.946	1.262	0.073	1.106
RI Poisson (robust SEs)	1.093	1.015	0.952	1.254	0.070	1.058
Shared frailty model	1.092	1.014	0.946	1.261	0.073	1.106
CLL (robust SEs)	1.088	1.010	0.871	1.358	0.113	1.705
RI CLL	1.089	1.011	0.953	1.244	0.068	1.023
Outcome: WHP (14%)						
Explanatory variable: Age (40-49 v 20-29)						
RI log-binomial	1.417	--	1.245	1.612	0.066	--
Zang & Yu	1.462	1.032	1.287	1.653	0.064	0.968
Modified Zang & Yu	1.421	1.003	1.264	1.588	0.058	0.882
RI Poisson (no robust SEs)	1.456	1.028	1.260	1.684	0.074	1.121
RI Poisson (robust SEs)	1.456	1.028	1.218	1.741	0.091	1.382
Shared frailty model	1.453	1.026	1.257	1.680	0.074	1.121
CLL (robust SEs)	1.506	1.063	1.145	1.979	0.139	2.114
RI CLL	1.457	1.028	1.276	1.663	0.068	1.024
Outcome: WHP (14%)						
Explanatory variable: Age (50-59 v 20-29)						

RI log-binomial	1.706	--	1.486	1.959	0.070	--
Zang & Yu	1.801	1.055	1.576	2.043	0.066	0.940
Modified Zang & Yu	1.783	1.045	1.565	2.017	0.065	0.919
RI Poisson (no robust SEs)	1.802	1.056	1.537	2.112	0.081	1.152
RI Poisson (robust SEs)	1.802	1.056	1.405	2.311	0.127	1.804
Shared frailty model	1.795	1.052	1.531	2.104	0.081	1.152
CLL (robust SEs)	1.796	1.052	1.298	2.485	0.166	2.353
RI CLL	1.797	1.053	1.559	2.072	0.072	1.029
Outcome: WHP (14%) Explanatory variable: Occupational activity						
RI log-binomial	1.844	--	1.608	2.116	0.070	--
Zang & Yu	1.928	1.045	1.680	2.206	0.070	0.993
Modified Zang & Yu	1.922	1.042	1.676	2.198	0.069	0.987
RI Poisson (no robust SEs)	1.856	1.007	1.601	2.152	0.075	1.076
RI Poisson (robust SEs)	1.857	1.007	1.590	2.168	0.079	1.129
Shared frailty model	1.854	1.005	1.599	2.149	0.075	1.076
CLL (robust SEs)	2.334	1.266	1.642	3.318	0.179	2.561
RI CLL	1.886	1.023	1.638	2.173	0.072	1.030
Outcome: WHP (14%) Explanatory variable: Distressing somatic symptoms (1 v 0)						
RI log-binomial	1.720	--	1.538	1.923	0.057	--
Zang & Yu	1.703	0.990	1.532	1.886	0.053	0.929
Modified Zang & Yu	1.458	0.848	1.360	1.557	0.035	0.606
RI Poisson (no robust SEs)	1.734	1.009	1.533	1.962	0.063	1.102
RI Poisson (robust SEs)	1.734	1.009	1.559	1.929	0.054	0.953
Shared frailty model	1.731	1.007	1.531	1.958	0.063	1.100
CLL (robust SEs)	1.895	1.102	1.672	2.148	0.064	1.120
RI CLL	1.702	0.990	1.529	1.894	0.055	0.957
Outcome: WHP (14%) Explanatory variable: Distressing somatic symptoms (2+ v 0)						
RI log-binomial	2.708	--	2.451	2.993	0.051	--
Zang & Yu	2.677	0.989	2.461	2.901	0.042	0.824
Modified Zang & Yu	2.326	0.859	2.175	2.477	0.033	0.652
RI Poisson (no robust SEs)	2.750	1.015	2.453	3.084	0.058	1.145
RI Poisson (robust SEs)	2.750	1.015	2.458	3.077	0.057	1.124
Shared frailty model	2.746	1.014	2.450	3.077	0.058	1.140
CLL (robust SEs)	3.418	1.262	2.839	4.115	0.095	1.858
RI CLL	2.800	1.034	2.534	3.094	0.051	1.000

Table 3. Relative risks and 95% CIs for the univariate associations between disabling ELP and risk factors

	Point Estimate		95% Confidence Intervals			
	RR	Ratio of RRs	lower	upper	SE(ln(RR))	Ratio of SEs
Outcome: ELP (6%) Explanatory variable: Sex (Female)						
RI log-binomial	1.828	--	1.476	2.263	0.109	--
Zang & Yu	1.855	1.014	1.494	2.294	0.109	1.004
Modified Zang & Yu	1.860	1.017	1.497	2.304	0.110	1.010
RI Poisson (no robust SEs)	1.826	0.999	1.463	2.278	0.113	1.036
RI Poisson (robust SEs)	1.826	0.999	1.433	2.326	0.124	1.134
Shared frailty model	1.808	0.989	1.450	2.254	0.113	1.033
CLL (robust SEs)	1.857	1.016	1.300	2.654	0.182	1.671
RI CLL	1.845	1.009	1.484	2.293	0.111	1.017
Outcome: ELP (6%) Explanatory variable: Age (30-39 v 20-29)						
RI log-binomial	1.523	--	1.159	2.001	0.139	--

Zang & Yu	1.508	0.990	1.161	1.943	0.131	0.944
Modified Zang & Yu	1.361	0.894	1.123	1.620	0.093	0.671
RI Poisson (no robust SEs)	1.530	1.005	1.157	2.023	0.143	1.023
RI Poisson (robust SEs)	1.530	1.005	1.165	2.009	0.139	0.998
Shared frailty model	1.532	1.006	1.159	2.026	0.142	1.022
CLL (robust SEs)	1.475	0.968	1.081	2.011	0.158	1.137
RI CLL	1.515	0.994	1.163	1.973	0.135	0.969
Outcome: ELP (6%) Explanatory variable: Age (40-49 v 20-29)						
RI log-binomial	3.226	--	2.494	4.174	0.131	--
Zang & Yu	3.036	0.941	2.425	3.756	0.112	0.849
Modified Zang & Yu	2.630	0.815	2.188	3.106	0.089	0.680
RI Poisson (no robust SEs)	3.251	1.008	2.494	4.237	0.135	1.029
RI Poisson (robust SEs)	3.251	1.008	2.539	4.162	0.126	0.959
Shared frailty model	3.243	1.005	2.489	4.225	0.135	1.027
CLL (robust SEs)	3.156	0.978	2.380	4.185	0.144	1.096
RI CLL	3.175	0.984	2.494	4.043	0.123	0.938
Outcome: ELP (6%) Explanatory variable: Age (50-59 v 20-29)						
RI log-binomial	3.769	--	2.869	4.952	0.139	--
Zang & Yu	3.477	0.922	2.750	4.328	0.116	0.831
Modified Zang & Yu	3.056	0.811	2.507	3.653	0.096	0.690
RI Poisson (no robust SEs)	3.787	1.005	2.853	5.027	0.144	1.037
RI Poisson (robust SEs)	3.787	1.005	2.841	5.049	0.147	1.054
Shared frailty model	3.772	1.001	2.843	5.005	0.144	1.036
CLL (robust SEs)	3.504	0.930	2.484	4.944	0.176	1.261
RI CLL	3.673	0.975	2.853	4.730	0.129	0.927
Outcome: ELP (6%) Explanatory variable: Occupational activity						
RI log-binomial	1.786	--	1.436	2.222	0.111	--
Zang & Yu	1.802	1.009	1.447	2.236	0.111	0.997
Modified Zang & Yu	1.798	1.007	1.446	2.230	0.111	0.992
RI Poisson (no robust SEs)	1.787	1.001	1.428	2.237	0.115	1.028
RI Poisson (robust SEs)	1.787	1.001	1.411	2.263	0.121	1.082
Shared frailty model	1.769	0.991	1.415	2.212	0.114	1.023
CLL (robust SEs)	2.181	1.221	1.629	2.921	0.149	1.338
RI CLL	1.792	1.004	1.438	2.233	0.112	1.008
Outcome: ELP (6%) Explanatory variable: Distressing somatic symptoms (1 v 0)						
RI log-binomial	1.926	--	1.591	2.332	0.098	--
Zang & Yu	1.902	0.987	1.579	2.281	0.094	0.962
Modified Zang & Yu	1.673	0.869	1.452	1.911	0.070	0.718
RI Poisson (no robust SEs)	1.923	0.998	1.577	2.344	0.101	1.036
RI Poisson (robust SEs)	1.923	0.998	1.581	2.339	0.100	1.024
Shared frailty model	1.914	0.994	1.570	2.332	0.101	1.035
CLL (robust SEs)	2.009	1.043	1.650	2.446	0.100	1.030
RI CLL	1.906	0.989	1.583	2.295	0.095	0.972
Outcome: ELP (6%) Explanatory variable: Distressing somatic symptoms (2+ v 0)						
RI log-binomial	3.226	--	2.713	3.837	0.088	--
Zang & Yu	3.149	0.976	2.681	3.679	0.081	0.913
Modified Zang & Yu	2.736	0.848	2.398	3.096	0.065	0.737
RI Poisson (no robust SEs)	3.224	0.999	2.687	3.868	0.093	1.051
RI Poisson (robust SEs)	3.224	0.999	2.770	3.751	0.077	0.874
Shared frailty model	3.200	0.992	2.669	3.836	0.093	1.047
CLL (robust SEs)	3.716	1.152	3.101	4.453	0.092	1.044
RI CLL	3.236	1.003	2.728	3.838	0.087	0.984

Table 4. Relative risks and 95% CIs for the associations between disabling low back pain and risk factors from the mutually adjusted models

	Point Estimate			**95% Confidence Intervals**			
	% Confounding	**RR**	**Ratio of RRs**	**lower**	**upper**	**SE(ln(RR))**	**Ratio of SEs**
Outcome: LBP (22%) Explanatory variable: Sex (Female)							
RI log-binomial	--	NC	--	NC	NC	--	--
Zang & Yu	-10.21	1.344	1.033	1.208	1.491	0.054	0.872
Modified Zang & Yu	-29.42	1.053	0.809	1.035	1.069	0.008	0.135
RI Poisson (no robust SEs)	-11.16	1.301	1.000	1.162	1.457	0.058	0.935
RI Poisson (robust SEs)	-11.16	1.301	--	1.153	1.468	0.062	--
Shared frailty model	-11.26	1.294	0.995	1.156	1.449	0.057	0.934
CLL (robust SEs)	-17.11	1.385	1.064	1.165	1.645	0.088	1.429
RI CLL	-11.98	1.302	1.000	1.178	1.439	0.051	0.828
Outcome: LBP (22%) Explanatory variable: Age (30-39 v 20-29)							
RI log-binomial	--	NC	--	NC	NC	--	--
Zang & Yu	4.56	1.272	1.026	1.141	1.413	0.055	1.075
Modified Zang & Yu	-10.80	1.044	0.842	1.025	1.061	0.009	0.174
RI Poisson (no robust SEs)	1.03	1.240	1.000	1.106	1.390	0.058	1.149
RI Poisson (robust SEs)	1.03	1.240	--	1.122	1.369	0.051	--
Shared frailty model	1.03	1.240	1.000	1.106	1.390	0.058	1.149
CLL (robust SEs)	1.65	1.262	1.018	1.128	1.412	0.057	1.132
RI CLL	0.66	1.234	0.996	1.122	1.358	0.049	0.957
Outcome: LBP (22%) Explanatory variable: Age (40-49 v 20-29)							
RI log-binomial	--	NC	--	NC	NC	--	--
Zang & Yu	8.94	1.637	1.059	1.474	1.812	0.053	0.908
Modified Zang & Yu	-24.52	1.083	0.700	1.068	1.097	0.007	0.119
RI Poisson (no robust SEs)	0.96	1.547	1.000	1.377	1.737	0.059	1.024
RI Poisson (robust SEs)	0.96	1.547	--	1.381	1.732	0.058	--
Shared frailty model	1.04	1.547	1.000	1.377	1.737	0.059	1.023
CLL (robust SEs)	2.24	1.610	1.041	1.397	1.855	0.072	1.249
RI CLL	0.37	1.543	0.998	1.403	1.698	0.049	0.839
Outcome: LBP (22%) Explanatory variable: Age (50-59 v 20-29)							
RI log-binomial	--	NC	--	NC	NC	--	--
Zang & Yu	9.11	1.620	1.059	1.434	1.821	0.061	1.030
Modified Zang & Yu	-23.64	1.082	0.707	1.063	1.098	0.008	0.136
RI Poisson (no robust SEs)	1.09	1.530	1.000	1.339	1.748	0.068	1.148
RI Poisson (robust SEs)	1.09	1.530	--	1.362	1.718	0.059	--
Shared frailty model	1.19	1.529	0.999	1.338	1.747	0.068	1.147
CLL (robust SEs)	2.71	1.556	1.017	1.352	1.792	0.072	1.215
RI CLL	0.19	1.516	0.991	1.363	1.687	0.054	0.917
Outcome: LBP (22%) Explanatory variable: Occupational activity							
RI log-binomial	--	NC	--	NC	NC	--	--
Zang & Yu	4.09	1.304	1.047	1.189	1.426	0.046	1.033
Modified Zang & Yu	-16.56	1.048	0.842	1.032	1.063	0.007	0.164
RI Poisson (no robust SEs)	0.70	1.245	1.000	1.135	1.365	0.047	1.052
RI Poisson (robust SEs)	0.70	1.245	--	1.140	1.359	0.045	--
Shared frailty model	0.55	1.243	0.999	1.134	1.364	0.047	1.051
CLL (robust SEs)	3.39	1.258	1.011	1.135	1.394	0.052	1.171
RI CLL	0.44	1.248	1.003	1.153	1.351	0.041	0.904
Outcome: LBP (22%) Explanatory variable: Distressing somatic symptoms (1 v 0)							
RI log-binomial	--	NC	--	NC	NC	--	--

Zang & Yu	2.17	1.613	1.014	1.479	1.756	0.044	0.829
Modified Zang & Yu	-16.04	1.081	0.679	1.068	1.093	0.006	0.110
RI Poisson (no robust SEs)	-2.56	1.591	1.000	1.441	1.757	0.050	0.954
RI Poisson (robust SEs)	-2.56	1.591	--	1.434	1.765	0.053	--
Shared frailty model	-2.51	1.591	1.000	1.441	1.756	0.050	0.954
CLL (robust SEs)	-4.59	1.572	0.988	1.435	1.722	0.047	0.881
RI CLL	-2.74	1.543	0.970	1.426	1.669	0.040	0.757
Outcome: LBP (22%)							
Explanatory variable: Distressing somatic symptoms (2+ v 0)							
RI log-binomial	--	NC	--	NC	NC	--	--
Zang & Yu	9.72	2.594	1.055	2.418	2.775	0.035	0.645
Modified Zang & Yu	-40.21	1.138	0.463	1.131	1.144	0.003	0.056
RI Poisson (no robust SEs)	-2.87	2.458	1.000	2.240	2.698	0.048	0.874
RI Poisson (robust SEs)	-2.87	2.458	--	2.210	2.735	0.054	--
Shared frailty model	-2.77	2.458	1.000	2.240	2.697	0.047	0.870
CLL (robust SEs)	-5.29	2.613	1.063	2.359	2.894	0.052	0.958
RI CLL	-3.54	2.436	0.991	2.261	2.624	0.038	0.698

Table 5. Relative risks and 95% CIs for the associations between disabling wrist/hand pain and risk factors from the mutually adjusted models

	Point Estimate			**95% Confidence Intervals**			
	% Confounding	**RR**	**Ratio of RRs**	**lower**	**upper**	**SE(ln(RR))**	**Ratio of SEs**
Outcome: WHP (14%)							
Explanatory variable: Sex (Female)							
RI log-binomial	-15.10	1.527	--	1.341	1.740	0.066	--
Zang & Yu	-13.55	1.648	1.079	1.430	1.894	0.072	1.079
Modified Zang & Yu	-39.37	1.147	0.751	1.108	1.182	0.016	0.245
RI Poisson (no robust SEs)	-13.99	1.555	1.018	1.346	1.797	0.074	1.109
RI Poisson (robust SEs)	-13.99	1.555	1.018	1.358	1.781	0.069	1.041
Shared frailty model	-14.05	1.542	1.010	1.336	1.780	0.073	1.102
CLL (robust SEs)	-22.75	1.855	1.214	1.453	2.368	0.125	1.875
RI CLL	-15.16	1.570	1.028	1.373	1.795	0.068	1.030
Outcome: WHP (14%)							
Explanatory variable: Age (30-39 v 20-29)							
RI log-binomial	-0.85	1.068	--	0.940	1.212	0.065	--
Zang & Yu	1.72	1.115	1.044	0.960	1.292	0.076	1.171
Modified Zang & Yu	-4.31	1.035	0.969	0.986	1.079	0.023	0.355
RI Poisson (no robust SEs)	0.17	1.094	1.025	0.948	1.264	0.073	1.135
RI Poisson (robust SEs)	0.17	1.094	1.025	0.941	1.272	0.077	1.187
Shared frailty model	0.18	1.094	1.025	0.947	1.263	0.073	1.135
CLL (robust SEs)	2.51	1.115	1.044	0.924	1.345	0.096	1.478
RI CLL	-0.26	1.086	1.017	0.954	1.236	0.066	1.019
Outcome: WHP (14%)							
Explanatory variable: Age (40-49 v 20-29)							
RI log-binomial	-2.31	1.384	--	1.222	1.567	0.063	--
Zang & Yu	5.56	1.543	1.115	1.331	1.784	0.075	1.176
Modified Zang & Yu	-20.51	1.129	0.816	1.088	1.167	0.018	0.281
RI Poisson (no robust SEs)	-1.18	1.439	1.040	1.246	1.663	0.074	1.161
RI Poisson (robust SEs)	-1.18	1.439	1.040	1.197	1.730	0.094	1.481
Shared frailty model	-1.09	1.437	1.038	1.244	1.660	0.074	1.160
CLL (robust SEs)	0.62	1.515	1.095	1.208	1.901	0.116	1.824
RI CLL	-1.48	1.435	1.037	1.263	1.630	0.065	1.025
Outcome: WHP (14%)							
Explanatory variable: Age (50-59 v 20-29)							
RI log-binomial	-6.10	1.602	--	1.405	1.827	0.067	--
Zang & Yu	10.61	1.992	1.243	1.699	2.325	0.080	1.195

Modified Zang & Yu	-33.05	1.193	0.745	1.155	1.228	0.016	0.234
RI Poisson (no robust SEs)	-1.80	1.769	1.104	1.510	2.073	0.081	1.208
RI Poisson (robust SEs)	-1.80	1.769	1.104	1.372	2.282	0.130	1.938
Shared frailty model	-1.71	1.764	1.101	1.505	2.067	0.081	1.207
CLL (robust SEs)	0.02	1.796	1.121	1.377	2.343	0.136	2.023
RI CLL	-2.63	1.750	1.092	1.527	2.005	0.069	1.036
Outcome: WHP (14%) Explanatory variable: Occupational activity							
RI log-binomial	-8.18	1.693	--	1.478	1.940	0.069	--
Zang & Yu	-5.33	1.825	1.078	1.585	2.095	0.071	1.024
Modified Zang & Yu	-39.00	1.173	0.692	1.137	1.205	0.015	0.215
RI Poisson (no robust SEs)	-6.78	1.731	1.022	1.492	2.007	0.076	1.089
RI Poisson (robust SEs)	-6.80	1.731	1.022	1.489	2.011	0.077	1.102
Shared frailty model	-6.83	1.727	1.020	1.489	2.003	0.076	1.089
CLL (robust SEs)	-16.74	1.943	1.148	1.531	2.467	0.122	1.752
RI CLL	-7.79	1.739	1.027	1.514	1.999	0.071	1.020
Outcome: WHP (14%) Explanatory variable: Distressing somatic symptoms (1 v 0)							
RI log-binomial	-5.29	1.628	--	1.457	1.821	0.057	--
Zang & Yu	-0.64	1.692	1.039	1.499	1.905	0.061	1.075
Modified Zang & Yu	-20.90	1.154	0.708	1.122	1.183	0.014	0.239
RI Poisson (no robust SEs)	-5.89	1.632	1.002	1.442	1.847	0.063	1.111
RI Poisson (robust SEs)	-5.89	1.632	1.002	1.467	1.816	0.054	0.956
Shared frailty model	-5.84	1.630	1.001	1.441	1.845	0.063	1.109
CLL (robust SEs)	-12.14	1.665	1.023	1.503	1.845	0.052	0.918
RI CLL	-6.12	1.598	0.981	1.437	1.776	0.054	0.950
Outcome: WHP (14%) Explanatory variable: Distressing somatic symptoms (2+ v 0)							
RI log-binomial	-8.38	2.482	--	2.245	2.743	0.051	--
Zang & Yu	8.41	2.902	1.170	2.614	3.214	0.053	1.032
Modified Zang & Yu	-45.34	1.271	0.512	1.252	1.289	0.008	0.147
RI Poisson (no robust SEs)	-6.71	2.566	1.034	2.286	2.879	0.059	1.152
RI Poisson (robust SEs)	-6.71	2.566	1.034	2.285	2.880	0.059	1.155
Shared frailty model	-6.63	2.564	1.033	2.286	2.876	0.059	1.147
CLL (robust SEs)	-15.86	2.876	1.159	2.504	3.303	0.071	1.384
RI CLL	-7.26	2.596	1.046	2.351	2.867	0.051	0.990

Table 6. Relative risks and 95% CIs for the associations between disabling elbow pain and risk factors from the mutually adjusted model

	Point Estimate			95% Confidence Intervals			
	% Confounding	RR	Ratio of RRs	lower	upper	SE(ln(RR))	Ratio of SEs
Outcome: ELP (6%) Explanatory variable: Sex (Female)							
RI log-binomial	-22.14	1.423	--	1.156	1.753	0.106	--
Zang & Yu	-4.23	1.725	1.027	1.385	2.143	0.111	1.008
Modified Zang & Yu	-37.52	1.123	0.669	1.078	1.162	0.019	0.172
RI Poisson (no robust SEs)	-4.85	1.700	1.012	1.358	2.128	0.115	1.036
RI Poisson (robust SEs)	-4.85	1.700	1.012	1.371	2.109	0.110	0.995
Shared frailty model	-4.38	1.691	1.007	1.353	2.114	0.114	1.030
CLL (robust SEs)	-14.68	1.861	1.108	1.468	2.359	0.121	1.094
Outcome: ELP (6%) Explanatory variable: Age (30-39 v 20-29)							
RI log-binomial	-0.21	1.520	--	1.160	1.992	0.138	--
Zang & Yu	2.06	1.539	1.012	1.172	2.012	0.138	1.000
Modified Zang & Yu	-19.14	1.101	0.724	1.040	1.151	0.026	0.188
RI Poisson (no robust SEs)	-0.07	1.529	1.006	1.158	2.019	0.142	1.028

RI Poisson (robust SEs)	-0.07	1.529	1.006	1.156	2.023	0.143	1.036
Shared frailty model	-0.22	1.529	1.006	1.158	2.018	0.142	1.027
CLL (robust SEs)	2.96	1.518	0.999	1.136	2.029	0.148	1.073
RI CLL	-0.70	1.504	0.989	1.163	1.945	0.131	0.951
Outcome: ELP (6%) Explanatory variable: Age (40-49 v 20-29)							
RI log-binomial	-2.82	3.135	--	2.435	4.037	0.129	--
Zang & Yu	7.50	3.264	1.041	2.559	4.131	0.122	0.948
Modified Zang & Yu	-53.56	1.221	0.390	1.189	1.247	0.012	0.094
RI Poisson (no robust SEs)	-1.46	3.203	1.022	2.467	4.160	0.133	1.034
RI Poisson (robust SEs)	-1.46	3.203	1.022	2.476	4.145	0.132	1.020
Shared frailty model	-1.30	3.201	1.021	2.466	4.155	0.133	1.032
CLL (robust SEs)	0.11	3.159	1.008	2.434	4.100	0.133	1.031
RI CLL	-3.25	3.072	0.980	2.437	3.873	0.118	0.916
Outcome: ELP (6%) Explanatory variable: Age (50-59 v 20-29)							
RI log-binomial	-3.02	3.656	--	2.799	4.775	0.136	--
Zang & Yu	8.40	3.769	1.031	2.912	4.827	0.129	0.946
Modified Zang & Yu	-59.51	1.238	0.339	1.207	1.261	0.011	0.082
RI Poisson (no robust SEs)	-2.63	3.688	1.009	2.788	4.877	0.143	1.046
RI Poisson (robust SEs)	-2.63	3.688	1.009	2.721	4.997	0.155	1.138
Shared frailty model	-2.43	3.681	1.007	2.784	4.866	0.142	1.045
CLL (robust SEs)	0.20	3.511	0.961	2.587	4.766	0.156	1.144
RI CLL	-4.41	3.511	0.961	2.761	4.465	0.123	0.900
Outcome: ELP (6%) Explanatory variable: Occupational activity							
RI log-binomial	-5.91	1.680	--	1.353	2.087	0.111	--
Zang & Yu	-4.23	1.725	1.027	1.385	2.143	0.111	1.008
Modified Zang & Yu	-37.52	1.123	0.669	1.078	1.162	0.019	0.172
RI Poisson (no robust SEs)	-4.85	1.700	1.012	1.358	2.128	0.115	1.036
RI Poisson (robust SEs)	-4.85	1.700	1.012	1.371	2.109	0.110	0.995
Shared frailty model	-4.38	1.691	1.007	1.353	2.114	0.114	1.030
CLL (robust SEs)	-14.68	1.861	1.108	1.468	2.359	0.121	1.094
RI CLL	-5.72	1.690	1.006	1.360	2.100	0.111	1.003
Outcome: ELP (6%) Explanatory variable: Distressing somatic symptoms (1 v 0)							
RI log-binomial	-6.18	1.807	--	1.493	2.187	0.097	--
Zang & Yu	-3.19	1.841	1.019	1.514	2.232	0.099	1.015
Modified Zang & Yu	-32.14	1.136	0.628	1.097	1.168	0.016	0.164
RI Poisson (no robust SEs)	-5.44	1.818	1.006	1.490	2.219	0.102	1.043
RI Poisson (robust SEs)	-5.44	1.818	1.006	1.480	2.234	0.105	1.078
Shared frailty model	-5.17	1.815	1.004	1.487	2.214	0.102	1.042
CLL (robust SEs)	-9.50	1.818	1.006	1.487	2.222	0.102	1.051
RI CLL	-6.47	1.782	0.986	1.484	2.140	0.093	0.958
Outcome: ELP (6%) Explanatory variable: Distressing somatic symptoms (2+ v 0)							
RI log-binomial	-6.77	3.008	--	2.528	3.578	0.089	--
Zang & Yu	0.72	3.172	1.054	2.666	3.759	0.088	0.990
Modified Zang & Yu	-55.48	1.218	0.405	1.195	1.237	0.009	0.100
RI Poisson (no robust SEs)	-5.57	3.044	1.012	2.532	3.659	0.094	1.060
RI Poisson (robust SEs)	-5.57	3.044	1.012	2.585	3.584	0.083	0.941
Shared frailty model	-5.21	3.033	1.008	2.525	3.642	0.093	1.055
CLL (robust SEs)	-11.40	3.292	1.094	2.777	3.902	0.087	0.980
RI CLL	-6.66	3.020	1.004	2.550	3.577	0.086	0.974

List of References

1. Waddell G. The Back Pain Revolution. 2nd ed. Edinburgh: Churchill Livingstone; 2004.

2. Palmer KT, Calnan M, Wainwright D, et al. Disabling musculoskeletal pain and its relation to somatization: a community-based postal survey. *Occup Med (Lond)*. 2005;55(8):612-7.

3. Palmer KT, Walker-Bone K, Griffin MJ, et al. Prevalence and occupational associations of neck pain in the British population. *Scand J Work Environ Health*. 2001;27(1):49-56.

4. Luime J, Koes B, Hendriksen I, et al. Prevalence and incidence of shoulder pain in the general population; a systematic review. *Scandinavian journal of rheumatology*. 2004;33(2):73-81.

5. Parsons S, Breen A, Foster NE, et al. Prevalence and comparative troublesomeness by age of musculoskeletal pain in different body locations. *Fam Pract*. 2007;24(4):308-16.

6. Dodd T. The prevalence of back pain in Great Britain 1996: a report on research for the Department of Health using the ONS Omnibus Survey: Stationary Office; 1996.

7. Murray CJ, Vos T, Lozano R, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet*. 2013;380(9859):2197-223.

8. Bevan S, Quadrello T, McGee R, et al. Fit for work? Musculoskeletal disorders in the European workforce. The Work Foundation, 2009.

9. Taylor K, Green N, Physio D. WHAT ARE THE PRODUCTIVITY LOSSES CAUSED BY MUSCULOSKELETAL DISORDERS?

10. Lotters F, Burdorf A, Kuiper J, et al. Model for the work-relatedness of low-back pain. *Scand J Work Environ Health*. 2003;29(6):431-40.

11. Palmer KT. Carpal tunnel syndrome: the role of occupational factors. *Best Pract Res Clin Rheumatol*. 2011;25(1):15-29.

12. Palmer KT, Harris EC, Coggon D. Compensating occupationally related tenosynovitis and epicondylitis: a literature review. *Occup Med (Lond)*. 2007;57(1):67-74.

13. Palmer KT, Smedley J. Work relatedness of chronic neck pain with physical findings--a systematic review. *Scand J Work Environ Health*. 2007;33(3):165-91.

14. Lang J, Ochsmann E, Kraus T, et al. Psychosocial work stressors as antecedents of musculoskeletal problems: A systematic review and meta-analysis of stability-adjusted longitudinal studies. *Social Science & Medicine*. 2012;75(7):1163-74.

15. Matsudaira K, Palmer KT, Reading I, et al. Prevalence and correlates of regional pain and associated disability in Japanese workers. *Occup Environ Med*. 2011;68(3):191-6.

16. Warnakulasuriya SS, Peiris-John RJ, Coggon D, et al. Musculoskeletal pain in four occupational populations in Sri Lanka. *Occup Med (Lond)*. 2012;62(4):269-72.

17. Freimann T, Coggon D, Merisalu E, et al. Risk factors for musculoskeletal pain amongst nurses in Estonia: a cross-sectional study. *BMC Musculoskeletal Disorders*. 2013;14(1):1-7.

18. Urquhart DM, Kelsall HL, Hoe VC, et al. Are psychosocial factors associated with low back pain and work absence for low back pain in an occupational cohort? *Clin J Pain*. 2013;29(12):1015-20.

19. Farioli A, Mattioli S, Quaglieri A, et al. Musculoskeletal pain in Europe: the role of personal, occupational, and social risk factors. *Scand J Work Environ Health*. 2014;40(1):36-46.

20. Solidaki E, Chatzi L, Bitsios P, et al. Work-related and psychological determinants of multisite musculoskeletal pain. *Scand J Work Environ Health*. 2010;36(1):54-61.

21. Coggon D, Ntani G, Palmer KT, et al. Patterns of multisite pain and associations with risk factors. *Pain*. 2013;154(9):1769-77.

22. Solidaki E, Chatzi L, Bitsios P, et al. Risk factors for new onset and persistence of multi-site musculoskeletal pain in a longitudinal study of workers in Crete. *Occup Environ Med*. 2013;70(1):29-34.

23. Sadeghian F, Raei M, Ntani G, et al. Predictors of incident and persistent neck/shoulder pain in Iranian workers: a cohort study. *PLoS One*. 2013;8(2):e57544.

24. Vargas-Prada S, Martinez JM, Coggon D, et al. Health beliefs, low mood, and somatizing tendency: contribution to incidence and persistence of musculoskeletal pain with and without reported disability. *Scand J Work Environ Health*. 2013;39(6):589-98.

25. Madan I, Reading I, Palmer KT, et al. Cultural differences in musculoskeletal symptoms and disability. *Int J Epidemiol*. 2008;37(5):1181-9.

26. Coggon D, Ntani G, Palmer KT, et al. Disabling musculoskeletal pain in working populations: is it the job, the person, or the culture? *Pain*. 2013;154(6):856-63.

27. Smedley J, Egger P, Cooper C, et al. Prospective cohort study of predictors of incident low back pain in nurses. *Bmj*. 1997;314(7089):1225-8.

28. Spurgeon A, Gompertz D, Harrington J. Modifiers of non-specific symptoms in occupational and environmental syndromes. *Occupational and environmental medicine*. 1996;53(6):361-6.

29. Bongers PM, de Winter CR, Kompier MA, et al. Psychosocial factors at work and musculoskeletal disease. *Scand J Work Environ Health*. 1993;19(5):297-312.

30. Coggon D, Ntani G, Harris EC, et al. Differences in risk factors for neurophysiologically confirmed carpal tunnel syndrome and illness with similar symptoms but normal median nerve function: a case–control study. *BMC musculoskeletal disorders*. 2013;14(1):240.

31. Waddell G, Feder G, Lewis M. Systematic reviews of bed rest and advice to stay active for acute low back pain. *Br J Gen Pract*. 1997;47(423):647-52.

32. Buchbinder R, Jolley D, Wyatt M. Population based intervention to change back pain beliefs and disability: three part evaluation2001 2001-06-23 07:00:00. 1516-20 p.

33. Vaughn LM, Jacquez F, Baker RC. Cultural health attributions, beliefs, and practices: Effects on healthcare and medical education. *mental*. 2009;10:11.

34. Schimmack U, Radhakrishnan P, Oishi S, et al. Culture, personality, and subjective well-being: integrating process models of life satisfaction. *J Pers Soc Psychol*. 2002;82(4):582-93.

35. Coggon D. Occupational medicine at a turning point. *Occup Environ Med*. 2005;62(5):281-3.

36. Coggon D, Ntani G, Palmer KT, et al. The CUPID (Cultural and Psychosocial Influences on Disability) study: methods of data collection and characteristics of study sample. *PLoS One*. 2012;7(7):e39820.

37. Waddell G, Newton M, Henderson I, et al. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain*. 1993;52(2):157-68.

38. Derogatis LR, Melisaratos N. The Brief Symptom Inventory: an introductory report. *Psychol Med*. 1983;13(3):595-605.

39. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30(6):473-83.

40. Schneider B. THE PEOPLE MAKE THE PLACE. *Personnel Psychology*. 1987;40(3):437-53.

41. Coggon D. Occupational medicine at a turning point. *Occupational and Environmental Medicine*. 2005;62(5):281-3.

42. Opdenakker M-C, Van Damme J, De Fraine DF, et al. The Effect of Schools and Classes on Mathematics Achievement. *School Effectiveness and School Improvement*. 2002;13(4):399-427.

43. Jones K, Johnston RJ, Pattie CJ. People, Places and Regions: Exploring the Use of Multi-Level Modelling in the Analysis of Electoral Data. *British Journal of Political Science*. 1992;22(03):343-80.

44. Arceneaux K, Nickerson DW. Modeling Certainty with Clustered Data: A Comparison of Methods. *Political Analysis*. 2009;17(2):177-90.

45. Duncan C, Jones K, Moon G. Smoking and deprivation: are there neighbourhood effects? *Soc Sci Med*. 1999;48(4):497-505.

46. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42(1):121-30.

47. Van de Vijver FJR, Van Hemert DA, Poortinga YH. Multilevel Analysis of Individuals and Cultures: Taylor & Francis; 2008.

48. Piantadosi S, Byar DP, Green SB. The ecological fallacy. *American Journal of Epidemiology*. 1988;127(5):893-904.

49. Stimson JA. Regression in Space and Time: A Statistical Essay. *American Journal of Political Science*. 1985;29(4):914-47.

50. Gonen M, Panageas KS, Larson SM. Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. *Radiology*. 2001;221(3):763-7.

51. Rosner B, Glynn RJ, Lee M-LT. Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach. *Biometrics*. 2003;59(4):1089-98.

52. Rosner B, Grove D. Use of the Mann-Whitney U-test for clustered data. *Stat Med*. 1999;18(11):1387-400.

53. Datta S, Satten GA. Rank-sum tests for clustered data. *Journal of American Statistical Association*. 2005(100):908-15.

54. Datta S, Satten GA. A signed-rank test for clustered data. *Biometrics*. 2008;64(2):501-7.

55. Rosner B, Glynn RJ, Lee ML. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*. 2006;62(1):185-92.

56. LIANG K-Y, ZEGER SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.

57. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38(4):963-74.

58. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics*. 1984;40(4):961-71.

59. Diggle PJ, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford, United Kingdom: Oxford University Press; 1994.

60. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? *Stat Med*. 2009;28(2):221-39.

61. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*. 2004;13(4):309-23.

62. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. *Stat Med*. 1996;15(16):1793-806.

63. Galbraith S, Daniel J, Vissel B. A study of clustered data and approaches to its analysis. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2010;30(32):10601-8.

64. Heo M, Leon AC. Comparison of statistical methods for analysis of clustered binary observations. *Stat Med*. 2005;24(6):911-23.

65. Yelland LN, Salter AB, Ryan P, et al. Analysis of binary outcomes from randomised trials including multiple births: when should clustering be taken into account? *Paediatric and Perinatal Epidemiology*. 2011;25(3):283-97.

66. Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling Using Stata: Taylor & Francis; 2005.

67. Biau DJ, Halm JA, Ahmadieh H, et al. Provider and center effect in multicenter randomized controlled trials of surgical specialties: an analysis on patient-level data. *Ann Surg*. 2008;247(5):892-8.

68. Bland JM. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Medical Research Methodology*. 2004;4.

69. Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of consulting and clinical psychology*. 1991;59(1):20.

70. Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. *Bmj*. 2005;330(7483):142-4.

71. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*. 1995;85(10):1378-83.

72. Chuang JH, Hripcsak G, Heitjan DF. Design and analysis of controlled trials in naturally clustered environments: implications for medical informatics. *J Am Med Inform Assoc*. 2002;9(3):230-8.

73. Blettner M, Sauerbrei W, Schlehofer B, et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol*. 1999;28(1):1-9.

74. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol*. 2003;3:21.

75. Zou GY, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Stat Methods Med Res*. 2013;22(6):661-70.

76. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*. 2004;159(7):702-6.

77. Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. *Am J Epidemiol*. 2011;174(8):984-92.

78. Bouwmeester W, Twisk JW, Kappen TH, et al. Prediction models for clustered data: Comparison of a random intercept and standard regression model. *BMC Medical Research Methodology*. 2013;13(1).

79. Denison HJ, Dodds RM, Ntani G, et al. How to get started with a systematic review in epidemiology: an introductory guide for early career researchers. *Archives of Public Health*. 2013;71(1):21-.

80. Bingenheimer JB, Raudenbush SW. Statistical and substantive inferences in public health: issues in the application of multilevel models. *Annu Rev Public Health*. 2004;25:53-77.

81. Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*. 1982;77(380):848-54.

82. Chu R, Thabane L, Ma J, et al. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC medical research methodology*. 2011;11(1):1.

83. Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. *BMC Medical Research Methodology*. 2013;13(1).

84. Barrios T, Diamond R, Imbens GW, et al. Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*. 2012;107(498):578-91.

85. Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*. 2008;62(8):752-8.

86. Donner A. A regression approach to the analysis of data arising from cluster randomization. *International Journal of Epidemiology*. 1985;14(2):322-6.

87. Kloek T. OLS Estimation in a Model Where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated. *Econometrica*. 1981;49(1):205-7.

88. Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. *Canadian Journal of Public Health*. 2001;92(2):150.

89. Lemeshow S, Letenneur L, Dartigues JF, et al. Illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the PAQUID study. *American Journal of Epidemiology*. 1998;148(3):298-306.

90. Maas CJ, Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*. 2004;46(3):427-40.

91. Park S, Lake ET. Multilevel modeling of a clustered continuous outcome: nurses' work hours and burnout. *Nurs Res*. 2005;54(6):406-13.

92. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*. 2005;2(2):152-62.

93. Bliese PD, Hanges PJ. Being Both Too Liberal and Too Conservative: The Perils of Treating Grouped Data as though They Were Independent. *Organizational Research Methods*. 2004;7(4):400-17.

94. Chuang J-H, Hripcsak G, Heitjan DF. Design and Analysis of Controlled Trials in Naturally Clustered Environments: Implications for Medical Informatics. *Journal of the American Medical Informatics Association : JAMIA*. 2002;9(3):230-8.

95. Desai M, Begg MD. A comparison of regression approaches for analyzing clustered data. *American Journal of Public Health*. 2008;98(8):1425-9.

96. Huang FL. Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education*. 2016;84(1):175-96.

97. Localio AR, Berlin JA, Ten Have TR, et al. Adjustments for center in multicenter studies: an overview. *Annals of internal medicine*. 2001;135(2):112-23.

98. Merlo J, Chaix B, Ohlsson H, et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Health*. 2006;60(4):290-7.

99. Astin AW, Denson N. Multi-campus studies of college impact: Which statistical method is appropriate? *Research in Higher Education*. 2009;50(4):354-67.

100. Cheong YF, Fotiu RP, Raudenbush SW. Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics*. 2001;26(4):411-29.

101. Dickinson LM, Basu A. Multilevel modeling and practice-based research. *The Annals of Family Medicine*. 2005;3(suppl 1):S52-S60.

102. Grieve R, Nixon R, Thompson SG, et al. Using multilevel models for assessing the variability of multinational resource use and cost data. *Health economics*. 2005;14(2):185-96.

103. Hedeker D, McMahon SD, Jason LA, et al. Analysis of clustered data in community psychology: with an example from a worksite smoking cessation project. *Am J Community Psychol*. 1994;22(5):595-615.

104. Moerbeek M, Van Breukelen GJP, Berger MPF. A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*. 2003;56(4):341-50.

105. Niehaus E, Campbell CM, Inkelas KK. HLM Behind the Curtain: Unveiling Decisions Behind the Use and Interpretation of HLM in Higher Education Research. *Research in Higher Education*. 2014;55(1):101-22.

106. Steenbergen MR, Jones BS. Modeling multilevel data structures. *american Journal of political Science*. 2002:218-37.

107. Galbraith S, Daniel JA, Vissel B. A study of clustered data and approaches to its analysis. *The journal of Neuroscience*. 2010;30(32):10601-8.

108. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology*. 2013;66(8):865-73.e4.

109. D'Errigo P, Tosti ME, Fusco D, et al. Use of hierarchical models to evaluate performance of cardiac surgery centres in the Italian CABG outcome study. *BMC Medical Research Methodology*. 2007;7.

110. Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. *Clin Trials*. 2005;2(2):163-73.

111. Mauny F, Viel JF, Handschumacher P, et al. Multilevel modelling and malaria: A new method for an old disease. *International Journal of Epidemiology*. 2004;33(6):1337-44.

112. Panageas KS, Schrag D, Russell Localio A, et al. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat Med*. 2007;26(9):2017-35.

113. Rose AJ, Backus BM, Gershman ST, et al. Predictors of aggressive therapy for nonmetastatic prostate carcinoma in Massachusetts from 1998 to 2002. *Medical Care*. 2007;45(5):440-7.

114. Scribner RA, Theall KP, Simonsen NR, et al. Misspecification of the effect of race in fixed effects models of health inequalities. *Social Science and Medicine*. 2009;69(11):1584-91.

115. Urbach DR, Austin PC. Conventional models overestimate the statistical significance of volume-outcome associations, compared with multilevel models. *Journal of Clinical Epidemiology*. 2005;58(4):391-400.

116. Xu Y, Lee CF, Cheung YB. Analyzing binary outcome data with small clusters: A simulation study. *Communications in Statistics: Simulation and Computation*. 2014;43(7):1771-82.

117. Yusuf B, Omigbodun O, Adedokun B, et al. Identifying predictors of violent behaviour among students using the conventional logistic and multilevel logistic models. *Journal of Applied Statistics*. 2011;38(5):1055-61.

118. Austin PC, Alte DA. Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently? *American heart journal*. 2003;145(1):27-35.

119. Subramanian SV, Jones K, Kaddour A, et al. Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*. 2009;38(2):342-60.

120. Bottomley C, Kirby MJ, Lindsay SW, et al. Can the buck always be passed to the highest level of clustering? *BMC Medical Research Methodology*. 2016;16(1).

121. Oltean H, Gagnier JJ. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. *BMC Medical Research Methodology*. 2015;15(1).

122. Diaz-Ordaz K, Froud R, Sheehan B, et al. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Medical Research Methodology*. 2013;13(1).

123. Goldstein H. Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika*. 1986;73(1):43-56.

124. Taylor PJ. An introduction to intraclass correlation that resolves some common confusions. *Cited on*. 2010:82.

125. Sainani K. The importance of accounting for correlated observations. *Pm r*. 2010;2(9):858-61.

126. Wendel-Vos GCW, Van Hooijdonk C, Uitenbroek D, et al. Environmental attributes related to walking and bicycling at the individual and contextual level. *Journal of Epidemiology and Community Health*. 2008;62(8):689-94.

127. Walters SJ. Therapist effects in randomised controlled trials: what to do about them. *J Clin Nurs*. 2010;19(7-8):1102-12.

128. Newman D, Newman I, Salzman J. Comparing OLS and HLM models and the questions they answer: Potential concerns for type VI errors. *Multiple Linear Regression Viewpoints*. 2010;36(1):1-8.

129. Goldstein H. Multilevel Statistical Models: Wiley; 2010.

130. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26(1):53-77.

131. Hox J. Multilevel Modeling: When and Why. In: Balderjahn I, Mathar R, Schader M, editors. Classification, Data Analysis, and Data Highways: Proceedings of the 21st Annual Conference of the Gesellschaft für Klassifikation e.V., University of Potsdam, March 12–14, 1997. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998. p. 147-54.

132. Jones K. Do multilevel models ever give different results? 2009.

133. Fan J, Gijbels I. Local polynomial modelling and its applications: monographs on statistics and applied probability 66: CRC Press; 1996.

134. Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*. 2002;1(4):223-31.

135. Snijders TAB, Bosker RJ. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling: Sage Publications; 1999.

136. Rodrıguez G, Elo I. Intra-class correlation in random-effects models for binary data. *The Stata Journal*. 2003;3(1):32-46.

137. Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*. 2002;9(4):599-620.

138. Yelland LN, Salter AB, Ryan P, et al. Analysis of binary outcomes from randomised trials including multiple births: when should clustering be taken into account? *Paediatr Perinat Epidemiol*. 2011;25(3):283-97.

139. Greenhalgh T. How to read a paper. Getting your bearings (deciding what the paper is about). *BMJ : British Medical Journal*. 1997;315(7102):243-6.

140. Vavken P, Dorotka R. A Systematic Review of Conflicting Meta-Analyses in Orthopaedic Surgery. *Clinical Orthopaedics and Related Research*. 2009;467(10):2723-35.

141. Walker E, Hernandez AV, Kattan MW. Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine*. 2008;75(6):431-9.

142. Easterbrook PJ, Gopalan R, Berlin JA, et al. Publication bias in clinical research. *The Lancet*. 1991;337(8746):867-72.

143. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat Med*. 1995;14(19):2057-79.

144. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting2010 2010-02-05 13:38:57.

145. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*. 1993;341(8842):418-22.

146. Clarke MJ, Stewart LA. Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? *Bmj*. 1994;309(6960):1007-10.

147. Lambert PC, Sutton AJ, Abrams KR, et al. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol*. 2002;55(1):86-94.

148. Simmonds MC, Higginsa JPT, Stewartb LA, et al. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials*. 2005;2(3):209-17.

149. Stewart LA, Tierney JF. To IPD or not to IPD?: Advantages and Disadvantages of Systematic Reviews Using Individual Patient Data. *Evaluation & the Health Professions*. 2002;25(1):76-97.

150. Ambrosone CB, Kropp S, Yang J, et al. Cigarette smoking, N-acetyltransferase 2 genotypes, and breast cancer risk: Pooled analysis and meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*. 2008;17(1):15-26.

151. Steinberg KK, Smith SJ, Stroup DF, et al. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *Am J Epidemiol*. 1997;145(10):917-25.

152. Valery PC, Williams G, Sleigh AC, et al. Parental occupation and Ewing's sarcoma: Pooled and meta-analysis. *International Journal of Cancer*. 2005;115(5):799-806.

153. Veglia F, Loft S, Matullo G, et al. DNA adducts and cancer risk in prospective studies: a pooled analysis and a meta-analysis. *Carcinogenesis*. 2008;29(5):932-6.

154. Gordon I, Boffetta P, Demers PA. A case study comparing a meta-analysis and a pooled analysis of studies of sinonasal cancer among wood workers. *Epidemiology*. 1998;9(5):518-24.

155. Tobias A, Saez M, Kogevinas M. Meta-analysis of results and individual patient data in epidemiological studies. *Journal of Modern Applied Statistical Methods*. 2004;3:176-85.

156. Walker-Bone K, Palmer KT, Reading I, et al. Occupation and epicondylitis: a population-based study. *Rheumatology (Oxford)*. 2012;51(2):305-10.

157. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-29.

158. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22(4):719-48.

159. Woolf B. On estimating the relation between blood group and disease. *Ann Hum Genet*. 1955;19(4):251-3.

160. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis*. 1985;27(5):335-71.

161. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-88.

162. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev*. 1987;9:1-30.

163. Riley RD, Lambert PC, Staessen JA, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med*. 2008;27(11):1870-93.

164. Riley RD, Dodd SR, Craig JV, et al. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med*. 2008;27(29):6111-36.

165. Heymsfield SB, van Mierlo CAJ, van der Knaap HCM, et al. Weight management using a meal replacement strategy: meta and pooling analysis from six studies. *International Journal of Obesity*. 2003;27(5):537-49.

166. Turner RM, Omar RZ, Yang M, et al. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2000;19(24):3417-32.

167. Black N. Why we need observational studies to evaluate the effectiveness of health care1996 1996-05-11 07:00:00. 1215-8 p.

168. Sankey SS, Weissfeld LA, Fine MJ, et al. An assessment of the use of the continuity correction for sparse data in meta-analysis. *Communications in Statistics-Simulation and Computation*. 1996;25(4):1031-56.

169. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004;23(9):1351-75.

170. Bravata DM, Olkin I. Simple pooling versus combining in meta-analysis. *Eval Health Prof*. 2001;24(2):218-30.

171. Walter SD. Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology*. 2000;53(9):931-9.

172. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. *Epidemiology*. 2015;26(4):466-72.

173. McNutt L-A, Wu C, Xue X, et al. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American journal of epidemiology*. 2003;157(10):940-3.

174. Santos CA, Fiaccone RL, Oliveira NF, et al. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. *BMC medical research methodology*. 2008;8(1):80.

175. Janani L, Mansournia MA, Nourijeylani K, et al. Statistical Issues in Estimation of Adjusted Risk Ratio in Prospective Studies. *Arch Iran Med*. 2015;18(10):713-9.

176. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Jama*. 1998;280(19):1690-1.

177. Dwivedi AK, Mallawaarachchi I, Lee S, et al. Methods for estimating relative risk in studies of common binary outcomes. *Journal of Applied Statistics*. 2014;41(3):484-500.

178. Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*. 2015;44(1):324-33.

179. WACHOLDER S. Binomial regression in GLIM: estimating risk ratios and risk differences. *American journal of epidemiology*. 1986;123(1):174-84.

180. Zocchetti C, Consonni D, Bertazzi PA. Estimation of prevalence rate ratios from cross-sectional data. *Int J Epidemiol*. 1995;24(5):1064-7.

181. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology*. 1987;125(5):761-8.

182. Lee J. Odds ratio or relative risk for cross-sectional data? *Int J Epidemiol*. 1994;23(1):201-3.

183. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology*. 2010;63(1):2-6.

184. Martuzzi M, Elliott P. Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. *Annals of Epidemiology*. 1998;8(1):52-5.

185. Nelder JA. Statistics in medical journals: some recent trends by Douglas G. Altman, Statistics in Medicine 2000; 19: 3275–3289. *Statistics in medicine*. 2001;20(14):2205-.

186. Perneger TV. Estimating the relative hazard by the ratio of logarithms of event-free proportions. *Contemporary Clinical Trials*. 2008;29(5):762-6.

187. Cummings P. Methods for estimating adjusted risk ratios. *Stata Journal*. 2009;9(2):175-96.

188. Penman AD, Johnson WD. Complementary log-log regression for the estimation of covariate-adjusted prevalence ratios in the analysis of data from cross-sectional studies. *Biometrical Journal*. 2009;51(3):433-42.

189. Cheah BC. Clustering standard errors or modeling multilevel data. *University of Columbia*. 2009:2-4.

190. Skov T, Deddens J, Petersen MR, et al. Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol*. 1998;27(1):91-5.

191. Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics*. 2005;6(1):39-44.

192. Petersen MR, Deddens JA. A comparison of two methods for estimating prevalence ratios. *BMC Medical Research Methodology*. 2008;8:9-.

193. Yelland LN, Salter AB, Ryan P. Relative risk estimation in randomized controlled trials: A comparison of methods for independent observations. *International Journal of Biostatistics*. 2011;7(1).