WORKPLACE PROJECT PORTFOLIO

Submitted in accordance with the requirements for the Masters of Biostatistics

(Biostatistics Collaboration of Australia)

Project A

Statistical model building for the Walk-to-school program, a cluster randomised controlled trial from Sydney, Australia

Project B

Hormonal contraception and smoking as risk factors for grade II or III cervical intraepithelial neoplasia in women aged 30-44 years: a case-control study in New South Wales, Australia

Huilan Xu

(SID 308074343)

The University of Sydney

School of Public Health

January 2011

PREFACE

Overview

This workplace project portfolio consists of two separate projects in which I was involved from August 2010 to December 2010. The first project involved model building for the Walk-to-School program, a cluster randomised controlled trial that was carried out by Dr Liming Wen, Research and Evaluation Manager at the Health Promotion Service, Sydney South West Area Health Service. The trial was designed to determine the efficacy of a coordinated and comprehensive Walk-to-School program as a strategy to increase walking frequency and duration on the students' journeys to and from school in the Central Sydney Area.

The second project was part of the Cervical Health Study, in which Associate Professor Freddy Sitas, Director of the Cancer Research Division and Professor Dianne O'Connell, Senior Epidemiologist and Manager, Cancer Epidemiology Research Unit, Cancer Council NSW were chief investigators. The objectives of this project focused on measuring the association between the use of hormonal contraception and smoking and the development of high grade (grade II or III) cervical intraepithelial neoplasia (CIN) in women aged 30-44 years.

Student's role

My role in the first project (project A) was to prepare the data for analysis, provide advice on how the data could be analysed, conduct the data analysis, write a report and provide interpretation of the final results. Professor Judy Simpson provided invaluable support and advice on the statistical analysis involved in the project.

For the second project (project B), my role was to conduct the data analysis and draft a manuscript for submission to a peer-reviewed journal. I was involved in each step of the data analysis for this project. Professor O'Connell and A/Professor Sitas provided invaluable advice and timely support during the course of the data analysis.

Reflections on Learning

Project A provided a good opportunity for me to review what I had learnt from the BCA courses, to develop new statistical knowledge and skills, and utilise them in practice. In order to find the optimal statistical models for the outcomes of the Walk-to-School program, I reviewed Poisson and negative binomial models and learnt new methods including zero-inflated Poisson and zero-inflated negative binomial models that were not covered in BCA courses. I also reviewed ordinal and multinominal regression models. In addition, I learnt to use many Stata commands that were new to me, such as “countfit” and “gologit2”. I now have additional skills in the use of the Stata software package.

Working on project B enabled me to work with a complex database and understand the importance of data management. The majority of my time was spent on merging different datasets, cleaning data and deriving variables for analysis (especially the definitions for cases and controls). Since this was a case-control study which was not covered extensively in BCA courses (which emphasised randomised controlled trials), I had to review the methods for case-control studies. I learnt that what was initially thought to be tedious data manipulation was actually very important and worthwhile. Once the analysis dataset was created, the process of data analysis to produce the final results was relatively straightforward. However, when conducting data analysis for epidemiological studies, statisticians should not only use appropriate statistical methods but they also need to be familiar with the context of the disease or health problem under study. Defining and understanding the risk factors and potential confounders was crucial for the data analysis and interpretation of the results. In addition, because of the matched design in this study, conditional logistic regression, which was just mentioned briefly in the BCA courses, was used to estimate odds ratios and obtain their 95% confidence intervals. I had to spend time reading about and understanding conditional logistic regression.

Both projects enhanced my understanding of the importance of the interpretation of study results. The final results should be interpreted in a clear, concise way, and over-interpretation should be avoided, otherwise misunderstandings may occur no matter how thoroughly and precisely the analysis has been conducted.

Teamwork

There was a team for project B, Professor O'Connell, A/Professor Sitas, Dr Canfell, Professor Banks, data manager Ms Luo, research assistant Ms Darlington-Brown and myself. From August to December 2010, we had regular study meetings every second week when we discussed the study design, the definition of cases and controls, and the results of the data management and analysis completed to date. At every meeting, I posed queries, and obtained advice, suggestions and feedback from the team members. Professor O'Connell also helped to keep me on the right track and to make sure that I could complete the project in time. It was a wonderful experience to work as part of a professional team.

Project A was directly supervised by the statistical supervisor Professor Judy Simpson and the project supervisor Dr Liming Wen. I met Professor Simpson regularly and discussed the statistical aspects of the project, which improved my communication skills. As a result of these meetings I sometimes needed to clarify some aspects of the data or analysis with my content supervisor, Dr Liming Wen. So Project A had a different teamwork model in which the whole team never met as a whole, but I provided the communication link.

Ethical considerations

The Walk-to-School program was approved by the ethics committees of Sydney South West Area Health Service, and the NSW Department of Education and Training.

The study materials for project B such as the questionnaire and consent form were approved by the Cancer Council NSW Ethics Committee, the NSW Department of

Health Ethics Committee and the Chief Health Officer in 2004. The study was also approved by the NSW Cancer Institute’s Ethics Committee allowing access to participants through the NSW Pap Test Register.

PROJECT A

Statistical model building for the Walk-to-school program, a cluster randomised controlled trial from Sydney, Australia

Table of contents

Project Title

Statistical model building for the Walk-to-School program, a cluster randomised controlled trial from Sydney, Australia

Location and dates

Health Promotion Service, Sydney South West Area Health Service:
August 2010 — November 2010

Context

This report is based on an enquiry from my supervisor who is the principal investigator for this study, Dr Liming Wen, regarding the optimal statistical methods for analysing the outcomes of the Walk-to-School program, a cluster randomized controlled trial. This study was conducted from October 2004 to June 2007. Summary measures were used previously to analyse the study. The percentage of children with each outcome was computed for each school and the mean percent and standard deviation for each group was then calculated. Independent t-tests were used to test for differences in mean proportions between the two study groups [1]. Obviously, there was some loss of information due to summarising the data in this way. Since the unit of analysis was the school (N=24) instead of the individual student (N=1975), and the cluster sizes were not equal, ranging from 22 to 249, statistical efficiency was reduced. So Dr Liming Wen suggested that I build appropriate models that take into account the cluster design and are suitable for the outcomes of the study.

Contribution of student

- Data manipulation and setting up an analysis dataset containing all the information necessary for the analysis.
- Background reading in statistical models for count data with excessive zeros and over-dispersion.
- Reviewing categorical data analysis and studying statistical methods that were not introduced in the BCA course.

- Conducting exploratory analyses for different models and doing model comparison.

Statistical issues involved

- Data cleaning and manipulation
- Selection of appropriate type of outcome (continuous, ordinal or nominal categorical data) based on their distribution
- Regression model building:

  Four regression models for count data were briefly described and discussed. They were Poisson, negative binomial, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB).

  Three logistic regression models for categorical outcomes were developed and compared. They were ordinal logistic regression ─ proportional odds model and partial proportional odds model ─ and multinomial logistic regression.

  Stata 10.1 was used for all data analysis.

Acknowledgements

I would like to thank Professor Judy Simpson for her kind, patient and very helpful supervision regarding analysing the data, choosing appropriate statistical methods and interpreting the results correctly; and thank Dr Liming Wen for his helpful advice about the grouping of variables and key predictors to be considered when building models.

Declaration by student

I declare that this project is my own work, with guidance provided by my project supervisor, Professor Judy Simpson, and that I have not previously submitted it for academic credit.

Signature ______________________ Date ____________________

Declaration by project supervisor

I confirm the above statements. Huilan worked diligently on this analysis and I believe she has learned a lot about analysing data that do not follow any of the standard distributions.

Signature ______________________ Date ____________________

Report

Introduction

There is evidence that children are walking less. The 2001 Household Travel Report indicated that between 1991 and 1999 there has been an increase in the share of education trips by school children as vehicle passengers (from 41% to 51%), while the share of walking trips decreased from 32% to 24% [2]. Walking to and from school may help school children to increase their daily physical activity and establish a habit of daily activity.

The Walk-to-School program was a cluster randomised controlled trial which was designed to determine the efficacy of a coordinated and comprehensive Walk-to-School program as a strategy to increase walking frequency and duration of a student's journey to or from school in the Central Sydney Area. A cluster randomised controlled trial was more appropriate than individual randomised controlled trial since this was a school-based intervention. The intervention was delivered by school teachers. In addition, students who attended the same school would share the same catchment area. The program was conducted from October 2004 to June 2007.

A total of 1,975 Year 3 and 4 students from 24 schools in the Central Sydney Area Health Service region were recruited on October 2004. The number of clusters was 24 while the cluster size ranged from 22 to 249. The interventions included student, staff and parent strategies. Control schools received nutrition and support for the area healthy canteen roll-out 'Fresh taste' (not related to journey to or from school). The outcome measures (assessed at baseline and one year follow-up) included frequency and duration of walking journey to or from school per week and to other destinations, knowledge of health benefits of walking, attitudes to walking and participation in walking and other physical activities.

The primary research question was:

Does a multi-strategic health promotion intervention increase walking frequency and duration on student’s journey to and from school?

In order to address the primary research question, the effect of the intervention on walking frequency and duration of a student’s journey to or from school need to be ascertained. So there were two outcomes of interest in the Walk-to-School program and the report for model building focused on the walking frequency (number of times in a week that a student walked to or from school).

The objective of this report is to find appropriate statistical models for walking frequency and duration of a student’s journey to or from school per week for the Walk-to-School program. The individual student is the unit of analysis and the standard errors are adjusted for the clustering.

Data description

The two main outcomes of the Walk-to-School program were walking frequency and walking duration, the duration of walking on a student’s journey to or from school during 5 weekdays. Walking frequency was measured as the number of times (out of 10) in a week that a student walked to or from school, while walking duration was measured as the total time of walking to or from school per week (5 school days a week). The main predictor was group (intervention, control). The other possible predictors were gender (boy, girl), distance to school (<1 km, ≥1 km), number of siblings (1, 2, ≥3), parent education (primary / some high school, completed high school, technical certificate / diploma, university / other tertiary degree), parent employment (employed full time, employed part time, other), number of cars (0, 1, ≥2), parent travel mode (car, other) and baseline walking frequency (0/week, 1-9/week, 10/week).

Descriptive statistics (frequency, mean and variance) of walking frequency were calculated for intervention and control groups respectively. Dot plots were used to show the distribution of walking duration for intervention and control groups.

Model building for count data

Methods

Conventionally, for count data like walking frequency during one week, Poisson regression is regarded as an appropriate approach while negative binomial regression is an approach when count data are over-dispersed (i.e., the variance is greater than the mean); when the count data contain excess zeros then zero-inflated Poisson regression could be used; zero-inflated negative binomial could be adopted when the count data contain excess zeros as well as over-dispersion.

The Stata command countfit (which supports Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models) was used to fit and compare count models [4]. To compare the four different models for walking frequency, the same variables were included in each of these models. To choose which variables to include, a forward selection process was used, where predictor variables were tested in the model in order of their unadjusted association with the outcome variable and only the predictors with $P<0.05$ were retained, except for the main predictor, intervention (group). Subsequently, the predictor variables which were not included in the model were given an extra chance to enter the final model one by one. But none of these variables were statistically significant. Since this was a randomized controlled trial, the outcome measures at baseline were not included in the models as they were balanced between the control and intervention groups.

The Stata command countfit gave a set of fit statistics for each of the four models. These included the log-likelihood, Schwarz's Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC). Likelihood ratio tests (log-likelihood difference test) were used to compare nested modes such as Poisson and negative binomial but were not used to compare models that were not nested, such as Poisson and zero-inflated Poisson. In this situation, BIC and AIC were used to compare models [5].

AIC and BIC are defined as

$$\text{AIC} = -2 \times \ln(\text{likelihood}) + 2 \times \text{k}$$

$$\text{BIC} = -2 \times \ln(\text{likelihood}) + \ln(\text{N}) \times \text{k}$$

where k = the number of covariance parameters in the model

N = number of subjects (observations).

Given two models fitted to the same data, the model with the smaller value of the information criterion is considered to be better.

The Stata command countfit also provided the Pearson goodness-of-fit test to check the overall fit as well as the quality of the fit of these four models. The Pearson chi-square statistic compares the observed distribution of the data with the distribution predicted by the model. It is calculated as [6]:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where O_i = an observed frequency

E_i = an model proposed frequency

n = the number of possible outcomes

The number of degrees of freedom (df) is n-p, where p is the number of parameters estimated by the model.

Results

Table 1 shows the characteristics of participants in the Walk-to-School program, which were used as potential predictors of the two outcomes, as well as the distribution of walking frequency at baseline by group.

Table 1 Characteristics of participants in the Walk-to-School program by group

	Control N=1002	Intervention N=973
	n (%)	n (%)
No of clusters	12	12
Mean cluster size	83.5	81.1
Range of cluster size	32 to 249	22 to 190
Gender		
boy	506 (50)	466 (48)
girl	488 (49)	496 (51)
missing	8 (1)	11 (1)
Distance to school		
< 1 km	439 (44)	447 (46)
≥ 1 km	342 (34)	375 (39)
missing	221 (22)	151 (15)
Number of sibling		
1	198 (20)	173 (18)
2	375 (37)	406 (42)
≥ 3	208 (21)	243 (25)
missing	221 (22)	151 (15)
Parent education		
primary	78 (8)	107 (11)
high school	153 (15)	183 (19)
technical certificate	187 (19)	208 (21)
university/other	349 (35)	314 (32)
missing	235 (23)	161 (16)
Parent employment		
employed full time	296 (30)	287 (29)
employed part time	200 (20)	228 (23)
other	272 (27)	297 (31)
missing	234 (23)	161 (17)
Number of cars		
0	59 (6)	53 (5)
1	371 (37)	368 (38)
≥ 2	330 (33)	376 (39)
missing	242 (24)	176 (18)
Parent travel mode		
car	329 (33)	362 (37)
other	177 (18)	153 (16)
missing	496 (49)	458 (47)
Baseline walking frequency		
0 / week	424 (42)	448 (46)
1-9 / week	449 (45)	368 (38)
10 / week	125 (12.5)	152 (15.5)
missing	4 (0.5)	5 (0.5)

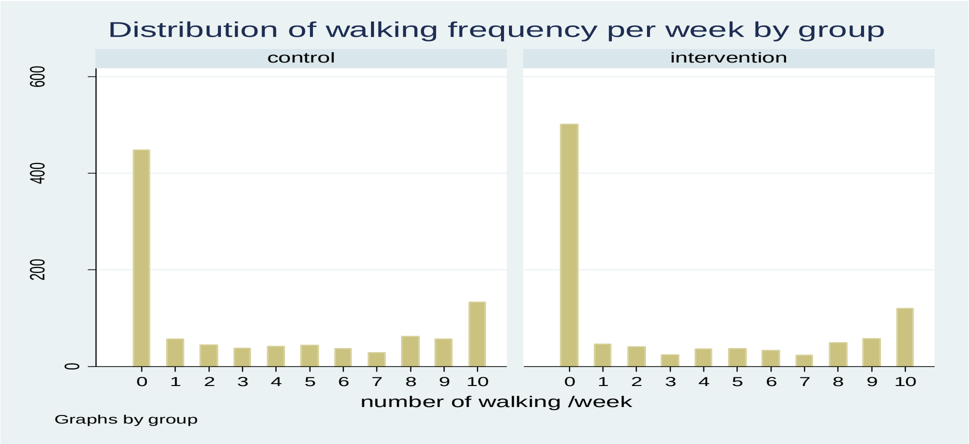
Note: The numbers of control and intervention groups were not always 1002 and 973 respectively due to missing values for some variables.

Table 2 and figure 1 display the distribution of walking frequency at the end of study by group. The number of participants in intervention and control are less than the number in Table 1 due to loss to follow-up. There are 3 features of this distribution: at one end, there is a high number of zeros in both the control (45%) and intervention (52%) groups; at the other end, there is a high proportion of tens (13% in control and 12% in intervention groups), with the other 42% in control and 36% in intervention groups distributed quite evenly between 1 and 9 (around 5% each). Apparently, the distribution is neither Poisson nor binomial.

Table 2 Descriptive statistics for walking frequency per week by group

Number of walking/week	Control N=1002	Intervention N=973
	Frequency (column %)	Frequency (column %)
0	448 (45)	501 (52)
1	57 (6)	46 (5)
2	45 (5)	41 (4)
3	38 (4)	24 (3)
4	42 (4)	36 (4)
5	44 (4)	37 (4)
6	37 (4)	33 (3)
7	29 (3)	23 (2)
8	62 (6)	49 (5)
9	57 (6)	58 (6)
10	133 (13)	120 (12)
Total	992 (100)	968 (100)
Missing	10	5

Figure 1 Distribution of walking frequency per week



The Poisson distribution is characterized by equal mean and variance. Table 3 shows that the mean walking frequency in both groups is much smaller than the variance which clearly indicates that over-dispersion exists and that the distribution is not Poisson. It also implies that zero-inflated Poisson, negative binomial regression or zero-inflated negative binomial (ZINB) might be more suitable for this outcome.

Table 3 Mean and variance of walking frequency by group

Group	**Number of walking per week**		
	Number of zero (%)	**Mean**	**Variance**
Control	448 (45)	3.44	15.38
Intervention	501 (52)	3.27	15.35

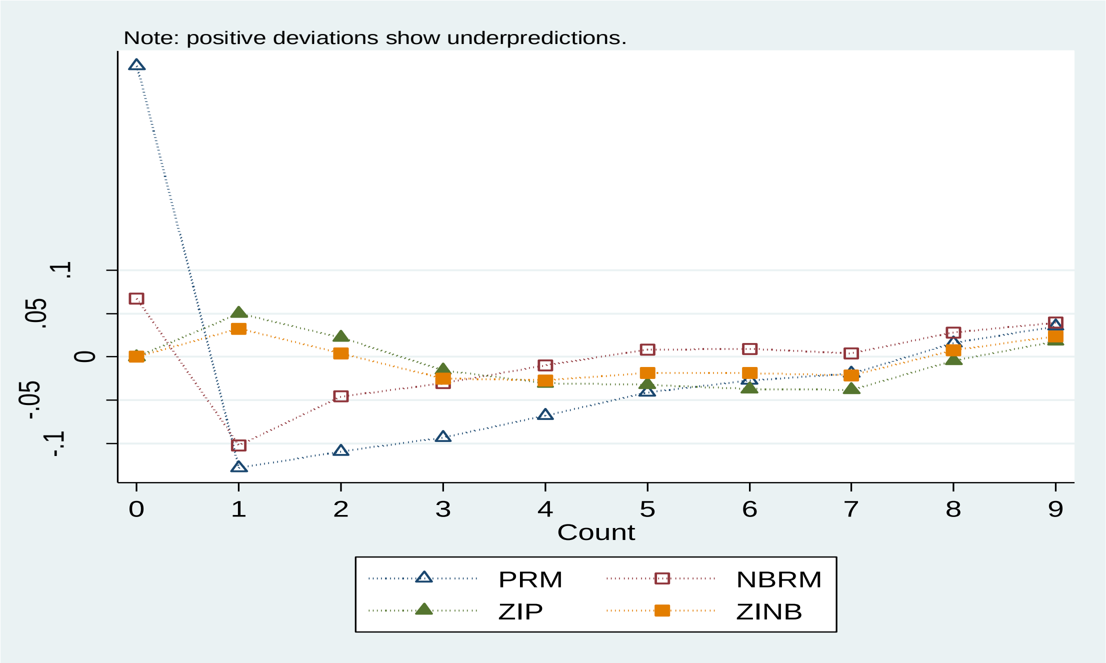
As shown in Table 4, the ZINB model has the highest log-likelihood and the lowest BIC and AIC which shows that the ZINB model gives a better fit than the other three models.

Table 4. Model fit characteristics: Log-likelihood, BIC and AIC

Model	**Log-likelihood**	**BIC**	**AIC**
Poisson	-2862.48	-1148.26	5.74
Negative binomial	-2081.98	-2702.36	4.18
Zero-inflated Poisson	-1957.48	-2937.53	3.93
Zero-inflated negative binomial	-1913.80	-3017.99	3.85

Figure 2 shows that the observed proportion minus the mean probability (i.e. predicted probability) of each count for each of the four models. It is again clear that the ZINB model gives the best fit to the data and Poisson provides the worst fit. However, the ZINB model still is not a good fit since the Pearson goodness-of-fit test shows the χ^2=99.78 with 10 df (P<0.001). In hindsight this is not really surprising, because 12% of the observed data are 10, whereas all the fitted models would predict fewer 10s than 9s. Therefore, all these four models are not suitable for walking frequency.

Figure 2 Difference between the observed proportions for each count and the mean probability from the four models



Walking frequency was therefore grouped into three categories: never walking (0/week), walking 1-9/week and walking every day (10/week). It could then be treated as ordinal categorical variable.

Model building for ordinal categorical data

Methods

Ordinal logistic regression: proportional odds model

The proportional odds model is the usual form of ordinal logistic regression provided by statistical software [6].

The models are:

$$\log(\frac{\pi_1+\pi_2}{\pi_0}) = \beta_{01}+\beta_1 group+\beta_2 distance+\beta_3 car\_1-\beta_4 car\_2+\beta_5 travelmode \quad (1)$$

$$\log(\frac{\pi_2}{\pi_0+\pi_1}) = \beta_{02}+\beta_1 group+\beta_2 distance+\beta_3 car\_1-\beta_4 car\_2+\beta_5 travelmode \quad (2)$$

where π_0, π_1 and π_2 are the probabilities of 0 /week (never walking), walking 1-9/week and walking 10/week. The intercept term β_0 depends on the way the categories are split into 2 sets (β_{01} in model (1), β_{02} in model (2)). Other betas with same subscript are same in model (1) and model (2).

One important assumption for the proportional odds model is that the effects of the covariates are the same for all splits of the categories. That is, the odds ratio of the effect of group (say) is the same for walking 1-10/week compared with never walking (0/week) as for walking every day (10/week) compared with not walking every day (0-9/week) (the coefficient of group in model (1) and (2) is the same, β_1). If this assumption is violated then proportional odds model is no longer appropriate. The Brant test of the parallel regression (proportional odds) assumption [7] can be used to test this assumption for each predictor separately.

Multinomial logistic regression model

Multinomial logistic regression can be used when the proportional odds assumption does not hold.

For walking frequency, the models are:

Walking 1-9 times/week vs. never walking (0 /week):

$$\log(\frac{\pi_1}{\pi_0}) = \beta_{01} + \beta_{11} group + \beta_{21} distance + \beta_{31} car\_1 + \beta_{41} car\_2 + \beta_{51} travelmode \quad (3)$$

Walking every day (10 /week) vs. never walking (0 /week):

$$\log(\frac{\pi_2}{\pi_0}) = \beta_{02} + \beta_{12} group + \beta_{22} distance + \beta_{32} car\_1 + \beta_{42} car\_2 + \beta_{52} travelmode \quad (4)$$

where group= $\begin{cases} 1 \text{ for intervention} \\ 0 \text{ for control} \end{cases}$, distance= $\begin{cases} 1 \text{ for } < 1 \text{ km} \\ 0 \text{ for } \geq 1 \text{ km} \end{cases}$, car_1= $\begin{cases} 1 \text{ for } 1 \text{ car} \\ 0 \text{ for other} \end{cases}$,

car_2= $\begin{cases} 1 \text{ for } \geq 2 \text{ car} \\ 0 \text{ for other} \end{cases}$, travel mode= $\begin{cases} 1 \text{ for by car} \\ 0 \text{ for other} \end{cases}$,

π_0, π_1 and π_2 are the probabilities of never walking, walking 1-9/week and walking 10/week. All the betas for each covariate in model (3) are different from those in model (4).

Compared with the ordinal model (6 parameters), the multinomial model has more parameters (12 parameters) and hence fewer degrees of freedom, so the statistical power is less than for the ordinal model. Besides, it does not take into account the order of the outcome. Therefore, another option, the partial proportional odds model will be considered.

Partial proportional odds model

The Stata command gologit2 can estimate models that are less restrictive than the proportional odds model but more parsimonious than the multinomial model [8]. In the partial proportional odds model, the effects of some covariates are the same for all categories if these covariates do not violate the proportional odds assumption, while others can differ if they violate the proportional odds assumption.

The partial proportional odds models are:

Walking 1-10 /week vs. never walking (0 /week)

$$\log(\frac{\pi_1 + \pi_2}{\pi_0}) = \beta_{01} + \beta_{11} group + \beta_2 distance + \beta_3 car\_1 + \beta_4 car\_2 + \beta_5 travelmode \quad (5)$$

Walking every day (10/ week) vs. not walking every day (0-9 /week)

$$\log(\frac{\pi_2}{\pi_0 + \pi_1}) = \beta_{02} + \beta_{12} group + \beta_2 distance + \beta_3 car\_1 + \beta_4 car\_2 + \beta_5 travelmode \quad (6)$$

Where group= $\begin{cases} 1 \text{ for intervention} \\ 0 \text{ for control} \end{cases}$, distance= $\begin{cases} 1 \text{ for} < 1 \text{ km} \\ 0 \text{ for} \geq 1 \text{ km} \end{cases}$, car_1= $\begin{cases} 1 \text{ for } 1 \text{ car} \\ 0 \text{ for other} \end{cases}$,

car_2= $\begin{cases} 1 \text{ for} \geq 2 \text{ car} \\ 0 \text{ for other} \end{cases}$, travel mode= $\begin{cases} 1 \text{ for by car} \\ 0 \text{ for other} \end{cases}$

π_0, π_1 and π_2 are the probabilities of never walking, walking 1-9/week and walking 10/week, respectively. The intercept β_0 and the coefficient for group β_1 in model (5) are different from model (6). Other betas with same subscript are same in model (5) and model (6).

To account for homogeneity within the clusters, the logistic robust cluster command was used in Stata for the proportional odds model, multinomial logistic regression model and partial proportional odds model.

Results

Walking frequency

Proportional odds model:

After conducting the Brant test of the parallel regression (proportional odds) assumption [7] for walking frequency, two important predictors, distance (P=0.036) and car (P=0.008), were found to violate the proportional odds assumption. Hence, the proportional odds model was not appropriate to analyse the grouped walking frequency.

Multinomial logistic regression model:

The results of the univariate and multivariate analysis for walking frequency are displayed in Table 5. Compared with never walking, students in the intervention group were less likely to walk 1-9 times per week (adjusted risk ratio (ARR) =0.62, 95%CI 0.42-0.92) or to walk everyday (ARR=0.77, 95%CI 0.48-1.21) than those in the control group. This means that this intervention had no effect in terms of increasing students' walking frequency to and from school.

All other predictors were associated with walking frequency (P<0.001). Compared with never walking, students who lived 1 km or more from school were less likely to walk 1-9 times per week (ARR=0.23, 95%CI 0.16-0.32) or to walk every day (ARR=0.05, 95%CI 0.03-0.10) than those who lived less than 1 km from school.

Students whose family had 2 or more cars were less likely to walk every day (ARR=0.20, 95%CI 0.08-0.53) than those whose family had no car. Overall, students whose family had a car were less likely to walk to or from school than those had no car (P<0.001).

Students whose parents did not travel to work by car were more likely to walk 1-9 times per week (ARR=1.76, 95%CI 1.38-2.25) or to walk every day (ARR=2.31, 95%CI 1.42-3.78) than those parents travelled to work by car.

Partial proportional odds model:

The results from the partial proportional odds model are quite similar to those from the multinomial logistic regression model (Table 6). Students in the intervention group were less likely to walk at all than those in the control group (adjusted odds ratio (AOR)=0.64, 95%CI 0.44-0.94). However, students in the intervention group were slightly more likely to walk every day than those in the control group (AOR=1.05, 95%CI 0.74-1.50), although the effect was not statistically significant. The overall effect of the intervention did not increase the walking frequency (P=0.03) which was similar to the result from the multinomial model.

All other predictors (distance, car ownership and parent travel mode) were associated with walking frequency (P<0.001). Students who lived 1 km or more from school were less likely to walk at all or to walk every day (OR=0.17, 95%CI 0.12-0.24) than those who lived less than 1 km away.

Students whose family had two or more cars were less likely to walk at all or to walk every day (AOR=0.40, 95%CI 0.23-0.69) than those whose family had no car. Overall, students whose family had a car were less likely to walk than those whose family had no car (P<0.001).

Students whose parents did not travel to work by car were more likely to walk at all or to walk every day (AOR=1.80, 95%CI 1.43-2.29) than those whose parents travelled to work by car.

Table 5. Risk ratios of walking frequency (multinomial logistic regression model)

Variable	Outcome Category[#]	RR*	95%CI	P	ARR**	95%CI	P
Group				0.24			0.054
control		1			1		
intervention	1-9/week	0.75	0.53-1.07		0.62	0.42-0.92	
intervention	10/week	0.81	0.46-1.42		0.77	0.48-1.21	
Distance to school				<0.001			<0.001
< 1 km		1			1		
≥ 1 km	1-9/week	0.22	0.16-0.31		0.23	0.16-0.32	
≥ 1 km	10/week	0.05	0.03-0.10		0.05	0.03-0.10	
Number of cars				<0.001			<0.001
0		1			1		
1	1-9/week	0.74	0.43-1.29		0.67	0.28-1.59	
≥2		0.45	0.26-0.78		0.52	0.25-1.09	
1	10/week	0.25	0.16-0.39		0.42	0.16-1.10	
≥2		0.10	0.05-0.17		0.20	0.08-0.53	
Parent travel mode				<0.001			<0.001
car		1			1		
Other	1-9/week	2.29	1.80-2.91		1.76	1.38-2.25	
Other	10/week	3.74	2.49-5.62		2.31	1.42-3.78	

# Reference category is 0/week

* Crude risk ratio

** Risk ratio adjusted for all other variables in the table.

Table 6 Odds ratios of walking frequency (partial proportional odds model)

Variable	Outcome Category (week)	OR*	95%CI	P	AOR	95%CI	P
Group				0.24			0.032
Control		1			1		
Intervention	1-10 vs. 0	0.80	0.55-1.16		0.64	0.44-0.94	
Intervention	10 vs. 0-9	0.80	0.55-1.16		1.05	0.74-1.50	
Distance				<0.001			<0.001
< 1 km		1			1		
≥ 1 km	1-10 vs. 0	0.17	0.12-0.24		0.17	0.12-0.24	
≥ 1 km	10 vs. 0-9	0.10	0.06-0.18		0.17	0.12-0.24	
Car				<0.001			<0.001
0		1			1		
1		0.36	0.24-0.53		0.59	0.32-1.08	
2		0.19	0.12-0.31		0.40	0.23-0.69	
Parent travel modes				<0.001			<0.001
car		1			1		
other		2.54	2.02-3.20		1.80	1.43-2.29	

* Crude odds ratio

** Odds ratio adjusted for the other variables in the table.

Duration of walking

Figure 3 displays the distribution of the duration of walking per week for the control and intervention groups. It is clear that duration of walking per week is not Normally distributed for both groups and there are many zeros in both groups, which indicates that even transformation will not work. Therefore, duration of walking was categorized into 4 groups: ≤ 30 minutes /week, > 30 minutes to 60 minutes /week, > 60 minutes to 120 minutes /week and > 120 minutes /week. Figure 4 is the bar chart of duration of walking per week by group.

Figure 3 Dotplot of duration of walking / week by group

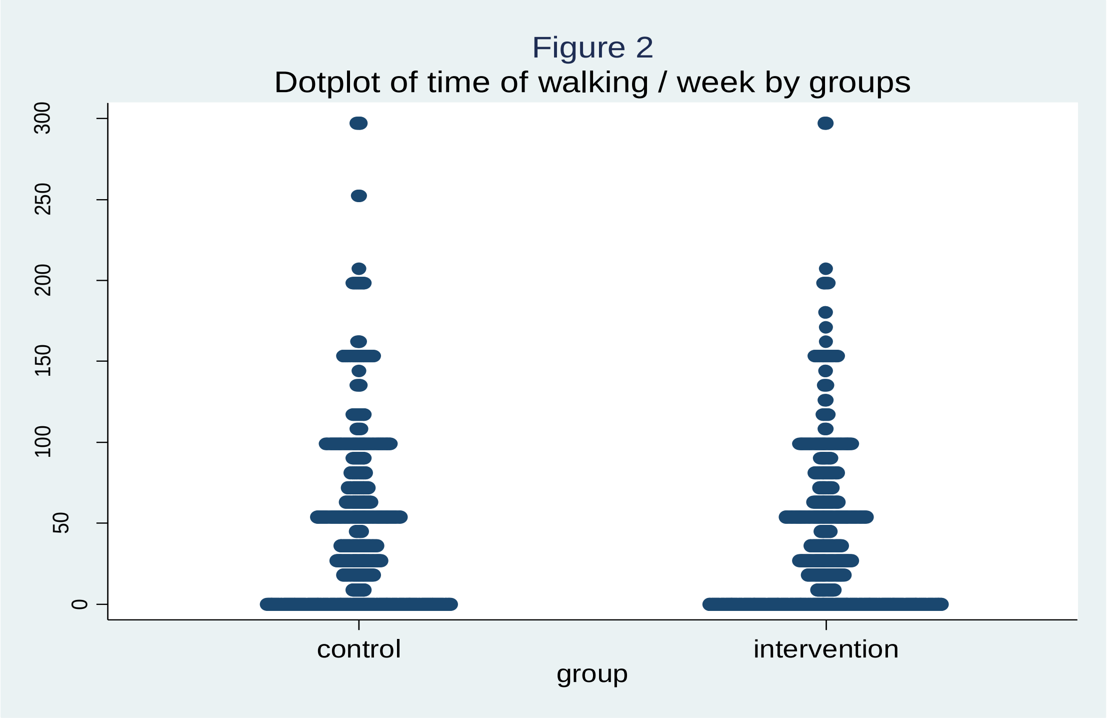
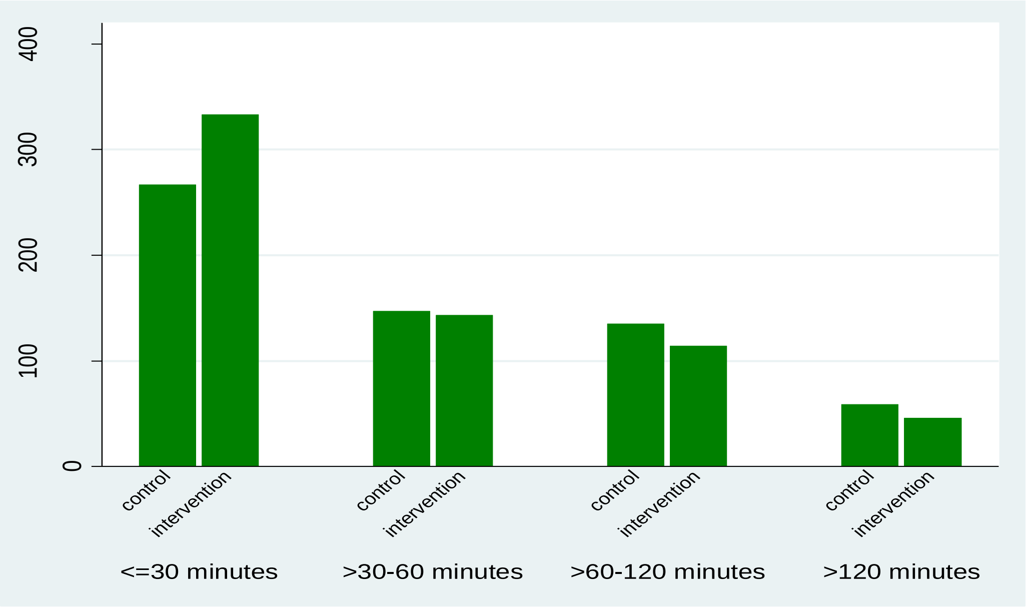


Figure 4 Bar chart of duration of walking / week by group



In multivariate analysis for duration of walking, the main predictor was group (intervention and control) and its effect was adjusted for distance to school (<1 km, ≥1 km), number of cars (0, 1, ≥2) and parent travel mode (car, other). Again, the logistic robust cluster command was used in Stata for model building.

Proportional odds model:

The Brant test showed that distance (P<0.001) violated the proportional odds assumption. Therefore, the proportional odds model is not appropriate for the duration of walking either.

Multinomial logistic regression model:

For the duration of walking to or from school per week, the intervention had no statistically significant effect overall (P=0.077) (Table 7). All other predictors were associated with walking duration (P<0.001). Students who lived 1 km and over away from school were less likely to walk longer than those who lived less than 1 km. Students whose families had one or more cars were less likely to walk longer than those whose families had no car. Students whose parents did not travel to work by car were more likely to walk longer than those whose parents travelled to work by car.

Table 7 Risk ratio of duration of walking (multinomial logistic regression model)

Variable	Outcome category#	RR*	95%CI	P	ARR	95%CI	P
Group				0.25			0.077
control		1			1		
Intervention	>30-60 minutes	0.78	0.53-1.16		0.79	0.56-1.12	
Intervention	>60-120 minutes	0.68	0.44-1.05		0.74	0.49-1.14	
Intervention	>120 minutes	0.63	0.36-1.08		0.47	0.25-0.88	
Distance to school				<0.001			<0.001
< 1 km		1			1		
≥ 1 km	>30-60 minutes	0.11	0.07-0.17		0.14	0.10-0.19	
≥ 1 km	>60-120 minutes	0.22	0.14-0.32		0.26	0.17-0.38	
≥ 1 km	>120 minutes	0.38	0.27-0.53		0.48	0.25-0.89	
Number of cars				<0.001			<0.001
0		1			1		_
1	>30-60 minutes	0.28	0.16-0.50		0.36	0.12-1.04	
≥ 2		0.18	0.10-0.32		0.25	0.09-0.70	
1	>60-120 minutes	0.23	0.12-0.42		0.32	0.11-0.90	
≥ 2		0.12	0.06-0.24		0.26	0.10-0.70	
1	>120 minutes	0.13	0.07-0.24		0.12	0.05-0.29	
≥ 2		0.04	0.02-0.09		0.08	0.03-0.22	
Parent travel mode							<0.001
car		1		<0.001	1		
other	>30-60 minutes	2.45	1.70-3.53		1.7	1.05-2.53	
other	>60-120 minutes	3.14	1.96-5.02		2.29	139-3.76	
other	>120 minutes	4.19	3.04-5.78		2.4	1.67-3.47	

# Reference category is 0/week

* Crude risk ratio

ARR: Risk ratio adjusted for the other variables in the table

Partial proportional odds model:

From the partial proportional odds model, the intervention had no effect in terms of increasing students’ duration of walking (AOR=0.71, 95%CI 0.52-0.98) (Table 8). In fact, if anything, the intervention appeared to have decreased walking duration. This effect was significant in this model since only one parameter was estimated for group, rather than three as in the multinomial model. All other predictors were associated with walking duration (P<0.001). Students who lived 1 km or more away from school were less likely to walk longer than those who lived less than 1 km away. Students whose families had one or more cars were less likely to walk longer than those whose families had no car. Students whose parents did not travel to work by car were more likely to walk longer than those whose parents travelled to work by car.

Table 8 Odds ratio of duration of walking (partial proportional odds model)

Variable	Outcome Category (minutes)	OR*	95%CI	P	OR**	95%CI	P
Group				0.046			0.036
control		1			1	_	
intervention		0.71	0.51-0.99		0.71	0.52-0.98	
Distance				<0.001			<0.001
< 1 km		1			1	_	
≥ 1 km	> 30 vs. ≤ 30	0.19	0.14-0.24		0.21	0.17-0.26	
≥ 1 km	> 60 vs. ≤ 60	0.46	0.33-0.64		0.56	0.38-0.83	
≥ 1 km	> 120 vs. ≤ 120	0.82	0.59-1.12		1.0	0.53-1.88	
Car				<0.001			<0.001
0		1			1	_	
1		0.30	0.22-0.42		0.36	0.23-0.55	
2		0.16	0.11-0.23		0.28	0.17-0.46	
Parent travel mode				<0.001			<0.001
car		1			1	_	
other		2.70	2.14-3.42		1.92	1.47-2.50	

* Crude risk ratio

** Risk ratio adjusted by variables in the table each other.

Comparisons of multinomial and partial proportional odds models

The multinomial logistic regression model and partial proportional odds model were compared in terms of goodness of fit, interpretation and parsimony. For walking frequency, the partial proportional odds model had slightly lower BIC and AIC (Table 9) which indicated that the partial proportional odds model fitted the data better.

Table 9 Model fit characteristics: BIC and AIC for walking frequency

Model	BIC	AIC
Multinomial model	1715.40	1774.30
Partial proportional odds model	1713.15	1752.41

As mentioned previously, the numbers of parameters in these two models were quite different. In the partial proportional odds model, there were 8 parameters while the multinomial model had 12 parameters. Therefore, the partial proportional odds model was more parsimonious than the multinomial model. And because it had fewer parameters, interpretation of the results of the partial proportional odds model was simpler than that of the multinomial model.

For walking duration, Table 10 shows that the partial proportional odds model was the more suitable model in terms of goodness of fit (lower BIC and AIC). In addition, the partial proportional odds model was more parsimonious and easier to interpret than the multinomial model due to fewer parameters.

Table 10 Model fit characteristics: BIC and AIC for duration of walking

Model	BIC	AIC
Multinomial model	1680.57	1764.16
Partial proportional odds model	1670.72	1717.16

Comparisons of partial proportional odds models with and without taking the clustering effect into account

Since this is a cluster randomized controlled trial, the outcome for each student is no longer independent of that for any other student. Students within one school are more likely to have similar outcomes. Therefore, the clustering effect should be taken into account when conducting the analysis. The intra-cluster correlation (ICC or) is a measure of the relatedness of clustered data. It accounts for the relatedness of clustered data by comparing the variance within clusters with the variance between clusters. It is calculated as [9]:

where is within clusters variance, is between clusters variance.

The Stata command loneway can estimate the ICC for two outcomes, using analysis of variance. The ICC for the raw data for walking frequency was 0.045 and for walking duration was 0.02 which implied that the within cluster variance was greater than the between cluster variance. Unfortunately, the partial proportional odds model used did not allow estimation of the ICC; instead clustering was adjusted for using the robust cluster (id) option in Stata to obtain robust standard errors based on the sandwich estimator.

Table 11 shows that the P values for the effect of intervention on walking frequency and walking duration were reduced from 0.03 to 0.0045 and 0.036 to 0.017 respectively if the clustering effect was not taken into account.

Table 11 Intervention effect in partial proportional odds models with and without adjusting for the clustering by school

Variable	**Outcome Category**	**SE* adjusted for ICC**			**SE not adjusted for ICC**		
		ARR**	**95%CI**	**P**	**ARR**	**95%CI**	**P**
Group	**walking frequency**			0.032			0.005
Control		1	_		1	_	
Intervention	1-10 vs. 0	0.64	0.44-0.94		0.64	0.49-0.85	
Intervention	10 vs. 0-9	1.05	0.74-1.50		1.05	0.71-1.56	
Group	**walking duration**			0.036			0.017
control		1	_		1	_	
intervention		0.71	0.52-0.98		0.71	0.54-0.94	

* SE: standard error

** ARR: Risk ratio adjusted for distance from home to school, number of cars in household and parent travel mode.

Conclusion

Based on the model building process and results above, we found that partial proportional odds models were appropriate for walking frequency and duration for the Walk-to-School program. In addition, the within-cluster correlation should be taken into account when built models since this was a cluster randomized controlled trial. The ICC for walking frequency was 0.045 and for walking duration was 0.02 which implied that the within cluster variance was greater than the between cluster variance. Although the ICCs were small, if the clustering effect was not taken into account when conducting the analysis, this would reduce the P value and narrow the confidence interval, resulting in false significant findings and misleading conclusions [10]. In this study, the P values for the effect of intervention on walking frequency and walking duration were reduced from 0.03 to 0.0045 and 0.036 to 0.017 when the clustering effect was not taken into account.

Since the distributions of walking frequency and duration were neither Poisson nor Normal, count models (Poisson regression, negative binomial regression, zero-inflated Poisson regression and zero-inflated negative binomial regression) and linear regression models were not appropriate. A reasonable approach to deal with data that were not Normally distributed (even after transformation) was to group the data into categories and treat them as ordinal categorical data.

However, standard ordinal logistic regression could not be adopted due to the proportional odds assumption not holding for some covariates.

Comparing multinomial logistic regression models and partial proportional odds models for two main outcomes in Walk-to-School, we found that the results were quite similar, but the overall goodness-of-fit showed that the partial proportional odds model was preferred. In addition, partial proportional odds models were more parsimonious and the results were easier to interpret than those of the multinomial models due to fewer parameters.

References

[1] Wen LM, et al: Increasing active travel to school: Are we on the right track? A cluster randomized controlled trial from Sydney, Australia. Preventive Medicine 2008; 47: 612-18.

[2] Transport Data Centre. Household travel survey, summary report 2001. Sydney Statistical Division, NSW Department of Transport, 2001.

[3] Jeph Herrin. CLTEST: Stata modules for performing cluster-adjusted chi-square and t-tests [Internet].2010 [Accessed 2010 Sep 1]. Available from: http://ideas.repec.org/c/boc/bocode/s424901.html

[4] Stata FAQ. How can I use countfit in choosing a count model? UCLA:Academic Technology Services, Statistical Consulting Group [Internet].2010 [Accessed 2010 Sep 5]. Available from: http://www.ats.ucla.edu/stat/stata/faq/countfit.htm

[5] Forbes A, Carlin J. Week 8: Module 4 Modelling longitudinal continuous outcomes. [unpublished lecture notes]. BSTA 5012: Longitudinal and correlated Data, Monash University & University of Melbourne.

[6] Dobson A. Week seven: Module 4 Binary Variables and Logistic Regression. [unpublished lecture notes]. BSTA 5008: Categorical Data Analysis and Generalized Linear Models, The University of Queensland.

[7] Stata FAQ. Stata Data Analysis Examples Ordinal Logistic Regression UCLA:Academic Technology Services, Statistical Consulting Group[Internet]. 2010 [Accessed 2010 Sep 6]. Available from: http://www.ats.ucla.edu/stat/stata/dae/ologit.htm

[8] Williams R. gologit2 documentation [Internet]. 2010 [cited Sep 20]. Available from: http://www.nd.edu/~rwilliam/gologit2/gologit2.pdf

[9] Campbell MJ. Cluster randomized trials in general (family) practice research. Statistical Methods in Medical Research 2000; 9: 81-94.

[10] Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement The implications of adopting a cluster design are still largely being ignored. British Medical Journal 1998; 317(31): 1171.

Appendix

The variables in the analysis dataset:

SID	Student ID number	
sid	School ID number	
group	Intervention group	0=control 1=intervention
swalkFb	Baseline walking frequency (number of times in a week that a student walked to or from school) from student survey	0=0 1=1- 9 times / week 2=10 times / week
swalkFf	Follow-up walking frequency from student survey	0=0 1=1- 9 times / week 2=10 times / week
pdwalkf	Follow-up duration of walking per week from parent survey	1= ≤ 30 minutes 2= > 30 – 60 minutes 3= > 60 – 120 minutes 4= > 120 minutes
gender	Student gender	0=boy 1=girl
dis	Distance from home to school	1= < 1 km 2= ≥ 1 km
sibling	Number of children in household	1=1 2=2 3= ≥ 3
edu	Parent’s education	1=primary / some high school 2=completed high school 3=technical certificate / diploma 4=university/other tertiary degree
emp	Parent’s employment status	1=employed full-time 2=employed part-time 3=other
car	Number of cars in household	0=0 1=1 2= ≥ 2
travelm	Parent travel mode	1=car 2=other

Stata code for modelling and model comparison:

Stata code for count models comparison

```
. countfit snwalkFf group dis car travelm, inf(dis car) noisily
```

Stata code for proportional odds model

```
. ologit snwalkFf dis, vce(cluster sid) or
. brant
. ologit snwalkFf car, vce(cluster sid) or
. brant
. ologit pdwalkf dis, vce(cluster sid) or
. brant
```

Stata code for multinomial model

```
. xi: snwalkFf i.group i.dis i.car i.travelm, vce(cluster sid) rrr
. estat ic
. xi: pdwalkf i.group i.dis i.car i.travelm, baseoutcome(1) vce(cluster sid) rrr
. estat ic
```

Stata code with gologit2

```
. xi: gologit2 snwalkFf i.group i.dis i.car i.travelm, autofit robust cluster(sid) or
. estat ic
. xi: gologit2 pdwalkf i.group i.dis i.car i.travelm, autofit robust cluster(sid) or
. estat ic
```

Stata code for estimating ICCs of walking frequency and duration of walking

```
. loneway snwalkFf sid
. loneway pdwalkf sid
```

PROJECT B

Hormonal contraception and smoking as risk factors for grade II or III cervical intraepithelial neoplasia in women aged 30-44 years: a case-control study in New South Wales, Australia

Table of contents

Number of Pap smears

Discussion

References

Tables

Appendix

Variables in analysis dataset

Stata code for data analysis

PROJECT B

Project Title

Hormonal contraception and smoking as risk factors for grade II or III cervical intraepithelial neoplasia in women aged 30-44 years: a case-control study in New South Wales, Australia

Location and dates

Cancer Epidemiology Research Unit, Cancer Council NSW:
August 2010-December 2010

Context

When I was seeking my second project for the work place portfolio, Professor Judy Simpson suggested I contact Professor Dianne O'Connell at the Cancer Epidemiology Research Unit, Cancer Council NSW to see if suitable projects were available. After meeting and talking to Professor O'Connell, the Cervical Health Study was chosen. Associate Professor Freddy Sitas and Professor O'Connell are chief investigators on the Cervical Health Study which is a nested case-control study within the cohort of women captured by the New South Wales Pap Test Register. The objectives of my project focused on measuring the association between the use of hormonal contraception and smoking and the development of high grade (grade II or III) cervical intraepithelial neoplasia (CIN) in women aged 30-44 years.

The statistical aspects of this project were performed under the supervision of Professor O'Connell, Senior Epidemiologist and Manager, Cancer Epidemiology Research Unit. A/Professor Freddy Sitas, Director of the Cancer Research Division, Cancer Council NSW, A/Professor Karen Canfell at the Cancer Research Division, Cancer Council NSW, and Professor Emily Banks, at the National Centre for Epidemiology and Population Health, The Australian National University provided epidemiological and content advice.

Contribution of student

From August to October 2010, I worked part time (2 days/week) at the Cancer Council NSW as a volunteer. I was involved in each step of the data analysis for this project, including defining cases and controls, merging different datasets, data cleaning and manipulation, creating an analysis dataset containing all the information necessary for the analysis, and conducting the statistical analysis. Also I conducted a literature review and drafted a manuscript for submission to a peer-reviewed journal.

Statistical issues involved

Since this is a case-control study, the major issues involved in this project were the definition of cases and controls, the creation of a variable reflecting the matching criteria, the definition of the exposure and confounding variables and the choice of analysis methods. As the controls were matched to cases on age group and date of Pap smear, conditional logistic regression was used.

Acknowledgements

I would like to thank Professor Dianne O’Connell for her kind, patient and very helpful supervision of the data analysis and report writing; A/Professor Freddy Sitas for allowing me to be involved in the study; A/Professor Karen Canfell and Professor Emily Banks for providing advice as the analysis progressed; and Qingwei Luo, Jessica Darlington-Brown and Sam Egger for advice and assistance.

Declaration by student

I declare that this project is my own work, with guidance provided by my project supervisor, Professor Dianne O’Connell, and that I have not previously submitted it for academic credit.

Signature ______________________ Date ____________________

Declaration by project supervisor

The declared contributions by this student are true and correct. To my knowledge, the involvement and effort of this student for this project is satisfactory for the requirements of a BCA Workplace Project.

Signature ______________________ Date ____________________

Manuscript

Hormonal contraception and smoking as risk factors for grade II or III cervical intraepithelial neoplasia in women aged 30-44 years: a case-control study in New South Wales, Australia

Abstract

Background: It has been recognised that human papillomavirus (HPV) infection is a necessary but not sufficient cause of cervical cancer. Hormonal contraception and smoking have been recognised as co-factors for the development of invasive and pre-invasive cervical cancer. Sexual behaviour and reproductive factors are also important co-factors for cervical cancer. In Australia, nearly half of women aged 35-49 years [1] are current hormonal contraceptive users and 25% of women aged 35-44 years [2] are current smokers. However, the relationship between these exposures and pre-invasive cervical cancer is unclear, and the population impact of these combined factors on the development of pre-invasive cervical cancer for young women has not been studied locally. A case-control study was conducted to measure the association in women aged 30-44 years between the use of hormonal contraception and smoking and the development of cervical intraepithelial neoplasia (CIN) grade II/III which are the lesions that precede invasive cervical cancer.

Methods: A total of 3555 women, 716 incidence cases with CIN II/III from January 2007 to February 2010, and 2839 controls without CIN II/III were selected from the NSW Pap Test Register (PTR). Cases and controls were matched by 5-year age band (30-34, 35-39, 40-44), and time (± 1 month) of index test. Conditional logistic regression was used to estimate odds ratios and 95% confidence intervals.

Results: Women who were current users of hormonal contraception were at higher risk for CIN II/III than never users (adjusted OR=1.61, 95%CI 1.02-2.53). Among current users of hormonal contraceptives the risk increased with increasing duration of use. The adjusted OR was 1.72 (95%CI 1.04-2.84) for women who were current-users with 10 to 14 years of use; while the adjusted OR was 2.04 (95%CI 1.24-3.36) for women who were current-users with 15 years or more of use. Current smoking was also significantly associated with CIN II/III (adjusted OR=1.54, 95%CI 1.19-1.99). Among women who were current smokers the risk was higher for those who smoked five or more cigarettes per day. Other risk factors associated with CIN II/III were age at first sexual intercourse and number of sexual partners in the last five years. There was no significant association between parity and CIN II/III.

Conclusions: Current hormonal contraceptive use and current smoking increase the risk of developing CIN II/III in women aged 30-44 years.

Introduction

Cervical cancer is one of the main causes of cancer mortality in women especially in those who do not have regular screening Pap smear tests. In Australia, as a result of the organised cervical screening program, the cervical cancer incidence rate of women aged 20 to 69 years decreased by approximately 50% from 1991 (the year the National Cervical Screening Program was introduced) to 2006 [3]. In New South Wales (NSW), Australia, the incidence rate of cervical cancer decreased by 25.1% and the mortality rate decreased by 21.6% between 1999 and 2008 [4]. However, despite screening there were still 739 new cases of cervical cancer in 2007 nationally [3], and in NSW, there were 248 new cases of cervical cancer and 101 deaths from cervical cancer in 2008 [4]. Also because of the high screening rate (58.1% biennial and 70% triennial) and high coverage in NSW, in 2005, there were 1106 women aged 30-34, 602 women aged 35-39 and 366 women aged 40-44 with high grade intraepithelial abnormalities (including cervical intraepithelial neoplasia (CIN) grade II or III) [5].

Current Australian cytology/histology conventions refer to CIN II/III as the lesions that precede invasive cancer. The risk of development of carcinoma in situ or worse increases 4 fold in those with mild dysplasia (approximately equivalent to CIN I), 14.5 fold in those with moderate dysplasia (~CIN II) and 46.5 fold in those with severe dysplasia (~CIN III) [6]. Although the human papillomavirus (HPV) has been recognised as the necessary cause of cervical cancer, only a small proportion of women who are infected with HPV develop cervical cancer [6], and other factors, such as hormonal contraceptive use, smoking, sexual behavior and reproductive factors are recognised as independent risk factors for the development of invasive cancer [7-22]. However, the relationship between these exposures and pre-invasive cervical cancer is unclear and some studies have provided conflicting results [23-26].

Moreover, the population impact of these factors on the development of pre-invasive cervical cancer for young women has not been studied locally. A case-control study was conducted in order to measure the association between the use of hormonal contraception and smoking and the development of CIN II/III in women of reproductive age and in whom the risk of these high grade lesions is high (i.e. those aged 30-44 years).

Subjects and methods

Setting

This case-control study was nested within the cohort of approximately 1.6 million women in the NSW Pap Test register (PTR) [27]. The PTR was established in 1996 and is a centralised database of NSW cytology results. It contains information on name, address, date of birth and cervical screening history of women who have had a Pap test, and each of their cytology and histology results.

Subjects

The study period was from January 2007 to February 2010. The initial cases were defined as women with an occurrence of CIN II/III during the study period. The date of the first abnormality was regarded as the date of entry into the study and this test was referred to as the index test. These cases were frequency-matched by 5-year age band and date of Pap smear test to three “control” women who had three consecutive normal results. Controls were selected at random from the clients meeting these criteria, except that controls with the closest age and closest test request date were favoured. For control women, the date of the test which was used to match them to the corresponding case was referred to as the index test date. Women were eligible if they were aged 30 to 44 years when they entered the study. Women with hysterectomy or oophorectomy were excluded since the cervix is generally removed and so the risk of CIN II/III is negligible.

Definition of cases and controls

Incident cases of CIN II/III were women with a CIN II/III smear cytology index test that was confirmed by a histology test within 3 months after the index test. Cases with CIN II/III cytology or positive histology within 5 years prior to the index test were excluded since they were considered to be prevalent cases.

Controls were women with a normal index smear cytology test and no CIN II/III cytology or histology test within 5 years prior to the index test.

Matching

Controls and cases were matched by 5-year age band (30-34, 35-39, 40-44) and time (± 1 month) of index test.

Data collection and measurements

The questionnaires and consent forms were mailed to women who were registered in the NSW PTR and were eligible for the study. A help line was established to help participants with queries about the study and consent and for assistance with questionnaire items. Non-respondents were followed up after two weeks by a repeat mailing.

The self-administered questionnaire sought information on demographic and relevant medical details, hormonal contraceptive use, history of smoking, alcohol consumption, reproductive and sexual history, use of hormone replacement therapy and cervical screening history. In addition, data from the Pap Test Register were used to ascertain previous frequencies of Pap tests and their results. This analysis particularly focused on the questions about hormonal contraceptive use, duration of hormonal contraceptive use, time since cessation of use of hormonal contraceptives and smoking status including duration of smoking and time since quitting smoking. Hormonal contraceptives included the combined pill, progesterone-only pill, injections, IUDs with hormones and implants. Current smokers were those who were smoking at the time of having the index Pap smear test or who had stopped smoking less than a year before the date of the index test. Parity was defined as the number of live births. Most of the questions used in the questionnaire have been used previously and validated in the UK Million Women Study [28].

Not currently smoking, having children, and having ever used oral contraceptives have been found to be associated with increased attendance for cervical screening [29]. Therefore it is important to adjust for the number of smears when assessing the potential risk factors for cervical disease. In Australia, it is recommended that cervical screening is carried out every second year; women with a smear result suggesting a low grade cervical lesion or a possible low grade squamous intraepithelial lesion (LSIL) are recommended to have a repeat cytology test at 12 months after the index smear; those aged over 30 years without a history of negative cytology in the preceding two to three years and with a low grade cervical lesion or a possible LSIL smear result are recommended to have a repeat cytology test within 6 months [30]. Hence, women with equivocal smears may have more smear results over a relatively short period of time and an increased number of smear tests overall. In order to take this into account tests conducted up to 1.5 years prior to the index test in this study were not included in the number of prior Pap tests. That is, the number of Pap smear tests was counted for the period 1.5 to 5 years prior to the index text.

Statistical analyses

All statistical analyses were performed using statistical software Stata 11. Odds ratios (ORs) and 95% confidence intervals (CIs) were estimated. Conditional logistic regression analysis with the matching variable based on age band (30-34, 35-39, 40-44) and time (± 1 month) of the index test was used to estimate ORs.

Hormonal contraception and smoking factors were the main exposures of interest, while sexual behaviours, reproductive factors and number of smears were potential confounders. Multivariable analyses of the hormonal contraception variables and smoking variables were conducted separately. Each of the effects of hormonal contraception use, duration of hormonal contraception use, time since stopping hormonal contraception, smoking status, duration of smoking and time since quitting were adjusted for parity (nulliparous, ≥1), age at first sexual intercourse (>21, 19-20, 17-18, <17 years), lifetime number of sexual partners (1-2, 3-5, 6-9, ≥10) or number of sexual partners in the last 5 years (0-1, 2, 3-5, ≥ 6) and number of Pap smears 1.5 to 5 years prior to the index cytology test. Body mass index (BMI) and a history of sexually transmitted diseases were not included in the multivariable analyses because adjustment for these factors did not change the estimated odds ratio. Also the number of sexual partners and age at first sexual intercourse which were included in the model are strongly associated with sexually transmitted diseases [31-34].

Three conditional logistic regression models were fitted in this study. The first model did not include any confounders and provided an estimate of the (matched) crude OR; the second model included smoking status, parity, age at first sexual intercourse, lifetime number of sexual partners and number of Pap smears 1.5 to 5 years prior to the index cytology test as confounders; the third model included all these confounders and the number of sexual partners in the last 5 years instead of lifetime number of sexual partners. The models for hormonal contraceptive use included smoking status (never smoked, ex-smoker, current smoker) and those for smoking included hormonal contraceptive use (never used, ex-user, current user). Subjects with missing data for any of the variables in the logistic regression models were excluded from the analysis [35]. The numbers of cases and controls included in the analysis ranged from 582 and 2304 to 716 and 2839 respectively.

Tests for trend for parity and for age at first sexual intercourse were performed by assigning an ordinal score (the median) to grouped values and then treating this score as continuous in the logistic regression models.

Results

A total of 12,202 consent forms and questionnaires were sent to potential cases and controls and 4349 consent forms and questionnaires were completed and returned. The overall response rate was 35.6%. The response rate for cases and controls were 38.4% (1371 out of 3567) and 34.5% (2978 out of 8635) respectively. Of these 4349 women, after excluding those who had any occurrence of a CIN II/III up to 5 years prior to entry into the study and those who had a hysterectomy or an oophorectomy, a total of 3555 women comprising 716 cases and 2839 controls were included in this analysis.

As shown in Table 1, the proportion of cases who were nulliparous (32%) was higher than for controls (22%); more cases had first sexual intercourse at 17-18 years of age (38%) than controls (33%); more cases had 10 or more lifetime sexual partners (36%) than controls (25%); cases also had more sexual partners in the last 5 years; and the median number of Pap smear tests for cases (1) was lower than that for controls (2). The mean age at the index test was not different in cases and controls as age was a matching variable.

Overall, the proportion of cases who were current users of hormonal contraceptives was higher than that for controls; and cases were more likely to be current smokers than controls (Table 2).

Hormonal contraceptive use

Table 3 shows the crude and adjusted odds ratios (ORs) for CIN II/III for each of the measures of hormonal contraceptive use from the three conditional logistic regression models. The results from model 2 and model 3 were quite consistent so the results from model 3 adjusting for the number of sexual partners in the last 5 years are discussed.

There were no significant differences in the risk of developing CIN II/III for ever/never users of hormonal contraceptives (model 1 p=0.29, model 2 p=0.53 and model 3 p=0.30). However, women who were current users of hormonal contraceptives were at higher risk of CIN II/III than never users (adjusted OR =1.61, 95%CI 1.02-2.53 from model 3) and previous users were at similar risk (adjusted OR=1.11, 95%CI 0.71-1.72 from model 3).

Risk of CIN II/III also varied with duration of hormonal contraceptive use. Among current users, the risk increased with increasing duration of use. From model 3, the odds ratio increased from 1.72 (95%CI 1.04-2.84) for current-users for 10-14 years to 2.04 (95%CI 1.24-3.36) for current-users for 15 years or more. Time since stopping use of hormonal contraceptives was also significantly associated with CIN II/III (p=0.003). However this was due to current users having increased odds of CIN II/III (adjusted OR=1.63, 95%CI 1.03-2.57) compared with those who had never used them. While risk of CIN II/III appeared to decrease with increasing number of years since stopping use, the odds ratios for ex-users compared to never users were not statistically significantly different from unity.

Smoking

The associations of the different measures of smoking with CIN II/III were assessed through bivariable and multivariable conditional logistic regression models (Table 4). The ORs estimated from model 3 were quite similar to those from model 2 and those from model 3 are discussed.

Overall, current smokers were at higher risk of developing CIN II/III but there was no increased risk for ex-smokers.

In multivariable analysis, the ORs estimated from model 3 indicated that women who were current smokers (adjusted OR=1.54, 95%CI 1.19-1.99), current smokers who smoked 5 cigarettes or more per day (adjusted OR=1.84, 95%CI 1.35-2.51) and current-smokers who had smoked for 10 or more years (adjusted OR=1.50, 95%CI 1.13-1.99) were at higher risk for CIN II/III than those who never smoked.

Possible effect modification

There were no significant interactions between smoking and hormonal contraceptive use when interaction terms were added to the logistic regression models (model 2: p=0.62, model 3: p=0.87).

Sexual and reproductive factors

Parity was associated with risk of CIN II/III in the bivariable (unadjusted) analysis (Table 5). However, after adjustment for potential confounders, parity was no longer associated with CIN II/III and there were no significant trends (model 2: p=0.33, model 3: p=0.57).

A trend of increasing risk of CIN II/III with decreasing age at “sexual debut” was found in the bivariable analysis (p=0.0001 for test of trend) and multivariable analysis (p=0.005 for test of trend in model 3). Women who were younger than 17 years at first sexual intercourse were at higher risk compared with those who were 21 years of age or over. After adjustment for lifetime number of sexual partners, age at first sexual intercourse was no longer associated with CIN II/III and there was no trend (p=0.50 for test of trend). This is due to the number of lifetime sexual partners increasing with younger age at first sexual intercourse.

Both lifetime number of sexual partners and number of sexual partners in the last 5 years were significantly associated with CIN II/III (p<0.0001) and the risk of CIN

II/III increased with the increasing number of sexual partners. Those with 3 or more partners in their lifetime and those who preferred not to answer were at higher risk of CIN II/III compared with those who had 1 or 2 partners. The odds ratio increased from 2.28 (95%CI 1.64-3.18) for women who had 3-5 lifetime sexual partners to 3.14 for those who had 10 or more lifetime sexual partners. Those with 2 or more partners in the last 5 years and those who preferred not to answer were also at higher risk compared with those who had 0 to 1 partner. The odds ratio increased from 1.64 (95%CI 1.20-2.25) for women who had 2 sexual partners in the last 5 years to 3.22 (95%CI 2.21-4.69) for those who had 6 or more sexual partners in the last 5 years.

Number of Pap smears

The number of Pap smears 1.5 to 5 years prior to the index test was significantly associated with CIN II/III ($p<0.0001$). The estimated ORs were almost identical from the bivariable and multivariable analyses (Table 5). With one additional Pap smear test, the odds of CIN II/III was reduced by 32% (OR=0.68, 95%CI 0.62-0.75 from model 3) reflecting the protective effect of regular screening.

Discussion

Our results were generally consistent with existing evidence. After adjusting for sexual behaviour, parity and number of Pap smears, women who were current users of hormonal contraceptives were at higher risk for CIN II/III than never users. Among current users of hormonal contraceptives the risk increased with increasing duration of use. Current smoking was also significantly associated with CIN II/III. Among current smokers the risk increased with increasing number of cigarettes smoked per day.

Previously some small studies found that oral contraceptive use was not a risk factor for CIN II/III and cervical cancer [23, 24]. A pooled analysis by the International Agency for Research on Cancer (IARC) of human papillomavirus (HPV) prevalence surveys found that oral contraceptive use was not associated with HPV prevalence, but rather might be involved in the transition from HPV infection to CIN [15].

Numerous studies have shown that smoking is an independent risk factor for CIN II/III and invasive cervical cancer [7-13]. But one cohort study found that smoking is an independent risk factor for HPV infection but not for CIN II/III [25]. A case-control study in Costa Rica also found that smoking and self-reported history of sexually transmitted diseases were not associated with invasive cervical cancer and this may be due to the low prevalence of smoking in this population [26]. It has been suggested that smoking could increase the risk of CIN II/III by increasing the risk of acquiring a cervical HPV infection [10]. Therefore, careful consideration of the confounding effect of sexual behaviour which is strongly correlated with smoking is required. Our study did consider the confounding effects of sexual behaviour that has been recognised as a proxy for HPV infection and reproductive factors. Further study is required to measure the association between smoking, hormonal use and HPV infection.

Our study also found that each of early age at first sexual intercourse and number of sexual partners was independently associated with CIN II/III. It has been recognised that sexual behaviour, including the number of sexual partners and age at first intercourse increased the risk of acquiring HPV infection [20, 22].

Some studies have found that high parity is associated with an increased risk of cervical cancer and CIN II/III [17-19]. However in our analysis parity was not associated with CIN II/III after adjusting for age at first sexual intercourse and number of sexual partners. A possible explanation is that nulliparous women were more sexually active than those with children. In this study, the median number of lifetime sexual partners for nulliparous women was 3; while for women who had 3 or more children was 2. Similarly, the median number of sexual partners in the last 5 years was 2 and 1 for nulliparous women and for women who had 3 or more children respectively.

Selecting cases with a new occurrence of CIN II/III is a strength of this study. The inclusion of prevalent cases would lead to the identification of factors associated with prolonged disease instead of factors associated with disease aetiology. Another strength is that histology tests were used to confirm the cytology test result to reduce the potential misclassification of CIN II/III. In order to minimise sampling bias the controls were selected from the same population as the cases. When the self-administrated questionnaires were sent to participants, they were not informed if they were considered to be a case or control for this study. We believe that this may have reduced the likelihood of recall bias. However, there may be some recall bias if women who were told by their doctors that the Pap test was abnormal searched their memories for possible causes more thoroughly than those with a normal result.

In Australia, it is recommended that women aged 20-69 years have a Pap test every second year. The overall biennial screening rate was 58.1%, the triennial screening rate was about 70% and 0.93% women opted off the Pap Test Register in 2005 [5]. Hence the Pap Test Register provided an ideal and representative sampling frame for this study. Therefore, the results from this study could possibly be generalised to all NSW women and indeed to women in Australia.

The main limitation of this study was the relatively low response rate (35.6%) with 38.4% of cases and 34.5% of controls participating.

In conclusion, in this case-control study, current use of hormonal contraceptives and current smoking increased the risk for CIN II/III. Among current users of hormonal contraceptives the risk increased with increasing duration of use. Among current smokers the risk increased with increasing number of cigarettes smoked per day. Therefore, from a public health perspective, for women who are current smokers quitting smoking will reduce their risk of developing CIN II/III. Current users of hormonal contraceptives should be advised to be diligent in having Pap smears in accordance with the national guidelines.

References

1. Yusuf F, Siedlecky S. Contraceptive use in Australia: evidence from the 1995 National Health Survey. Australian and New Zealand Journal of Obstetrics and Gynaecology 1999; 39(1): 58-62.

2. National Health Survey: Summary of Results 2004-2005. [Accessed 2010 November 5]; Available from: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/3B1917236618A042CA25711F00185526/$File/43640_2004-05.pdf

3. Cancer in Australia: an overview, 2010. Cancer series no. 60. Cat. no. CAN 56. Canberra: Australian Institute of Health and Welfare & Australasian Association of Cancer Registries, 2010.

4. Tracey E, Kerr T, Dobrovic A, Currow D. Cancer in New South Wales: Incidence and Mortality 2008. New South Wales Central Cancer Registry Cancer Institute NSW. [Accessed 2010 November 5]; Available from: http://www.cancerinstitute.org.au/cancer_inst/publications/CIM2008/CIM_2008_full.pdf

5. Alam N, Banks C, Chen W, Baker D, Kwaan G, Bishop J. Cervical Cancer Screening in New South Wales: Annual Statistical Report 2005. Sydney: Cancer Institute NSW, January 2008.

6. Miller AB. Cervical cancer screening programmes. Managerial guidelines. WHO, Geneva, 1992.

7. Collins S, Rollason TP, Young LS, Woodman CB. Cigarette smoking is an independent risk factor for cervical intraepithelial neoplasia in young women: a longitudinal study. European Journal of Cancer 2010; 46(2):405-11.

8. Kapeu AS, Luostarinen T, Jellum E, Dillner J, Hakama M, Koskela P et al. Is smoking an independent risk factor for invasive cervical cancer? A nested case-control study within Nordic biobanks. American Journal of Epidemiology 2009; 169(4):480-88.

9. International Collaboration of Epidemiological Studies of Cervical Cancer. Carcinoma of the cervix and tobacco smoking: collaborative reanalysis of individual data for 13.541 women with carcinoma of the cervix and 23,017 women without carcinoma of the cervix from 23 epidemiological studies. International Journal of Cancer 2006; 118: 1481-95.

10. Sarian LO, Hammes LS, Longatto-Filho A, Guarisi R, Derchain AFM, Roteli-Martins C et al. Increased risk of oncogenic human papillomavirus infections and incident high-grade cervical intraepithelial neoplasia among smokers experience from the Latin American screening study. Sexually Transmitted Diseases 2009; 36(4):241-48.

11. Ylitalo N, Sørensen P, Josefsson A, Frisch M, Sparén P, Pontén J et al. Smoking and oral contraceptives as risk factors for cervical carcinoma in situ. International Journal of Cancer 1999; 81(3):357-65.

12. Plummer M, Herrero R, Franceschi S, Meijer CJLM, Snijders P, Bosch FX et al. IARC Multi-centre Cervical Cancer Study Group. Smoking and cervical cancer: pooled analysis of the IARC multi-centric case--control study. Cancer Causes and Control 2003; 14(9):805-14.

13. Coker AL, Rosenberg AJ, McCann MF, Hulka BS. Active and passive cigarette smoke exposure and cervical intraepithelial neoplasia. Cancer Epidemiology Biomarkers and Prevention 1992; 1:349-56.

14. International Collaboration of Epidemiological Studies of Cervical Cancer. Cervical cancer and hormonal contraceptives: collaborative reanalysis of individual data for 16,573 women with cervical cancer and 35,509 women without cervical cancer from 24 epidemiological studies. Lancet 2007; 370: 1609-21.

15. Vaccarella S, Herreto R, Dai M, Snijders PJF, Meijer CJLM, Thomas JO et al. Reproductive factors, oral contraceptive use, and human papillomavirus infection: pooled analysis of the IARC HPV prevalence surveys. Cancer Epidemiology Biomarkers and Prevention 2006; 15(11): 2148-53.

16. Lacey JV, Brinton LA, Abbs FM, Barnes WA, Gravitt PE, Greenberg MD et al. Oral contraceptives as risk factors for cervical adenocarcinomas and squamous cell carcinomas. Cancer Epidemiology Biomarkers and Prevention 1999; 8: 1079-85.

17. International Collaboration of Epidemiological Studies of Cervical Cancer. Cervical carcinoma and reproductive factors: collaborative reanalysis of individual data on 16,563 women with cervical carcinoma and 33,542 women without cervical carcinoma from 25 epidemiological studies. International Journal of Cancer 2009; 119:1108-24.

18. Muñoz N, Franceschi S, Bosetti C, Moreno V, Herrero R, Smith JS et al. International Agency for Research on Cancer. Multicentric Cervical Cancer Study Group. Role of parity and human papillomavirus in cervical cancer: the IARC multicentric case-control study. Lancet 2002; 359(9312):1093-101.

19. International Collaboration of Epidemiological Studies of Cervical Cancer. Cervical carcinoma and sexual behavior: collaborative reanalysis of individual data on 15,461 women with cervical carcinoma and 29,164 women without cervical carcinoma from 21 epidemiological studies. Cancer Epidemiology Biomarkers and Prevention 2009; 18(4):1060-9.

20. Deacon JM, Evans CD, Yule R, Desai M, Binns W, Taylor C et al. Sexual behaviour and smoking as determinants of cervical HPV infection and of CIN3 among those infected: a case-control study nested within the Manchester cohort. British Journal of Cancer 2000; 83(11):1565-72.

21. Green J, Gonzalez AB, Sweetland S, Beral V, Chilvers C, Crossley B et al. Risk factors for adenocarcinoma and squamous cell carcinoma of the cervix in women aged 20-44 years: the UK National Case-Control Study of Cervical Cancer. British Journal of Cancer 2003; 89: 2078-86.

22. Moreno V, Munoz N, Bosch FX, De Sanjose S, Gonzalez LC, Tafur L et al. Risk factors for progression of cervical intraepithelial neoplasm grade III to invasive cervical cancer. Cancer Epidemiology Biomarkers and Prevention 1995; 4: 459-67.

23. Hannaford PC, Selvaraj S, Elliott AM, Angus V, Iversen L, Lee AJ. Cancer risk among users of oral contraceptives: cohort data from the Royal College of General Practitioner’s oral contraception study. British Medical Journal 2007; 335(7621): 651.

24. Cuzick J, Singer A, De Stavola BL, Chomet J. Case-control study of risk factors for cervical intraepithelial neoplasia in young women. European Journal of Cancer 1990; 26(6): 684-90.

25. Syrjänen K, Shabalova I, Petrovichev N, Kozachenko V, Zakharova T, Pajanidi J et al. Smoking is an independent risk factor for oncogenic human papillomavirus (HPV) infections but not for high-grade CIN. European Journal of Epidemiology 2007; 22(10):723-35.

26. Stone KM, Zaidi A, Rosero-Bixby L, Oberle MW, Reynolds G., Larsen S et al. Sexual behavior, sexually transmitted diseases, and risk of cervical cancer. Epidemiology 1995; 6(4): 409-14.

27. Cervical Cancer Screening in NSW: *Annual Statistical Reports 2005 Factsheet*. [Accessed 2010 November 6]; Available from: http://www.cancerinstitute.org.au/cancer_inst/publications/pdfs/sf-2008-02_ csp-annual-report-2005-factsheet.pdf

28. The Million Women Study. Questionnaires. [Accessed 2010 November 5]; Available from: http://www.millionwomenstudy.org/questionnaires/

29. Canfell K, Banks E. Oral contraceptives, hormone replacement therapy, and cancers of the female reproductive system. Chapter in: When Cancer Crosses Disciplines. eds. Robotin M, Olver I, Girgis A, Imperial College Press; 2009.

30. Screening to Prevent Cervical Cancer: Guidelines for the Management of Asymptomatic Women with Screen Detected Abnormalities. [Accessed 2010 November 6]; Available from: http://www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/wh39.pdf

31. Michael RT, Wadsworth J, Feinleib J, Johnson AM, Laumann EO, Wellings K. Private Sexual Behavior, Public Opinion, and Public Health Policy Related toSexually Transmitted Diseases: A US-British Comparison. American Journal of Public Health 1998; 88(5):749-54.

32. Aral SO, Soskoline V, Joesoef RM, O'Reilly KR. Sex partner recruitment as risk factor for STD: Clustering of risky models. Sexually Transmitted Diseases 1991; 18(1):10-17.

33. Greenberg J, Magder L, Aral S. Age at first coitus: A marker for risky sexual behaviour in women. Sexually Transmitted Diseases 1992; 19(6):331-34.

34. Kenney JW, Reinholt C, Angelini PJ. Sexual Abuse, Sex Before Age **16,** and High-Risk Behaviors of Young Females with Sexually Transmitted Diseases. JOGNN 1998; 27(1): 54-63.

35. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association 1996; 91(433): 222-30.

Tables

Table 1 Characteristics of cases with CIN* II/III (high grade cervical lesions) and controls in NSW

Variable	Cases(N=716) n (%)	Controls (N=2839) n (%)
Age groups		
30-34	305 (43)	1208 (43)
35-39	267 (37)	968 (34)
40-44	144 (20)	663 (23)
Age mean, SD (years)	35.75, 3.96	35.98, 4.13
Parity		
0	215 (32)	591 (22)
1	130 (19)	500 (19)
2	205 (31)	974 (37)
≥ 3	122 (18)	597 (22)
Age at first sexual intercourse (years)		
≥ 21	84 (12)	519 (19)
19-20	101 (15)	433 (16)
17-18	261 (38)	901 (33)
< 17	238 (35)	860 (32)
Lifetime no. sexual partners		
1-2	78 (11)	793 (28)
3-5	150 (21)	592 (21)
6-9	138 (20)	430 (15)
≥ 10	256 (36)	686 (25)
prefer not answer	84 (12)	305 (11)
No. sexual partners in the last 5 years		
0-1	399 (56)	2188 (77)
2	77 (11)	222 (8)
3-5	134 (19)	243 (9)
≥ 6	72 (10)	108 (4)
prefer not answer	30 (4)	60 (2)
	Median	**Median**
No. Pap smears 1.5 to 5 years prior to index test	1	2

* CIN: cervical intraepithelial neoplasia

Note: The numbers of cases and controls were not always 716 and 2839 respectively due to missing values for some variables: the lowest were 672 and 2662 for parity.

Table 2 Hormonal contraceptive use and smoking behaviour of cases with CIN* II/III (high grade cervical lesions) and controls in NSW

Variable	Cases(N=716) n (%)	Controls (N=2839) n (%)
Ever HC**		
never used	38 (5)	180 (6)
ever used	678 (95)	2659 (94)
HC using status		
never used	38 (5)	180 (6)
ex-user	413 (58)	1858 (66)
current user	265 (37)	801 (28)
Duration of using HC (years)		
never used	38 (5)	180 (6)
ex-user < 10	198 (28)	964 (34)
ex-user 10 - 14	125 (18)	575 (21)
ex-user ≥ 15	83 (12)	300 (11)
current-user < 10	49 (7)	218 (8)
current-user 10 - 14	103 (15)	307 (11)
current-user ≥ 15	110 (15)	264 (9)
Time since stopping (years)		
never used	38 (6)	180 (7)
≥ 10	103 (15)	496 (19)
5 - 9	98 (15)	491 (19)
1 - 4	162 (24)	638 (24)
current user	265 (40)	801 (31)
Smoking status		
never smoked	329 (47)	1619 (57)
ex-smoker	208 (29)	822 (29)
current smoker	169 (24)	381 (14)
Number of cigarettes / day		
never smoked	329 (47)	1619 (57)
ex-smoker < 5	98 (14)	385 (14)
ex-user ≥ 5	109 (15)	435 (15)
current-smoker < 5	50 (7)	141 (5)
current-smoker ≥ 5	119 (17)	240 (9)
Smoking duration (years)		
never smoked	329 (47)	1619 (57)
ex-smoker < 10	98 (14)	409 (14)
ex-smoker ≥ 10	110 (16)	410 (15)
current-smoker < 10	22 (3)	47 (2)
current-smoker ≥ 10	147 (21)	333 (12)
Time since quitting (years)		
never smoked	329 (47)	1619 (58)
≥ 10	67 (9)	309 (11)
5-9	62 (9)	241 (9)
1-4	75 (11)	255 (9)
current smoker	169 (24)	381 (13)

* CIN: cervical intraepithelial neoplasia

**HC: hormonal contraceptives

Note: The numbers of cases and controls were not always 716 and 2839 respectively due to missing values for some variables: the lowest were 666 and 2606 for time since stopping (years).

Table 3 Association between high grade cervical lesions (CIN* II/III) and hormonal contraceptive use

Variable	OR[1]	95%CI	AOR[2]	95%CI	AOR[3]	95%CI
Ever HC**						
never used	1		1		1	
ever used	1.21	0.84-1.74	1.15	0.74-1.78	1.26	0.81-1.94
	P=0.29	n=3555	P=0.53	n=3135	P=0.30	n=3156
HC using status						
never used	1		1		1	
ex-user	1.05	0.73-1.53	1.00	0.64-1.56	1.11	0.71-1.72
current-user	1.59	1.08-2.33	1.50	0.95-2.37	1.61	1.02-2.53
	P<0.0001	n=3555	P=0.0004	n=3135	P=0.0008	n=3156
Duration of use (years)						
never used	1		1		1	
ex-user < 10	0.98	0.66-1.44	0.97	0.61-1.53	1.02	0.65-1.62
ex-user 10 - 14	1.04	0.70-1.56	1.06	0.65-1.70	1.20	0.74-1.93
ex-use r ≥ 15	1.29	0.84-1.99	1.11	0.66-1.84	1.35	0.81-2.25
current-user < 10	1.05	0.65-1.69	0.98	0.56-1.72	1.00	0.58-1.76
current-user 10 - 14	1.60	1.05-2.45	1.61	0.97-2.66	1.72	1.04-2.84
current-user ≥ 15	1.96	1.29-2.98	1.85	1.13-3.05	2.04	1.24-3.36
	P<0.0001	n=3514	P=0.0003	n=3100	P=0.0001	n=3122
Time since stopping (years)						
never used	1		1		1	
≥ 10	1.00	0.66-1.52	0.87	0.53-1.42	0.97	0.60-1.59
5 - 9	0.94	0.62-1.42	0.96	0.59-1.57	1.08	0.66-1.76
1 - 4	1.22	0.82-1.81	1.19	0.74-1.90	1.34	0.84-2.14
current user	1.61	1.10-2.36	1.51	0.96-2.38	1.63	1.03-2.57
	P=0.0001	n=3272	P=0.001	n=2886	P=0.003	n=2906

OR: odds ratio AOR: adjusted odds ratio

CI: confidence interval

* CIN: cervical intraepithelial neoplasia

**HC: hormonal contraceptives

1. Crude odds ratio
2. Odds ratio adjusted for smoking status (never smoked, ex-smoker, current smoker), parity (nulliparous, ≥1), age at first sexual intercourse (≥21, 19-20, 17-18, <17), lifetime number of sexual partners (1-2, 3-5, 6-9, ≥10) and number of Pap smears in 1.5 to 5 years prior to index test.
3. Odds ratio adjusted for smoking status (never smoked, ex-smoker, current smoker), parity (nulliparous, ≥1), age at first sexual intercourse (≥21, 19-20, 17-18, <17), number of sexual partners in last 5 years (0-1, 2, 3-5, ≥ 6) and number of Pap smears in 1.5 to 5 years prior to index test.

Table 4 Association between high grade cervical lesions (CIN* II/III) and smoking behaviour

Variable	OR[1]	95%CI	AOR[2]	95%CI	AOR[3]	95%CI
Smoking status						
never smoked	1		1		1	
ex-smoker	1.25	1.03-1.52	0.97	0.77-1.21	1.08	0.87-1.35
current smoker	2.18	1.75-2.72	1.56	1.21-2.02	1.54	1.19-1.99
	P<0.0001	n=3528	P=0.0006	n=3135	P=0.004	n=3156
Number of cigarettes /day						
never smoked	1		1		1	
ex-smoker < 5	1.26	0.98-1.63	0.95	0.72-1.26	1.05	0.79-1.39
ex-smoker ≥ 5	1.32	1.04-1.66	1.07	0.82-1.40	1.18	0.90-1.54
current-smoker < 5	1.61	1.13-2.31	1.05	0.70-1.57	1.01	0.67-1.53
current-smoker ≥ 5	2.53	1.94-3.30	1.84	1.35-2.49	1.84	1.35-2.51
	P<0.0001	n=3525	P=0.002	n=3133	P=0.003	n=3153
Duration (years)						
never smoked	1		1		1	
ex-smoker < 10	1.18	0.91-1.51	0.96	0.73-1.27	1.06	0.80-1.39
ex-smoker ≥ 10	1.42	1.12-1.79	1.07	0.82-1.40	1.19	0.91-1.56
current-smoker < 10	2.27	1.32-3.89	1.51	0.82-2.76	1.51	0.82-2.78
current-smoker ≥ 10	2.15	1.69-2.74	1.51	1.14-2.00	1.50	1.13-1.99
	P<0.0001	n=3524	P=0.03	n=3131	P=0.06	n=3152
Time since quitting (years)						
never smoked	1		1		1	
≥ 10	1.09	0.81-1.47	0.90	0.65-1.24	0.99	0.72-1.37
5-9	1.25	0.92-1.70	1.02	0.72-1.43	1.21	0.86-1.69
1-4	1.45	1.09-1.94	1.02	0.74-1.41	1.10	0.79-1.52
current smoker	2.18	1.75-2.72	1.57	1.21-2.03	1.54	1.19-2.00
	P<0.0001	n=3507	P=0.004	n=3114	P=0.02	n=3135

OR: odds ratio AOR: adjusted odds ratio

CI: confidence interval

* CIN: cervical intraepithelial neoplasia

1. Crude odds ratio
2. Odds ratio adjusted for HC use (never used, ex-user, current user), parity (nulliparous, ≥1), age at first sexual intercourse (≥21, 19-20, 17-18, <17), lifetime number of sexual partners (1-2, 3-5, 6-9, ≥10) and number of Pap smears in 1.5 to 5 years prior to index test.
3. Odds ratio adjusted for HC use (never used, ex-user, current user), parity (nulliparous, ≥1), age at first sexual intercourse (≥21, 19-20, 17-18, <17), number of sexual partners in last 5 years (0-1, 2, 3-5, ≥ 6) and number of Pap smears in 1.5 to 5 years prior to index test.

Table 5 Association between high grade cervical lesions (CIN* II/III) and sexual behaviour and parity

Variable	OR[1]	95%CI	AOR[2]	95%CI	AOR[3]	95%CI
Parity						
0	1		1		1	
1	0.76	0.59-0.99	0.93	0.70-1.23	1.04	0.78-1.39
2	0.61	0.48-0.77	0.83	0.64-1.08	0.95	0.73-1.25
≥ 3	0.60	0.46-0.79	0.78	0.57-1.05	0.85	0.62-1.15
	P=0.0001#	n=3334	P=0.33#	n=3135	P=0.57#	n=3156
Age at first sexual intercourse (years)						
≥ 21	1		1		1	
19-20	1.45	1.06-2.00	1.15	0.80-1.66	1.42	0.99-2.03
17-18	1.86	1.42-2.45	1.26	0.91-1.74	1.75	1.28-2.38
< 17	1.77	1.35-2.33	1.13	0.80-1.60	1.60	1.16-2.21
	P<0.0001#	n=3397	P=0.50#	n=3135	P=0.005#	n=3156
No. smears 1.5 to 5 years prior to index test	0.69	0.63-0.75	0.69	0.62-0.76	0.68	0.62-0.75
	P<0.0001	n=3555	P<0.0001	n=3135	P<0.0001	n=3156
Lifetime no. sexual partners						
1-2	1		1		_	_
3-5	2.58	1.92-3.46	2.28	1.64-3.18	_	_
6-9	3.28	2.42-4.44	2.97	2.10-4.21	_	_
≥ 10	3.75	2.85-4.94	3.14	2.24-4.40	_	_
prefer not answer	2.69	1.92-3.77	2.30	1.52-3.46	_	_
	P<0.0001	n=3512	P<0.0001	n=3135	_	_
No. sexual partners in the last 5 years						
0-1	1		1		1	
2	1.86	1.41-2.47	_	_	1.64	1.20-2.25
3-5	2.99	2.35-3.81	_	_	2.53	1.91-3.34
≥ 6	3.68	2.66-5.10	_	_	3.22	2.21-4.69
prefer not answer	2.73	1.72-4.33	_	_	2.30	1.20-4.41
	P<0.0001	n=3533	_	_	P<0.0001	n=3156

OR: odds ratio AOR: adjusted odds ratio CI: confidence interval

* CIN: cervical intraepithelial neoplasia # p value for test of trend

1. Crude odds ratio

2. Odds ratio adjusted for HC use (never used, ex-user, current user), smoking status (never smoked, ex-smoker, current smoker), parity (nulliparous,1, 2, ≥3), age at first sexual intercourse (≥21, 19-20, 17-18, <17), lifetime number of sexual partners (1-2, 3-5, 6-9, ≥10) and number Pap smears in 1.5 to 5 years prior to index test.

3. Odds ratio adjusted for HC use (never used, ex-user, current user), smoking status (never smoked, ex-smoker, current smoker), parity (nulliparous, 1, 2, ≥3), age at first sexual intercourse (≥21, 19-20, 17-18, <17), number of sexual partners in last 5 years (0-1, 2, 3-5, ≥ 6) and number Pap smears in 1.5 to 5 years prior to index test.

Appendix

Variables in analysis dataset

cc	case or control	0=control
		1= case
ehc	ever using hormonal contraception	0=never use
		1=ever use
hcs	hormonal contraception using status	0=no
		1=current user
		2=ex-user
dhc	duration of using hormonal contraception	0= never use
		1= ex-user <10 years
		2= ex-user 10-14 years
		3= ex-user ≥5 years
		4= current-user <10 years
		5= current-user 10-14 years
		6= current-user ≥5 years
lasthc	time since last using pill	0= never user
		1= >10 years
		2= 5-9 year
		3= 1-4 years
		4= current user
smk	smoking status	0=never smoke
		1= current smoker
		2= ex-smoker
nsmk	number of cigarettes	0= never smoker
		1= ex-smoker <5/day
		2= ex-smoker ≥5/day
		3= current-smoker <5/day
		4= current-smoker ≥5/day
smkqt	time since quitting	0= never smoker
		1= ≥10 years
		2= 5-9 years
		3=1-4 years
		4= current smoker
dsmk	duration of smoking	0= never smoker
		1= ex-smoker <10 years
		2= ex-smoker ≥10 years
		3= current-smoker <10 years
		4= current-smoker ≥10 years
sexp	lifetime number of sexual partners	1=1-2
		2=3-5
		3=6-9
		4= ≥10
		5= prefer not to answer

sexp5	number of sexual partners in the last 5 years	1=0-1 2=2 3=3-5 4= ≥6 5= prefer not to answer
sexage	age at first sexual intercourse	1= ≥21 years 2=19-20 years 3= 17-18 years 4= <17 years
parity	number of children	0=nulliparous 1=1 2=2 3= ≥3
parity2g	number of children 2 groups	0=nulliparous 1= ≥1
nopap	number of Pap smears 1.5-5 years prior to index test	
age	age at reference test	
bmi	body mass index	
std	sexually transmitted disease	0=no 1=yes

STATA code for data analysis

Model 2:

```
xi: clogit cc i.ehc i.smk i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.hcs i.smk i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.dhc i.smk i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.lasthc i.smk i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.smk i.hcs i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.nsmk i.hcs i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.dsmk i.hcs i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.smkqt i.hcs i.parity2g i.agesex i.sexp nopap, group(matching) or
xi: clogit cc i.smk i.hcs i.parity i.agesex i.sexp nopap, group(matching) or
```

Model 3:

```
xi: clogit cc i.ehc i.smk i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.hcs i.smk i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.dhc i.smk i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.lasthc i.smk i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.smk i.hcs i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.nsmk i.hcs i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.dsmk i.hcs i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.smkqt i.hcs i.parity2g i.agesex i.sexp5 nopap, group(matching) or
xi: clogit cc i.smk i.hcs i.parity i.agesex i.sexp5 nopap, group(matching) or
```