# How to adjust for baseline heterogeneity in Count Data occurring from Experimental studies

Matthew Morton

April 16, 2022

# Contents


**1 Introduction** **15**
- 1.1 Research objectives . . . 17
- 1.2 Motivating data sets . . . 18
  - 1.2.1 Belcap study . . . 18
  - 1.2.2 Polyps data set . . . 18
  - 1.2.3 Falls dataset . . . 19
- 1.3 Structure of thesis . . . 20

**2 Background to experimental and observational studies** **22**
- 2.1 Experimental studies . . . 22
  - 2.1.1 Types of clinical trial . . . 22
  - 2.1.2 Phases of a clinical trial . . . 23
  - 2.1.3 Randomisation techniques . . . 24
- 2.2 Observational studies . . . 25
  - 2.2.1 Cohort studies . . . 26
  - 2.2.2 Case-control studies . . . 26

**3 Hypothetical example of issues with baseline heterogeneity** **28**
- 3.1 Hypothetical situation . . . 28
- 3.2 The models when no heterogeneity is present . . . 28
- 3.3 The models when heterogeneity is present . . . 29
- 3.4 How heterogeneity can influence the findings . . . 29

**4 Methodology** **31**
- 4.1 Generalised Linear Model . . . 31
  - 4.1.1 Exponential Family of Distributions . . . 31
  - 4.1.2 Example: The Bernoulli Distribution belongs to the Exponential Family of Distributions . . . 31
  - 4.1.3 Example: The Poisson Distribution belongs to the Exponential Family of Distributions . . . 32
  - 4.1.4 Structure of a Generalised Linear Model . . . 32
  - 4.1.5 Iterative methods for model parameter estimation . . . 33
  - 4.1.6 Model selection criteria . . . 33
  - 4.1.7 Residual Analysis . . . 34
- 4.2 Six methods for adjusting baseline heterogeneity in the outcome variable . . . 35

4.2.1 Method 1: Poisson regression using an offset term . . . 36
4.2.2 Method 2: Poisson regression using baseline measurements as a continuous covariate . . . 37
4.2.3 Method 3: Poisson regression using baseline measurements as a categorical covariate . . . 37
4.2.4 Method 4: Poisson regression with a random effect . . 38
4.2.5 Method 5: Adaptation of the conditional linear mixed model . . . 38
4.2.6 Method 6: Mantel-Haenszel Approach . . . 39

**5 Analysing the motivating data sets 41**
5.1 Belcap data set . . . 41
5.1.1 Exploratory Data Analysis . . . 41
5.1.2 Poisson Regression ignoring Baseline Heterogeneity . . 46
5.1.3 Model 1: Poisson Regression using baseline measurements as an offset . . . 49
5.1.4 Method 2: Poisson Regression using baseline measurements as a continuous covariate . . . 54
5.1.5 Method 3: Poisson Regression using baseline measurements as a categorical covariate . . . 58
5.1.6 Method 4: Poisson Regression using baseline measurements as a random effect . . . 63
5.1.7 Method 5: Adaptation of the conditional linear mixed model . . . 66
5.1.8 Method 6: The Mantel-Haenszel Approach . . . 69
5.2 Analysing the Polyps data set . . . 71
5.2.1 Exploratory Data Analysis . . . 72
5.2.2 Poisson Regression ignoring Baseline Heterogeneity . . 75
5.2.3 Method 1: Poisson Regression using baseline measurements as an offset . . . 79
5.2.4 Method 2: Poisson Regression using baseline measurements as a continuous covariate . . . 81
5.2.5 Method 3: Poisson Regression using baseline measurements as a categorical covariate . . . 83
5.2.6 Method 4: Poisson Regression using baseline measurements as a random effect . . . 87
5.2.7 Method 5: Adaptation of the conditional linear mixed model . . . 88

5.2.8 Method 6: The Mantel-Haenszel Approach . . . . . 90
5.3 The Falls dataset . . . . . 90
5.3.1 Exploratory Data Analysis . . . . . 90
5.3.2 Poisson Regression ignoring Baseline Heterogeneity . . 95
5.3.3 Method 1: Poisson Regression using baseline measurements as an offset . . . . . 98
5.3.4 Method 2: Poisson Regression using baseline measurements as a continuous covariate . . . . . 101
5.3.5 Method 3: Poisson Regression using baseline measurements as a categorical covariate . . . . . 104
5.3.6 Method 4: Poisson Regression using baseline measurements as a random effect . . . . . 107
5.3.7 Method 5: Adaptation of the conditional linear mixed model . . . . . 109
5.3.8 Method 6: The Mantel-Haenszel Approach . . . . . 111
5.4 How successful are the proposed methods at dealing with baseline heterogeneity . . . . . 111
5.4.1 Results from the Belcap study . . . . . 111
5.4.2 Results from the Polyps study . . . . . 112
5.4.3 Results from the Falls study . . . . . 114

**6 Which method is best at adjusting for baseline heterogeneity 116**
6.1 A simulation algorithm based on the Polyps study . . . . . 116
6.1.1 Results from simulation based on the Polyps study . . 118
6.1.2 Influence of a smaller sample size of each replication on the simulation results . . . . . 119
6.1.3 Influence of true treatment effect on the simulation results . . . . . 121
6.1.4 Relationship between the RMSE and true risk ratio . . 125
6.2 A simulation based on the Belcap study . . . . . 127
6.2.1 Results from simulation based on the Belcap study . . 129
6.2.2 Influence of true treatment effect on the simulation results . . . . . 131
6.2.3 Relationship between the bias and true risk ratio . . . 133
6.3 A simulation based on a cluster randomised trial where there are more clusters than treatments . . . . . 134
6.4 A simulation based on data with a third time point . . . . . 135
6.4.1 Results from simulation based on three time points . . 137

6.5 A simulation created from the Offset Method . . . . . 138

**7 Discussion and Conclusions** **139**
7.1 Demonstrating the need to adjust for baseline heterogeneity . 139
7.2 Finding the best method to adjust for baseline heterogeneity . 140
7.3 Limitations of this thesis . . . . . 141
7.4 Overall conclusions and Further work . . . . . 142

**8 Appendicies** **143**
8.1 Appendix 1: Polyps simulation . . . . . 143
8.2 Appendix 2: Belcap simulation . . . . . 151
8.3 Appendix 3: Third time point . . . . . 176

# List of Figures


1 Boxplot showing the spread in the outcome variable at baseline split by treatment group. . . . . 42
2 Boxplot showing the spread in the outcome variable at the end split by treatment group. . . . . 43
3 Scatter plot showing average DMFS count before and after the study for all treatment groups. . . . . 44
4 Scatter plot showing the standardised residual for each observation. . . . . 47
5 Scatter plot showing the Cook's distance of each observation. . 48
6 Scatter plot showing the standardised residual for each observation from method 1. . . . . 51
7 Scatter plot showing the Cook's distance of each observation from method 1. . . . . 52
8 Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the offset adjustment and when no adjustment is done. . . . . 53
9 Scatter plot showing the standardised residual for each observation from method 2. . . . . 55
10 Scatter plot showing the Cook's distance of each observation from method 2. . . . . 56
11 Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the continuous adjustment and when no adjustment is done. . . . . 58
12 Scatter plot showing the standardised residual for each observation from method 3. . . . . 60
13 Scatter plot showing the Cook's distance of each observation from method 3. . . . . 61
14 Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the categorical adjustment and when no adjustment is done. . . . . 63
15 Scatter plot showing the standardised residual for each observation from method 4. . . . . 64
16 Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the random effect adjustment and when no adjustment is done. . . . . 66

17 Scatter plot showing the standardised residual for each observation from method 5. . . . . 68
18 Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for method 5 and when no adjustment is done. . . . . 69
19 Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the Mantel-Haenszel method and when no adjustment is done. . . . . 71
20 Boxplot showing the spread in the outcome variable at baseline split by treatment group. . . . . 72
21 Boxplot showing the spread in the outcome variable at baseline split by treatment group with outliers removed. . . . . 73
22 Boxplot showing the spread in the outcome variable after 3 months split by treatment group with outliers removed. . . . . 74
23 Scatter plot showing the standardised residual for each observation from the model having not adjusted for baseline heterogeneity. . . . . 77
24 Scatter plot showing the Cook's distance of each observation from the model having not adjusted for baseline heterogeneity. 78
25 Scatter plot showing the standardised residual for each observation from method 1. . . . . 80
26 Scatter plot showing the Cook's distance of each observation from method 1. . . . . 81
27 Scatter plot showing the standardised residual for each observation from method 2. . . . . 82
28 Scatter plot showing the Cook's distance of each observation from method 2. . . . . 83
29 Scatter plot showing the standardised residual for each observation from method 3. . . . . 85
30 Scatter plot showing the Cook's distance of each observation from method 3. . . . . 86
31 Scatter plot showing the standardised residual for each observation from method 4. . . . . 88
32 Scatter plot showing the standardised residual for each observation from method 5. . . . . 89
33 Boxplot showing the spread in baseline falls split by treatment group. . . . . 91

34 Boxplot showing the spread in baseline falls (capped at 100) split by treatment group. . . . . 92
35 Boxplot showing the spread in follow up falls split by treatment group. . . . . 93
36 Boxplot showing the spread in follow up falls (capped at 100) split by treatment group. . . . . 94
37 Scatter plot showing the standardised residual for each observation from the model having not adjusted for baseline heterogeneity. . . . . 96
38 Scatter plot showing the Cook's distance of each observation from the model having not adjusted for baseline heterogeneity. 97
39 Scatter plot showing the standardised residual for each observation from method 1. . . . . 99
40 Scatter plot showing the Cook's distance of each observation from method 1. . . . . 100
41 Scatter plot showing the standardised residual for each observation from method 2. . . . . 102
42 Scatter plot showing the Cook's distance of each observation from method 2. . . . . 103
43 Scatter plot showing the standardised residual for each observation from method 3. . . . . 105
44 Scatter plot showing the Cook's distance of each observation from method 3. . . . . 106
45 Scatter plot showing the residuals for each observation from method 4. . . . . 108
46 Scatter plot showing the residual for each observation from method 5. . . . . 110
47 Scatter plot showing the relationship between Risk Ratio and RMSE for no adjustment and all 6 trial methods. . . . . 126
48 Scatter plot showing the relationship between Risk Ratio and RMSE for no adjustment and all 6 trial methods. . . . . 133

# List of Tables


1 Variables in Brazil data set. . . . . 18
2 Variables in Polyps data set. . . . . 19
3 Variables in Falls data set. . . . . 20
4 Advantages and disadvantages of cohort studies. . . . . 26
5 Advantages and disadvantages of case-control studies. . . . . 27
6 Effect of baseline heterogeneity when the true risk ratio is 1.4 and $\gamma = 0.2$. . . . . 30
7 Average DMFS score at baseline. . . . . 42
8 Average DMFS-end split by treatment group. . . . . 45
9 Average DMFS-end split by ethnicity. . . . . 45
10 Average DMFS-end split by gender. . . . . 46
11 Coefficients and standard errors from model with no adjustment for heterogeneity. . . . . 46
12 Coefficients and standard errors from the negative binomial model with no adjustment for heterogeneity. . . . . 49
13 Coefficients and standard errors from model with the natural log of baseline measurements as an offset. . . . . 50
14 Coefficients and standard errors from the negative binomial model with an offset adjustment for heterogeneity. . . . . 52
15 Coefficients and standard errors from model where baseline measurements are used as a continuous covariate. . . . . 54
16 Coefficients and standard errors from the negative binomial model with a continuous adjustment for heterogeneity. . . . . 57
17 Sample sizes of the baseline DMFS groups. . . . . 59
18 Coefficients and standard errors from model where baseline measurement groups are used as a categorical covariate. . . . . 59
19 Coefficients and standard errors from a negative binomial model where baseline measurement groups are used as a categorical covariate. . . . . 62
20 Coefficients and standard errors from model where baseline measurement are used as a random effect. . . . . 64
21 Coefficients and standard errors from the negative binomial model with a random adjustment for heterogeneity. . . . . 65
22 Coefficients and standard errors from conditional model. . . . . 67
23 Mantel-Haenszel estimates of the risk ratio for each experimental group . . . . . 70

24 Average polyps count at baseline. . . . . 73
25 Average number of polyps after 3 months split by treatment group. . . . . 75
26 Average number of polyps after 3 months split by gender. . . . 75
27 Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made. . . . . 76
28 Coefficients and standard errors from the negative binomial model having not adjusted for baseline heterogeneity. . . . . 78
29 Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made. . . . . 79
30 Coefficients and standard errors from model where baseline measurements are used as a continuous covariate. . . . . 82
31 Sample sizes of the Count-b groups. . . . . 84
32 Coefficients and standard errors from model where baseline measurement groups are used as a categorical covariate. . . . . 84
33 Coefficients and standard errors from the negative binomial model with a categorical adjustment for heterogeneity. . . . . 87
34 Coefficients and standard errors from model where baseline measurements are used as a random effect. . . . . 87
35 Coefficients and standard errors from model with the natural log of baseline measurements as an offset. . . . . 89
36 Average number of falls during the baseline period. . . . . 92
37 Average number of falls during the followup period. . . . . 94
38 Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made. . . . . 95
39 Coefficients and standard errors from the negative binomial model having not adjusted for baseline heterogeneity. . . . . 97
40 Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made. . . . . 98
41 Coefficients and standard errors from the negative binomial model having an offset adjustment for baseline heterogeneity. . 100
42 Coefficients and standard errors from model where baseline measurements are used as a continuous covariate. . . . . 101
43 Coefficients and standard errors from negative binomial model where baseline measurements are used as a continuous covariate. 103
44 Sample sizes of the baseline fall groups. . . . . 104
45 Coefficients and standard errors from model where baseline measurement groups are used as a categorical covariate. . . . . 105

46 Coefficients and standard errors from negative binomial model where baseline measurement groups are used as a categorical covariate. . . . . 107
47 Coefficients and standard errors from model where baseline measurements are used as a random effect. . . . . 107
48 Coefficients and standard errors from the negative binomial model where baseline measurements are used as a random effect. 108
49 Coefficients and standard errors from the Conditional method. 109
50 Coefficients and standard errors from the Conditional method accounting for overdispersion. . . . . 110
51 Model criterion for models (analysing the belcap data) with and without adjustment for baseline heterogeneity and having adjusted for overdispersion. . . . . 111
52 Risk ratios (treatment versus control) from models with and without adjustment for baseline heterogeneity and having adjusted for overdispersion. . . . . 112
53 Model criterion for models (analysing the polyps data) with and without an offset term for the Polyps data set. . . . . 113
54 Risk ratios (treatment versus control) from models with and without adjustment for baseline heterogeneity. . . . . 113
55 Model criterion for models (analysing the falls data) with and without an adjustment for baseline heterogeneity. . . . . 114
56 Risk ratios (Intervention versus Standard care) from models with and without adjustment for baseline heterogeneity. . . . . 114
57 Estimated risk ratios for each method from the first 5 iterations of the simulation study with one treatment group. The true risk ratio is 0.5499. . . . . 118
58 Average estimated risk ratio, bias and mean squared error for each of the six trial methods having run 1000 replicates each with sample size 250. . . . . 119
59 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the six trial methods having run 1000 iterations with sample size 100. . . . . 120
60 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the six trial methods when the true risk ratio is 0.25. 121
61 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the six trial methods when the true risk ratio is 0.5. 122

62 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 0.75. . . . . 122
63 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1. 123
64 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.25. . . . . 123
65 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.5.124
66 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.75. . . . . 124
67 Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 2. 125
68 Results from a simulation with a replication number of 1,000 (each with a sample size of 750 observations) where the true risk ratios for ALL, ESD, MW, OHE, OHY are 0.5724, 0.8226, 0.6615, 0.7043, 0.7348 respectively . . . . . 130
69 Simulation with a replication number of 1,000. Each replicate has a sample size of 750. The true risk ratio for ALL is varied 132
70 2 clusters allocated to each treatment group . . . . . 134
71 3 clusters allocated to each treatment group . . . . . 135
72 Average estimated risk ratio, bias and mean squared error for each of the six trial methods having run 1000 replicates with sample size 250. . . . . 137
73 Simulation based on the Offset method . . . . . 138

**Abstract**

When analysing the results from experimental and observational studies, the main aim is often to estimate what effect the treatment is having on the participant. For this reason, it is important that the estimate for the treatment effect is not influenced (biased) by having imbalanced (heterogeneous) treatment groups. Randomisation is often used in experimental studies to obtain homogenous treatment groups i.e. the distribution of all covariates (except treatment group) is the same between treatment groups. However, when the sample size is small, randomisation may not successfully obtain entirely homogeneous groups. Heterogeneous groups are often an issue in observational studies as randomisation cannot be used. This thesis aims to demonstrate the need to adjust for baseline heterogeneity by showing the potential consequences of not. The thesis also aims to find a method to successfully adjust for baseline heterogeneity.

A hypothetical example is drawn up to demonstrate the potential bias in the results if no adjustment for the baseline heterogeneity is made. An explanation is given on how failing to adjust for heterogeneity, could lead to false conclusions to the extent that a harmful drug could be wrongly authorised. This thesis examines the properties of six potential methods for adjusting for baseline heterogeneity which include 5 parametric methods (using an Offset, Continuous Covariate, Categorical Covariate, Random Effect and a Conditional model) and a non-parametric method (Mantel-Haenszel). The ability of these methods to adjust for heterogeneity is assessed by using them to analyse three datasets containing baseline heterogeneity. Furthermore, a detailed simulation study is undertaken to analyse the bias and RMSE of each of the methods.

The treatment effects obtained from the different methods (for adjusting for baseline heterogeneity) differ in the analysis of the example datasets. The AIC and BIC demonstrate that adjusting for baseline heterogeneity is required. However, the AIC and BIC cannot convincingly separate the parametric methods (AIC and BIC is not available for the Mantel Haenszel method). In order to distinguish between the 6 different trial methods, simulation studies are used. The RMSE is then calculated for a range of different scenarios (different sample sizes, risk ratios, number of treatment groups and time points). The Continuous method consistently performs well. For this reason, the

Continuous method appears to be the preferred method to use.

# 1 Introduction

The issues caused by baseline heterogeneity are a major problem in Observational studies such as Cohort studies but the issues also occur in Experimental studies in Medicine and Life sciences such as Clinical Trials. Details of why this is an issue is explained in detail later in the thesis. The importance of well managed clinical trials has possibly never been more apparent than with the occurrence of Covid-19. Given this, the ability to analyse data from such studies accurately and reliably is essential.

The data collected from these studies is often longitudinal. This means data for the outcome variable is taken at various different times (for each individual) including the start. The final measurement (known as the end point) can be thought of as the starting value plus some evolution over time. The evolution part is known as the "longitudinal effect" and the starting point is known as the "cross sectional effect". Typically, it is the longitudinal effect that is studied.

Longitudinal studies are particularly susceptible to baseline heterogeneity. This is because heterogeneity can occur in the baseline measurement of the outcome as well as other variables which could potentially influence the outcome. Clinical trials use "randomisation" to adjust for baseline heterogeneity in variables which could influence the outcome variable. It is not always possible to remove all heterogeneity this way. Kent et al. [21] have done research into how to assess any unaccounted for heterogeneity and how best to report it.

One common thought is that balanced studies (equal numbers are analysed at each time point) produce substantially higher levels of validity for the longitudinal effect. A presentation given in February 2020 by Verbeke shows that the decrease in validity caused by unbalanced data is typically quite minor. This is also supported in a paper by Verbeke et al. [45]).

When carrying out analysis of longitudinal data from both observational studies and clinical trials, it is thought vital to adjust for any significant baseline variables. The need to adjust for such variables is outlined in the CONSORT and EMA guidelines [12]. Failing to do this adjustment can cause some bias to occur in the longitudinal effects, with the exception of the growth

curve model for completely balanced data (see Verbeke et al. [44]). The term bias can have different meanings depending on the situation. In this thesis, bias is used to describe the precision of the longigitudinal estimates.The bias caused by leaving out significant baseline variables in generalised linear models is shown by Neuhaus [27]. Gail et al. [15] demonstrate the bias for non-linear models. A further example of just how severe the bias can be, is given in Chao et al. [7] where binary clustered data is used.

In the case of linear models (e.g. ANOVA) or linear mixed models with balanced longitudinal data (equal numbers are analysed at each time point), the longitudinal effects are protected from bias when omitting baseline effects [44]. This is due to the orthogonality properties of the model. The orthogonal properties (absence of correlation between model parameters) in linear models such as ANOVA are explained in the paper by Winer [47]. It is worth noting that the baseline measurement of the outcome variable is not one of the baseline effects which can safely be omitted. Interestingly, in linear mixed models, the orthogonal properties protect the standard errors from bias where this is not the case for linear models such as ANOVA [44].

Palta and Yao [32] also did research into whether this useful finding (i.e no bias) holds when omitting baseline covariates in linear models when the data is correlated. A compound covariance structure and normality of covariates were assumed. These conditions are highly unlikely to hold in most situations, hence the results have little practical meaning and are therefore not considered here.

There is clear agreement, as pointed out in the CONSORT guidelines, on the need to adjust for the baseline measurement of the outcome variable but what type of adjustment should be used? Stephen Senn in a guest post on the website "Error Statistics" [38] examines the merit of analysing the difference between outcome and baseline measurements. The results presented show that an ANCOVA (covariate adjustment) approach is better than the difference method. Research by Hernández et al. [18] shows adjusting for a significant baseline covariate increases the power of the analysis on the longitudinal effect (i.e. treatment effect). This finding is also supported by Kahan et al [22].

There is support for carrying out a covariate adjustment for the baseline

measurement of the outcome variable. But what type of covariate adjustment is best? Verbeke et al. [45] suggest using a conditional linear mixed model as this does not cause any bias in the longitudinal effect. Another benefit is that this model can be interpreted as an extension of the classical paired t-test. This is a major benefit given the popularity of paired t-tests (see Verbeke and Fieuws, [43]). The main disadvantage of conditional linear mixed models is that typically all information about cross sectional effects is lost while carrying out the conditioning. This is of little concern as it is the longitudinal effects which are of more interest. Neuhaus and Kalbfleisch [28] tried to extend these findings to clustered data with generalised linear models. However, a covariance structure (which is very unlikely to occur) must be assumed for there to be no bias.

From reviewing the current literature, there is a clear problem with baseline heterogeneity in longitudinal studies. The vast majority of current research focuses on scenarios where binary or continuous data is collected. There does not appear to have been any research on the issues of baseline heterogeneity in count data. Thus, this thesis is going to focus on count data and how to account for any possible baseline heterogeneity. Three longitudinal datasets from clinical trials are used to demonstrate the adjustments.

## 1.1 Research objectives

- **Demonstrate the need to adjust for baseline heterogeneity:** Multiple papers have demonstrated issues with baseline heterogeneity and the need to account for it [44] [45] [7]. This thesis aims to demonstrate the need for adjustment in the case of count data.
- **Find a suitable method to perform the adjustment:** Six statistical methods are investigated for their ability to adjust for baseline heterogeneity. Five of the methods are a form of parametric analysis where the other method investigated is non-parametric.
- **Use of statistical software:** Demonstrate how to employ the necessary statistical techniques with the statistical software R.

## 1.2 Motivating data sets

This thesis uses data from 3 clinical trials to achieve the above research objectives. These publicly available datasets are explained below.

### 1.2.1 Belcap study

This community randomised trial recruited pre-school children in Belo Horizonte (Brazil) and looked at preventing caries (decay or cavities). Four different interventions were randomised to four different schools. A fifth school then received all four intervention and finally a sixth school acted as a control. All children from the same school receive the same intervention. The outcome variable was the number of tooth surfaces with decay, missing teeth or surfaces with fillings (DMFS). This outcome measure was taken at baseline and at the end of the study which was two years later (see Böhning et al, [6]). Table 1 below shows what variables are included in the data set.

**Table 1:** Variables in Brazil data set.

| **Variable** | **Description** |
|---|---|
| Treat | OHE = Oral Health Education<br>ESD = Enrichment of school diet with rice bran<br>MW = Mouth Wash with 0.2% sodium fluoride<br>OHY= Oral Hygiene<br>ALL = All interventions received<br>CONTROL = No intervention |
| Ethnicity | 1=dark, 2=white, 3=black |
| Gender | 0=female, 1=male |
| DMFS-beg | Number of tooth surfaces with decay,<br>missing or filling at the start of study |
| DMFS-end | Number of tooth surfaces with decay,<br>missing or filling at the end of study |

### 1.2.2 Polyps data set

A polyp is a clump of cells which, in the case of bowel polyps, form on the inner lining of the large intestine or rectum. They are very common, affecting around 25% of people at some point in their lives. These polyps are rarely cancerous. However, if they are not removed, there is a chance that they will

turn cancerous. It is thought by doctors that a specific form of bowel polyp, known as an Adenoma, is what bowel cancer stems from (see nhs.uk, [29]). For this reason, it is always recommended to get polyps treated. The data set here is from a clinical trial comparing an experimental and control group for the removal of polyps. It follows that the outcome measure here is the number of polyps present. It is measured at baseline and after 3 months of treatment. Table 2 below shows what variables are included in the data set.

**Table 2:** Variables in Polyps data set.

| **Variable** | **description** |
|---|---|
| Treat | 0 = Placebo, 1 = Experimental |
| Gender | 0 = Female, 1= Male |
| Age | ranges 13 -50 years old |
| Count-b | Number of polyps at the start of the study |
| Count-3 | Number of polyps after three months of treatment |

### 1.2.3 Falls dataset

"Parkinson's disease" is a neurological condition which, among other things, affects people's balance and thus, increases the risks of falling over. It is thought that about 0.5% of people will suffer from this disease. Typically, symptoms start to appear once the sufferer is aged over 50 however, about 20% of cases show symptoms under the age of 40 [30].

The falls suffered by Parkinson's patients can often lead to severe injuries [17] [31] [34] [46] and in some cases lead to psychological difficulties [20] [50]. Currently there is no cure for Parkinson's disease. However, a considerable amount of research has gone into how to provide better care for patients with Parkinson's [4] [24].

One such study involved a randomised clinical trial which took place in the South West of England [16]. The eligibility criteria for this trial required participants to have a diagnosis of Parkinson's, have suffered at least 2 falls in the last year, have a mobilising ability and either be a resident or enrolled with a GP practice in Devon. The trial enrolled 130 participants on the study which was split into three 10 week periods.

The first block of ten weeks (known as the baseline period) involved all participants' received standard care, during which the number of falls suffered was recorded. This measurement was used as the baseline measurement.

During the second block of 10 weeks (intervention period), participants received the treatment which was randomly allocated to them. This resulted in 64 receiving the intervention and 66 continuing with just standard care. The participants in the intervention group received standard care as well as, one group exercise session and two home exercises in each of the 10 weeks. Each participant recorded the number of falls they had during this period. This measurement was known as the intervention measurement.

The final block of 10 weeks was known as the outcome period. During the outcome period, all participants only received standard care. All participants recorded the number of falls they suffered during this period. This measurement was known as the outcome measurement. Table 3 below shows the variables collected from this trial.

**Table 3:** Variables in Falls data set.

| **Variable** | **description** |
|---|---|
| ID | Number code to identify each participant |
| Baseline falls | The number of falls during the baseline period |
| Follow up falls | The number of falls while recieving treatment |
| Intervention | 0 = Control, 1 = Intervention |
| Log baseline | Natural log of the baseline count |

## 1.3 Structure of thesis

Chapter 2 looks at different aspects of clinical trials and cohort studies. This includes different randomization techniques. Chapter 3 gives a hypothetical situation where heterogeneity could occur and the potential harm which failing to account for it could do. Chapter 4 looks at the methodology behind GLMs and describes the five methods, proposed in this paper, for dealing with baseline heterogeneity. In chapter 5, the six trial methods are applied to the three data sets (Belcap, Polyps, Falls). Chapter 6 performs a simulation study to try and further understand the usefulness of the 6 differ-

ent methods. Finally, chapter 7 has a discussion of the findings of this thesis as well as a description of possible future work.

# 2 Background to experimental and observational studies

Statistical studies often take the form of either an experimental or an observational study. The choice between the two could be influenced by a variety of issues such as time constraints or ethical matters. The key difference between the types of study is how the treatment is allocated. Exposure to the treatment of interest occurs naturally within observational studies which means the investigator has no control over treatment allocation (see Timothy and Legg, [42]). In an experimental study, the investigator controls the method of allocation to either the experimental or control group.

## 2.1 Experimental studies

The most common type of experimental study in medicine is a clinical trial. It is these trials which are used to check new treatments or drugs for effectiveness and safety in humans. All drugs or treatments must undergo "successful" clinical trials before they can be licensed for general use, thus the vital importance of well conducted clinical trials is readily apparent. Most clinical trials randomly allocate participants into groups, typically called an "experimental group" and a "control group" but there may be more than one experimental group where this is seen as beneficial. Analysis of the outcome variable is performed after the trial to assess the performance of the new drug or treatment. The random allocation of participants to groups is important as it can help form homogeneous groups. There are different types of clinical trials which have no randomisation or the amount of randomisation is limited. This could lead to some baseline heterogeneity appearing and damaging the validity of the results if not accounted for appropriately (see Piantadosi, [33]).

### 2.1.1 Types of clinical trial

There are different types of trial depending on the trial objectives and the resources available. The reliability of conclusions made from clinical trials also depends on which type of trial has been conducted.

The type of trial that generally provides the most accurate results is known as an "Individual randomised trial" (IRT) with a large sample size which could extend into the thousands. Due to the law of large numbers,

key characteristics have similar or identical distributions between treatment groups when the sample size is large. Hence very little if any baseline heterogeneity is left unaccounted for. This type of trial is typically a phase 3 trial. Due to the high level of randomisation, the methods studied later in this thesis are not required in these large IRTs.

There are also IRTs with small sample sizes, perhaps only double figures (the Polys data set in section 2.2 is an example of this). In these trials some randomisation takes place, however, the sample size is too small makes it unlikely the law of large numbers will work. Thus, not all key characteristics have similar distributions. This means there could be some baseline heterogeneity unaccounted for. The methods studied later in the thesis may then be of use in improving the reliability of the results.

Another type of trial is a "Community randomised trial" (CRT). This is where the treatment is randomised to a whole community, meaning potential key characteristics are not specifically randomised. An example of such a trial is given in section 2.1 (the Belcap study). This limited amount of randomisation can lead to baseline heterogeneity between treatment groups to be problematic. This issue may be dealt with using one of the methods described later in the thesis. Given equal sample sizes, IRTs have less baseline heterogeneity than CRTs.

Finally, there are trials with no randomisation at all. These are the most likely to suffer from heterogeneity at baseline and it is of definite importance that adjustments for this heterogeneity are made in these trials.

### 2.1.2 Phases of a clinical trial

Clinical trials can differ greatly depending upon what "phase" of the whole process they belong to.

**Phase 1 trials** involve very few patients and is the first occasion that the treatment is used in humans. Typically, this phase uses healthy participants, however, when the illness is terminal such as some cancers, participants with the illness may be involved. This phase has the primary aims of finding out if the drug or treatment is safe along with identifying any side effects. These trials have too small a sample size to use randomisation (see Shamley and

Wright, [39]).

**Phase 2 trials** are slightly larger than Phase 1 trials and look into the efficiency of the treatment, as well as possible side effects, in greater detail. It is worth noting that Phase 2 trials may be large enough to involve randomisation (see Shamley and Wright, [39]).

**Phase 3 trials** attract the most attention. They tend to involve the most participants, the number of which can extend into the thousands. It is now that the new drug or treatment is compared to the standard treatment or placebo. A new treatment providing better outcomes than the standard is clearly beneficial. Should the new treatment be cheaper, it would only need to perform equally as well as the standard for it to be viewed as an improvement. After a drug or treatment has passed the third Phase, it is licensed and released onto the market. Given the much larger size of Phase 3 trials and the fact that the drug then goes onto the market, randomisation is essential whenever possible (see Shamley and Wright, [39]).

**Phase 4 trials** are used to examine the long term benefits or risks of the treatment and follow on from Phase 3. Randomisation is rare in phase 4, information on benefits and adverse events are often collected in a more routine manner (see Shamley and Wright, [39]).

Trials can sometimes span 2 phases. For example, there are trials which are known as Phase 1/2 which cover the aims of both a Phase 1 and a Phase 2 trial.

### 2.1.3 Randomisation techniques

As discussed above, all phase 3 trials try to use randomisation, however, there are numerous possible techniques for carrying out this process. Below are some examples.

**Simple Randomisation** is the most basic method. This involves random numbers being generated every time a participant is enrolled in a trial. For example, if the random number generated is even, the participant is allocated to the experimental group, if odd, he or she is allocated to the control group. This method is flawed as no patient characteristics are accounted

for so, in an extreme case, all the males could end up in one group and the females in the other group. This would clearly bias the results. Simple randomisation can also fail to produce balanced group sizes, particularly if the sample size is small (see Suresh, [41]).

**Blocked Randomisation** is an improvement on simple randomisation. This involves choosing a participant "block size", typically between 2 and 8 participants and randomising each block separately. Taking a block size of 6 participants as an example, a list of randomly generated single figure numbers is produced. The numbers greater than the block size are ignored. The first three numbers (half of the block size) are read off and these correspond to the individuals allocated to the experimental group. For example, if the first three numbers are 2, 5 and 6, the second, fifth and sixth participants in the first block are allocated to the experimental group. The first, third and fourth participants are then allocated to the control group. This process is then repeated for each block (see Altman and Bland, [3]). This method produces equal group sizes unless the trial stops recruiting mid-block, but, this degree of imbalance is likely to be insignificant. The issue of key characteristics not being accounted for has again, not been solved by this method.

**Stratified Randomisation** accounts for perceived key characteristics. Strata are formed for each category of the key characteristics. If the characteristics are say gender and age, age maybe split into categories, of say, under 60 and 60 or above. The number of strata is given by multiplying the number of categories for each characteristic together. For the gender and age example, there are 4 ($2\times2$) strata. These are; male and under 60, male and 60 or above, female and under 60, female and 60 or above (see Weir and Lees, [48]). Randomisation is performed (blocked randomisation is preferred) on each stratum individually. This method produces groups which are balanced in terms of size and the key characteristics (age and gender in the above example).

## 2.2 Observational studies

There are different types of observational study. The different types vary in the way they are set up and in what circumstances they may be used. A description of two types of observational study, known as cohort and case-control studies are described below. Cohort studies suffer from baseline het-

erogeneity more frequently than case-control studies.

### 2.2.1 Cohort studies

Cohort studies can be conducted prospectively or retrospectively. They don't produce as high a level of medical evidence as randomised clinical trials however, they still provide a decent standard [40].

**Prospective cohort studies** take a group of people that are assumed to be free from the condition of interest. The participants are split into groups according to exposure from a particular factor (a common factor used is smoking) and are then followed over time to see whether exposure to the factor is associated with future occurrence of the disease [40].

**Reterospective cohort studies** differ as they use data from the past. These studies are also called historical cohort studies. The key to conducting these studies is that the exposure groups are defined according to the historical exposure. This historical exposure is analysed to see if it is associated with current occurrence of the disease [40].

Table 4 below shows the advantages and disadvantages of cohort studies when compared with other observational studies [40].

**Table 4:** Advantages and disadvantages of cohort studies.

| **Advantages** | **Disadvantages** |
|---|---|
| Able to assess causality | Possible selection bias |
| Look at how an exposure is associated with multiple outcomes | may need large sample size (expensive) |
| Suitable for rare exposures | Long follow up, may be hard to maintain, people drop out * |
| Able to calculate relative risk | recall bias and less control on variables ** |

* Only applies to prospective studies ** Only applies to retrospective studies

### 2.2.2 Case-control studies

Case-control studies are often retrospective. They differ from retrospective cohort studies as in case-control studies the cases and controls are deter-

mined by the outcome not the exposure. Thus, when setting up the study a set of cases are found and then controls are found to match the cases. The matching of cases and controls is done so the two groups are identical in every way except for exposure to the potential risk factor being studied (smoking is a common one). This means that baseline heterogeneity is not an issue in these studies and therefore, the methods presented in this thesis are not needed for case-control studies [40].

Table 5 below shows the advantages and disadvantages of case-control studies when compared with other observational studies [40].

**Table 5:** Advantages and disadvantages of case-control studies.

| **Advantages** | **Disadvantages** |
|---|---|
| Good for rare outcomes | Susceptible to recall bias |
| Quick as existing records used | Difficult to validate information |
| Small sample size so cheaper | Control of external variables may not be possible |
| Examine multiple risk factors | Risk of disease cannot be established |

# 3 Hypothetical example of issues with baseline heterogeneity

This section describes a purely hypothetical situation where an issue with heterogeneity could occur. A potential issue caused by failing to account for heterogeneity is shown here.

## 3.1 Hypothetical situation

A new cream has been developed and is believed to reduce the number of spots (following 3 days of treatment) caused by being infected by Chicken Pox. A randomised clinical trial is set up to investigate whether this new cream is successful. Let Y denote the number of spots at the end of the study (after 3 days) and B the number of spots at the start of the study. Also, let T be an indicator variable taking the value 1 if the experimental cream is used or the value 0 if the control treatment is used.

Now under randomisation, it is expected that E(B|T=1) = E(B|T=0). Now if E(log(B+1)|T=1) = E(log(B+1)|T=0) occurs, the baseline value can be safely ignored in the analysis. The above equality is not always the case though and it is this situation that this thesis focuses on.

## 3.2 The models when no heterogeneity is present

Given the number of spots can be treated as a count, the Poisson models shown below are assumed to hold. Equation 1 is for the experimental group and equation 2 is for the control group.

$$\log(E(Y)) = \alpha + \beta \times T + \gamma \times E(\log(B+1)) \tag{1}$$

$$\log(E(Y)) = \alpha + \gamma \times E(\log(B+1)) \tag{2}$$

A unit increase is added to the baseline value to remove any issues caused by zero counts leading to undefined values when the natural logarithm is taken. Conditional on the baseline value and E(log(B+1)|T=1) = E(log(B+1)|T=0), the difference between equation 1 and 2 gives

$$\log(E(Y|T=1)) - \log(E(Y|T=0)) = \beta. \tag{3}$$

Now, as the difference between to logarithms can be written as a ratio, the term $\beta$ can be interpreted as the log relative risk between the treatment groups.

## 3.3 The models when heterogeneity is present

Now if E(log(B+1)|T=1) $\neq$ E(log(B+1)|T=0) the baseline value cannot be ignored and the models differ for the two treatment groups. The model for the experimental group (T=1) is

$$\log(E(Y)) = \alpha + \beta + \gamma \times E(\log(B+1)|T=1), \tag{4}$$

and the model for the control group (T=0) is

$$\log(E(Y)) = \alpha + \gamma \times E(\log(B+1)|T=0). \tag{5}$$

Now, the larger the difference between E(log(B+1)|T=1) and E(log(B+1)|T=0), the more heterogeneity there is and the more of an issue the heterogeneity causes. The true log relative risk is given by $\beta$ however, substracting equation 5 from 4 gives,

$$\beta + \gamma \times [E(\log(B+1)|T=1) - E(\log(B+1)|T=0)] = \beta + \gamma \times \delta. \tag{6}$$

where $\delta$ = E(log(B+1)|T=1) - E(log(B+1)|T=0). Thus, $\delta$ is a measure of the baseline heterogeneity present and how it affects the true relative risk.

## 3.4 How heterogeneity can influence the findings

Suppose that the hypothetical experimental cream studied here actually has a log relative risk of 0.4. This would mean that log($\beta$) equals 0.4 and the relative risk would approximately equal 1.5. Thus, a person receiving the experimental cream would, on average, suffer 1.5 times the number of spots as an individual on the control treatment. This is a situation where the experimental treatment is worse than the control and should therefore not be released onto the market. What would an analysis ignoring baseline

heterogeneity conclude? The estimated relative risks, when not adjusting for baseline, from various scenarios (varying amounts of imbalance between control and experimental groups) are displayed in Table 6 below. Note the value of $\gamma$ is taken to be 0.2 in these calculations.

**Table 6:** Effect of baseline heterogeneity when the true risk ratio is 1.4 and $\gamma = 0.2$.

| $\delta$ | **Estimated relative risk** |
|---|---|
| $-1$ | 1.2214 |
| $-2$ | 1 |
| $-3$ | 0.8187 |
| $-4$ | 0.6703 |
| $-5$ | 0.5488 |

Table 6 shows that in this situation, an imbalance of 3 or more (with the experimental group having the lower baseline) between the treatment groups at baseline would have led to an estimated relative risk below 1. This could then in turn lead to a harmful drug being released onto the market. This demonstrates the need for methods to be able to isolate the $\delta$ term from the $\beta$ term in equation 6.

# 4 Methodology

The main aim of this project is to demonstrate that baseline heterogeneity in the outcome variable, in longitudinal count data, should be accounted for. This will be achieved by analysing three data sets, firstly ignoring baseline heterogeneity, then accounting for it and comparing the results. Six different ways of adjusting for the heterogeneity will be studied to give an idea of the effectiveness of each method. A "Generalised Linear Model" (GLM) is the model type used to model these data sets, with and without adjustment for heterogeneity. Below is a description of GLMs and how they are built.

## 4.1 Generalised Linear Model

In this paper, the regression model used is known as a "Generalised Linear Model" (GLM). This is used because simpler regression models can only cope successfully with normally distributed outcome variables. GLMs can deal with normally distributed outcome variables, as well as binary outcome variables and other types (see Agresti, [2]).

### 4.1.1 Exponential Family of Distributions

GLMs are a group of models which can model outcome variables which belong to "the exponential family of distributions". A distribution belongs to the exponential family if the distribution can be written in the form (see Forbes, [13]),

$$f(y;\theta) = \exp(a(y)b(\theta) + c(y) + d(\theta)),$$

or equivalently written as,

$$f(y;\theta) = c'(y)d'(\theta)\exp(a(y)b(\theta)), \tag{7}$$

where $f$ is the density function, $c$ and $a$ are functions of the data, $d$ and $b$ are functions of $\theta$.

### 4.1.2 Example: The Bernoulli Distribution belongs to the Exponential Family of Distributions

Assume that $y$ has a Bernoulli distribution with parameter $p$. Below is the probability mass function for the Bernoulli distribution.

$$f(y|p) = p^y(1-p)^{1-y}$$

This can then be written in the following form,

$$f(y|p) = (1-p)e^{y\log(\frac{p}{1-p})}. \tag{8}$$

Thus giving,

$$c'(y) = 1, d'(p) = 1-p, a(y) = y, b(p) = \log(\frac{p}{1-p}).$$

The above shows that the Bernoulli distribution belongs to the exponential family of distributions, therefore a GLM can be used to model Bernoulli data.

### 4.1.3 Example: The Poisson Distribution belongs to the Exponential Family of Distributions

Assume that $y$ has a Poisson distribution with parameter $\theta$. Below is the probability mass function for the Poison distribution.

$$f(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

This can then be written in the following form.

$$f(y|\theta) = \frac{e^{y\log(\theta)}e^{-\theta}}{y!}. \tag{9}$$

Thus giving,

$$c'(y) = \frac{1}{y!}, d'(\theta) = e^{-\theta}, a(y) = y, b(\theta) = \log(\theta).$$

The above proves that the Poisson distribution belongs to the exponential family of distributions, therefore a GLM can be used to model Poisson data.

### 4.1.4 Structure of a Generalised Linear Model

All GLMs, regardless of distribution, have the same basic structure. They consist of 3 components which are the random component (determined by the distribution), the systematic component and the link function. The systematic component is also known as the linear predictor and is formed from

the explanatory variables. The link function is then used to link together the mean of the random component and the systematic component. In the case where the outcome variable (Y) has a Poisson distribution, the components of the GLM with p explanatory variables are (see Forbes, [14]),

- $$\text{random component} : Y|X \text{with distribution Po}(\mu)$$
- $$\text{systematic component} : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$
- $$\text{link function} : log(\mu)$$
- $$\text{model} : \mu = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}$$

### 4.1.5 Iterative methods for model parameter estimation

In GLMs, a method known as "maximum likelihood estimation" is used to estimate model parameters. This method involves maximising the likelihood function, however, this is often not feasible in reality in the sense that closed form solutions are not available. Thus, a type of numerical method is often used instead. Two common methods are the Newton-Raphson method (see Collett, [8]) and iterative weighted least squares (see Nelder and Wedderburn, [26]), and it is normally one of these two which is employed by computer stats packages such as "R", which is used in chapter 5 of this thesis.

### 4.1.6 Model selection criteria

There are various different ways to compare different GLMs. This subsection describes the 3 different methods used in this thesis. A formal test often used is the likelihood ratio test (LRT) which involves comparing the change in likelihood between models (see Hosmer et al., [19]). This difference in likelihood is cross referenced with a Chi-squared distribution to determine whether the addition of a variable significantly improves the model. The LRT can also be used to see whether the removal of a variable significantly harms the model. The degrees of freedom for the LRT is the difference in the number of parameters between the two models. A downside of the LRT is

that it can only compare nested models i.e. models which are subsets of each other. In the case of non-nested models, other methods such as the "Akaike Information Criterion" and the "Bayesian Information Criterion" are often used where the model with the lowest value is deemed the superior model (see Agressti, [1]). Given below are the formulas for AIC and BIC.

$$AIC = -2l + 2p \tag{10}$$

$$BIC = -2l + p\log(n), \tag{11}$$

where, $n$ is the sample size, $p$ is the number of parameters in the model and $l$ is the log-likelihood.

### 4.1.7 Residual Analysis

Once the superior model has been found, the model assumptions need to be checked before any interpretation can be performed. The assumptions are checked using residual analysis. As the name suggests, this involves using residuals. These are the differences between the predicted values (under the superior model) and the true values in the data set.

Many types of adjusted residuals exist and in this thesis standardised residuals are used unless stated otherwise. Standardised residuals are residuals that have been divided by the standard deviation of the residuals.

A first step is to produce a plot of standardised residuals and look at whether the plot shows any patterns. A plot which demonstrates a pattern suggests that the model has not been specified properly whereas, a plot with no pattern implies the correct type of model has been chosen.

The way standardised residuals are calculated, it is expected 5% of the standardised residuals will have magnitude greater than 1.96. If more than 5% of the standardised residuals have magnitude greater than 1.96, there is potentially an issue with outliers in the data. The topic of dealing with outliers is not considered here.

The linear predictor needs to be checked for whether any important variables have been missed out or any variables incorrectly specified. This is

done by plotting standardised residuals against the variables. Should the plots show any pattern then the linear predictor may have been wrongly specified.

Finally, any observations which have a large influence on the model parameters need to be identified. Many statistics exist for this too but the simplest two are used here which are known as leverages and cook's distances. These statistics are then plotted against observation number to see whether any observation has too much influence on the model parameters.

## 4.2 Six methods for adjusting baseline heterogeneity in the outcome variable

When including variables, there are numerous different ways this can be done. For example, the variable could be included as a main effect, within an interaction term or random effect. How the variable is included in the model also depends on whether the variable is numeric or not. This thesis is looking at how to include baseline measurements (of the outcome variable) in the model, in order to account for baseline heterogeneity for count data.

When conducting analysis of count data, there are two broad types of method that can be used known as Parametric methods and non-Parametric methods. As the name suggests, the Parametric methods use parameters to explain relationships in the data and typically take the form of regression models. This thesis is going to look at using 4 different Poisson regression models (overdispersion is tested using a negative binomial model). These models will differ in the way that the baseline count is included (via an Offset, Continuous covariate, Categorical covariate and a Random effect). These 4 variations of the Poisson model are being examined due to them using the most commonly used ways of including a baseline measurement in the analysis. On top of this, an adaptation of the conditional linear mixed model used by Verbeke is examined to see whether the same findings (absence of bias in treatment effects) hold in the scenario of count data. These don't cover all parametric methods and to do so would be infeasible. The above 5 do however, allow a varied analysis of the relative merits of the different ways (Offset, Continuous covariate, Categorical covariate, Random effect and a Time by Treatment interaction) of including the baseline measurement.

Finally, it was desirable to include a non-Parametric method to gain an understanding of the usefulness of such a method at accounting for baseline heterogeneity. The Mantel-Haneszel method is used here because it is the most commonly used non parametric method. Below is a description of the notation and in the preceding subsections, a description of each method is given.

The variable $Y_{ijs}$ is used to represent the outcome value for the $j^{th}$ individual, in the $i^{th}$ treatment group and the $s^{th}$ stratum of the baseline measurement for that individual.

The variable $T_i$ is an indicator as to which treatment group the individual belongs to.

$\log(B+1)_{ij}$ where B is the baseline measurement of the outcome variable for the $j^{th}$ individual, in the $i^{th}$ treatment group.

The continuous baseline values could be split into groups (strata) and the strata used in the model. This is done using the variable $I_s$ as an indicator variable which indicates which stratum the individual belongs to.

$\alpha$, $\beta_i$, $\gamma_s$ are regression coefficients representing a constant, the effect of the $s^{th}$ treatment and the effect of being in the $s^{th}$ stratum respectively.

### 4.2.1 Method 1: Poisson regression using an offset term

An offset term is a way of including an extra variable which influences the intercept, depending on the offset's value. The offset is included in the model in the same way as a numeric variable except that an offset has a regression coefficient fixed at one (see Dobson and Barnett, [11]).

Typically, Poisson regression is used to model count data, however, it is also possible to model rates (e.g. rates of occurrence). When modelling rates, the exposure times of the objects being observed need to be taken into account. It is sometimes difficult to maintain equal exposure times hence, an offset term representing exposure is used to establish equal exposure within the regression model (see Cummings, [9]).

The use of an offset term is investigated to determine whether an offset can adjust for baseline heterogeneity. The variable trialled as the offset is the baseline measurement for the outcome variable. Hence, this methodology looks to adjust for baseline heterogeneity within the outcome variable. Note, in the regression model the natural logarithm is taken, so there is a potential issue with baseline measurements taking the value zero. This is overcome by adding 1 to all baseline observations.

The model equation is shown below,

$$\log(E(Y_{ij})) = \alpha + \beta_i + \log(B_{ij} + 1), \tag{12}$$

where, i denotes the treatment group and j the individual. Note that the control group is shown using i=0 and $\beta_0 = 0$.

### 4.2.2 Method 2: Poisson regression using baseline measurements as a continuous covariate

This method includes the baseline measurements of the outcome variable as a continuous (numeric) variable. This is very similar to the use of an offset term. The difference is that when the adjustment is via a numeric variable, there is an estimated regression coefficient rather than the fixed parameter (value one) for the offset. Using a numeric variable for the adjustment tests whether a continuous baseline is appropriate. The model equation is shown below,

$$\log(E(Y_{ij})) = \alpha + \beta_i + \gamma \times \log(B_{ij} + 1), \tag{13}$$

where, i denotes the treatment group and j the individual. Note that the control group is shown using i=0 and $\beta_0 = 0$.

### 4.2.3 Method 3: Poisson regression using baseline measurements as a categorical covariate

The possibility of using the baseline measurements as a single categorical variable in order to adjust for baseline heterogeneity will also be investigated. This will involve creating a category for each unique baseline measurement which in practice could lead to far too many categories. A solution to this could be to group similar measurements together. The decision on what is

similar would probably be made in consultation with an expert in the field, such as a doctor in a medical situation. The model equation is shown below,

$$\log(E(Y_{ijs})) = \alpha + \beta_i + \gamma_s \tag{14}$$

where, i denotes the treatment group and j the individual. Note that the control group is shown using i=0 and $\beta_0 = 0$. $\gamma_s$ is the regression coefficient for the sth stratum

### 4.2.4 Method 4: Poisson regression with a random effect

This technique makes use of a mixed Poisson model. Mixed regression models contain both fixed and random effects. The adjustment for baseline heterogeneity in the outcome variable will be via a random effect term. This random effect term will have a normal distribution with zero mean and variance $\sigma_B^2$. The model equation is shown below,

$$\log(E(Y_{ijs})) = \alpha + \beta_i + \gamma_s. \tag{15}$$

where, i denotes the treatment group and j the individual. Note that the control group is shown using i=0 and $\beta_0 = 0$. $\gamma_s$ is the random effect for stratum s. The $\gamma_s$ each have a normal distribution with mean 0 and variance $\sigma_B^2$.

### 4.2.5 Method 5: Adaptation of the conditional linear mixed model

Research has taken place into how a conditional approach can account for baseline heterogeneity in normally distributed outcome data [44] [45]. This methodology cannot be directly transferred into our situation as the count data is not normally distributed. Thus, an adapted form is proposed here. This involves turning the count data which contains more than 1 time point (i.e the data is also longitudinal) into long format. This means each participant has as many rows of data as there are time points. Hence, if there are 2 time points, each participant has 2 rows of data. As part of turning the data into long format an indicator variable depicting the time point is formed. The equation of the adapted model is shown below,

$$\log(E(\mathrm{Y}_{ijt})) = \alpha + \alpha_j + \beta_i + \delta * t + \delta_i * t + \beta * \delta_i * t \tag{16}$$

In the above equaation $\alpha$ is the fixed intercept, $\alpha_j$ is the random intercept, $\beta_i$ is the fixed treatment effect, $\delta$ is the fixed time effect, $\delta_j$ is the random time effect and $\delta^*\beta_i$ is the interaction between treatment and time.

### 4.2.6 Method 6: Mantel-Haenszel Approach

The Mantel-Haenszel Approach will be the final method investigated and is an example of a non parametric method. All the methods discussed above are parametric methods. The Mantel-Haenszel method originally dates back to 1959 (see Mantel et Haenszel, [23]) but, a more modern general version will be used here (see Woodward, [49]). Risk ratios can be computed for each stratum using the formula,

$$RR_s = \frac{\sum_j \frac{Y_{1js} n_{0s}}{n_s}}{\sum_j \frac{Y_{0js} n_{1s}}{n_s}}. \tag{17}$$

Here, $n_{1s}$ is the number of individuals in the experimental treatment group within the $s^{th}$ stratum. $n_{0s}$ is the number of individuals in the control treatment group within the $s^{th}$ stratum. Also $n_{0s} + n_{1s} = n_s$. This does assume there is one experimental group. In the case of more experimental groups, a ratio is produced for each experimental group separately.

From here it is possible to derive the Mantel-Haenszel estimate of the overall risk ratio using the formula below. Note that the summations are taken before the ratio.

$$RR_{MH} = \frac{\sum_s [\sum_j \frac{Y_{1js} n_{0s}}{n_s}]}{\sum_s [\sum_j \frac{Y_{0js} n_{1s}}{n_s}]}. \tag{18}$$

Like with all statistics, an indication of the variability of the statistic is required. This comes in the form of the statistic's variance. The issue with the Mantel-Haenszel risk ratio is that no exact formula for the variance exists. Robins et al [35] proposed an estimator of the variance for an odds ratio which is consistent for both sparse and large strata sizes. It is claimed by Greenland and Robins [36] that the methodology used for the odds ratio can be extended to give an estimator of the variance for the log risk ratio.

Adapting the notation to match what is used here, the estimator of the variance for the log risk ratio is given below.

$$VAR(log(RR_{MH})) = \frac{\sum_s [\sum_j \frac{n_{1s}n_{0s}(Y_{1js}+Y_{0js})-Y_{1js}Y_{0js}n_s}{n_s^2}]}{\sum_s [\sum_j \frac{Y_{0js}n_{1s}}{n_s}] \sum_s [\sum_j \frac{Y_{1js}n_{0s}}{n_s}]}. \tag{19}$$

It is common to extend this to form a 95% confidence interval (CI) for the log risk ratio. The lower and upper bounds of this interval are given by the expression,

$$log(RR_{MH}) - 1.96 * \sqrt{VAR(log(RR_{MH}))} \tag{20}$$

$$log(RR_{MH}) + 1.96 * \sqrt{VAR(log(RR_{MH}))}. \tag{21}$$

This CI can be converted into a CI for the risk ratio by taking the exponential of the lower and upper bounds.

# 5 Analysing the motivating data sets

This chapter looks at applying the methods outlined in section 4.2 to the 2 data sets outlined in sections 1.2.1 and 1.2.2. This will lead to an examination as to whether these methods successfully adjust for baseline heterogeneity. Five of the methods to be used involve a Poisson regression model (the Negative Binomial versions are also explored to account for any overdispersion in the data). This choice of model assumes the outcome variables in both data sets have a Poisson distribution. As both outcome variables contain count data, this seems a reasonable assumption. The Poisson variables will contain all explanatory variable deemed to be significant using a test known as "The Likelihood Ratio Test". Given it is baseline heterogeneity that this project is focusing on, the baseline measurements are used in the models to adjust for baseline heterogeneity. Finally, every Poisson model is checked for overdispersion via the use of a Negative Binomial model.

## 5.1 Belcap data set

This data set has data on the number of tooth surfaces and the number of teeth with decay, missing or presence of filling (DMFS) at the start and end of the study. Along with this the participant's treatment group, ethnicity and gender were recorded.

This data set will be analysed taking DMFS-end as the outcome variable. Before modelling the data, exploratory data analysis is performed to better understand the make-up of the data set. Also, any potential outliers or errors in the data set may be picked up at this point.

### 5.1.1 Exploratory Data Analysis

Figure 1 below shows a boxplot of damaged tooth surfaces with DMFS at baseline, split by treatment group. The different treatment groups have reasonably similar spreads of data with very few outliers. The outliers which are present do not appear to be unreasonable measurements so they are left in the data set.

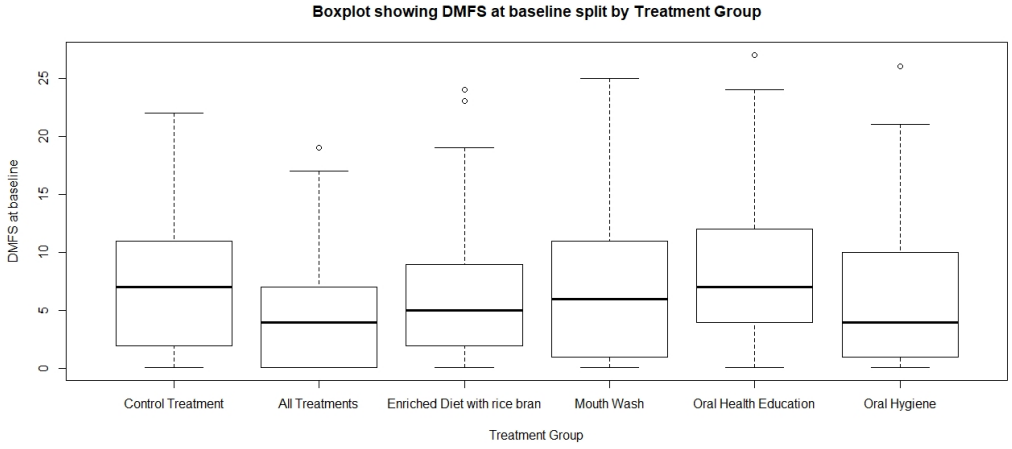


**Figure 1:** Boxplot showing the spread in the outcome variable at baseline split by treatment group.

Table 7 below shows the average DMFS scores for each treatment group at baseline. The standard deviation in scores is also given for each treatment group. Table 7 shows that the average DMFS score does vary slightly between treatment group which implies that in future analysis baseline heterogeneity may be an issue. The presence of this heterogeneity makes this dataset a good example for demonstrating how well the proposed 6 trial methods deal with this imbalance.

**Table 7:** Average DMFS score at baseline.

| **Treatment** | **n** | **Mean DMFS count** | **SD** |
|---|---|---|---|
| Control | 136 | 7.2994 | 5.6772 |
| ALL | 127 | 4.3535 | 4.3610 |
| ESD | 132 | 6.1136 | 5.4419 |
| MW | 155 | 6.6974 | 6.1806 |
| OHE | 124 | 7.7387 | 5.7890 |
| OHY | 123 | 5.5024 | 5.3823 |

Figure 2 below shows a boxplot of tooth surfaces with DMFS at the end of the study split by treatment group. The experimental groups have fairly

similar spread but the "ALL" group does appear to have the healthier values. The "Control" group has a wider spread than the experimental groups, indicating that there are still some participants in this group with many tooth surfaces with DMFS.

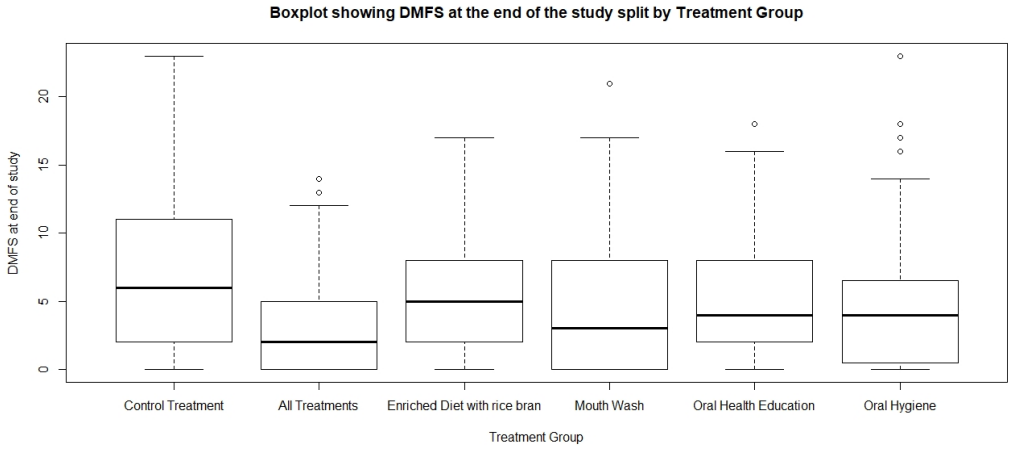


**Figure 2:** Boxplot showing the spread in the outcome variable at the end split by treatment group.

Examining the data shows most participants have lower DMFS values at the end than at the start. This implies that some tooth surfaces which had decay, in the sense of small lesions on the tooth surface, are now clear of problems.

Figure 3 below is a scatter plot which has a line with unit slope added to it. All the points lie below the line which means the average DMFS count has decreased for all treatment groups by the end of the study. The vertical gap between the line and the point shows how much the average DMFS in that particular group has changed. It is clear that the ALL group has the lowest average DMFS at the end but this could be partly due to this group having the smallest average at baseline. The treatment groups which have had the biggest effect are OHE and MW. The treatment group with the smallest change is the Control group.

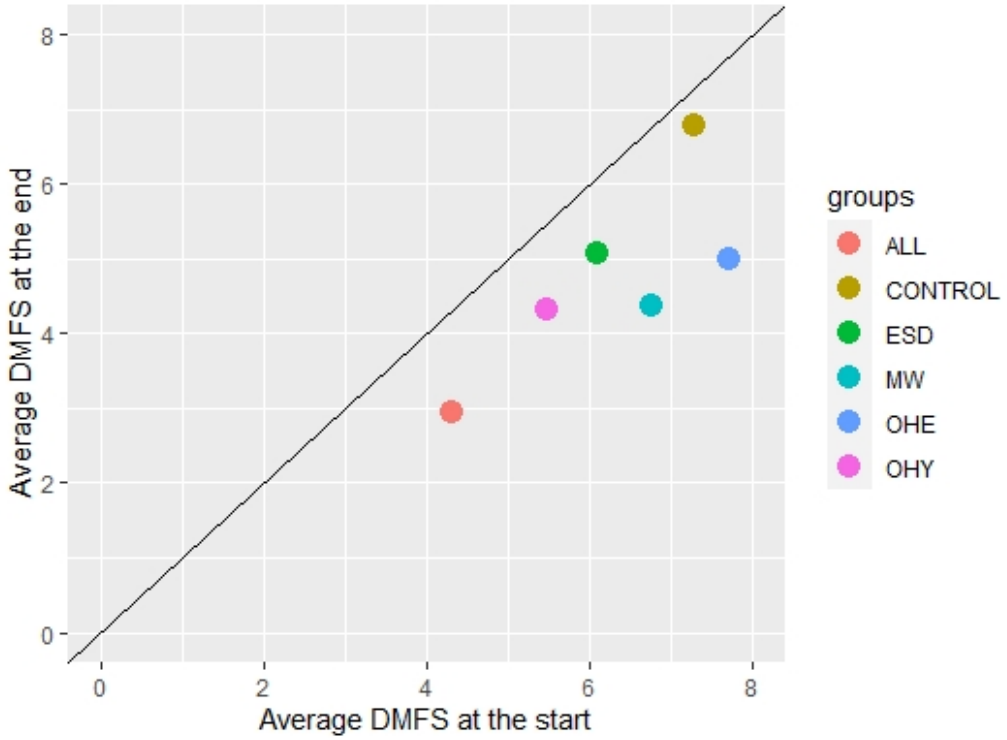


**Figure 3:** Scatter plot showing average DMFS count before and after the study for all treatment groups.

Table 8 shows how the average DMFS-end varies between treatment groups. Note that some of the 95 percent confidence intervals (CI) do not overlap between treatment groups. The P-value from a one way anova is also vertually 0. This implies that there is a difference between the treatment groups. This means treatment group would be expected to be significant at the 5 percent significance level when carrying out the modelling, ignoring all other explanatory variables.

**Table 8:** Average DMFS-end split by treatment group.

| Treat | n | Mean DMFS-end | SD | 95% CI |
|---|---|---|---|---|
| OHE | 124 | 4.9839 | 4.2846 | (4.2297 , 5.7380) |
| ALL | 127 | 2.9528 | 3.4037 | (2.3608 , 3.5447) |
| CONTROL | 136 | 6.7794 | 5.5320 | (5.8497 , 7.7092) |
| ESD | 132 | 5.0682 | 4.3027 | (4.3342 , 5.8022) |
| MW | 155 | 4.3806 | 4.5773 | (3.6600 , 5.1013) |
| OHY | 123 | 4.3089 | 4.5071 | (3.5124 , 5.1055) |

* A one way anova produces a p-value of vartually 0.

Table 9 shows how the average DMFS-end varies between different ethnicities. Note that all of the 95 percent CI overlap ( the p-value from a one way anova is 0.3650) hence, ethnicity is not expected to be significant at the 5 percent significance level when carrying out the modelling, ignoring all other explanatory variables.

**Table 9:** Average DMFS-end split by ethnicity.

| Ethnicity | n | Mean DMFS-end | SD | 95% CI |
|---|---|---|---|---|
| dark | 302 | 4.7848 | 4.5775 | (4.2685 , 5.3010) |
| white | 383 | 4.9269 | 4.6579 | (4.4604 , 5.3934) |
| black | 112 | 4.1161 | 4.3158 | (3.3168 , 4.9154) |

* A one way anova gives a p-value of 0.3650

Table 10 shows how the average DMFS-end varies between males and females. Note that the 95 percent CI just overlap however the t-test produces a highly significant p-value. There is therfore some evidence to suggest that gender is significant at the 5 percent significance level.

**Table 10:** Average DMFS-end split by gender.

| Gender | n | Mean DMFS-end | SD | 95% CI |
|---|---|---|---|---|
| Female | 389 | 4.3059 | 4.2755 | (3.8810 , 4.7308) |
| Male | 408 | 5.1912 | 4.8239 | (4.7231 , 5.6593) |

* A 2 sample t-test (due to equal variances and normality) leads to a p-value of 0.0063

### 5.1.2 Poisson Regression ignoring Baseline Heterogeneity

Poisson regression models are produced where only significant explanatory variables are used (likelihood ratio test is used to test significance). This results in a model which contains the variables treatment- group and gender. Table 11 below shows the model coefficients and their standard errors.

**Table 11:** Coefficients and standard errors from model with no adjustment for heterogeneity.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 1.8120 | 0.0386 | $< 0.0001$ |
| Gender (Male) | 0.1715 | 0.0328 | $< 0.0001$ |
| School (All) | $-0.8234$ | 0.0613 | $< 0.0001$ |
| School (ESD) | $-0.2797$ | 0.0508 | $< 0.0001$ |
| School (MW) | $-0.4148$ | 0.0507 | $< 0.0001$ |
| School (OHE) | $-0.2952$ | 0.0520 | $< 0.0001$ |
| School (OHY) | $-0.4456$ | 0.0545 | $< 0.0001$ |

The model equation produced from this methodology is below.

$$log(E(\text{DMFS-end})) = 1.81197 + 0.17149 * (\text{Male}) - 0.82335 * (\text{ALL})$$

$$-0.27970*(\text{ESD})-0.41477*(\text{MW})-0.29518*(\text{OHE})-0.44555*(\text{OHY}). \quad (22)$$

The terms Male represents a dummy variable for gender which takes the value 1 if the person is Male and 0 otherwise. The same logic is used for the other dummy variables (ALL, ESD, MW, OHE, OHY).

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 3493.2 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 4 below shows a "funnelling" meaning that the points are less concentrated on the right hand side of the plot. This suggests there could be something wrong with the model.

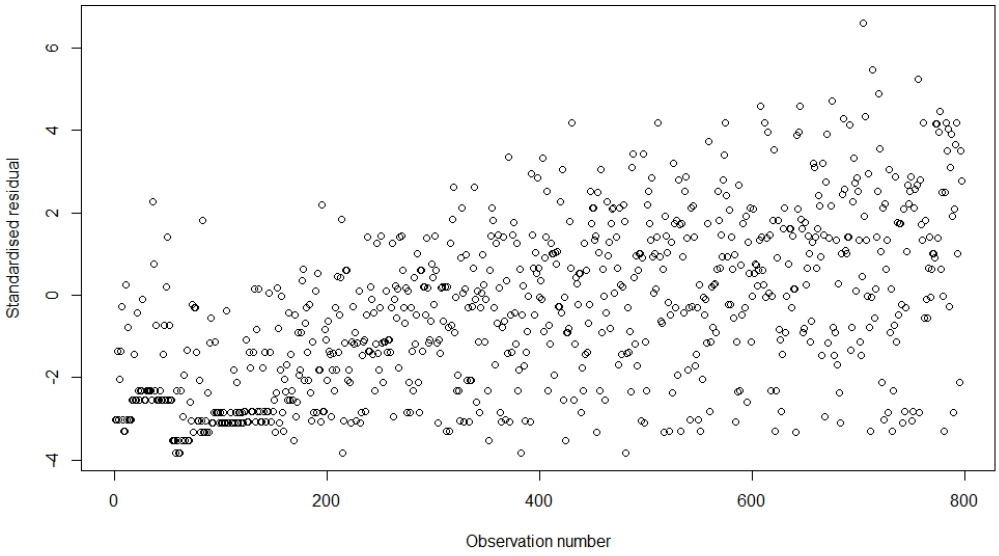


**Figure 4:** Scatter plot showing the standardised residual for each observation.

The Cook's distance of every observation is calculated and Figure 5 shows a plot of the Cook's distances. None of the observations have a Cooks's distance greater than 1 hence, none of the observations are influential.

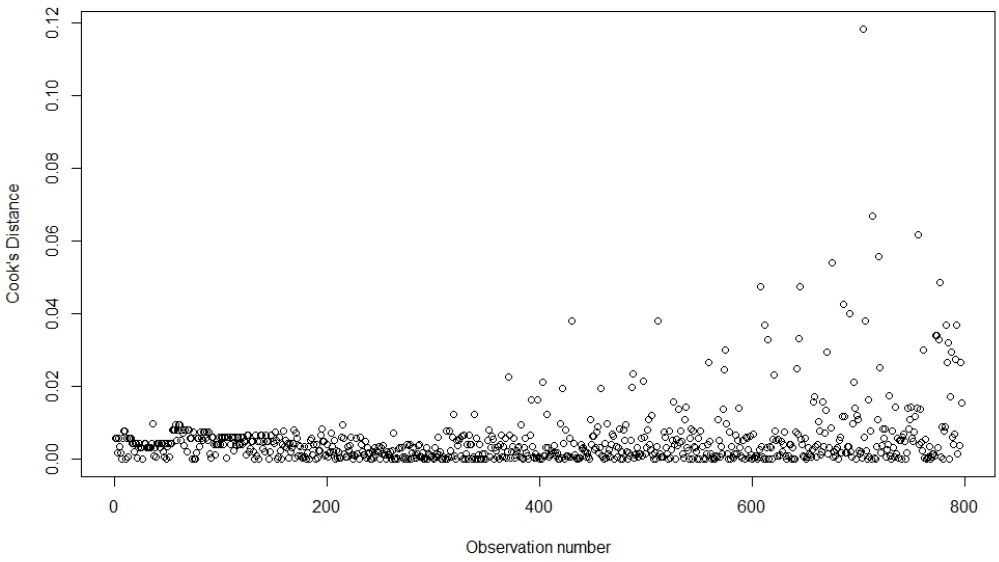


**Figure 5:** Scatter plot showing the Cook's distance of each observation.

Finally, the model is checked for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of virtually 0, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The test to look for overdispersion used here is the likelihood ratio test. This is a valid test as the poisson model and the negative binomial model are "nested". The poisson distribution is a special case of the negative binomial. In the case of the poisson the parameter theta is equal to infinity. The coefficients for this model are given in Table 12 below.

**Table 12:** Coefficients and standard errors from the negative binomial model with no adjustment for heterogeneity.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 1.749 79 | 0.0986 | $< 0.0001$ |
| Gender (Male) | 0.1745 | 0.0764 | 0.0500 |
| School (All) | $-0.8757$ | 0.1342 | $< 0.0001$ |
| School (ESD) | $-0.2509$ | 0.1270 | 0.0365 |
| School (MW) | $-0.4078$ | 0.1235 | 0.0012 |
| School (OHE) | $-0.3184$ | 0.1301 | 0.0296 |
| School (OHY) | $-0.5589$ | 0.1328 | 0.0012 |

Notice that the exponential of the coefficients represents the relative risk for that category compared with the reference category. The reference category for gender and treatment are female and control respectively. Thus the negative sign for all treatments show that all the treatments are more beneficial than the control treatment. Looking at the ALL category, participants have 0.42 times the risk (slightly less than half) of DMFS than participants in the CONTROL.

### 5.1.3 Model 1: Poisson Regression using baseline measurements as an offset

Poisson regression models are produced here with the DMFS-beg variable included as an offset term. The role of the offset term is to adjust for baseline heterogeneity. The inclusion of the offset caused gender to cease being significant at the 5 percent significance level. Significance is assessed via the likelihood ratio test. Thus, the coefficients and standard errors for the final model having adjusted for baseline heterogeneity are shown in Table 13.

**Table 13:** Coefficients and standard errors from model with the natural log of baseline measurements as an offset.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | −0.2008 | 0.0329 | $< 0.0001$ |
| School (All) | −0.3885 | 0.0613 | $< 0.0001$ |
| School (ESD) | −0.1361 | 0.0508 | 0.0074 |
| School (MW) | −0.3603 | 0.0506 | $< 0.0001$ |
| School (OHE) | −0.3593 | 0.0520 | $< 0.0001$ |
| School (OHY) | −0.2072 | 0.0545 | $< 0.0001$ |

The equation for the model produced is shown below.

$$\log(E((DMFS - end))) = -0.2008 - 0.3885 * (\text{ALL}) - 0.1361 * (\text{ESD})$$

$$-0.3603 * (\text{MW}) - 0.3593 * (\text{OHE}) - 0.2072 * (\text{OHY}) + \log(\text{DMFS-beg}). \quad (23)$$

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 3106.6 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 6 below shows a downward trend in standardised residual which further indicates possible misspecification.

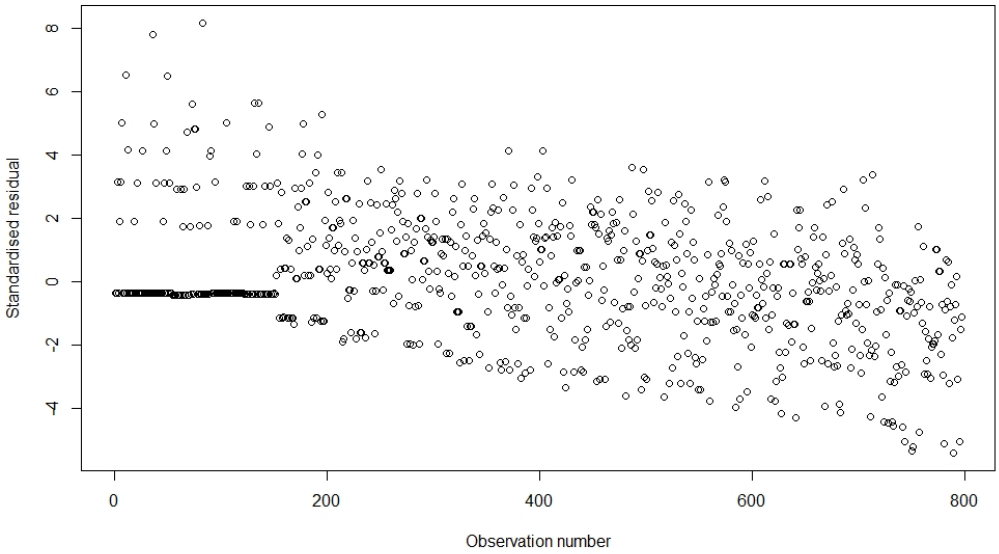


**Figure 6:** Scatter plot showing the standardised residual for each observation from method 1.

The Cook's distance of every observation is calculated and Figure 7 shows a plot of the Cook's distances. All the Cook's distances are below 1 hence, none of the observations are deemed to be influential.

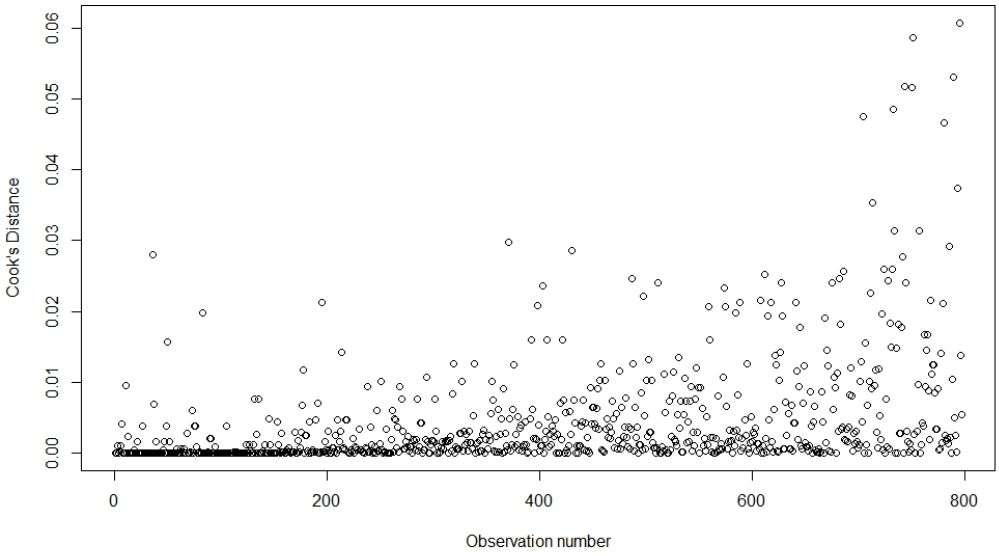


**Figure 7:** Scatter plot showing the Cook's distance of each observation from method 1.

Finally, the model is checked for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 14 below.

**Table 14:** Coefficients and standard errors from the negative binomial model with an offset adjustment for heterogeneity.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | −0.1293 | 0.0743 | 0.0819 |
| School (All) | −0.3601 | 0.1157 | 0.0019 |
| School (ESD) | −0.0199 | 0.1074 | 0.8533 |
| School (MW) | −0.3446 | 0.0011 | 0.0011 |
| School (OHE) | −0.3034 | 0.0055 | 0.0055 |
| School (OHY) | −0.1503 | 0.1803 | 0.1803 |

The negative sign for all the treatments shows that the treatments are more beneficial than the control. Looking at the ALL category, participants have 0.6868 times the risk of DMFS than participants in the CONTROL.

Figure 8 below shows a scatter graph comparing the risk ratios produced by ignoring baseline heterogeneity and those produced via method 1. The points above the superimposed line show the risk ratios which were increased when the offset term was used to adjust for heterogeneity. This applies to all Treatment Groups other than OHE. The risk ratio for OHE lies below the superimposed line meaning the offset adjustment caused the risk ratio to decrease.

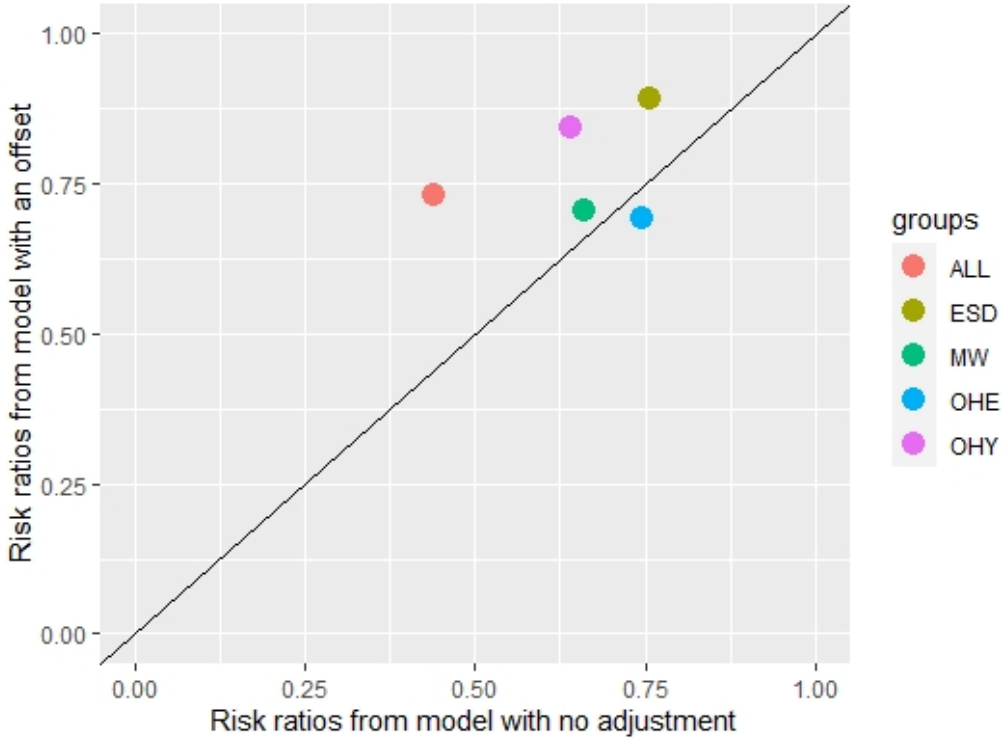


**Figure 8:** Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the offset adjustment and when no adjustment is done.

An increase in risk ratio, from using the offset term to adjust for baseline

heterogeneity, for all Treatment Groups excluding OHE means the benefit of the treatments (excluding OHE) is less than originally thought. Whereas the decrease in risk ratio for the OHE Treatment Group means the benefit of this treatment is higher than originally thought.

### 5.1.4 Method 2: Poisson Regression using baseline measurements as a continuous covariate

In this subsection, the baseline measurements are included in the Poisson regression as a continuous covariate. This is potentially another way of dealing with any heterogeneity in the data set. The coefficients from the model are shown below in Table 15 along with their standard errors.

**Table 15:** Coefficients and standard errors from model where baseline measurements are used as a continuous covariate.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 0.4795 | 0.0611 | $< 0.0001$ |
| Gender (Male) | 0.0666 | 0.0331 | 0.0441 |
| School (All) | $-0.5029$ | 0.0620 | $< 0.0001$ |
| School (ESD) | $-0.1736$ | 0.0509 | $< 0.0001$ |
| School (MW) | $-0.3530$ | 0.0509 | $< 0.0001$ |
| School (OHE) | $-0.3423$ | 0.0521 | $< 0.0001$ |
| School (OHY) | $-0.2613$ | 0.0547 | $< 0.0001$ |
| DMFS-beg+1 | 0.6863 | 0.0219 | $< 0.0001$ |

The equation produced from this methodology is shown below.

$$\log(E((DMFS{-}end))) = 0.4795{+}0.0666{*}(\text{Male}){-}0.5029{*}(\text{ALL}){-}0.1736{*}(\text{ESD})$$

$$-0.3530{*}(\text{MW}){-}0.3423{*}(\text{OHE}){-}0.2613{*}(\text{OHY}){+}0.6863{*}\log(\text{DMFS-beg} + 1)$$

.

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 2267.0 which is much higher than the

residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 9 below does not show any trend (no issue with non-constant variances).

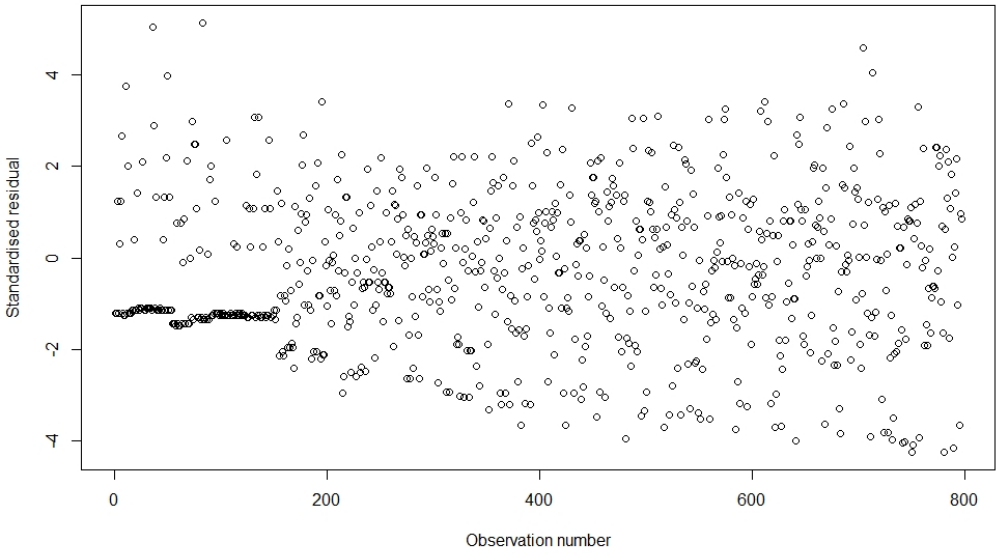


**Figure 9:** Scatter plot showing the standardised residual for each observation from method 2.

The Cook's distance of every observation is calculated and Figure 10 shows a plot of the Cook's distances. None of the Cook's distances have a value greater than 1 hence, there are no influential observations.

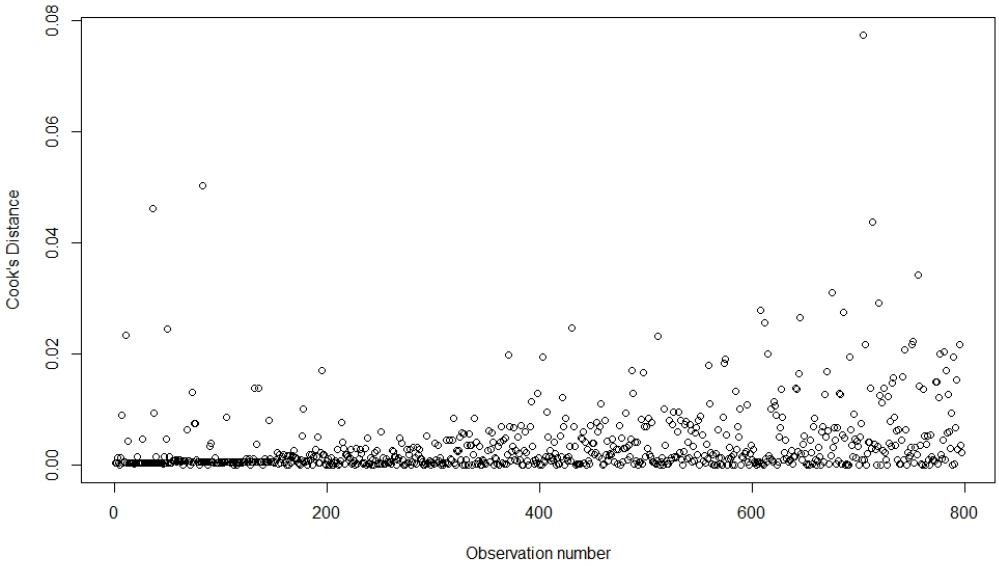


**Figure 10:** Scatter plot showing the Cook's distance of each observation from method 2.

Finally, the model is checked for overdispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value smaller than 0.05, hence, there is sufficient evidence of model overdispersion. Thus, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 16 below. It is worth noting that gender ceased to be significant in this analysis.

**Table 16:** Coefficients and standard errors from the negative binomial model with a continuous adjustment for heterogeneity.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 0.3786 | 0.0743 | $< 0.0001$ |
| School (All) | $-0.4664$ | 0.1157 | $< 0.0001$ |
| School (ESD) | $-0.0933$ | 0.1074 | 0.3573 |
| School (MW) | $-0.3575$ | 0.1058 | 0.0003 |
| School (OHE) | $-0.2982$ | 0.1092 | 0.0038 |
| School (OHY) | $-0.2262$ | 0.1121 | 0.0328 |
| DMFS beg + 1 | 0.7423 | 0.1121 | $< 0.0001$ |

Figure 11 below shows a scatter graph comparing the risk ratios produced by ignoring baseline heterogeneity and those produced via method 2. The points above the superimposed line show the risk ratios which were increased when the continuous adjustment was used to adjust for heterogeneity. This applies to all Treatment Groups other than OHE. The risk ratio for OHE lies below the superimposed line meaning the continuous adjustment caused the risk ratio to decrease.

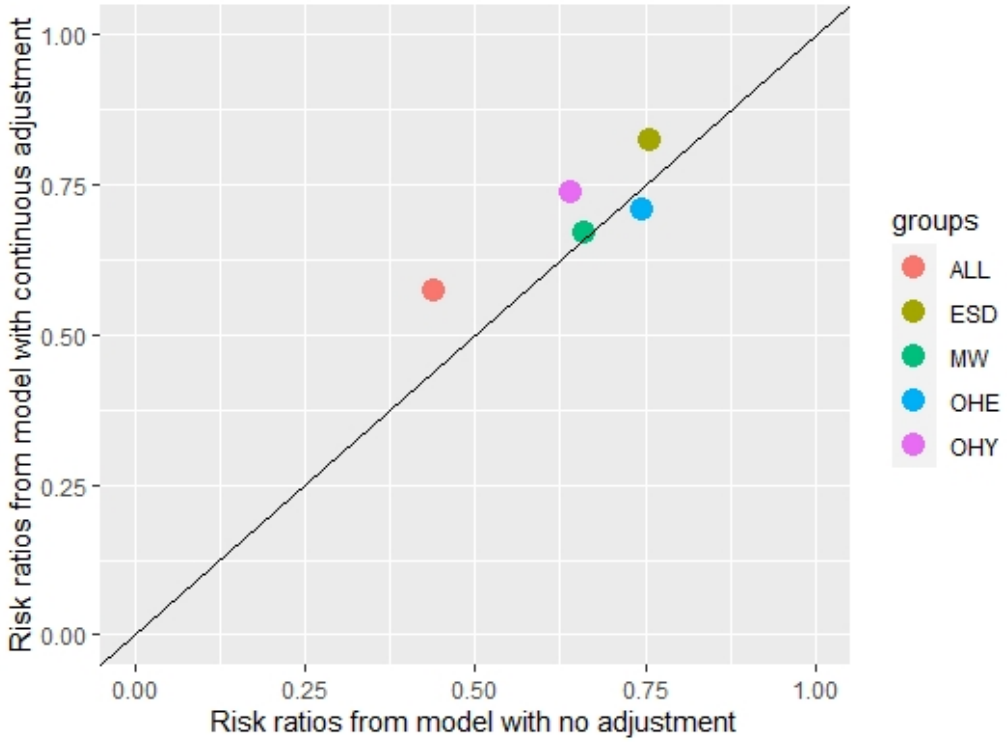


**Figure 11:** Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the continuous adjustment and when no adjustment is done.

### 5.1.5 Method 3: Poisson Regression using baseline measurements as a categorical covariate

In this subsection, the baseline measurements are included in the Poisson regression as a categorical covariate. This means that the values of the baseline variable have to be grouped. Ideally, in a medical situation like this, the decision of which values to group together would be made in consultation with a medical professional such as a doctor (or dentist in the case of teeth). This was not possible here, so the values for baseline have been grouped to try and produce even groups in terms of size. Table 17 below shows the groups produced for baseline and the sample sizes in each group.

**Table 17:** Sample sizes of the baseline DMFS groups.

| Baseline values in group | Sample Size |
|---|---|
| 0 | 151 |
| 1-2 | 118 |
| 3-5 | 144 |
| 6-8 | 133 |
| 9-12 | 130 |
| 13+ | 121 |

The grouping of baseline measurements into groups and then using these groups within the model, is potentially another way of dealing with any heterogeneity in the data set. The coefficients from this model are shown below in Table 18 along with their standard errors.

**Table 18:** Coefficients and standard errors from model where baseline measurement groups are used as a categorical covariate.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 0.1376 | 0.0949 | 0.1469 |
| Gender (Male) | 0.0797 | 0.0331 | 0.0160 |
| School (All) | −0.5390 | 0.0624 | $< 0.0001$ |
| School (ESD) | −0.1915 | 0.0510 | 0.0002 |
| School (MW) | −0.3415 | 0.0511 | $< 0.0001$ |
| School (OHE) | −0.3410 | 0.0521 | $< 0.0001$ |
| School (OHY) | −0.2581 | 0.0549 | $< 0.0001$ |
| DMFS-beg (1-2) | 1.0159 | 0.1044 | $< 0.0001$ |
| DMFS-beg (3-5) | 1.5459 | 0.0958 | $< 0.0001$ |
| DMFS-beg (6-8) | 1.8635 | 0.0941 | $< 0.0001$ |
| DMFS-beg (9-12) | 2.0131 | 0.0933 | $< 0.0001$ |
| DMFS-beg (13+) | 2.2023 | 0.0927 | $< 0.0001$ |

The equation produced from this methodology is shown below.

$$\log(E((DMFS{-}end))) = 0.1376{+}0.0797{*}(\text{Male}){-}0.5390{*}(\text{ALL}){-}0.1915{*}(\text{ESD})$$
$$-0.3415{*}(\text{MW}){-}0.3410{*}(\text{OHE}){-}0.2581{*}(\text{OHY}){+}1.0159{*}(\text{1-2}){+}1.5459{*}(\text{3-5})$$

$$+1.8635 * (6\text{-}8) + 2.0131 * (9\text{-}12) + 2.2023 * (13+)$$

.

The deviance for the model is 2295.9 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 12 below does not show any trend (no issue with non-constant variances).

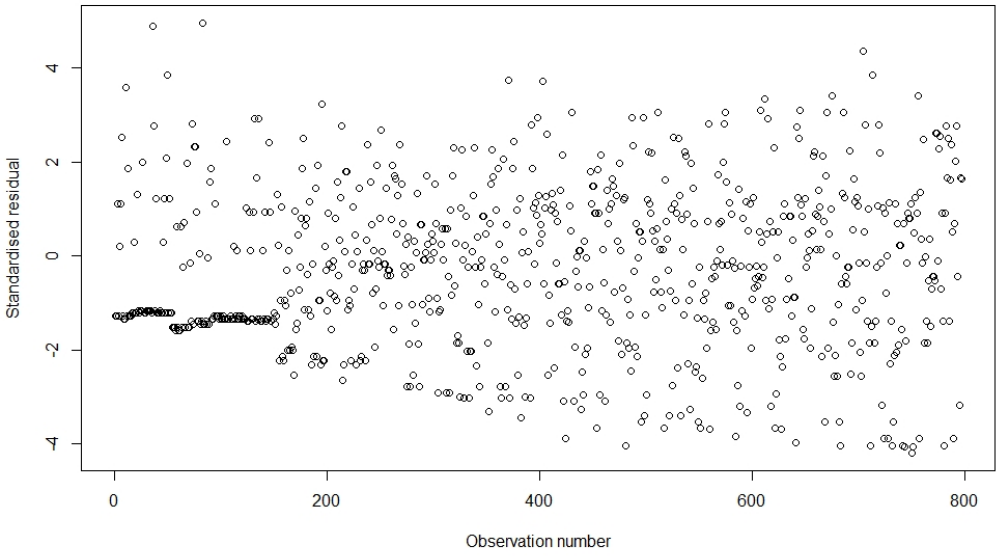


**Figure 12:** Scatter plot showing the standardised residual for each observation from method 3.

The Cook's distance of every observation is calculated and Figure 13 shows a plot of the Cook's distances. None of the observations have a Cook's distance greater than 1 hence, there are no issues with influential observations.

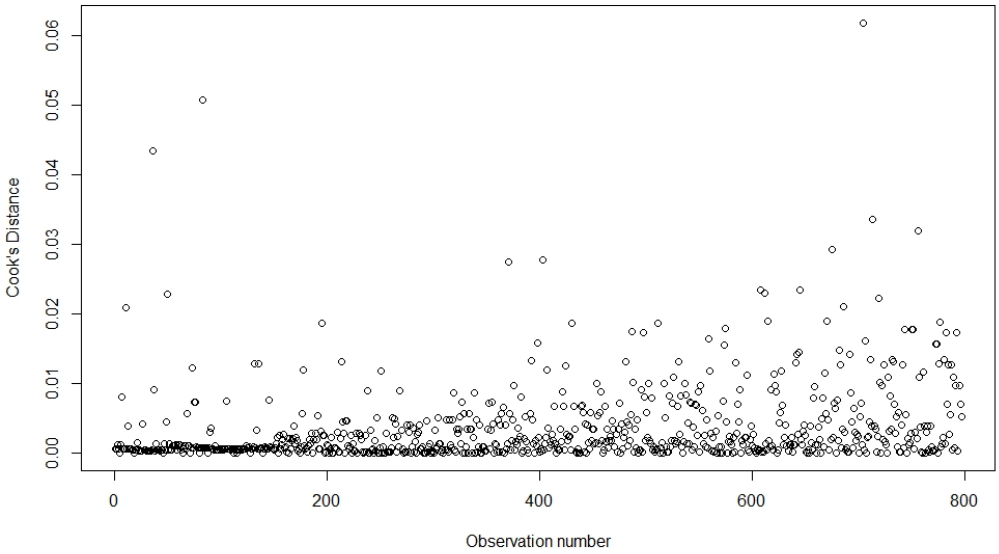


**Figure 13:** Scatter plot showing the Cook's distance of each observation from method 3.

Finally, the model is checked for overdispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value smaller than 0.05, hence, there is sufficient evidence of model overdispersion. Thus, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 19 below. It is worth noting that gender ceased to be significant in this analysis.

**Table 19:** Coefficients and standard errors from a negative binomial model where baseline measurement groups are used as a categorical covariate.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 0.1533 | 0.1234 | 0.2143 |
| School (All) | −0.4949 | 0.1109 | $< 0.0001$ |
| School (ESD) | −0.1247 | 0.1013 | 0.2180 |
| School (MW) | −0.3648 | 0.0996 | 0.0003 |
| School (OHE) | −0.2940 | 0.1028 | 0.0042 |
| School (OHY) | −0.2203 | 0.1059 | 0.0375 |
| DMFS-beg (1-2) | 1.0070 | 0.1343 | $< 0.0001$ |
| DMFS-beg (3-5) | 1.5389 | 0.1248 | $< 0.0001$ |
| DMFS-beg (6-8) | 1.8703 | 0.1246 | $< 0.0001$ |
| DMFS-beg (9-12) | 2.0170 | 0.1248 | $< 0.0001$ |
| DMFS-beg (13+) | 2.2081 | 0.1254 | $< 0.0001$ |

Figure 14 below shows a scatter graph comparing the risk ratios produced by ignoring baseline heterogeneity and those produced via method 3. The points above the superimposed line show the risk ratios which were increased when the continuous adjustment was used to adjust for heterogeneity. This applies to all Treatment Groups other than OHE. The risk ratio for OHE lies below the superimposed line meaning the continuous adjustment caused the risk ratio to decrease.

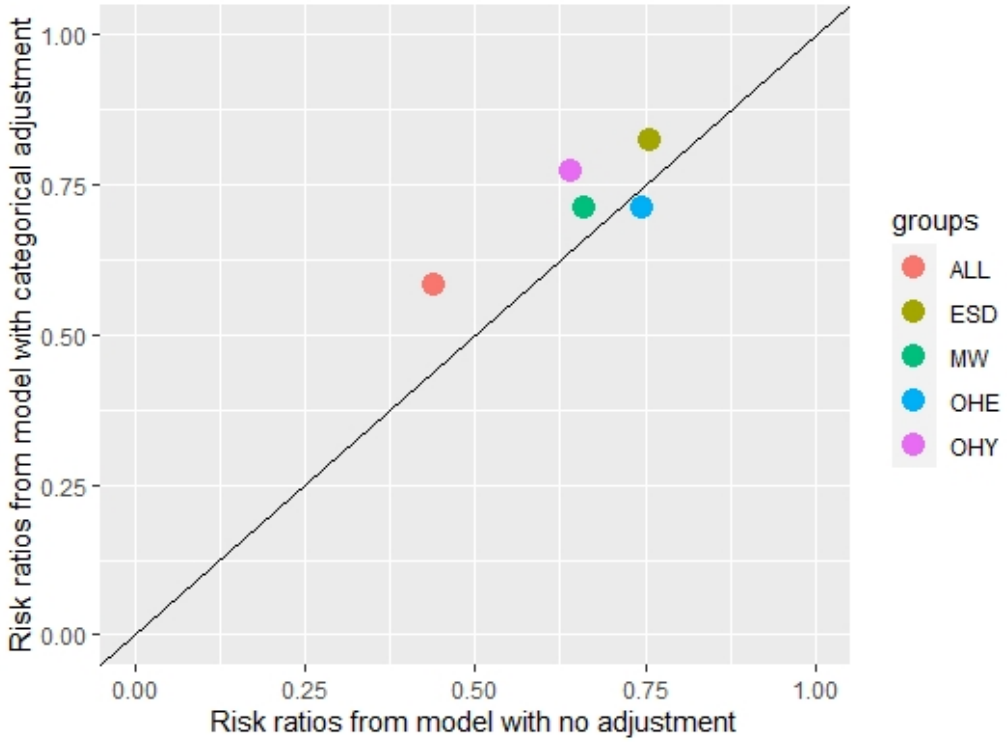


**Figure 14:** Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the categorical adjustment and when no adjustment is done.

### 5.1.6 Method 4: Poisson Regression using baseline measurements as a random effect

In this subsection, the baseline measurements are used as a random effect in the Poisson model. The role of this random effect is to account for baseline heterogeneity. When fitting this model, the log likelihood is approximated using the Adaptive Gauss-Hermite approximation with twenty points per axis. Table 20 below shows the coefficients from this GLMM model.

**Table 20:** Coefficients and standard errors from model where baseline measurement are used as a random effect.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 1.7934 | 0.2572 | $< 0.0001$ |
| Gender (Male) | 0.0712 | 0.0331 | 0.0313 |
| School (All) | $-0.5768$ | 0.0623 | $< 0.0001$ |
| School (ESD) | $-0.1914$ | 0.0510 | 0.0002 |
| School (MW) | $-0.3498$ | 0.0511 | $< 0.0001$ |
| School (OHE) | $-0.3531$ | 0.0521 | $< 0.0001$ |
| School (OHY) | $-0.2709$ | 0.0548 | $< 0.0001$ |

The deviance for the model is 4538.3 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 15 below does not show any trend (no issue with non-constant variances).

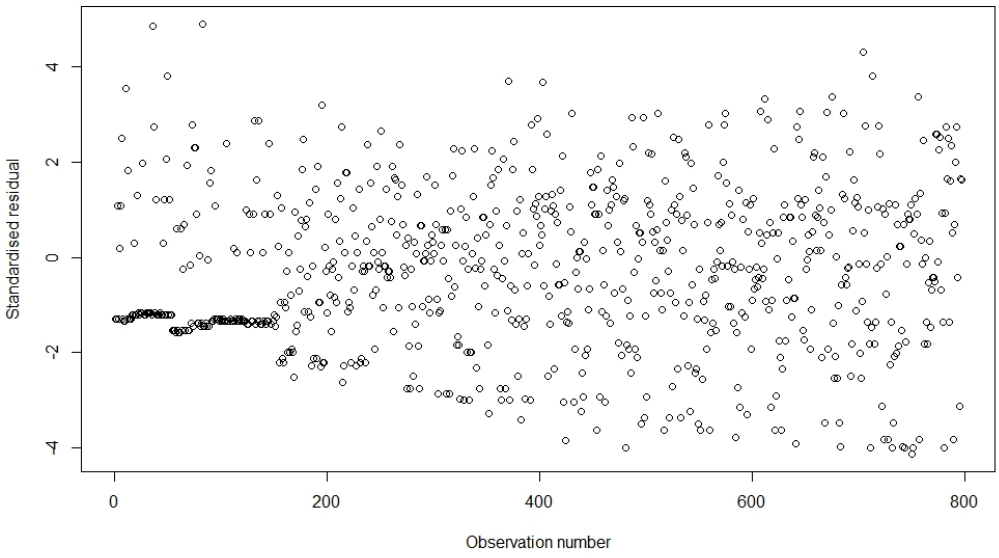


**Figure 15:** Scatter plot showing the standardised residual for each observation from method 4.

In the models used in methods 1-3 are all examples of generalised linear models (GLMs) where method 4 is a generalised linear mixed model (GLMM). There are concerns about the validity of Cook's distances for GLMMs, for example, the statistical software package R gives a warning should Cook's distances be calculated for such a model. As this thesis is mainly concerned with adjusting for baseline heterogeneity and not residuals analysis, an alternative method is not looked for. Thus, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 21 below. It is worth noting that gender ceases to be significant.

**Table 21:** Coefficients and standard errors from the negative binomial model with a random adjustment for heterogeneity.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 1.8175 | 0.2658 | $< 0.0001$ |
| School (All) | $-0.5712$ | 0.1140 | $< 0.0001$ |
| School (ESD) | $-0.1305$ | 0.1057 | 0.2168 |
| School (MW) | $-0.3809$ | 0.1022 | 0.0002 |
| School (OHE) | $-0.3268$ | 0.1066 | 0.0022 |
| School (OHY) | $-0.2503$ | 0.1096 | 0.0223 |

Figure 16 below shows a scatter graph comparing the risk ratios produced by ignoring baseline heterogeneity and those produced via method 4. The points above the superimposed line show the risk ratios which were increased when the continuous adjustment was used to adjust for heterogeneity. This applies to all Treatment Groups other than OHE. The risk ratio for OHE lies below the superimposed line meaning the continuous adjustment caused the risk ratio to decrease.

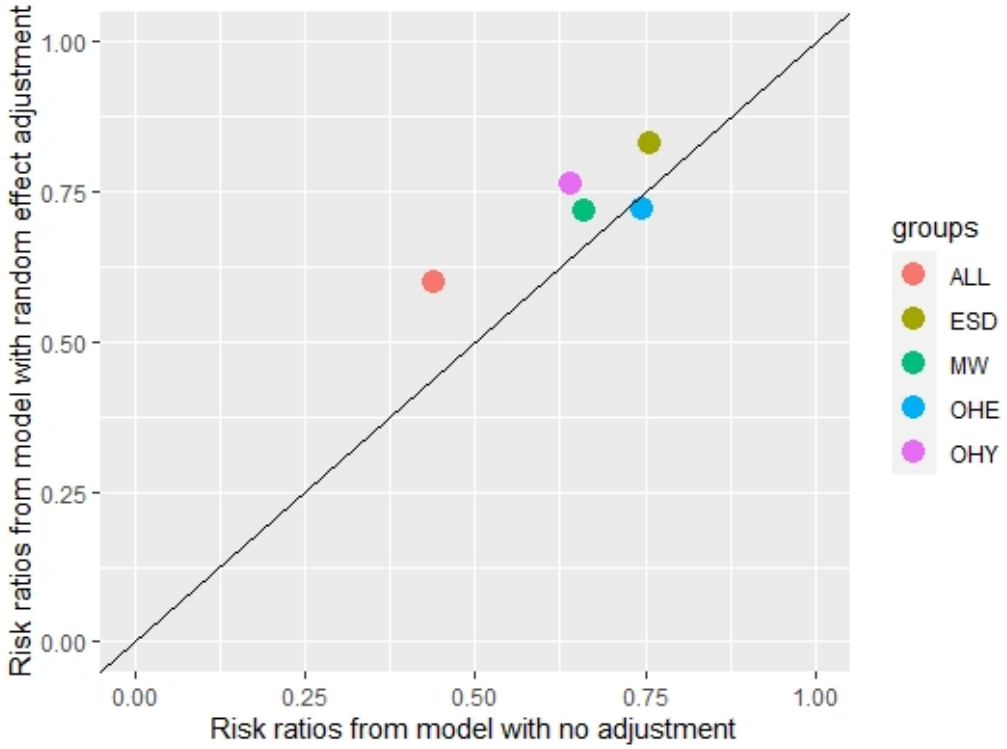


**Figure 16:** Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the random effect adjustment and when no adjustment is done.

### 5.1.7 Method 5: Adaptation of the conditional linear mixed model

For this method, the data has to be converted into long format. For the Belcap study, this means every participant has two rows of data, one for baseline and the other for the end point. An indicator variable is also formed to indicate whether the data is from baseline or the end point. Likelihood ratio tests are used to assess the significance of any covariates. Table 22 below shows the coefficients and standard errors from this model.

**Table 22:** Coefficients and standard errors from conditional model.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 1.8731 | 0.0741 | $< 0.0001$ |
| School (All) | $-0.4780$ | 0.1095 | $< 0.0001$ |
| School (ESD) | $-0.1537$ | 0.1060 | 0.1471 |
| School (MW) | $-0.0833$ | 0.1018 | 0.4133 |
| School (OHE) | 0.0814 | 0.1064 | 0.4441 |
| School (OHY) | $-0.2873$ | 0.1090 | 0.0084 |
| time | $-0.3197$ | 0.0709 | $< 0.0001$ |
| time : School (All) | $-0.4474$ | 0.1130 | 0.0001 |
| time : School (ESD) | $-0.1005$ | 0.1026 | 0.3274 |
| time : School (MW) | $-0.5004$ | 0.1014 | $< 0.0001$ |
| time : School (OHE) | $-0.3600$ | 0.1034 | 0.0005 |
| time : School (OHY) | $-0.2352$ | 0.1079 | 0.0292 |

The deviance for the model is 8665.6 which is much higher than the residual degrees of freedom. Thus, there is strong evidence this model doesn't fit the data well. Figure 17 below does not show any clear trend (no issue with non-constant variances).

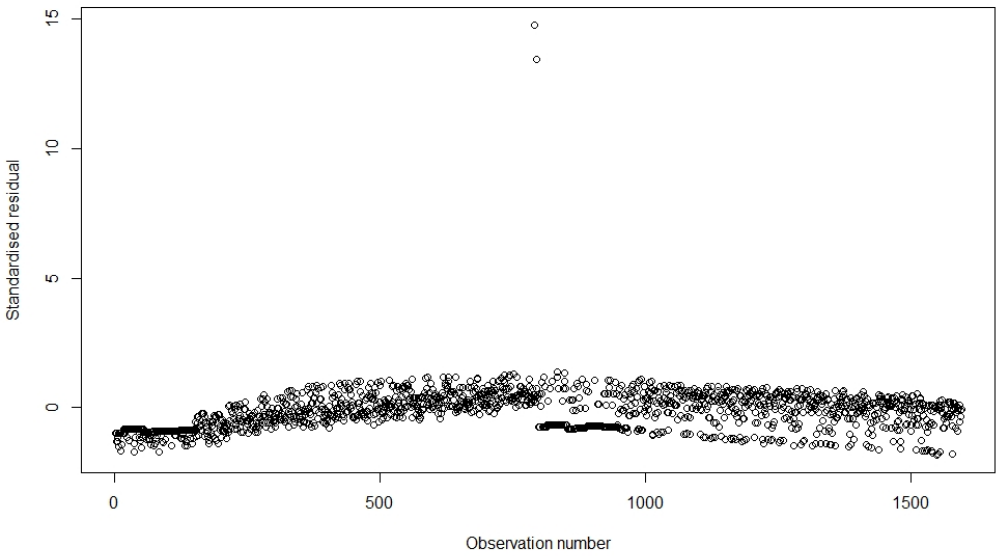


**Figure 17:** Scatter plot showing the standardised residual for each observation from method 5.

This method is also an example of a GLMM which means Cook's distances are not available. Like with method 4 above, the only way left to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value above 0.05, hence, the Poisson model is suitable.

Figure 18 below shows a scatter graph comparing the risk ratios produced by ignoring baseline heterogeneity and those produced via method 4. The points above the superimposed line show the risk ratios which were increased when the continuous adjustment was used to adjust for heterogeneity. This applies to all Treatment Groups other than OHE. The risk ratio for OHE lies below the superimposed line meaning the continuous adjustment caused the risk ratio to decrease.

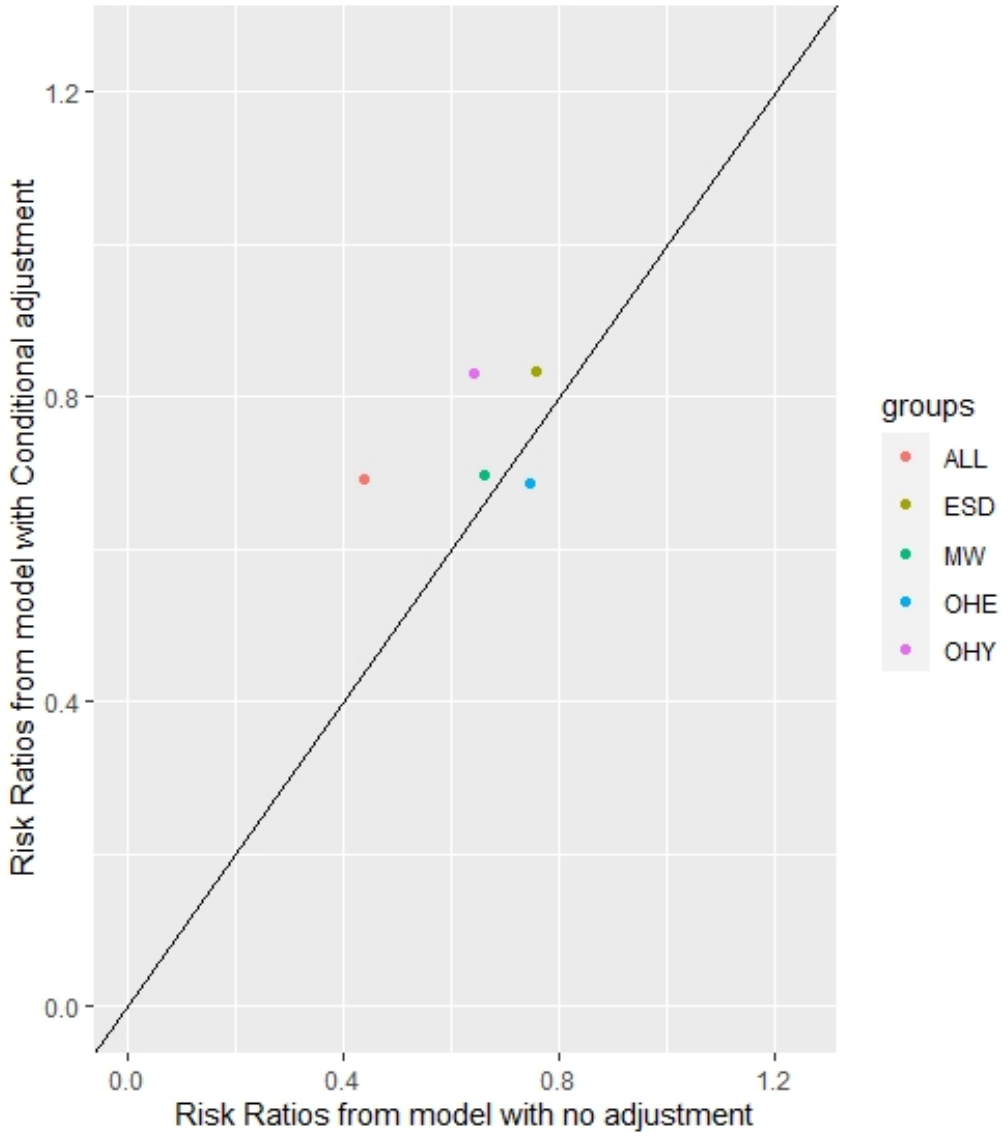


**Figure 18:** Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for method 5 and when no adjustment is done.

### 5.1.8 Method 6: The Mantel-Haenszel Approach

The Mantel-Haenszel approach is a non-parametric method so, there is no model equation produced like with the other methods. Instead, the Mantel-

Haenszel formula (given in section 4.2.5) is used to produce a risk ratio for being in an experimental group relative to the control group. In this example, there are 5 experimental groups hence, there will be 5 risk ratios. Table 23 below shows the estimated risk ratios for each experimental group.

**Table 23:** Mantel-Haenszel estimates of the risk ratio for each experimental group

| **Experimental Group** | **Risk Ratio** |
|---|---|
| ALL | 0.6307 |
| ESD | 0.8827 |
| MW | 0.8828 |
| OHE | 0.8143 |
| OHY | 1.0164 |

Table 23 suggests that the treatment "ALL" gives the best results i.e. lowest risk of DMFS where the treatment "OHY" gives the worst results. The risk ratio for "OHY" being greater than 1 implies that this treatment is worse than the control however, the difference between 1 and 1.0164 is minimal and probably would not meet statistical significance. It is also worth noting that the treatments "OHE", "ESD","MW" all have similar risk ratios implying these treatments are almost equally effective.

Figure 19 below shows a scatter graph comparing the risk ratios produced by ignoring baseline heterogeneity and those produced via method 5. The points above the superimposed line show the risk ratios which were increased when the continuous adjustment was used to adjust for heterogeneity. This applies to all Treatment Groups other than OHE. The risk ratio for OHE lies below the superimposed line meaning the Mantel-Haenszel adjustment caused the risk ratio to decrease.

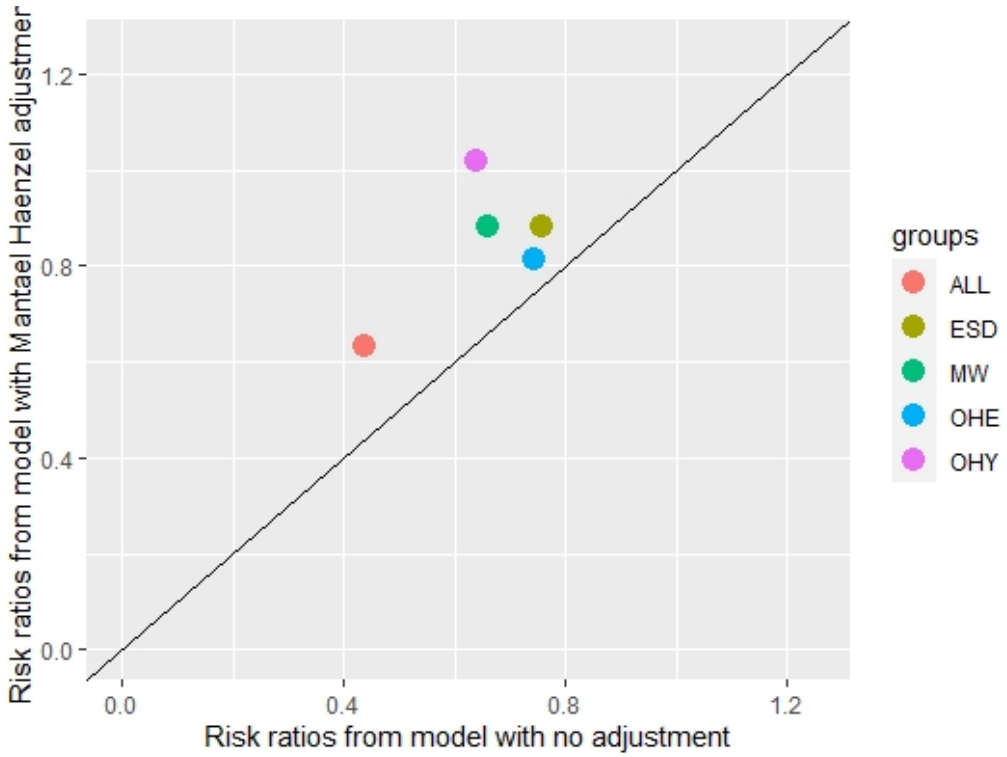


**Figure 19:** Scatter plot showing the risk ratio, for each experimental treatment relative to the control, for the Mantel-Haenszel method and when no adjustment is done.

## 5.2 Analysing the Polyps data set

Data is collected from a study looking at whether a treatment influences the number of polyps after three months. The data is analysed to see whether the experimental treatment has a significantly larger difference on the number polyps than the placebo. Along with the count of polyps after three months, the count is taken at baseline and the participant's treatment group, age and gender is recorded. Poisson regression models with and without an offset term are produced in order to assess whether the offset term successfully adjusts for heterogeneity.

### 5.2.1 Exploratory Data Analysis

Looking at Figure 20, there is clearly one outlier in each treatment group which could heavily bias any results. For this reason, these two observations are removed from the analysis.

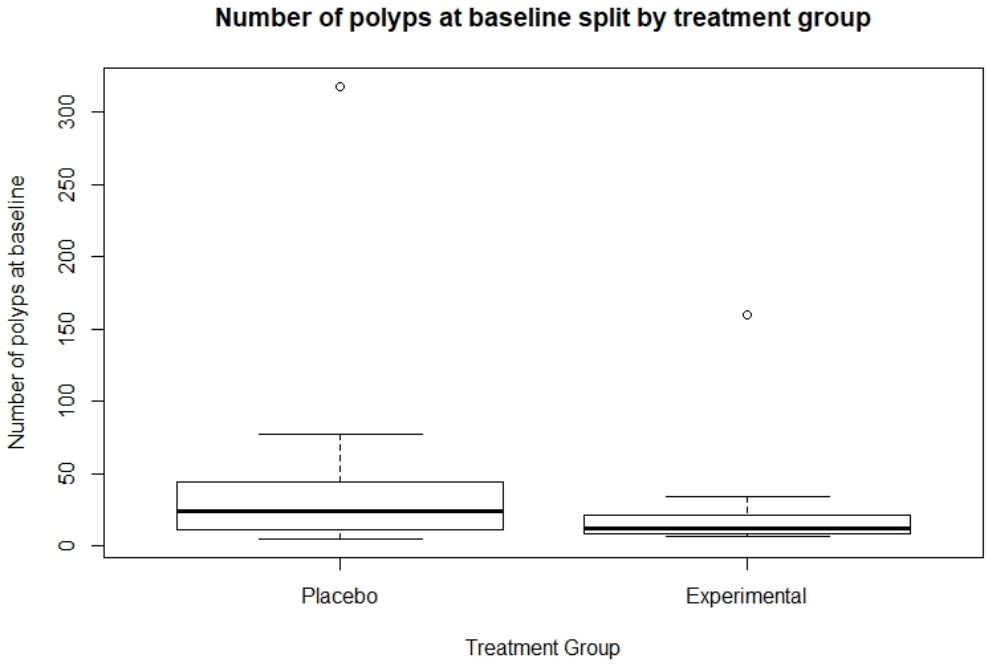


**Figure 20:** Boxplot showing the spread in the outcome variable at baseline split by treatment group.

Figure 21 is an identical boxplot to Figure 20 except for the outliers already mentioned having been removed. There is now a new observation classed as an outlier. This observation is not removed as it is perfectly feasible that this is an accurate measurement and was not previously flagged as an outlier. Figure 21 also shows a much smaller spread of data for the experimental group. In addition, participants in the experimental group appear to have a lower count of polyps at baseline.

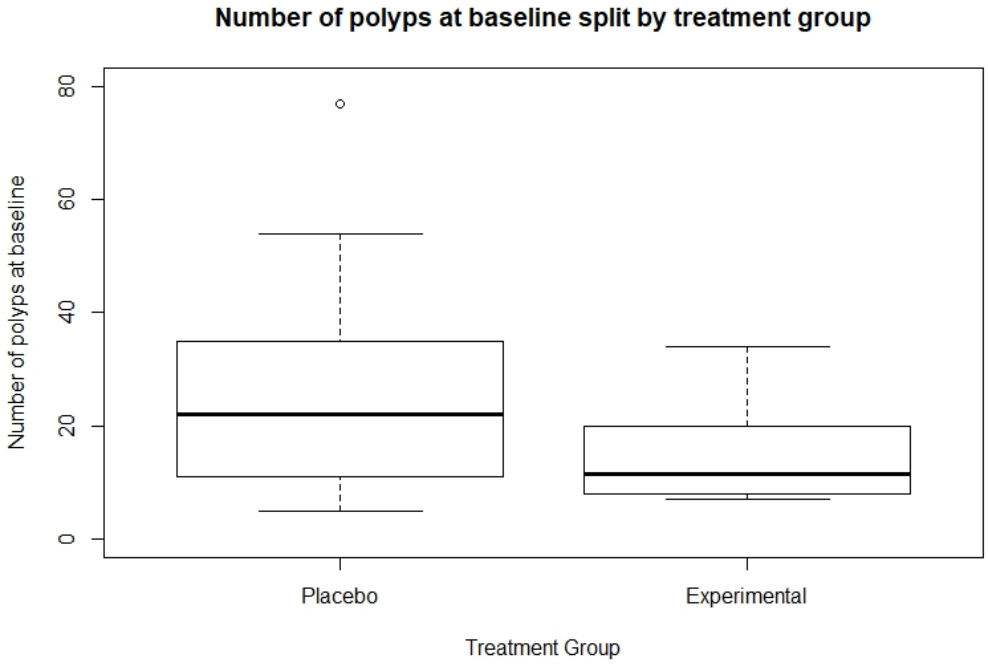


**Figure 21:** Boxplot showing the spread in the outcome variable at baseline split by treatment group with outliers removed.

Table 24 below shows the average polyps count at baseline along with the standard deviation. Table 24 shows that the average polyps count is a lot higher in the placebo group than the experimental group. This is a very good example of where baseline heterogeneity could potentially cause an issue in future analysis.

**Table 24:** Average polyps count at baseline.

| **Treatment** | **n** | **Mean polyps count** | **SD** |
|---|---|---|---|
| Placebo | 10 | 27.5 | 22.9068 |
| Experimental | 10 | 14.8 | 8.6769 |

The boxplot in Figure 22 below examines the number of polyps after 3

months. There is a much smaller spread of data and a lower count of polyps for the experimental group. The outlier being flagged is not thought to be an issue as the observation is perfectly feasible hence, no additional action is needed.

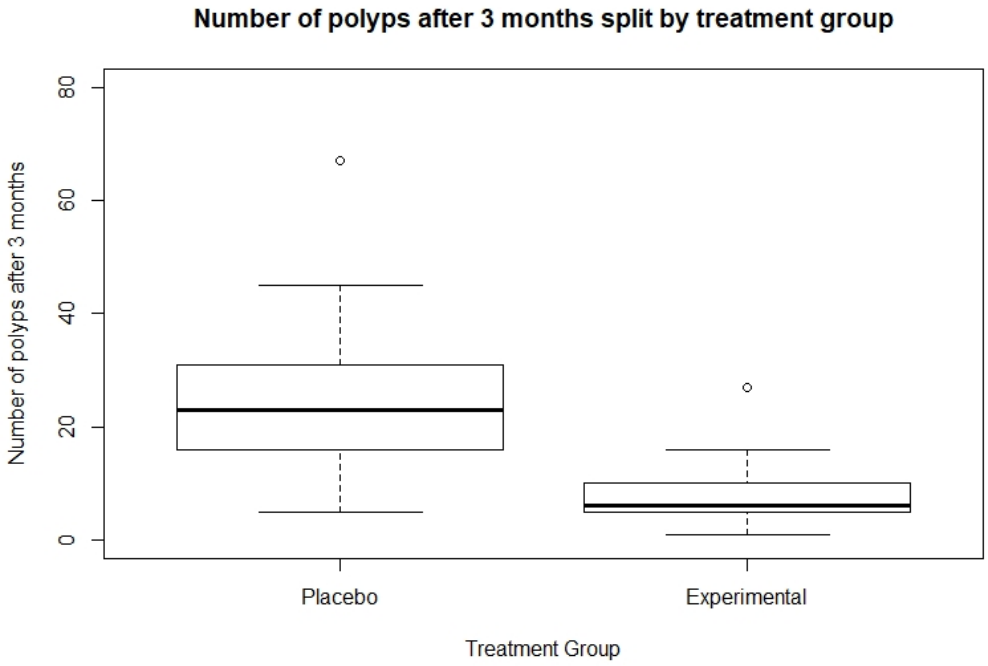


**Figure 22:** Boxplot showing the spread in the outcome variable after 3 months split by treatment group with outliers removed.

The CIs in Table 25 do not overlap (the p-value from Welch's t-test is also significant) which suggests that the average number of polyps varies between treatment groups. Thus, it is likely that treatment group would be found significant in future modelling.

**Table 25:** Average number of polyps after 3 months split by treatment group.

| Treatment | n | Mean polyps count | SD | 95% CI |
|---|---|---|---|---|
| Placebo | 10 | 26.7 | 18.4153 | (15.2861 , 38.1139) |
| Experimental | 10 | 8.9 | 7.5028 | (4.2485 , 13.5515) |

* a Welch t-test (Welch due to non equal variances) produces a p value of 0.0153

Table 26 below shows that females on average have the higher polyps count after three months of treatment. Despite this the CI for the means overlap so, the difference appears to be insignificant. The Welch's t-test also gives a p-value of 0.775 which is insignificant.Thus, gender is likely to be insignificant in future modeling.

**Table 26:** Average number of polyps after 3 months split by gender.

| Gender | n | Mean polyps count | SD | 95% CI |
|---|---|---|---|---|
| Female | 9 | 19.1111 | 23.3101 | (3.8818 , 34.3404) |
| Male | 11 | 16.7273 | 8.7646 | (11.5477 , 21.9068) |

* a Welch t-test (Welch due to non equal variances) produces a p value of 0.7775

Performing Pearson's correlation coefficient, on the variables age and polyps count after three months of treatment gives a value -0.086. Thus, resulting in a p-value of 0.7177 meaning there is no evidence of a correlation between age and polyps count after three months. On this basis, age would probably be insignificant in future modeling.

### 5.2.2 Poisson Regression ignoring Baseline Heterogeneity

Poisson regression models are produced using the likelihood ratio test for significance testing. This results in a model which only contains the variable treatment group (the only significant explanatory variable). The coefficients from the model are shown below in Table 27 along with their standard errors.

**Table 27:** Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 3.2847 | 0.0612 | $< 0.0001$ |
| Treatment (Experimental) | $-1.0986$ | 0.1224 | $< 0.0001$ |

Below is the equation for this model.

$$log(\text{E(3 month polyps count)}) = 3.2847 - 1.0986 * \text{(experimental treatment)}$$

.

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 158.73 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 23 below does not show any trend (no issue with non-constant variances).

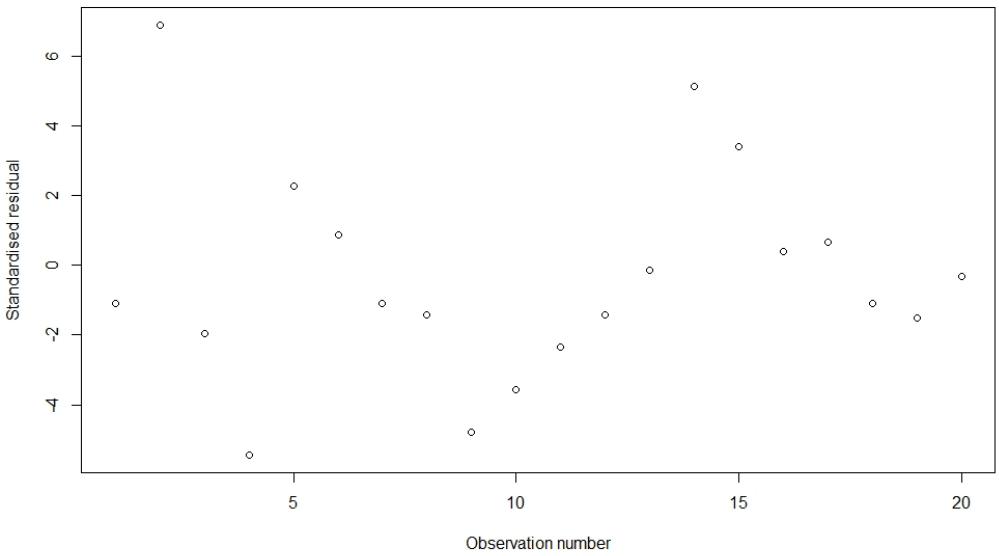


**Figure 23:** Scatter plot showing the standardised residual for each observation from the model having not adjusted for baseline heterogeneity.

The Cook's distance of every observation is calculated and Figure 24 shows a plot of the Cook's distances. The points which lie above the superimposed line are deemed to have too great an influence on the model parameters. Thus, these observations are removed from the data set and the model is recalculated.

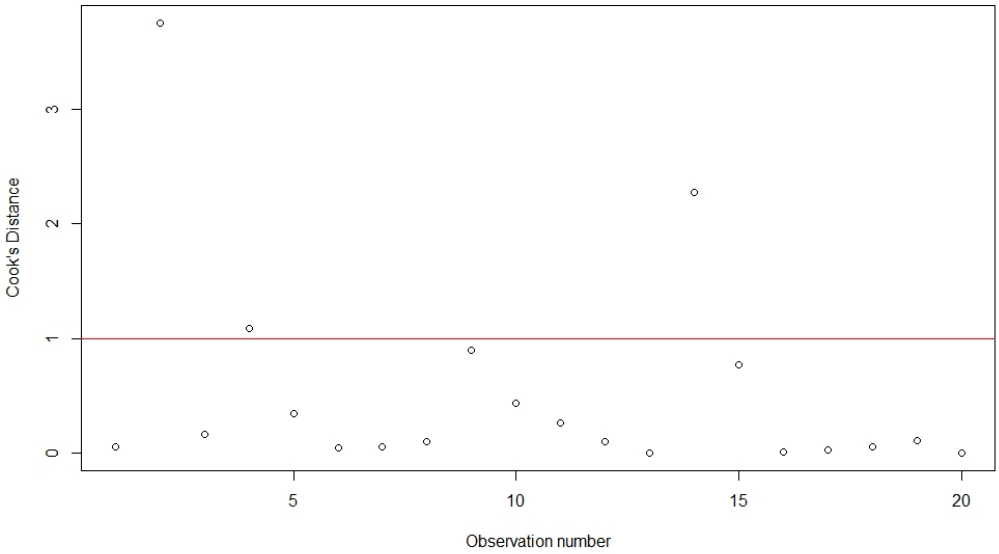


**Figure 24:** Scatter plot showing the Cook's distance of each observation from the model having not adjusted for baseline heterogeneity.

The deviance for the model (with the influential points removed) is now 59.561 which is much closer to the residual degrees of freedom. The last check is for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 28 below.

**Table 28:** Coefficients and standard errors from the negative binomial model having not adjusted for baseline heterogeneity.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 3.1936 | 0.1646 | $< 0.0001$ |
| Treatment (Experimental) | $-1.2636$ | 0.2505 | $< 0.0001$ |

The negative sign for experimental treatment shows that the experimental

treatment is more beneficial than the control. Participants taking the experimental treatment are estimated to have 0.28 times (just over a quarter) the risk of polyps than participants in the control group.

### 5.2.3 Method 1: Poisson Regression using baseline measurements as an offset

A Poisson regression models is produced here using the count-b variable as an offset term. The role of the offset term is to adjust for baseline heterogeneity. Note only significant variables are included and the significance is assessed using the likelihood ratio test. The coefficients and standard errors, for the model having adjusted for baseline heterogeneity via an offset term, are shown below in Table 29.

**Table 29:** Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | $-0.0652$ | 0.0612 | 0.2860 |
| Treatment (Experimental) | $-0.5087$ | 0.1224 | $< 0.0001$ |

Below is the equation for this model.

$$log(\text{E(3 month polyps count)}) = -0.0652 - 0.5087 * (\text{experimental treatment}) + \log(\text{baseline polyps count} + 1)$$

.

Note that the deviance for the model is 25.94 which is higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 25 below does not show any trend (no issue with non-constant variances).

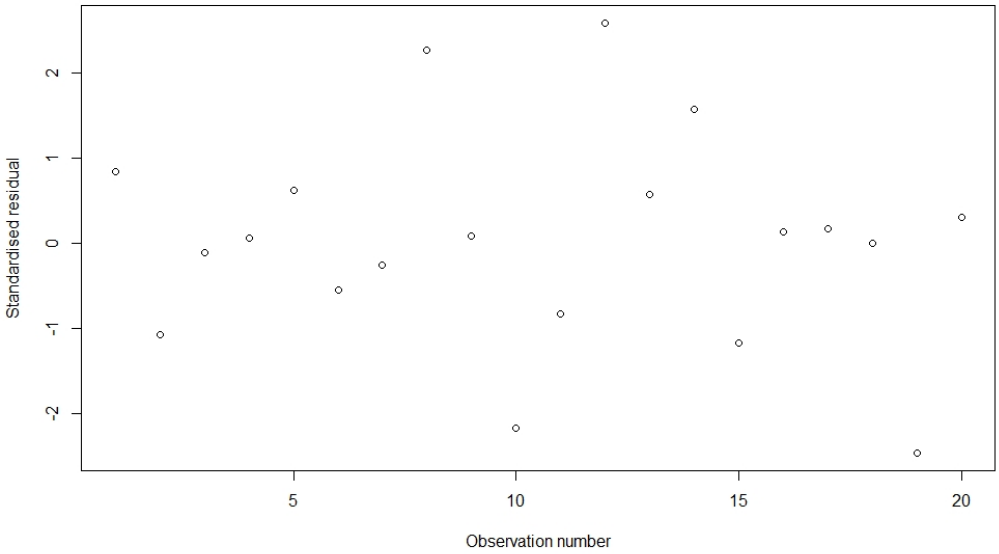


**Figure 25:** Scatter plot showing the standardised residual for each observation from method 1.

The Cook's distance of every observation is calculated and Figure 26 shows a plot of the Cook's distances. None of the observations have a Cook's distance greater than 1. Thus, none of the observations are having a large influence on the model.

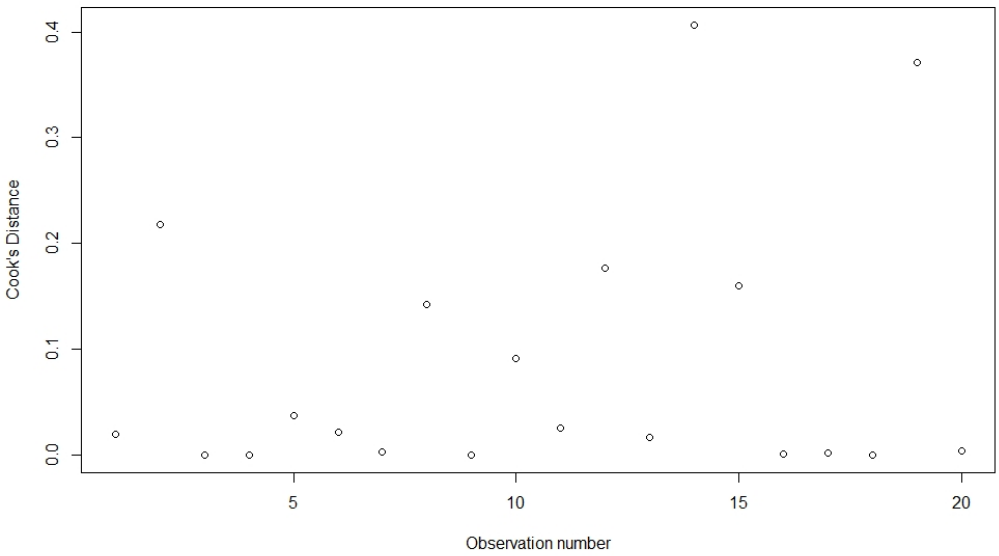


**Figure 26:** Scatter plot showing the Cook's distance of each observation from method 1.

The last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of 0.837. Thus the model written above is optimal.

### 5.2.4 Method 2: Poisson Regression using baseline measurements as a continuous covariate

In this subsection, the baseline measurements are included in the Poisson regression as a continuous covariate. This is potentially another way of dealing with any heterogeneity in the data set. The coefficients from the model are shown below in Table 30 along with their standard errors.

**Table 30:** Coefficients and standard errors from model where baseline measurements are used as a continuous covariate.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 0.3120 | 0.2957 | 0.2910 |
| Treatment (Experimental) | $-0.5846$ | 0.1353 | $< 0.0001$ |
| log(Count-b + 1) | 0.8950 | 0.0811 | $< 0.0001$ |

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 24.288 which is slightly higher than the residual degrees of freedom. Figure 27 below does not show any trend (no issue with non-constant variances).

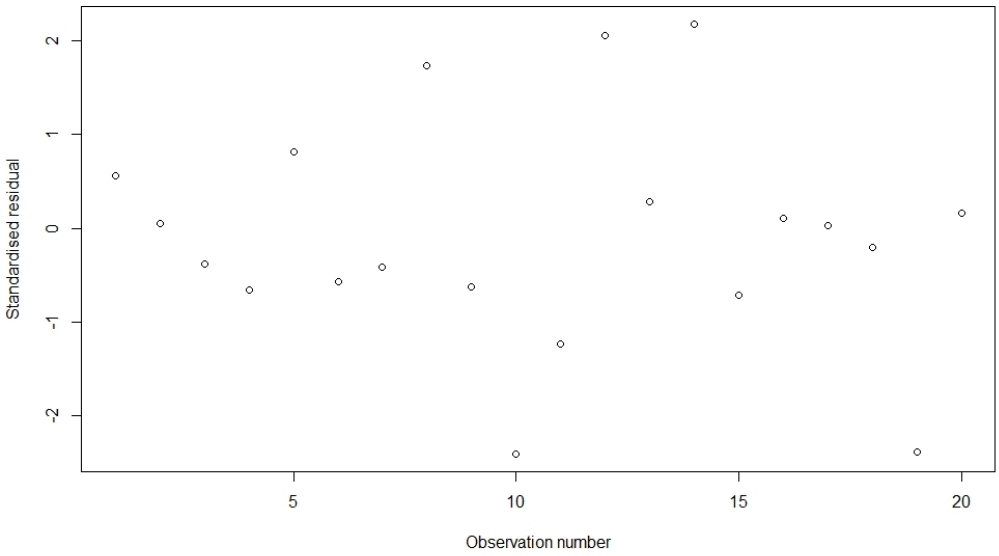


**Figure 27:** Scatter plot showing the standardised residual for each observation from method 2.

The Cook's distance of every observation is calculated and Figure 28 shows a plot of the Cook's distances. No observation has a Cook's distance

greater than 1 which means no observation is having too great an influence on the model parameters. Thus, these observations are removed from the data set and the model is recalculated.

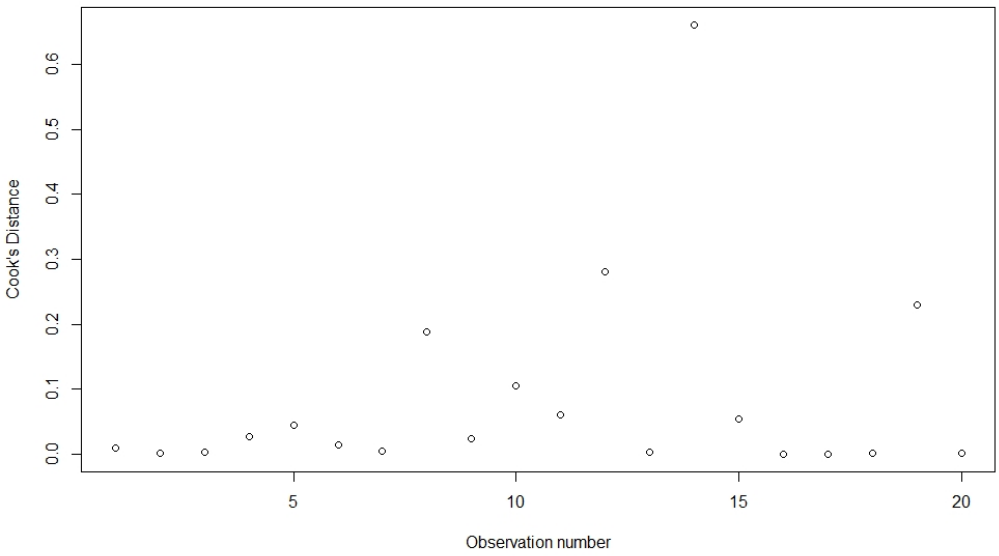


**Figure 28:** Scatter plot showing the Cook's distance of each observation from method 2.

Given only one iteration of residual analysis is being allowed, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution. The algorithm used in R-studio to maximise the log likelihood fails to converge hence, no comparison is possible here. For this reason the Poisson model is assumed to be the better model.

### 5.2.5 Method 3: Poisson Regression using baseline measurements as a categorical covariate

In this subsection, the baseline measurements are included in the Poisson regression as a categorical covariate. This means that the values of the baseline variable have to be grouped. Ideally, in a medical situation like this, the decision of which values to group together would be made in consultation

with a medical professional such as a doctor (or dentist in the case of teeth). This was not possible here, so the values for baseline have been grouped to try and produce even groups in terms of size. Table 31 below shows the groups produced for baseline and the sample sizes in each group.

**Table 31:** Sample sizes of the Count-b groups.

| **Baseline values in group** | **Sample Size** |
|---|---|
| 0-7 | 4 |
| 8-11 | 4 |
| 12-20 | 5 |
| 21-34 | 4 |
| 35+ | 3 |

The grouping of baseline measurements into groups and then using these groups within the model, is potentially another way of dealing with any heterogeneity in the data set. The coefficients from this model are shown below in Table 32 along with their standard errors.

**Table 32:** Coefficients and standard errors from model where baseline measurement groups are used as a categorical covariate.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 1.9678 | 0.2189 | $< 0.0001$ |
| Treatment (Experimental) | $-0.6209$ | 0.1416 | $< 0.0001$ |
| Count-b (8-11) | 0.5685 | 0.2778 | 0.0407 |
| Count-b (12-20) | 0.8254 | 0.2502 | 0.0010 |
| Count-b (21-34) | 1.5041 | 0.2357 | $< 0.0001$ |
| Count-b (35+) | 1.8965 | 0.2343 | $< 0.0001$ |

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 36.213 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 29 below does not show any trend (no issue with non-constant variances).

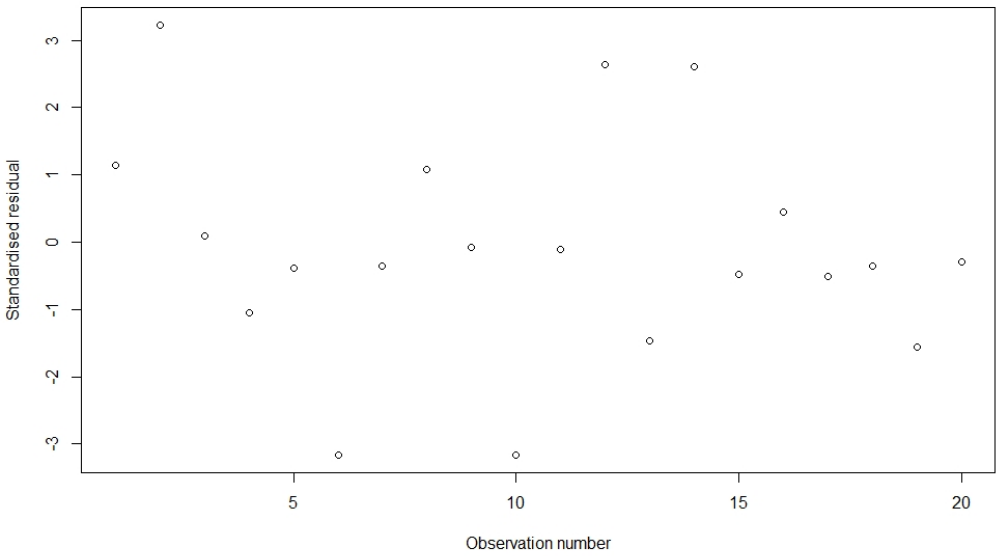


**Figure 29:** Scatter plot showing the standardised residual for each observation from method 3.

The Cook's distance of every observation is calculated and Figure 30 shows a plot of the Cook's distances. The points which lie above the superimposed line are deemed to have too great an influence on the model parameters. Thus, these observations are removed from the data set and the model is recalculated.

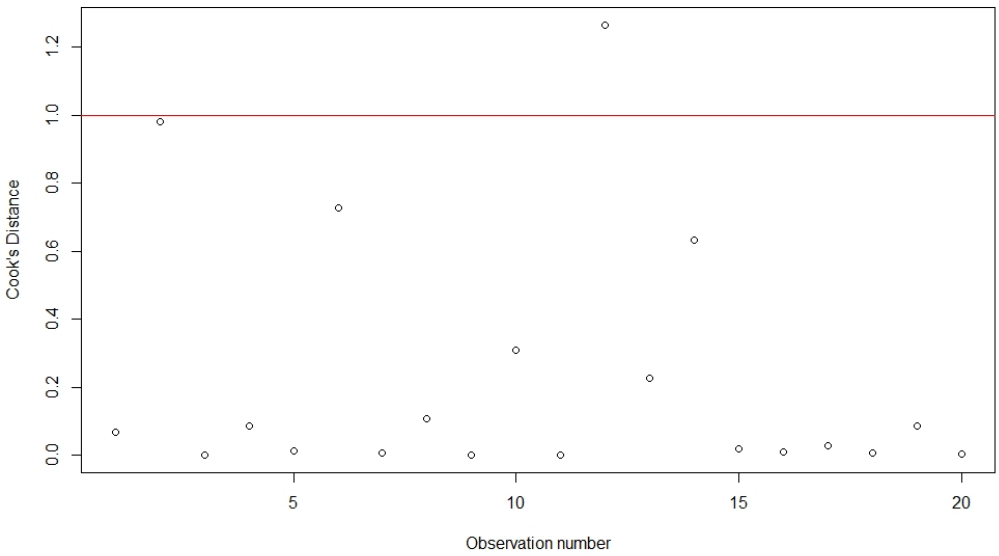


**Figure 30:** Scatter plot showing the Cook's distance of each observation from method 3.

The deviance for the model (with the influential points removed) is now 28.235 which is still higher than the new residual degrees of freedom. Given only one iteration of residual analysis is being allowed, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 33 below.

**Table 33:** Coefficients and standard errors from the negative binomial model with a categorical adjustment for heterogeneity.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 1.9151 | 0.2422 | $< 0.0001$ |
| Treatment (Experimental) | $-0.4652$ | 0.1857 | 0.0122 |
| Count-b (8-11) | 0.0165 | 0.3905 | 0.9663 |
| Count-b (12-20) | 0.7938 | 0.2775 | 0.0042 |
| Count-b (21-34) | 1.5093 | 0.2663 | $< 0.0001$ |
| Count-b (35+) | 1.9492 | 0.2752 | $< 0.0001$ |

### 5.2.6 Method 4: Poisson Regression using baseline measurements as a random effect

In this subsection, the baseline measurements are used as a random effect in the Poisson model. The role of this random effect is to account for baseline heterogeneity. When fitting this model, the log likelihood is approximated using the Adaptive Gauss-Hermite approximation with twenty points per axis. Table 34 below shows the coefficients from this GLMM model.

**Table 34:** Coefficients and standard errors from model where baseline measurements are used as a random effect.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 2.9428 | 0.3050 | $< 0.0001$ |
| Treatment (Experimental) | $-0.6408$ | 0.1411 | $< 0.0001$ |

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 57.5 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 31 below does not show any trend (no issue with non-constant variances).

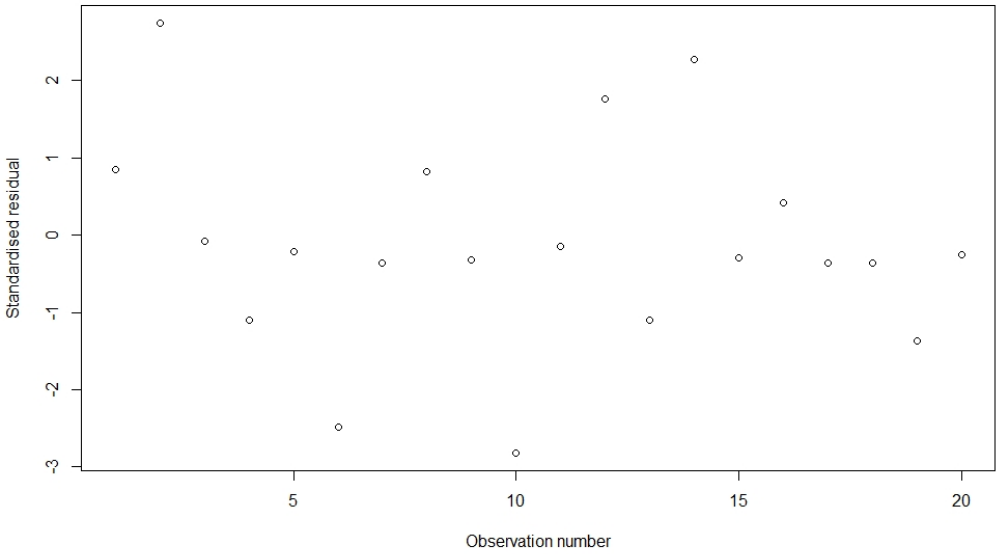


**Figure 31:** Scatter plot showing the standardised residual for each observation from method 4.

Given only one iteration of residual analysis is being allowed and Cook's distances aren't valid for this type of model, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of .0871. Thus, there is insufficient evidence of overdispersion so the Poisson model is sufficient.

### 5.2.7 Method 5: Adaptation of the conditional linear mixed model

For this method, the data has to be converted into long format. For the Belcap study, this means every participant has two rows of data, one for baseline and the other for the end point. An indicator variable is also formed to indicate whether the data is from baseline or the end point. Likelihood ratio tests are used to assess the significance of any covariates. Table 35 below shows the coefficients and standard errors from this model.

**Table 35:** Coefficients and standard errors from model with the natural log of baseline measurements as an offset.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 3.1138 | 0.2085 | $< 0.0001$ |
| Treatment | $-0.4994$ | 0.2992 | 0.0951 |
| Time | $-0.0652$ | 0.0852 | 0.4437 |
| Treatment : Time | $-0.5087$ | 0.1575 | 0.0012 |

Figure 32 below does not show any clear trend (no issue with non-constant variances).

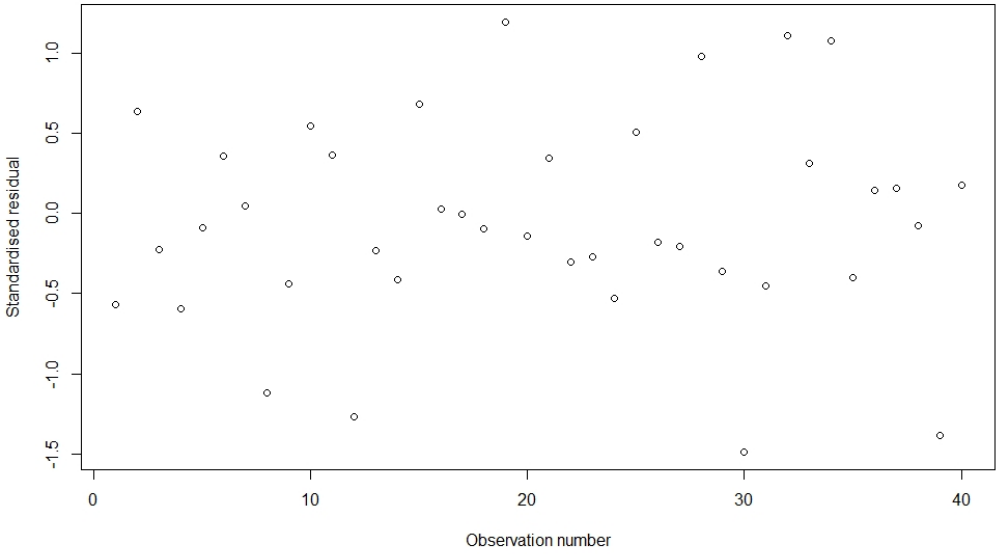


**Figure 32:** Scatter plot showing the standardised residual for each observation from method 5.

This method is also an example of a GLMM which means Cook's distances are not available. Like with method 4 above, the only way left to deal with

the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of almost 1. Thus, there is insufficient evidence of overdispersion so the Poisson model is sufficient.

### 5.2.8 Method 6: The Mantel-Haenszel Approach

In this example, the polyps count after 3 months is used as the outcome variable and the same strata (categories) are used as in section 5.2.5. Using the formula for the Mantel-Haenszel Approach given in section 4.2.5, an estimate of 0.3993 is produced for the risk ratio of being in the experimental group relative to the control group. This estimate suggests that the risk of polyps in the experimental group is 0.3993 times the risk of the control group.

## 5.3 The Falls dataset

Data is collected on the number of falls suffered by Parkinson's patients during a baseline period and a follow up period. The treatments being compared is Standard care against some intervention. The patients' treatment allocation is also recorded along with their baseline and follow up counts. Poisson regression models with and without an adjustment are produced in order to show an adjustment is needed and then find which adjustment is best.

### 5.3.1 Exploratory Data Analysis

Looking at Figure 33 below shows that there are a lot of outliers in the baseline data. None of these outliers appear to be that extreme and therefore no action is taken. It is clear that the data from the intervention group has a greater interquartile range which means the centre of the data is more spread out.

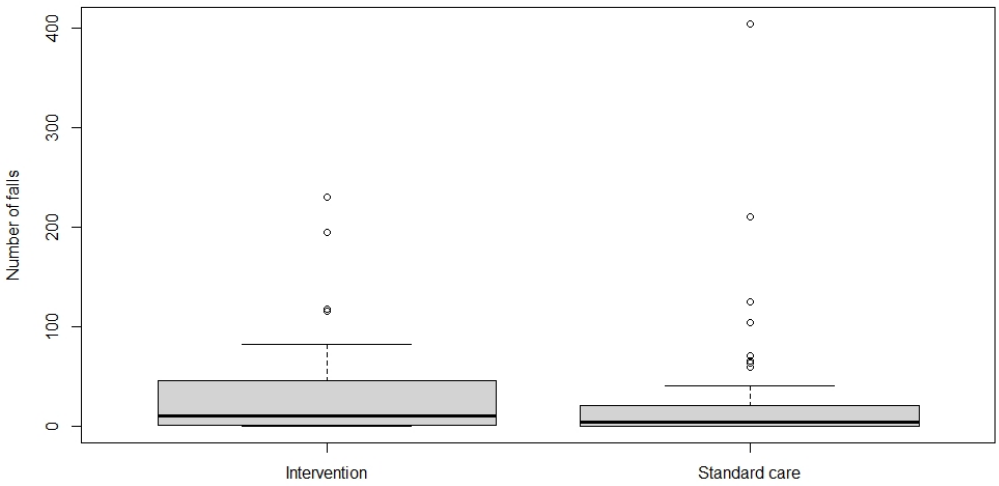


**Figure 33:** Boxplot showing the spread in baseline falls split by treatment group.

Figure 34 now shows the same boxplot but with the y axis limited to 100. This zoomed in view of the boxplot improves the readability of the graph and it is now clear that the Standard care group has the smaller median.

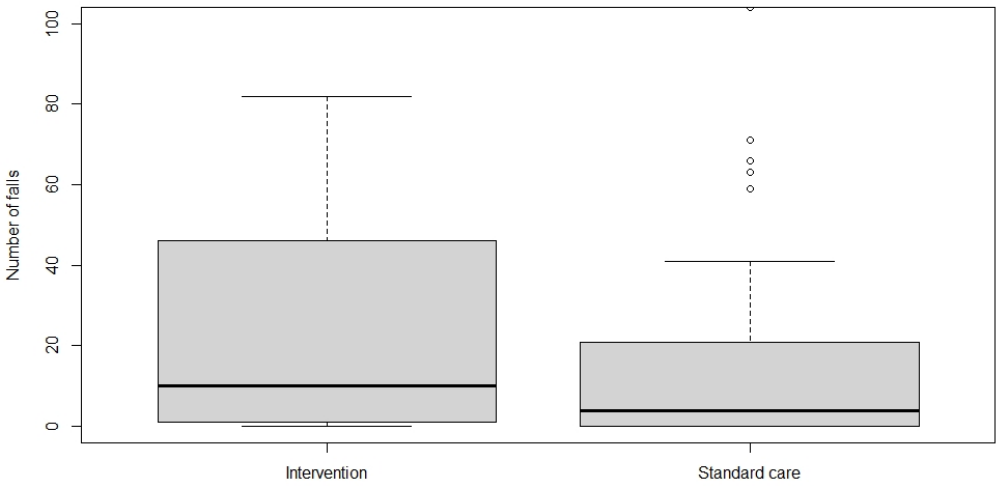


**Figure 34:** Boxplot showing the spread in baseline falls (capped at 100) split by treatment group.

Table 36 below shoes the average (and standard deviation) falls count at baseline for those receiving Standard care and the Intervention. From this table, the two treatment types have similar number of patients and the patients in the Standard care group were suffering fewer falls on average during the baseline period. Finally, the standard deviation shows that there was greater variation in the number of falls suffered by patients in the Standard care group. This could be caused by some of the large outliers within this group.

**Table 36:** Average number of falls during the baseline period.

| **Treatment** | **n** | **Mean falls count** | **SD** |
|---|---|---|---|
| Standard Care | 63 | 25.4762 | 59.9572 |
| Intervention | 61 | 30.3934 | 45.4603 |

Moving on to look at the follow up data, Figure 35 below shows a boxplot

of the follow up counts. Again, there are numerous outliers however, they are not that extreme so no action is taken. It is hard to read many of the features of the graph as the outliers are damaging the scaling.

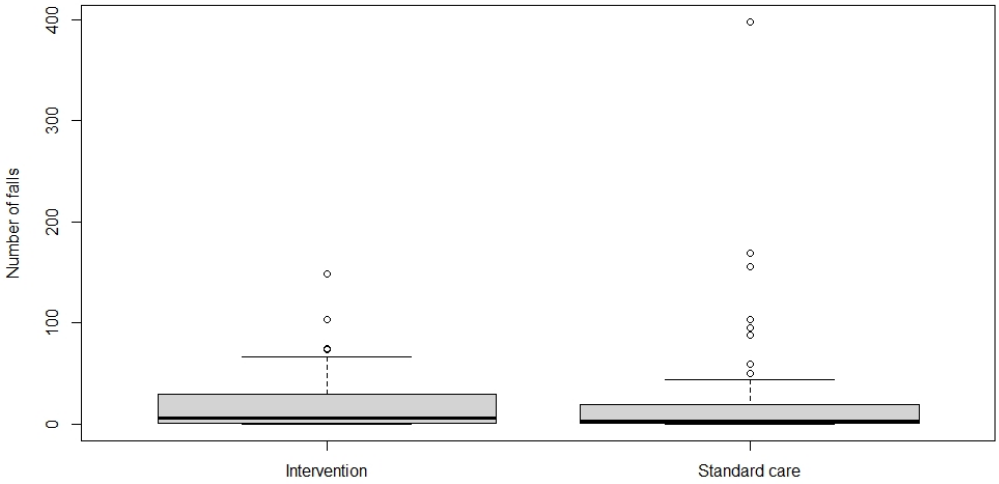


**Figure 35:** Boxplot showing the spread in follow up falls split by treatment group.

Thus, Figure 36 below shows the same graph but with the y axis restricted to 100. Again the Standard care group has the smaller interquartile range and median count of falls.

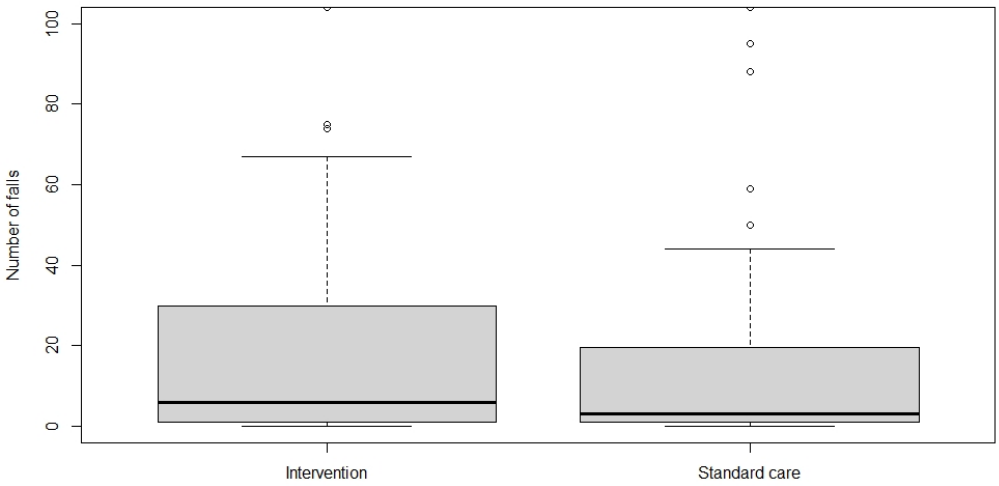


**Figure 36:** Boxplot showing the spread in follow up falls (capped at 100) split by treatment group.

Table 37 below shoes the average (and standard deviation) falls count at baseline for those receiving Standard care and the Intervention. The patients in the Standard care group were suffering fewer falls on average during the follow up period. The difference between the groups is smaller now though. Finally, the standard deviation shows that there was greater variation in the number of falls suffered by patients in the Standard care group. This could be caused by some of the large outliers within this group.

**Table 37:** Average number of falls during the followup period.

| Treatment | n | Mean falls count | SD |
|---|---|---|---|
| Standard Care | 63 | 24.9683 | 58.9182 |
| Intervention | 61 | 19.5410 | 28.6389 |

### 5.3.2 Poisson Regression ignoring Baseline Heterogeneity

Poisson regression models are produced using the likelihood ratio test for significance testing. This results in a model which only contains the variable treatment group (the only significant explanatory variable). The coefficients from the model are shown below in Table 38 along with their standard errors.

**Table 38:** Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 3.2176 | 0.0252 | $< 0.0001$ |
| Intervention | $-0.2451$ | 0.0384 | $< 0.0001$ |

Below is the equation for this model.

$$log(\text{E}(3 \text{ month polyps count})) = 3.2176 - 0.2451 * (\text{Intervention})$$

.

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 6007.5 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 37 below does not show any trend however, there are some very big residuals.

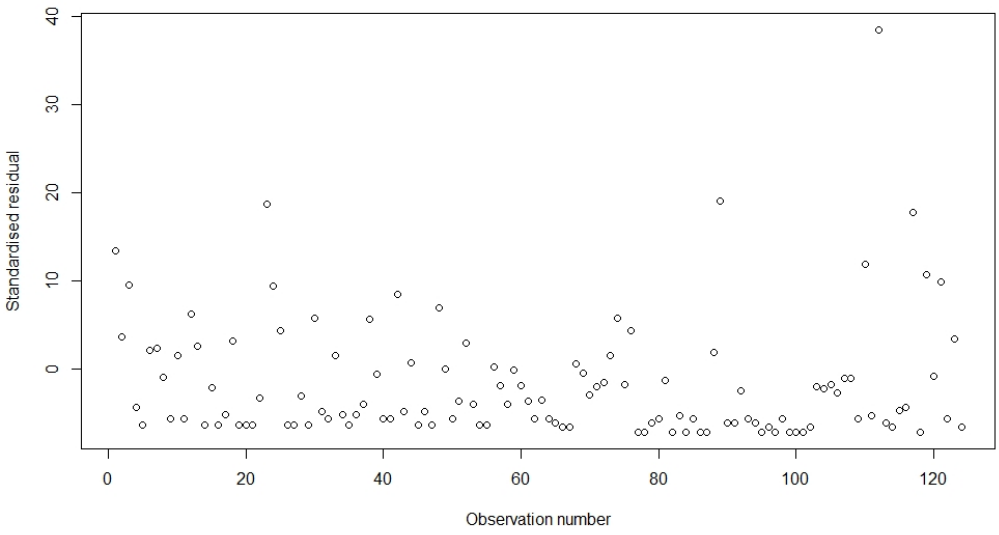


**Figure 37:** Scatter plot showing the standardised residual for each observation from the model having not adjusted for baseline heterogeneity.

The Cook's distance of every observation is calculated and Figure 38 shows a plot of the Cook's distances. The points which lie above the superimposed line are deemed to have too great an influence on the model parameters. Thus, these observations are removed from the data set and the model is recalculated.

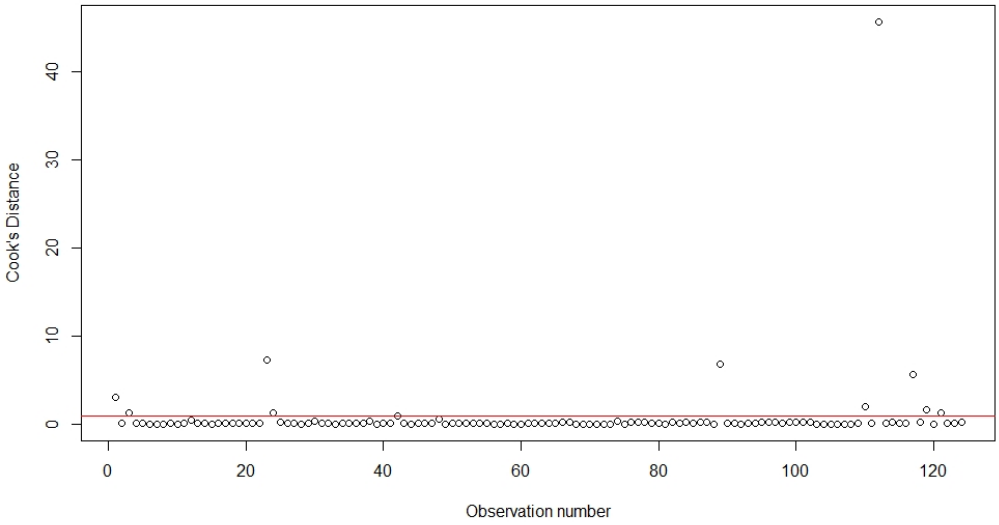


**Figure 38:** Scatter plot showing the Cook's distance of each observation from the model having not adjusted for baseline heterogeneity.

The deviance for the model (with the influential points removed) is now 2048.1 which is still a lot higher than the residual degrees of freedom. The last check is for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 39 below.

**Table 39:** Coefficients and standard errors from the negative binomial model having not adjusted for baseline heterogeneity.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 2.2902 | 0.2024 | $< 0.0001$ |
| Intervention | 0.3388* | 0.2854 | 0.2350 |

*This fails to achieve statistical significance.

The lack of significance suggests that there was no difference between

receiving the intervention and the standard care.

### 5.3.3 Method 1: Poisson Regression using baseline measurements as an offset

A Poisson regression model is produced here using the baseline falls count as an offset term. The role of the offset term is to adjust for baseline heterogeneity. Note only significant variables are included and the significance is assessed using the likelihood ratio test. The coefficients and standard errors, for the model having adjusted for baseline heterogeneity via an offset term, are shown below in Table 40.

**Table 40:** Coefficients and standard errors from model where no adjustment for baseline heterogeneity is made.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | −0.0586 | 0.0252 | 0.02 |
| Treatment (Experimental) | −0.4154 | 0.0384 | $< 0.0001$ |

Below is the equation for this model.

$log(\text{E(followup falls)}) = -0.0586 - 0.4154 * \text{(experimental treatment)} + \log(\text{baseline falls} + 1)$

.

Note that the deviance for the model is 220.62 which is higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 39 below does not show any trend (no issue with non-constant variances).

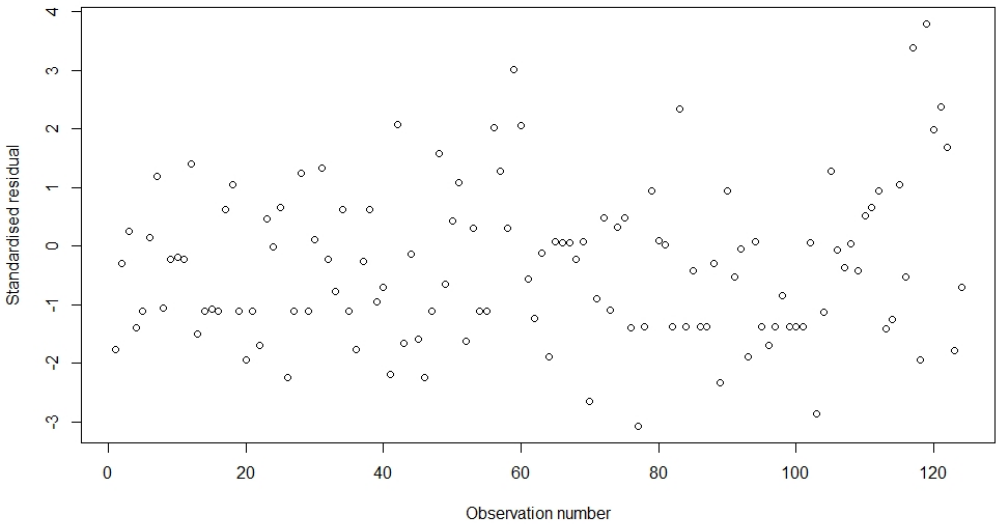


**Figure 39:** Scatter plot showing the standardised residual for each observation from method 1.

The Cook's distance of every observation is calculated and Figure 40 shows a plot of the Cook's distances. None of the observations have a Cook's distance greater than 1. Thus, none of the observations are having a large influence on the model.

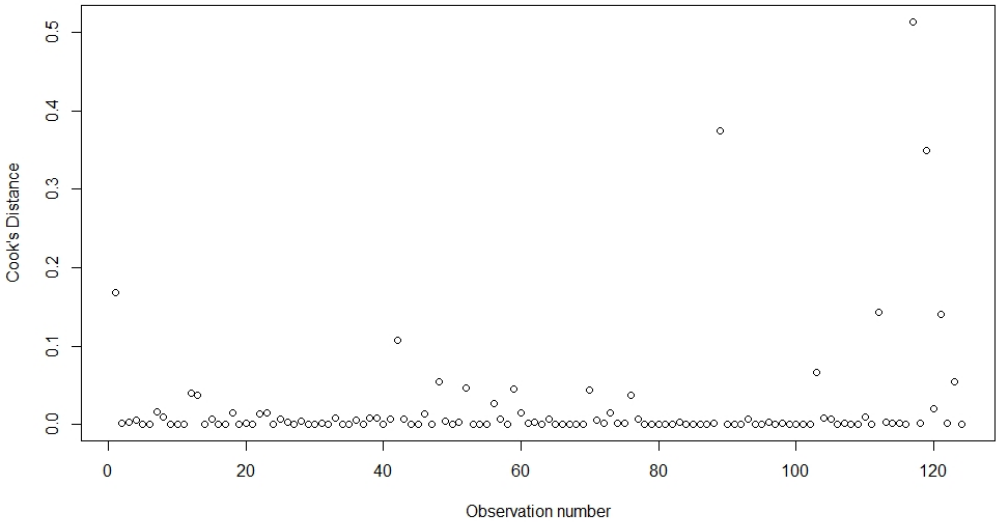


**Figure 40:** Scatter plot showing the Cook's distance of each observation from method 1.

The last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 41 below.

**Table 41:** Coefficients and standard errors from the negative binomial model having an offset adjustment for baseline heterogeneity.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | −0.1020 | 0.0436 | 0.0193 |
| Intervention | −0.3791 | 0.0615 | $< 0.0001$ |

Below is the equation for this model.

$$log(\text{E}(\text{followup falls})) = -0.1020 - 0.3791 * (\text{intervention})$$

$$+ \log(\text{baseline falls} + 1)$$

.

### 5.3.4 Method 2: Poisson Regression using baseline measurements as a continuous covariate

In this subsection, the baseline measurements are included in the Poisson regression as a continuous covariate. This is potentially another way of dealing with any heterogeneity in the data set. The coefficients from the model are shown below in Table 42 along with their standard errors.

**Table 42:** Coefficients and standard errors from model where baseline measurements are used as a continuous covariate.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | −0.2300 | 0.0777 | 0.0031 |
| Intervention | −0.4066 | 0.0386 | $< 0.0001$ |
| log(baseline + 1) | 1.0383 | 0.0163 | $< 0.0001$ |

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 215.04 which is higher than the residual degrees of freedom. Figure 41 below does not show any trend (no issue with non-constant variances).

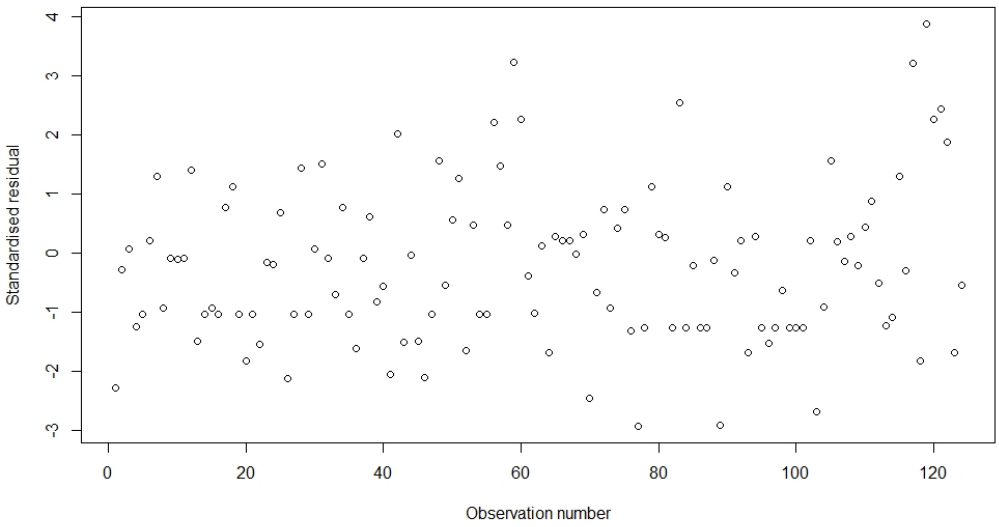


**Figure 41:** Scatter plot showing the standardised residual for each observation from method 2.

The Cook's distance of every observation is calculated and Figure 42 shows a plot of the Cook's distances. No observation has a Cook's distance greater than 1 which means no observation is having too great an influence on the model parameters.

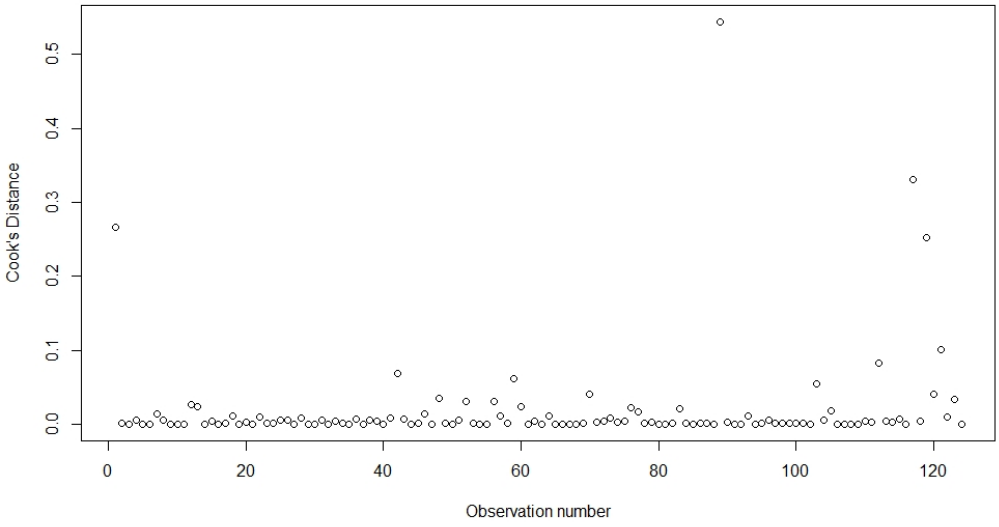


**Figure 42:** Scatter plot showing the Cook's distance of each observation from method 2.

Given only one iteration of residual analysis is being allowed, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 43 below.

**Table 43:** Coefficients and standard errors from negative binomial model where baseline measurements are used as a continuous covariate.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | −0.3293 | 0.1029 | 0.0014 |
| Intervention | −0.3965 | 0.0613 | $< 0.0001$ |
| log(baseline + 1) | 1.0638 | 0.0259 | $< 0.0001$ |

Below is the equation for this model.

$$log(\text{E(followup falls)}) = -0.3293 - 0.3965 * (\text{experimental treatment})$$

$$+1.0638 * \log(\text{baseline falls} + 1)$$

.

### 5.3.5 Method 3: Poisson Regression using baseline measurements as a categorical covariate

In this subsection, the baseline measurements are included in the Poisson regression as a categorical covariate. This means that the values of the baseline variable have to be grouped. Ideally, in a medical situation like this, the decision of which values to group together would be made in consultation with a medical professional such as a doctor. This was not possible here, so the values for baseline have been grouped to try and produce even groups in terms of size. Table 44 below shows the groups produced for baseline and the sample sizes in each group.

**Table 44:** Sample sizes of the baseline fall groups.

| Baseline values in group | Sample Size |
|---|---|
| 0-1 | 38 |
| 2-4 | 23 |
| 5-14 | 16 |
| 15-45 | 30 |
| 46+ | 17 |

The grouping of baseline measurements into groups and then using these groups within the model, is potentially another way of dealing with any heterogeneity in the data set. The coefficients from this model are shown below in Table 45 along with their standard errors.

**Table 45:** Coefficients and standard errors from model where baseline measurement groups are used as a categorical covariate.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | −0.9776 | 0.2892 | 0.0003 |
| Intervention | −0.3641 | 0.0385 | $< 0.0001$ |
| Count-b (2-4) | 2.0827 | 0.3151 | $< 0.0001$ |
| Count-b (5-14) | 3.3511 | 0.3009 | $< 0.0001$ |
| Count-b (15-45) | 4.3674 | 0.2910 | $< 0.0001$ |
| Count-b (46+) | 5.8145 | 0.2896 | $< 0.0001$ |

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 813.4 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 43 below does not show any trend but there is 1 exceptionally large residual

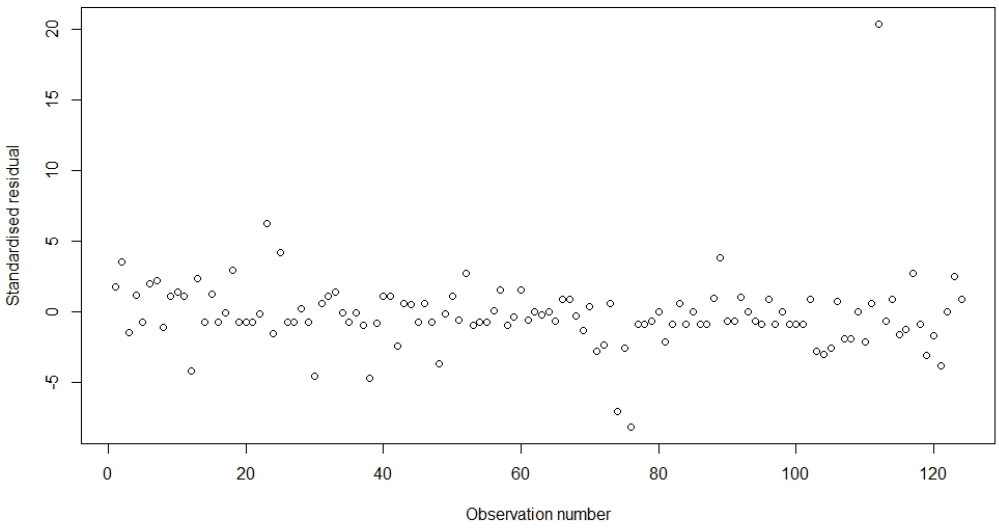


**Figure 43:** Scatter plot showing the standardised residual for each observation from method 3.

The Cook's distance of every observation is calculated and Figure 44 shows a plot of the Cook's distances. The points which lie above the superimposed line are deemed to have too great an influence on the model parameters. Thus, these observations are removed from the data set and the model is recalculated.

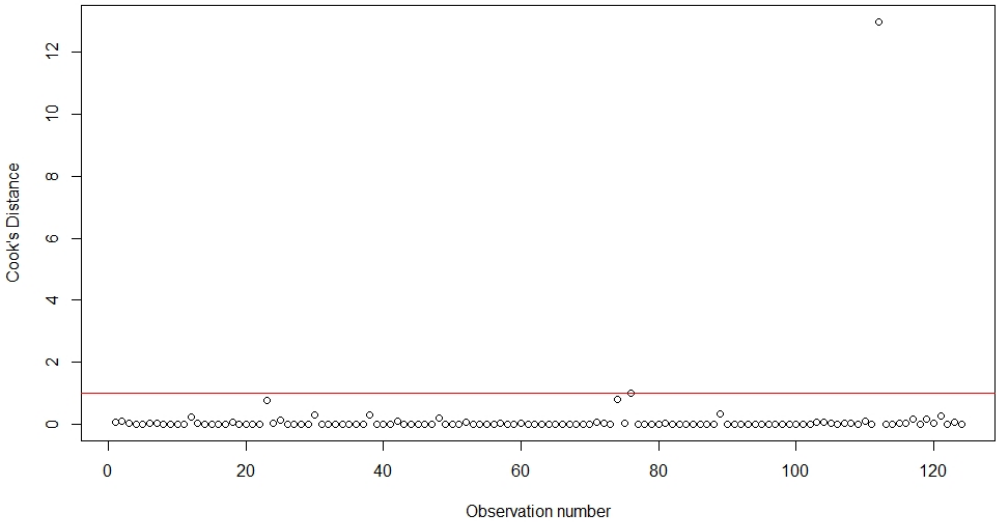


**Figure 44:** Scatter plot showing the Cook's distance of each observation from method 3.

The deviance for the model (with the influential points removed) is now 339.03 which is still higher than the new residual degrees of freedom. Given only one iteration of residual analysis is being allowed, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 46 below.

**Table 46:** Coefficients and standard errors from negative binomial model where baseline measurement groups are used as a categorical covariate.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | −1.1126 | 0.2958 | 0.0002 |
| Intervention | −0.0776* | 0.0870 | 0.3728 |
| Count-b (2-4) | 2.1434 | 0.3246 | $< 0.0001$ |
| Count-b (5-14) | 3.3206 | 0.3131 | $< 0.0001$ |
| Count-b (15-45) | 4.3783 | 0.2994 | $< 0.0001$ |
| Count-b (46+) | 5.6521 | 0.3031 | $< 0.0001$ |

*Failed to reach statistical significance.

### 5.3.6 Method 4: Poisson Regression using baseline measurements as a random effect

In this subsection, the baseline measurements are used as a random effect in the Poisson model. The role of this random effect is to account for baseline heterogeneity. When fitting this model, the log likelihood is approximated using the Adaptive Gauss-Hermite approximation with twenty points per axis. Table 47 below shows the coefficients from this GLMM model.

**Table 47:** Coefficients and standard errors from model where baseline measurements are used as a random effect.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 2.1487 | 0.8884 | 0.0156 |
| Intervention | −0.3641 | 0.0385 | $< 0.0001$ |

Before interpreting any regression model, it is good practice to perform a residual analysis to ensure the model assumptions are not violated. Noting that the deviance for the model is 1262.4 which is much higher than the residual degrees of freedom. Thus, there is some evidence of model misspecification. Figure 45 below does not show any trend however, there is 1 extremely big residual.

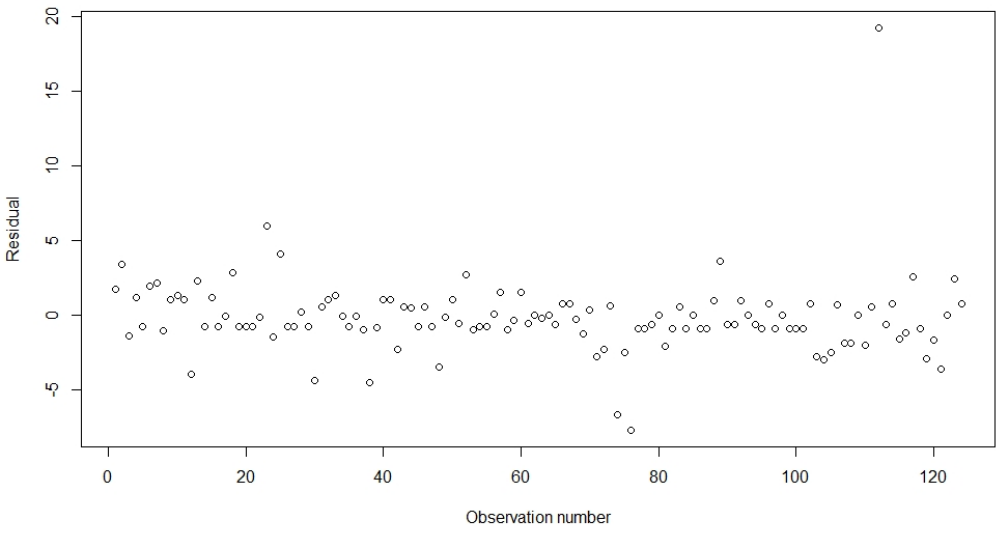


**Figure 45:** Scatter plot showing the residuals for each observation from method 4.

Given only one iteration of residual analysis is being allowed and Cook's distances aren't valid for this type of model, the last option to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 48 below.

**Table 48:** Coefficients and standard errors from the negative binomial model where baseline measurements are used as a random effect.

| Covariate name | Coefficient | Standard error | P-value |
|---|---|---|---|
| Intercept | 2.0527 | 0.8862 | 0.0205 |
| Intervention | −0.1488* | 0.1035 | 0.1505 |

*This fails to reach statistical significance.

### 5.3.7 Method 5: Adaptation of the conditional linear mixed model

For this method, the data has to be converted into long format. For the Belcap study, this means every participant has two rows of data, one for baseline and the other for the end point. An indicator variable is also formed to indicate whether the data is from baseline or the end point. Likelihood ratio tests are used to assess the significance of any covariates. Table 49 below shows the coefficients and standard errors from this model.

**Table 49:** Coefficients and standard errors from the Conditional method.

| **Covariate name** | **Coefficient** | **Standard error** |
|---|---|---|
| Intercept | 1.6147 | 0.2122 |
| Intervention | 0.3918* | 0.3019 |
| Time | −0.1458 | 0.0603 |
| Intervention : Time | −0.3463 | 0.0786 |

*Failed to reach statistical significance.

Figure 46 below shows some huge residuals so there is clearly some model misspecification.

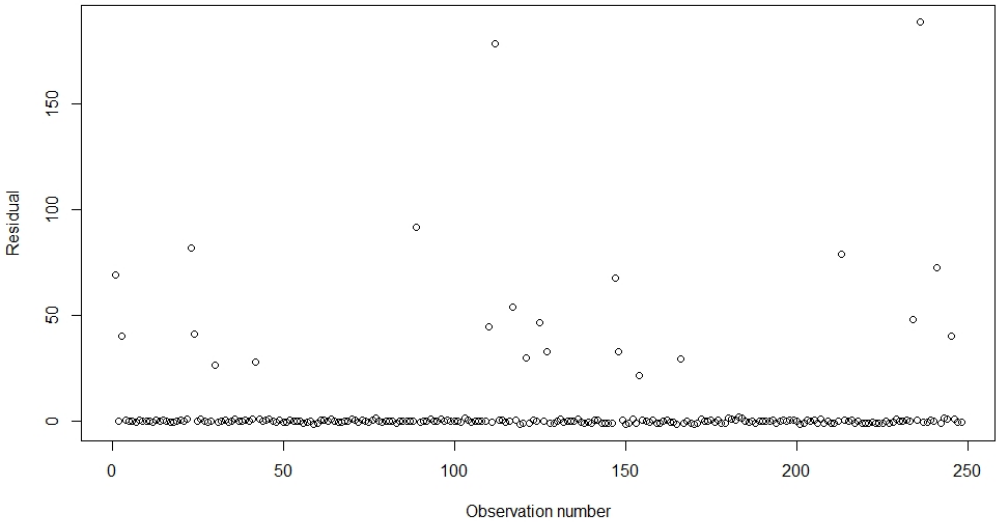


**Figure 46:** Scatter plot showing the residual for each observation from method 5.

This method is also an example of a GLMM which means Cook's distances are not available. Like with method 4 above, the only way left to deal with the large deviance is to look for potential over-dispersion. This is done by refitting the model under a negative binomial distribution and comparing this to the Poisson model. This results in a p-value of below 0.05, hence, the negative binomial model (over-dispersion adjusted for) is the superior model. The coefficients for this model are given in Table 50 below.

**Table 50:** Coefficients and standard errors from the Conditional method accounting for overdispersion.

| **Covariate name** | **Coefficient** | **Standard error** | **P-value** |
|---|---|---|---|
| Intercept | 1.8515 | 0.2281 | $< 0.0001$ |
| Intervention | 0.4190* | 0.3236 | 0.1940 |
| Time | $-0.0740$* | 0.0452 | 0.016 |
| Intervention : Time | $-0.4029$ | 0.0597 | $< 0.0001$ |

*Failed to reach statistical significance.

### 5.3.8 Method 6: The Mantel-Haenszel Approach

In this example, the follow up falls count is used as the outcome variable and the same strata (categories) are used as in section 5.3.5. Using the formula for the Mantel-Haenszel Approach given in section 4.2.5, an estimate of 0.6962 is produced for the risk ratio of being in the Intervention group relative to the Standard care group. This estimate suggests that the risk of falls in the intervention group is 0.6962 times the risk of the Standard care group.

## 5.4 How successful are the proposed methods at dealing with baseline heterogeneity

The two data sets have been analysed with and without adjusting for baseline heterogeneity. The question of interest is whether any of the 5 proposed methods used in chapter 4 are successful at adjusting for baseline heterogeneity. This is judged, for the 4 parametric models, using model selection criterion known as "AIC" and "BIC" (see Agressti, [1]). The success of the non-parametric method (Mantael-Haenszel) is judged by performing it with and without adjustment for baseline heterogeneity.

### 5.4.1 Results from the Belcap study

The results from using "AIC" and "BIC" for the Belcap data set are shown below in Table 51.

**Table 51:** Model criterion for models (analysing the belcap data) with and without adjustment for baseline heterogeneity and having adjusted for overdispersion.

| **Type of Adjustment** | **AIC** | **BIC** |
|---|---|---|
| No Adjustment | 4210 | 4248 |
| Offset | 3919 | 3952 |
| Continuous Covariate | 3876 * | 3914 * |
| Categorical Covariate | 3880 | 3936 |
| Random Effect | 3955 | 3992 |

*this method fits the data the best

Table 51 above shows that the AIC and BIC are lower, for all adjustment methods when compared to the no adjustment method. Thus, the four other models for adjusting baseline heterogeneity shown in this table, fit the data better than the model with no adjustment. Note the AIC and BIC for the Conditional model cannot be meaningfully compared with the corresponding values of the no adjustment method. This is because different datasets are used, where the Conditional model has a much larger sample size which inflates the AIC and BIC. The remaining question is how similar are the estimates of the risk ratios (statistic of interest) produced by these parametric and non-parametric methods. Table 52 below shows the estimated risk ratios from all 6 proposed methods and the model with no adjustment for baseline heterogeneity.

**Table 52:** Risk ratios (treatment versus control) from models with and without adjustment for baseline heterogeneity and having adjusted for overdispersion.

| | **No Adjust** | **Method 1** | **Method 2** | **Method 3** | **Method 4** | **Method 5** | **Method 6** |
|---|---|---|---|---|---|---|---|
| ALL | 0.4390 | 0.7302 | 0.5739 | 0.5834 | 0.5998 | 0.6916 | 0.6307 |
| ESD | 0.7560 | 0.8926 | 0.8246 | 0.8257 | 0.8310 | 0.8336 | 0.8827 |
| MW | 0.6605 | 0.7042 | 0.6688 | 0.7107 | 0.7195 | 0.6977 | 0.8828 |
| OHE | 0.7444 | 0.6934 | 0.7091 | 0.7111 | 0.7201 | 0.6869 | 0.8143 |
| OHY | 0.6405 | 0.8431 | 0.7371 | 0.7725 | 0.7638 | 0.8289 | 1.0164 |

Table 52 above shows that these 6 methods are not similar estimates of the risk ratios. This then begs the question, "Which method produces the best estimate?" This question is investigated in chapter 6.

### 5.4.2 Results from the Polyps study

The results from using "AIC" and "BIC" for the Polyps data set are shown below in Table 53.

**Table 53:** Model criterion for models (analysing the polyps data) with and without an offset term for the Polyps data set.

| Type of Adjustment | AIC | BIC |
|---|---|---|
| No Adjustment | 249 | 251 |
| Offset | 116 * | 118 * |
| Continuous Covariate | 117 | 120 |
| Categorical Covariate | 120 | 126 |
| Random Effect | 150 | 153 |

*This model fits the data the best

Table 53 above shows that the AIC and BIC are lower, for all adjustment methods when compared to the no adjustment method. Thus, the four other models for adjusting baseline heterogeneity shown in this table, fit the data better than the model with no adjustment. Note the AIC and BIC are not shown for the Conditional model as they cannot be meaningfully compared with the corresponding values of the no adjustment method. This is because different datasets are used. The Conditional model has a much larger sample size which inflates the AIC and BIC. The remaining question is how similar are the estimates of the risk ratios (statistic of interest) produced by these parametric and non-parametric methods. Table 54 below shows the estimated risk ratios from all 6 proposed methods and the model with no adjustment for baseline heterogeneity.

**Table 54:** Risk ratios (treatment versus control) from models with and without adjustment for baseline heterogeneity.

| | No Adjust | Method 1 | Method 2 | Method 3 | Method 4 | Method 5 | Method 6 |
|---|---|---|---|---|---|---|---|
| Experimental | 0.2826 | 0.6194 | 0.5548 | 0.6280 | 0.5269 | 0.6193 | 0.3993 |

Table 54 above shows that these 6 methods are not similar in terms of estimating the risk ratios. This then begs the question, "Which method produces the best estimate?" This question is investigated within chapter 6.

### 5.4.3 Results from the Falls study

The results from using "AIC" and "BIC" for the Polyps data set are shown below in Table 55.

**Table 55:** Model criterion for models (analysing the falls data) with and without an adjustment for baseline heterogeneity.

| **Type of Adjustment** | **AIC** | **BIC** |
|---|---|---|
| No Adjustment | 767.1399 | 775.3485 |
| Offset | 617.4739 | 625.9347 |
| Continuous Covariate | 613.3379 | 624.619 |
| Categorical Covariate | 582.4625* | 602.0906* |
| Random Effect | 656.8407 | 668.1218 |

*This method fits the data the best

Table 55 above shows that the AIC and BIC are lower, for all adjustment methods when compared to the no adjustment method. Thus, the four other models for adjusting baseline heterogeneity shown in this table, fit the data better than the model with no adjustment. Note the AIC and BIC are not shown for the Conditional model as they cannot be meaningfully compared with the corresponding values of the no adjustment method. This is because different datasets are used. The Conditional model has a much larger sample size which inflates the AIC and BIC. The remaining question is how similar are the estimates of the risk ratios (statistic of interest) produced by these parametric and non-parametric methods. Table 56 below shows the estimated risk ratios from all 6 proposed methods and the model with no adjustment for baseline heterogeneity.

**Table 56:** Risk ratios (Intervention versus Standard care) from models with and without adjustment for baseline heterogeneity.

| | **No Adjust** | **Method 1** | **Method 2** | **Method 3** | **Method 4** | **Method 5** | **Method6** |
|---|---|---|---|---|---|---|---|
| Intervention | 1.4033* | 0.6845 | 0.6727 | 0.9253* | 0.8617* | 0.9287* | 0.6962 |

* Are judged to not be significantly different to 1.

Table 56 above shows that these 6 methods are not similar in terms of

estimating the risk ratios. This then begs the question, "Which method produces the best estimate?" This question is investigated within chapter 6.

# 6 Which method is best at adjusting for baseline heterogeneity

The results in chapter 5 demonstrate the need for some form of adjustment when analysing data sets such as those introduced in sections 1.2.1 and 1.2.2. The next issue to solve is which method is best at adjusting for heterogeneity. The concept of which method is best can get ambiguous as there are different criteria which could be used. A common requirement for a good method is that of zero Bias meaning on average, the estimates have zero error. This can be confusing though as the error from the individual estimates produced could have large positive or negative values which then cancel once averaged. Thus, making a poor method look reasonable. In this chapter the Bias and Mean Squared Error (MSE) are going to be used when judging the results from the different methods used in chapter 5. MSE is calculated by summing the variance and the bias squared.

To be able to do this, a data set needs to be formed where the true risk ratio between the experimental and control groups is known. In reality, this is never possible as clinical trials are only using a sample of the population, and if a different sample was taken, slightly different data would be obtained which could then lead to different results being produced. The way around this is to perform a simulation study. Some key terms which need explaining are that of replication value and sample size. Replication value refers to how many times is the simulation is repeated and sample size refers to how many observations are simulated in each replication.

## 6.1 A simulation algorithm based on the Polyps study

The first simulation will be of a situation with one treatment group such as the Polyps study. Below is an outline of the process. See appendix 1 for the R code.

Part 1: Set the parameters for the simulation

Step 1: Produce a Logistic regression model where the treatment group is the outcome variable and the baseline value is the predictor. This is done using the original Polyps data set.

Step 2: Produce a null Poisson model with baseline value as the outcome. This is also done using the original Polyps data set.

Step 3: Produce a Poisson model with end data as the outcome, treatment group as a categorical predictor and log baseline value as a continuous predictor. The point of this model is to understand the relationship between the experiment and control treatment having adjusted for baseline heterogeneity (via the log baseline term). This model is produced using the original Polyps data set too.

Part 2: Perform the simulation

Step 4: Simulate baseline data from a Poisson distribution where the parameter is equal to the linear predictor from step 2. The example in section 6.1.1 simulates 250 values.

Step 5: For each of the simulated baseline values from step 4, predict the probability of it belonging to the experimental group using the Logistic model from step 1.

Step 6: For each of the simulated baseline values (from step 4) separately, simulate a draw from a Bernoulli distribution where the "Success" probability is equal to the probability predicted in step 5. If the value drawn from the Bernoulli distribution is 1 then the baseline value belongs to the experimental group. Should the value drawn from the Bernoulli distribution be 0 then the baseline value belongs to the control group.

Step 7: Now, combine the simulated baseline data and the indicator for treatment group together forming a data set with 2 columns.

Step 8: Simulate end of study data from a Poisson distribution where the parameter is given by the linear predictor of the model in step 3. This uses the baseline data and the indicator of treatment group from the data set produced in step 7.

Step 9: Combine the end of study data with the data set from step 7. This forms the full simulated data set.

Step 10: Apply the 6 trial methods being studied to the simulated data set.

Step 11: Repeat steps 4 - 10 multiple times. The example shown in section 6.1.1 uses a replication value of 1000.

Part 3: Analyse the results from the simulation

Step 12: Calculate the biases and mean squared errors for the estimates of the risk ratio (for the treatment group) produced by each of the six methods being studied.

### 6.1.1 Results from simulation based on the Polyps study

Using a replication value of 1,000 (each with a sample size of 250) for the simulation laid out in section 6.1 produces the results shown below in Table 57 (only the first 5 replicates are shown).

**Table 57:** Estimated risk ratios for each method from the first 5 iterations of the simulation study with one treatment group. The true risk ratio is 0.5499.

| **No Adjustment** | **Method 1** | **Method 2** | **Method 3** | **Method 4** | **Method 5** | **Method 6** |
|---|---|---|---|---|---|---|
| 0.5203 | 0.5610 | 0.5540 | 0.5494 | 0.5491 | 0.5501 | 0.5495 |
| 0.5287 | 0.5450 | 0.5427 | 0.5409 | 0.5408 | 0.5463 | 0.5411 |
| 0.5274 | 0.5513 | 0.5470 | 0.5449 | 0.5448 | 0.5591 | 0.5446 |
| 0.5116 | 0.5455 | 0.5393 | 0.5365 | 0.5362 | 0.5437 | 0.5364 |
| 0.5270 | 0.5434 | 0.5406 | 0.5407 | 0.5406 | 0.5526 | 0.5408 |

Table 58 below, now shows the average point estimate produced for the Risk Ratio along with estimates for the bias and mean squared error produced by each of the six trial methods being studied.

**Table 58:** Average estimated risk ratio, bias and mean squared error for each of the six trial methods having run 1000 replicates each with sample size 250.

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.5253 | −0.0246 | 0.0211 | 0.0324 |
| Offset | 0.5554 | 0.0055 | 0.0179 | 0.0187 |
| Continuous | 0.5509 | 0.0010 | 0.0176 | 0.0176 |
| Categorical | 0.5482 | −0.0017 | 0.0181 | 0.0181 |
| Random Effect | 0.5473 | −0.0026 | 0.0180 | 0.0182 |
| Mantel-Haenszel | 0.5483 | −0.0016 | 0.0181 | 0.0181 |
| Conditional Method | 0.5554 | 0.0055 | 0.0179 | 0.0187 |

*The Conditional methodology failed to converge on 89 replicates. Thus, the analysis is based on only the 911 replicates for which all methods achieved convergence

Table 58 above shows that the variability in the estimates produced for each of the six types of adjustment are similar (to 4dp). The assessment of which method is best is made using the root mean squared error (RMSE). This shows that the results produced using no adjustment for heterogeneity are the worst. Thus, some adjustment for baseline heterogeneity is required. The continuous adjustment has the lowest RMSE. Hence, the continuous adjustment appears the better method for adjusting baseline heterogeneity.

### 6.1.2 Influence of a smaller sample size of each replication on the simulation results

In Section 6.1.2, the continuous method is shown to be the best method. The difference between the trial methods is very small though and there is very little variability between replications. Given this, Table 59 below repeats the simulation with 1000 replications however the sample size for each is reduced to 100.

**Table 59:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the six trial methods having run 1000 iterations with sample size 100.

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.5256 | −0.0230 | 0.0340 | 0.0410 |
| Offset | 0.5462 | −0.0037 | 0.0295 | 0.0297 |
| Continuous | 0.5503 | 0.0004 | 0.0292 | 0.0292 |
| Categorical | 0.5431 | 0.0068 | 0.0312 | 0.0319 |
| Random effect | 0.5408 | −0.0091 | 0.0312 | 0.0325 |
| Mantel Haenszel | 0.5497 | −0.0002 | 0.0279 | 0.0279 |
| Conditional method | 0.5432 | −0.0067 | 0.0316 | 0.0323 |

*The Conditional methodology failed to converge on 155 iterations. Thus, the analysis is based on only the 845 iterations for which all methods achieved convergence

Table 59 above shows that all six trial methods have a similar RMSE which is also lower than the RMSE for the no adjustment method. This again shows the need to adjust for baseline heterogeneity. The Mantel-Haenszel and has the lowest RMSE and could therefore be viewed the better methods. It is worth noting that the continuous and offset adjustments produced very similar levels of RMSE.

### 6.1.3 Influence of true treatment effect on the simulation results

The above subsection compared the relative performance of the 6 trial methods at adjusting baseline heterogeneity for a specific Risk Ratio (0.5499). But, what happens for treatments which cause other Risk Ratios? Is the best method for adjusting baseline heterogeneity influenced by the magnitude of the Risk Ratio?

These questions are answered by performing the same simulation but with different Risk Ratios. The new Risk Ratios to be examined are 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2. Tables 60 - 67 show the average point estimate along with the Bias and Mean Squared Error for each of the five trial methods (at adjusting baseline heterogeneity) being studied and for when no adjustment is performed.

**Table 60:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the six trial methods when the true risk ratio is 0.25.

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.2392 | −0.0108 | 0.0207 | 0.0234 |
| Offset | 0.2531 | 0.0031 | 0.0187 | 0.0189 |
| Continuous | 0.2511 | 0.0011 | 0.0184 | 0.0184 |
| Categorical | 0.2502 | 0.0002 | 0.0193 | 0.0193 |
| Random Effect | 0.2489 | −0.0011 | 0.0191 | 0.0191 |
| Conditional method | 0.2531 | 0.0031 | 0.0187 | 0.0190 |
| Mantel Haenszel | 0.2505 | 0.0005 | 0.0196 | 0.0196 |

**Table 61:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the six trial methods when the true risk ratio is 0.5.

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.4797 | −0.0203 | 0.0346 | 0.0401 |
| Offset | 0.050 39 | 0.0039 | 0.0297 | 0.0300 |
| Continuous | 0.5000 | 0.0000 | 0.0291 | 0.0291 |
| Categorical | 0.4965 | −0.0035 | 0.0314 | 0.0316 |
| Random Effect | 0.4948 | −0.0052 | 0.0315 | 0.0319 |
| Conditional method | 0.5039 | −0.0039 | 0.0298 | 0.0300 |
| Mantel Haenszel | 0.4965 | −0.0035 | 0.0313 | 0.0315 |

**Table 62:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 0.75.

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.7124 | −0.0376 | 0.0421 | 0.0565 |
| Offset | 0.7567 | 0.0067 | 0.0331 | 0.0338 |
| Continuous | 0.7504 | 0.0004 | 0.0336 | 0.0336 |
| Categorical | 0.7448 | −0.0052 | 0.0365 | 0.0369 |
| Random Effect | 0.7419 | −0.0081 | 0.0365 | 0.0374 |
| Conditional method | 0.7567 | 0.0067 | 0.0331 | 0.0338 |
| Mantel Haenszel | 0.7451 | −0.0049 | 0.0361 | 0.0365 |

**Table 63:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.

| Method | Average estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|
| No adjustment | 0.9598 | −0.0402 | 0.0668 | 0.0779 |
| Offset | 1.0197 | 0.0197 | 0.0516 | 0.0553 |
| Continuous | 1.0094 | 0.0094 | 0.0512 | 0.0520 |
| Categorical | 1.0059 | 0.0059 | 0.0556 | 0.0559 |
| Random Effect | 1.0023 | 0.0023 | 0.0554 | 0.0555 |
| Conditional method | 1.0196 | 0.0196 | 0.0517 | 0.0553 |
| Mantel Haenszel | 1.0060 | 0.0060 | 0.0555 | 0.0558 |

**Table 64:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.25.

| Method | Average estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|
| No adjustment | 1.1939 | −0.0561 | 0.0656 | 0.0863 |
| Offset | 1.2568 | 0.0068 | 0.0592 | 0.0596 |
| Continuous | 1.2484 | −0.0016 | 0.0564 | 0.0564 |
| Categorical | 1.2448 | −0.0052 | 0.0601 | 0.0603 |
| Random Effect | 1.2414 | −0.0086 | 0.0596 | 0.0602 |
| Conditional method | 1.2568 | 0.0068 | 0.0593 | 0.0596 |
| Mantel Haenszel | 1.2448 | −0.0052 | 0.0599 | 0.0601 |

**Table 65:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.5.

| Method | Average estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|
| No adjustment | 1.4262 | −0.0738 | 0.0763 | 0.1061 |
| Offset | 1.5111 | 0.0111 | 0.0594 | 0.0605 |
| Continuous | 1.4999 | −0.0001 | 0.0590 | 0.0590 |
| Categorical | 1.4965 | −0.0035 | 0.0646 | 0.0647 |
| Random Effect | 1.4923 | −0.0077 | 0.0641 | 0.0646 |
| Conditional method | 1.5111 | 0.0111 | 0.0595 | 0.0605 |
| Mantel Haenszel | 1.4965 | −0.0031 | 0.0647 | 0.0647 |

**Table 66:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 1.75.

| Method | Average estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|
| No adjustment | 1.6875 | −0.0625 | 0.0945 | 0.1133 |
| Offset | 1.7817 | 0.0317 | 0.0709 | 0.0776 |
| Continuous | 1.7706 | 0.0206 | 0.0722 | 0.0750 |
| Categorical | 1.7636 | 0.0136 | 0.0786 | 0.0798 |
| Random Effect | 1.7593 | 0.0093 | 0.0782 | 0.0787 |
| Conditional method | 1.7816 | 0.0316 | 0.0708 | 0.0776 |
| Mantel Haenszel | 1.7633 | 0.0133 | 0.0782 | 0.0794 |

**Table 67:** Average estimated Risk Ratio, Bias and Mean Squared Error for each of the five trial methods when the true risk ratio is 2.

| Method | Average estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|
| No adjustment | 1.9124 | −0.0876 | 0.1085 | 0.1394 |
| Offset | 2.0323 | 0.0323 | 0.0788 | 0.0852 |
| Continuous | 2.0153 | 0.0153 | 0.0778 | 0.0792 |
| Categorical | 2.0065 | 0.0065 | 0.0807 | 0.0810 |
| Random Effect | 2.0011 | 0.0011 | 0.0813 | 0.0813 |
| Conditional method | 2.0323 | 0.0323 | 0.0788 | 0.0852 |
| Mantel Haenszel | 2.0073 | 0.0073 | 0.0804 | 0.0807 |

Tables 60-67 above show that the results produced using no adjustment for heterogeneity are the worst (the no adjustment method consistently has the highest RMSE). Thus, some adjustment for baseline heterogeneity is required. The continuous adjustment obtains the lowest RMSE in each case and is therefore viewed the better method for adjusting baseline heterogeneity. It is worth noticing that the RMSE is often very similar for the Random, Categorical and Mantel-Haenszel methods suggesting these methods perform equally well to each other. Finally, the Offset and Conditional methods have identical RMSEs in all but one case and their performance in relation to the other methods is more favourable at lower risk ratios.

### 6.1.4 Relationship between the RMSE and true risk ratio

This section provides a visual comparison (Figure 47 below) of the RMSE at varying risk ratios.

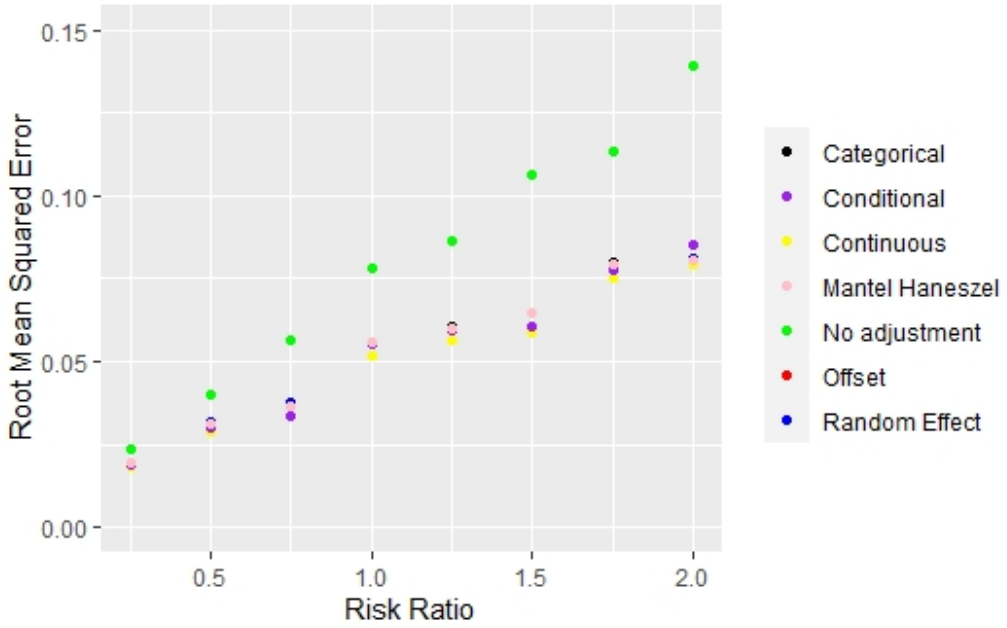


**Figure 47:** Scatter plot showing the relationship between Risk Ratio and RMSE for no adjustment and all 6 trial methods.

Figure 33 above shows the RMSE increases as the true risk ratio increases for all the methods being studied. The stark difference in results, between not adjusting and adjusting for baseline heterogeneity, shows just how vital it is that some form of adjustment is made for baseline heterogeneity. The penalty of not adjusting is also larger as the risk ratio increases. The RMSEs for the categorical, random effect and Mantel-Haenszel methods are extremely similar. This is not surprising as they all involve forming the same stratums in the analysis. It is also very clear, just how successful the Continuous method is at adjusting for baseline heterogeneity.

## 6.2 A simulation based on the Belcap study

Section 6.1 above examines the effectiveness of the 6 trial methods when there is only 1 experimental group however, many studies in real life have multiple experimental groups such as the Belcap study. Thus in this section, the situation with more than one treatment group is considered. The effectiveness of the 6 trial methods are again assessed via their bias and root mean squared error (RMSE). Hence another simulation is required and a modified algorithm, to account for the extra experimental treatments, is outlined below. See appendix 2 for the R code.

Part 1: Set the parameters for the simulation

Step 1: Produce a Multinomial regression model where the treatment group is the outcome variable and the baseline value is the predictor. This is done using the original Belcap data set.

Step 2: Produce a null Poisson model with baseline value as the outcome. This is also done using the original Belcap data set.

Step 3: Produce a Poisson model with end data as the outcome, treatment group as a categorical predictor and log baseline value as a continuous predictor. The point of this model is to understand the relationship between the different treatment groups having adjusted for baseline heterogeneity (via the log baseline term). This model is produced using the original Belcap data set too.

Part 2: Perform the simulation

Step 4: Simulate baseline data from a Poisson distribution where the parameter is equal to the linear predictor from step 2. The example in section 5.2.1 simulates 500 values.

Step 5: For each of the simulated baseline values from step 4, predict the probability of it belonging to each treatment group using the Mulitnomial model from step 1.

Step 6: For each of the simulated baseline values (from step 4) separately,

simulate a draw from a Multinomial distribution, where the probability of being in any of the treatment groups is equal to the probability predicted in step 5. The order in which the probabilities are given matters. One way of stating the probabilities is control first, then the first experimental, then the second experimental and so on. This ordering is used during the example in section 6.2.1. Every draw from the distribution gives you a sequence of numbers. The length of the sequence is equal to the number of treatment groups. All the numbers in the sequence are zero except one which has value 1. The position of the value 1 in the sequence determines which group the baseline value belongs to. For example, if the 1 comes first in the sequence, the baseline value belongs to the treatment group specified first in the Multinomial distribution.

Step 7: Now, combine the simulated baseline data and the indicator for treatment group together forming a data set with 2 columns.

Step 8: Simulate end of study data from a Poisson distribution where the parameter is given by the linear predictor of the model in step 3. This uses the baseline data and the indicator of treatment group from the data set produced in step 7.

Step 9: Combine the end of study data with the data set from step 7. This forms the full simulated data set.

Step 10: Apply all 7 methods (no adjustment and 6 trial methods) being studied to the simulated data set.

Step 11: Repeat steps 4 - 10 multiple times. The example shown in section 6.2.1 uses a replication value of 1,000.

Part 3: Analyse the results from the simulation

Step 12: Calculate the biases and root mean squared errors for the estimates of the risk ratio (for the treatment group) produced by each of the seven methods.

### 6.2.1 Results from simulation based on the Belcap study

The Belcap study is going to be used as an example. This data set has 6 treatment groups, 5 experimental and 1 control. The results from applying the simulation outlined in section 6.2 are shown below in Table 68.

**Table 68:** Results from a simulation with a replication number of 1,000 (each with a sample size of 750 observations) where the true risk ratios for ALL, ESD, MW, OHE, OHY are 0.5724, 0.8226, 0.6615, 0.7043, 0.7348 respectively

| **Method** | **Treatment** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|---|
| ALL | No Adjustment | 0.5643 | −0.0288 | 0.0345 | 0.0449 |
| | Offset | 0.6181 | 0.0250 | 0.0393 | 0.0466 |
| | Continuous | 0.5943 | 0.0012 | 0.0347 | 0.0347* |
| | Categorical | 0.5906 | −0.0025 | 0.0347 | 0.0348 |
| | Random Effect | 0.5896 | −0.0035 | 0.0346 | 0.0348 |
| | Mantel Haenszel | 0.5915 | −0.0016 | 0.0349 | 0.0349 |
| | Conditional method | 0.6272 | 0.0341 | 0.0422 | 0.0543 |
| ESD | No Adjustment | 0.8113 | −0.0149 | 0.0429 | 0.0454 |
| | Offset | 0.8382 | 0.0120 | 0.0414 | 0.0431 |
| | Continuous | 0.8266 | 0.0003 | 0.0397 | 0.0397* |
| | Categorical | 0.8252 | −0.0010 | 0.0397 | 0.0397 |
| | Random Effect | 0.8247 | −0.0015 | 0.0396 | 0.0397 |
| | Mantel Haenszel | 0.8244 | −0.0018 | 0.0402 | 0.0402 |
| | Conditional method | 0.8425 | 0.0163 | 0.0438 | 0.0467 |
| MW | No Adjustment | 0.6977 | −0.0054 | 0.0393 | 0.0397 |
| | Offset | 0.7081 | 0.0051 | 0.0372 | 0.0375 |
| | Continuous | 0.7036 | 0.0005 | 0.0354 | 0.0354* |
| | Categorical | 0.7026 | −0.0005 | 0.0362 | 0.0362 |
| | Random Effect | 0.7024 | −0.0006 | 0.0362 | 0.0362 |
| | Mantel Haenszel | 0.7025 | −0.0006 | 0.0368 | 0.0368 |
| | Conditional method | 0.7097 | 0.0067 | 0.0388 | 0.0393 |
| OHE | No Adjustment | 0.7136 | 0.0052 | 0.0408 | 0.0412 |
| | Offset | 0.7055 | −0.0029 | 0.0377 | 0.0378 |
| | Continuous | 0.7087 | 0.0003 | 0.0357 | 0.0357* |
| | Categorical | 0.7098 | 0.0013 | 0.0364 | 0.0364 |
| | Random Effect | 0.7099 | 0.0015 | 0.0364 | 0.0365 |
| | Mantel Haenszel | 0.7100 | 0.0016 | 0.0367 | 0.0367 |
| | Conditional method | 0.7044 | −0.0040 | 0.0394 | 0.0396 |
| OHY | No Adjustment | 0.7401 | −0.0225 | 0.0438 | 0.0493 |
| | Offset | 0.7794 | 0.0168 | 0.0453 | 0.0483 |
| | Continuous | 0.7623 | −0.0004 | 0.0414 | 0.0414* |
| | Categorical | 0.7595 | −0.0031 | 0.0418 | 0.0419 |
| | Random Effect | 0.7588 | −0.0038 | 0.0418 | 0.0419 |
| | Mantel Haenszel | 0.7601 | −0.0025 | 0.0418 | 0.0419 |
| | Conditional method | 0.7856 | 0.0229 | 0.0483 | 0.0535 |

* best result for this treatment group

The findings shown by Table 68 above vary slightly from the case of just 1 treatment group. It appears the inclusion of extra treatment groups causes the Conditional method to perform worse than the no adjustment methods. The Offset method was also reasonably close to the no adjustment method. On this basis, it is safe to say these are the worst three methods at adjusting for baseline heterogeneity. The difference between the no adjustment method and the remaining trial methods is fairly large like in the case of one treatment group. The Continuous method performs better for all the treatment groups however, like in the one treatment group scenario, the Continuous, Categorical and Mantel-Haenszel methods are fairly similar. It is also worth noting that the Random Effect also has similar performance to the top 3 methods with the extra treatment groups where, this was not the case in the one treatment group scenario. The final conclusion to draw from these findings is that the Continuous method is the best choice for adjusting baseline heterogeneity.

### 6.2.2 Influence of true treatment effect on the simulation results

Section 6.2.1 above examined how the different methods performed with the risk ratios from the Belcap study. This section looks at how varying the true risk ratio for one of the treatment groups affects the performance of the method with no adjustment for baseline heterogeneity and the five methods being studied.

Given that only one treatment group's true risk ratio is being changed, only estimates for this treatment group will be affected. The "ALL" treatment group is the one being modified and when presenting the results in Table 69 below, only the results for this group will be presented.

**Table 69:** Simulation with a replication number of 1,000. Each replicate has a sample size of 750. The true risk ratio for ALL is varied

| True Risk Ratio | Method | Average Estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|---|
| 0.5 | No Adjustment | 0.4735 | −0.0265 | 0.0315 | 0.0412 |
| | Offset | 0.5197 | 0.0197 | 0.0335 | 0.0389 |
| | Continuous | 0.4993 | −0.0007 | 0.0307 | 0.0307* |
| | Categorical | 0.4962 | −0.0038 | 0.0309 | 0.0312 |
| | Random Effect | 0.4956 | −0.0044 | 0.0309 | 0.0312 |
| | Mantel Haenszel | 0.4961 | −0.0039 | 0.0315 | 0.0317 |
| | Conditional Method | 0.5277 | 0.0277 | 0.0354 | 0.0450 |
| 1 | No Adjustment | 0.9826 | −0.0174 | 0.0519 | 0.0547 |
| | Offset | 1.0173 | 0.0173 | 0.0528 | 0.0556 |
| | Continuous | 1.0021 | 0.0021 | 0.0472 | 0.0473* |
| | Categorical | 1.0003 | 0.0003 | 0.0477 | 0.0477 |
| | Random Effect | 0.9999 | −0.0001 | 0.0477 | 0.0477 |
| | Mantel Haenszel | 1.0007 | 0.0007 | 0.0480 | 0.0480 |
| | Conditional Method | 1.0230 | 0.0230 | 0.0563 | 0.0608 |
| 1.5 | No Adjustment | 1.4899 | −0.0101 | 0.0718 | 0.0725 |
| | Offset | 1.5122 | 0.0122 | 0.0666 | 0.0677 |
| | Continuous | 1.5023 | 0.0023 | 0.0614 | 0.0615* |
| | Categorical | 1.5026 | 0.0026 | 0.0630 | 0.0631 |
| | Random Effect | 1.5023 | 0.0023 | 0.0630 | 0.0631 |
| | Mantel Haenszel | 1.5023 | 0.0023 | 0.0636 | 0.0636 |
| | Conditional Method | 1.5159 | 0.0159 | 0.0710 | 0.0728 |
| 2 | No Adjustment | 2.0088 | 0.0088 | 0.1015 | 0.1018 |
| | Offset | 1.9879 | −0.0121 | 0.0900 | 0.0908 |
| | Continuous | 1.9964 | −0.0036 | 0.0842 | 0.0843* |
| | Categorical | 1.9997 | −0.0003 | 0.0868 | 0.0868 |
| | Random Effect | 1.9999 | −0.0001 | 0.0869 | 0.0869 |
| | Mantel Haenszel | 1.9988 | −0.0012 | 0.0872 | 0.0872 |
| | Conditional Method | 1.9851 | −0.0149 | 0.0955 | 0.0967 |

* best result for this risk ratio

Table 69 above shows that regardless of risk ratio, the no adjustment method performs very badly which demonstrates the need for some form of adjustment for heterogeneity. The conditional method is not suitable for adjusting for baseline heterogeneity given that it often performs worse than the no adjustment method. The best method (based on having the lowest RMSE) is the continuous method for all risk ratios.

### 6.2.3 Relationship between the bias and true risk ratio

This section provides a visual comparison (Figure 48 below) of the RMSE at varying risk ratios.

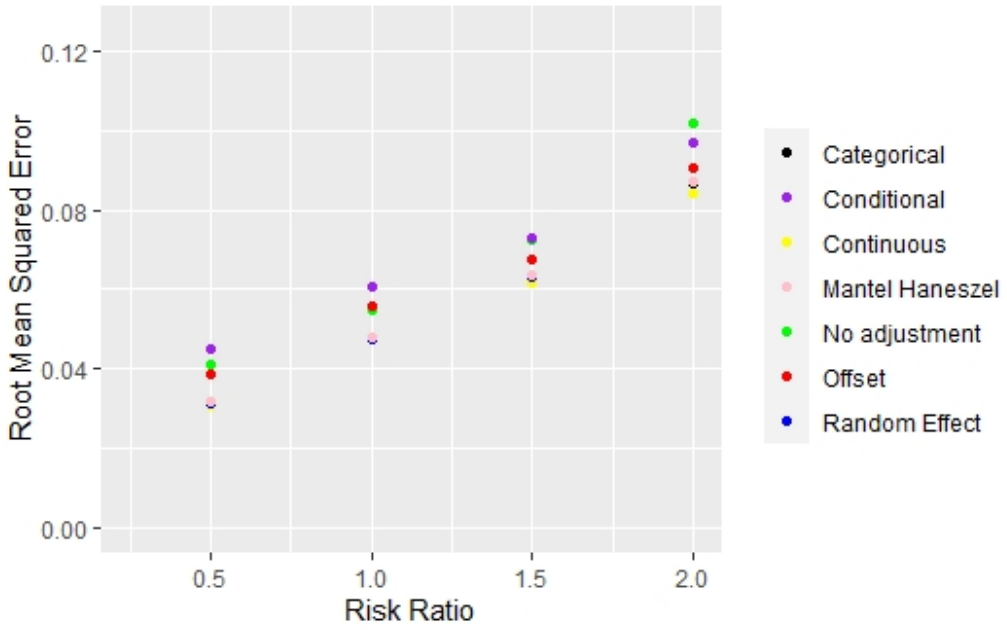


**Figure 48:** Scatter plot showing the relationship between Risk Ratio and RMSE for no adjustment and all 6 trial methods.

Figure 34 above shows the RMSE increases as the true risk ratio increases for all the methods being studied. The RMSEs for the categorical and random effect are identical to each other. The Mantel-Haenszel method is very similar to the random effect and categorical method. It is not surprising that these three methods (categorical, random effect and Mantel-Haenszel) as they all involve forming the same stratums in the analysis. The continuous method gives the smallest RMSE at all risk ratios but it is only once the risk ratio hits 1.5 that the difference between the continuous method and the others is large enough to show up in Figure 34.

## 6.3 A simulation based on a cluster randomised trial where there are more clusters than treatments

In this section, the situation where there are more clusters than treatments is considered. In the situation of the Belcap study, this means more than one school being allocated to each treatment. Table 70 below shows the results for a treatment group with two clusters allocated to it and the true risk ratio for the treatment is 0.5.

**Table 70:** 2 clusters allocated to each treatment group

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.4917 | −0.0082 | 0.0180 | 0.0198 |
| Offset | 0.5066 | 0.0066 | 0.0171 | 0.0183 |
| Continuous | 0.5002 | 0.0002 | 0.0163 | 0.0163 |
| Categorical | 0.4992 | −0.0008 | 0.0167 | 0.0167 |
| Random Effect | 0.4989 | −0.0011 | 0.0167 | 0.0167 |
| Mantel Haenszel | 0.4992 | −0.0008 | 0.0167 | 0.0167 |
| Conditional method | 0.5091 | 0.0091 | 0.0178 | 0.0200 |

Table 70 shows the similar finding of the continuous method being best at adjusting for baseline heterogeneity. Interestingly, there is virtually no difference in the success of the categorical, random effect and Mantel-Haenszel methods. Like before, the conditional method is not successful at all.

Table 71 below repeats the same simulation as above yet there are now three clusters being allocated to each treatment group.

**Table 71:** 3 clusters allocated to each treatment group

| Method | Average estimate | Bias | Standard Deviation | RMSE |
|---|---|---|---|---|
| No adjustment | 0.4897 | −0.0103 | 0.0220 | 0.0243 |
| Offset | 0.5046 | 0.0046 | 0.0231 | 0.0236 |
| Continuous | 0.4987 | −0.0013 | 0.0220 | 0.0221 |
| Categorical | 0.4977 | −0.0023 | 0.0228 | 0.0229 |
| Random Effect | 0.4974 | −0.0026 | 0.0227 | 0.0229 |
| Mantel Haenszel | 0.4977 | −0.0023 | 0.0229 | 0.0231 |
| Conditional method | 0.5058 | 0.0058 | 0.0248 | 0.0255 |

Comparing Tables 70 and 71, it is noticeable that the RMSEs are higher in Table 71. This means that the results are less accurate in the case where there are more clusters allocated to a treatment. This is due to the greater within treatment group variation.

## 6.4 A simulation based on data with a third time point

All the simulations above agree that the adaptation of the conditional linear model proposed by Verbeke compares unfavourably with the other methods being studied here. All the above situations for the simulations have one key similarity, they all assume data is only collected at the start and end of the study. This is by no means always the case in reality (the Falls dataset is an example of an exception). One of the potential upsides of how the conditional method as defined here was that it could deal with longitudinal data with multiple time points. For this reason another simulation is performed to analyse the performance of the different methods being studied when there are interim follow ups. For simplicity and the purpose of comparability, the methods used to analyse the data from this simulation are used as defined in section 4.2. In addition, the scenario is only extended to have three time points and only one treatment group. Below is an outline of the simulation process. See appendix 3 for the r code.

Part 1: Set the parameters for the simulation

Step 1: Produce a Logistic regression model where the treatment group is the outcome variable and the baseline value is the predictor. This is done

using the original Polyps data set.

Step 2: Produce a null Poisson model with baseline value as the outcome. This is also done using the original Polyps data set.

Step 3: Produce a Poisson model with end data as the outcome, treatment group as a categorical predictor and log baseline value as a continuous predictor. The point of this model is to understand the relationship between the experiment and control treatment having adjusted for baseline heterogeneity (via the log baseline term). This model is produced using the original Polyps data set too.

Part 2: Perform the simulation

Step 4: Simulate baseline data from a Poisson distribution where the parameter is equal to the linear predictor from step 2. The example in section 6.1.1 simulates 250 values.

Step 5: For each of the simulated baseline values from step 4, predict the probability of it belonging to the experimental group using the Logistic model from step 1.

Step 6: For each of the simulated baseline values (from step 4) separately, simulate a draw from a Bernoulli distribution where the "Success" probability is equal to the probability predicted in step 5. If the value drawn from the Bernoulli distribution is 1 then the baseline value belongs to the experimental group. Should the value drawn from the Bernoulli distribution be 0 then the baseline value belongs to the control group.

Step 7: Now, combine the simulated baseline data and the indicator for treatment group together forming a data set with 2 columns.

Step 8: Simulate data from a Poisson distribution where the parameter is given by the linear predictor of the model in step 3. This uses the baseline data and the indicator of treatment group from the data set produced in step 7. Label the simulated data as "Time 1".

Step 9: Combine the end of study data with the data set from step 7.

Step 10: Repeat step 8 but use "Time 1" data in place of the baseline data. Label the simulated data as "End data ".

Step 11: Add the "End data" to the dataset in step 9 to form the final dataset

Step 12: Apply the 5 trial methods being studied to the simulated data set.

Step 13: Repeat steps 4 - 10 multiple times. The example shown in section 6.4.1 uses a replication number of 1,000.

Part 3: Analyse the results from the simulation

Step 14: Calculate the biases and mean squared errors for the estimates of the risk ratio (for the treatment group) produced by each of the six methods being studied.

### 6.4.1 Results from simulation based on three time points

Table 72 below shows the results from implementing the above simulation. Here a replication number of 1,000 is used and each run has a sample size of 250. The true risk ratio by the end of the study was 0.2767.

**Table 72:** Average estimated risk ratio, bias and mean squared error for each of the six trial methods having run 1000 replicates with sample size 250.

| **Method** | **Average Estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.2652 | −0.0115 | 0.0223 | 0.0251 |
| Offset | 0.2801 | 0.0034 | 0.0223 | 0.0226 |
| Continuous | 0.2761 | −0.0006 | 0.0219 | 0.0219 |
| Categorical | 0.2748 | −0.0019 | 0.0221 | 0.0222 |
| Random Effect | 0.2734 | −0.0033 | 0.0220 | 0.0222 |
| Mantel Haenszel | 0.2747 | −0.0020 | 0.0222 | 0.0222 |
| Conditional Method | 0.2803 | 0.0036 | 0.0223 | 0.0226 |

*The Conditional methodology failed to converge on 277 iterations. Thus, the analysis is based on only the 723 replicates for which all methods achieved convergence

Table 72 above shows that the inclusion of a third time point hasn't changed the pecking order of the different methods. Thus, all the conclusions made above appear to hold true for the scenario of three time points.

## 6.5 A simulation created from the Offset Method

All the simulations carried out in the previous sections all find the Continuous method to be best at adjusting for baseline heterogeneity. The only concern with this is that all the simulations were created using results from a Continuous method. Thus, a simulation is produced which replicates the process outlined in Section 6.2 except for the outcome data being generated using the Offset method. Table 73 below shows the results of this simulation.

**Table 73:** Simulation based on the Offset method

| **Method** | **Average estimate** | **Bias** | **Standard Deviation** | **RMSE** |
|---|---|---|---|---|
| No adjustment | 0.5339 | −0.0592 | 0.0326 | 0.0675 |
| Offset | 0.5849 | −0.0082 | 0.0206 | 0.0221 |
| Continuous | 0.5943 | 0.0012 | 0.0205 | 0.0205 |
| Categorical | 0.5877 | −0.0054 | 0.0221 | 0.0228 |
| Random Effect | 0.5875 | −0.0056 | 0.0221 | 0.0228 |
| Mantel Haenszel | 0.5879 | −0.0052 | 0.0223 | 0.0229 |
| Conditional method | 0.6966 | 0.1035 | 0.0223 | 0.1059 |

Table 73 shows the Continuous method is still best at adjusting for baseline heterogeneity. Like before, the conditional method is still not successful at all.

# 7 Discussion and Conclusions

The goals of this thesis are to demonstrate the need to adjust for baseline heterogeneity when analysing Count Data from Experimental. The thesis also tries to find a successful statistical method for carrying out this adjustment. The statistical methods considered consist of 5 parametric methods (Offset, Continuous, Categorical, Random Effect, and Conditional) and a Non-parametric method (Mantel-Haenszel). The analysis within this thesis is based on 3 datasets (Belcap, Polyps, Falls) and many simulations which are created to match the properties of one of the 3 datasets.

## 7.1 Demonstrating the need to adjust for baseline heterogeneity

This goal is addressed by analysing the 3 datasets using the 5 parametric methods, a non-parametric method and a method which does not allow for baseline heterogeneity. The method that does not allow for baseline heterogeneity is included to provide a comparison between analysis with and without allowing for baseline heterogeneity. The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) statistics are used to examine how well the statistical model used within each method fits the data. The AIC and BIC are not available for Non-parametric methods and the values produced for the Conditional method cannot be meaningfully compared with the other values. This leaves values for AIC and BIC from 4 of the Parametric Methods (Offset, Continuous, Categorical and Random Effect) as well as the method with no allowance for heterogeneity. Comparing these values, the AIC and BIC are much higher when not allowing for heterogeneity meaning it does not fit the data very well. Thus, it is clear that a method which allows for baseline heterogeneity is required. This supports the findings of Senn [38].

A purely hypothetical situation is also drawn up to demonstrate how severe the consequences of not correctly allowing for baseline heterogeneity can be. This includes a potentially unsafe drug being wrongly authorised for general use.

## 7.2 Finding the best method to adjust for baseline heterogeneity

The AIC and BIC for the Offset, Continuous, Categorical and Random Effect methods are all fairly similar, hence, they do not differ significantly in terms of how well they fit the data. Interestingly, the estimates for the treatment effect(s) vary substantially between the 5 parametric methods and the non-parametric method. For this reason some way of distinguishing between all these methods is required. This is done using simulation studies based upon each of the original datasets separately. The baseline data are simulated to have a Poisson distribution with the simulated mean equal to the mean of the baseline data in the original dataset. The data is then split into the required number of treatment groups in such a way that the resulting treatment groups are a close match to the treatment groups in the original datasets. The outcome data (for the simulated datasets) is then created according to the model produced by using the Continuous method to analyse the original dataset (e.g. Belcap / Polyps / Falls). Once the simulated datasets are produced, the 5 parametric methods and the non-parametric method are used to analyse each simulated dataset. The method with no adjustment is also used to analyse the simulated datasets to further emphasise how important it is to allow for baseline heterogeneity. From this analysis, an estimate for each treatment effect is produced for each of the methods. As the simulations are repeated 1000 times, 1000 estimates of each treatment effect are produced for each method (i.e.7000 estimates in total). These estimates of the treatment effect are then summarised by the Root Mean Squared Error (RMSE) for each method separately. Thus at the end of the analysis, one RMSE is obtained for each method.

Applying the above simulation methodology to the Polyps dataset leads to all 5 parametric methods and the non-parametric method having smaller RMSEs than the method with no adjustment for baseline heterogeneity. Thus providing more evidence of how important it is to allow for baseline heterogeneity. The results appear to be reasonably robust to changes in the sample size of each simulation and changes in the assumed treatment effect. The RMSE's for the Categorical, Continuous and the Mantel-Haenszel methods are very similar. This suggests that any one of these methods is a reasonable choice for adjusting baseline heterogeneity in the simple case of 1 treatment group and 2 time points.

The simulation strategy is also applied to the Belcap dataset in order to see whether the relative performance of the parametric and non-parametric methods vary with the number of treatment groups. Note, the Polyps dataset contains only 2 treatment groups where, the Belcap study contains 6 treatment groups. The simulations show that the Continuous method performs the best (i.e. lowest RMSE) in this situation. The Conditional method performs very badly in the simulation with many treatment groups, to the extent that not allowing for baseline heterogeneity produces better estimates for the treatment effect(s). This finding was not expected given that research by Verbeke shows that using a Conditional Mixed model successfully allows for baseline heterogeneity. One potential explanation for the contrary results is that the data being analysed here is Count Data whereas, the data analysed by Verbeke was normally distributed.

Repeating the simulation strategy for the Falls dataset which has 2 treatment groups and 3 time points leads to the same results as those produced when simulating the Belcap dataset. The Continuous method has the lowest RMSE and the Conditional method has the highest RMSE. This result was not expected as the Conditional method is the only method being studied which makes use of the data from the additional time point. One piece of anlysis not considered here would have been to look whether treating the count data as having an approximate normal distribution leads to better results. This approximation could have been viewed as being valid due to the large sample sizes (hence the Central Limit Theorem holds) in the simulations.

One concern regarding the Continuous method producing the best estimates of the treatment effect(s) (smallest RMSE), is that the simulations were designed based on the results from applying the Continuous method to the relevant example dataset. Given this, another simulation is performed based on the Offset method to check whether the same result is obtained. The extra simulation produced the same result as before i.e. the Continuous method has the smallest RMSE. This may not be too surprising as the offset term is the same as a continuous term with coefficient equal to 1.

## 7.3 Limitations of this thesis

One limitation of this thesis is that only Count Data is considered. This means the findings of this research are only useful in the case of Count Data.

In addition, the models being used in the simulation studies do not take account of overdispersion which frequently occurs in Count Data. Another limitation is that only experimental studies were looked at and the issue of baseline heterogeneity is arguably a larger problem in observational studies as methods like randomisation are not used to create balanaced groups.

The parametric models considered in this thesis were not extended to include any polynomial terms which could potentially lead to better model fits. The definition of the strata being used in the Categorical and Random Effect methods is very subjective. The models can be made artificially good or bad based on the definition of the strata being used. Creating only 1 stratum makes the Categorical and Random Effect artificially bad as all individuals are in the same stratum and are therefore treated as being the same. At the other extreme, creating strata such that the strata all contain only 1 individual makes the Random Effect model and the Categorical model artificially good. For this reason a compromise between the above two situations is required. In this thesis, strata are created based on the Quintiles of the data. Thus, individuals with baseline values below the first quintile are in the first strata and so on.

## 7.4 Overall conclusions and Further work

From the analysis in this thesis, it is clear that baseline heterogeneity should be allowed for in the case of Count Data. This is the same as with other types of data such as Normal data. The simulation studies further demonstrate this need along with showing that the Continuous method is the better method for dealing with baseline heterogeneity. On the basis of this, it would be sensible to use the Continuous method when analysing Count Data from Experimental studies.

In the future, this research could be extended to look at using a non-linear term for the continuous adjustment or to look at data with different distributions (e.g. Binomial). Another possibility would be to look at Repeated Measures data which is either complete or incomplete. The effect of over-dispersion on the results in this thesis should also be looked into in any future research. This could be of particular value as overdispersion is common in Count Data. Finally it would also be useful to see wether the findings here are the same for count data coming from observational studies.

# 8 Appendicies

## 8.1 Appendix 1: Polyps simulation

```
# Load original polys data

setwd("C:/PhD")

polyps_edited<-read.csv('polyps edited.csv')

attach(polyps_edited)

# Poisson model (null model) to baseline data to get a mean Poisson value

library(plyr)

model_base<-glm(count_b~1,family = "poisson")

summary(model_base) # mean poisson value = exp(3.05164) = 21.15

# Logistic model to demonstrate correlation between baseline value and tr
# This is used for treatment allocation in the simulation

model_logistic<-glm(treatment~count_b,family = "binomial")

summary(model_logistic)

#########################################################################

# start loop to control number of simulations

overview<-NULL

number_iterations<-1000

sample_size_each_iteration <-100
```

```
for(a in 1:number_iterations){

  tryCatch({ #stops loop ending if error occurs. the error that can occur
##########################################################################

# Simulate baseline data

baseline<-rpois(sample_size_each_iteration,21.15)

prob_in_treat<-exp(summary(model_logistic)$coefficients[1,1]+summary(mode
                (1+exp(summary(model_logistic)$coefficients[1,1]+summary(

treat_sim<-NULL

for(i in 1:length(prob_in_treat)){

  t_sim<-rbinom(1,1,prob_in_treat[i])

  treat_sim<-rbind(treat_sim,t_sim)

}

sim_data_base_treat<-data.frame(cbind(baseline,treat_sim))

##########################################################################

# Simulate Outcome data

outcome_data<-NULL

for(k in 1:length(prob_in_treat)){

  if(sim_data_base_treat$V2[k]==0){

  outcome<-rpois(1,exp(0.4924+0.8536*log(baseline[k])))

  }else{

  outcome<-rpois(1,exp(0.4924-0.5981+0.8536*log(baseline[k])))
```

```
  }

  outcome_data<-rbind(outcome_data,outcome)

}

sim_data_base_treat_out<-cbind(sim_data_base_treat,outcome_data)

names(sim_data_base_treat_out)<-c("baseline","treat","outcome_adj")

#########################################################################

# Create baseline value categories


cate<-NULL

for (i in 1:length(baseline)){

  if(sim_data_base_treat_out$baseline[i]<=quantile(baseline,probs = 0.2))

    category<-"A"}else if (sim_data_base_treat_out$baseline[i]<=quantile(

      category<-"B"}else if (sim_data_base_treat_out$baseline[i]<=quantil

        category<-"C"}else if (sim_data_base_treat_out$baseline[i]<=quant

          category<-"D"} else{

            category<-"E"
          }

  cate<-rbind(cate,category)

}

final_data<-cbind(sim_data_base_treat_out,cate)

#########################################################################
# Poisson model for simulated dataset, ignoring baseline heterogeneity
```

```
model_no<-glm(outcome_adj~treat,data=final_data,family = "poisson")

sum_no<-summary(model_no)

# Apply the four parametric methods to the simulated dataset

model_cont<-glm(outcome_adj~treat+log(baseline),data=final_data,family =

sum_cont<-summary(model_cont)

model_off<-glm(outcome_adj~treat+offset(log(baseline)),data=final_data,fa

sum_off<-summary(model_off)

model_cat<-glm(outcome_adj~treat+cate,data=final_data,family = "poisson")

sum_cat<-summary(model_cat)

suppressMessages(library(lme4))

model_ran<- glmer(outcome_adj~treat+(1|cate),data=final_data,family=poiss

sum_ran<-summary(model_ran)

#########################################################################
#verbeke method

suppressMessages(attach(sim_data_base_treat_out))

longdata<-data.frame(outcome=c(baseline,outcome_adj),
                     parti=factor(rep(paste('p', 1:100, sep=''), 2)),
                     time=as.factor(rep(0:1, each=100)),
                     treat_group=rep(treat, 2))

suppressMessages(library(glmmADMB))

model_ver <- glmmadmb(outcome ~ treat_group*time+(1+time|parti), family =

sum_ver<-summary(model_ver)
```

```
if (ver_coef==sum_ver$coefficients[4,1]) stop(paste("error␣on␣iteration",
  }, error=function(e){cat("ERROR␣:",conditionMessage(e), "\n")})

#####################################################################
# Mantel-Haneszel

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$`1`

experiment_cat<-split(experiment,experiment$cate)

exp_a<-experiment_cat$A

exp_b<-experiment_cat$B

exp_c<-experiment_cat$C

exp_d<-experiment_cat$D

exp_e<-experiment_cat$E

control<-split_treat$`0`

control_cat<-split(control,control$cate)

con_a<-control_cat$A

con_b<-control_cat$B

con_c<-control_cat$C

con_d<-control_cat$D

con_e<-control_cat$E

numerator_a<-sum(exp_a$outcome_adj)/length(exp_a$outcome_adj)
```

```
numerator_b<-sum(exp_b$outcome_adj)/length(exp_b$outcome_adj)

numerator_c<-sum(exp_c$outcome_adj)/length(exp_c$outcome_adj)

numerator_d<-sum(exp_d$outcome_adj)/length(exp_d$outcome_adj)

numerator_e<-sum(exp_e$outcome_adj)/length(exp_e$outcome_adj)

denominator_a<-sum(con_a$outcome_adj)/length(con_a$outcome_adj)

denominator_b<-sum(con_b$outcome_adj)/length(con_b$outcome_adj)

denominator_c<-sum(con_c$outcome_adj)/length(con_c$outcome_adj)

denominator_d<-sum(con_d$outcome_adj)/length(con_d$outcome_adj)

denominator_e<-sum(con_e$outcome_adj)/length(con_e$outcome_adj)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e

weight_a<-(sum(con_a$outcome_adj)*length(exp_a$outcome_adj))/(length(con_
weight_b<-(sum(con_b$outcome_adj)*length(exp_b$outcome_adj))/(length(con_
weight_c<-(sum(con_c$outcome_adj)*length(exp_c$outcome_adj))/(length(con_
weight_d<-(sum(con_d$outcome_adj)*length(exp_d$outcome_adj))/(length(con_
weight_e<-(sum(con_e$outcome_adj)*length(exp_e$outcome_adj))/(length(con_


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e

mant_hans<-final_numerator/final_denominator

#########################################################################

# results from 5 methods
```

```
sample_size_experiment<-length(experiment$baseline)

sample_size_control<-length(control$baseline)

no_adj_coef<-sum_no$coefficients[2,1]

no_adj_se<-sum_no$coefficients[2,2]

off_coef<-sum_off$coefficients[2,1]

off_se<-sum_off$coefficients[2,2]

cont_coef<-sum_cont$coefficients[2,1]

cont_se<-sum_cont$coefficients[2,2]

cat_coef<-sum_cat$coefficients[2,1]

cat_se<-sum_cat$coefficients[2,2]

ran_coef<-sum_ran$coefficients[2,1]

ran_se<-sum_ran$coefficients[2,2]

ver_coef<-sum_ver$coefficients[4,1]

ver_se<-sum_ver$coefficients[4,2]

test_coef<-sum_ver$coefficients[2,1]+sum_ver$coefficients[4,1]

alt_coef<-sum_ver$coefficients[2,1]

full_coef<-sum_ver$coefficients[2,1]+sum_ver$coefficients[3,1]+sum_ver$co

iteration<-cbind(sample_size_experiment,sample_size_control,no_adj_coef,o
             cont_coef,cat_coef,ran_coef,ver_coef,test_coef,alt_coef,full
             log(mant_hans),no_adj_se,off_se,cont_se,cat_se,ran_se,ver_se

overview<-rbind(overview,iteration)
```

```
print(a)

}

overview<-as.data.frame(overview)
```

## 8.2 Appendix 2: Belcap simulation

```
# Load original polys data

setwd("C:/PhD")

belcap<-read.csv('belcap with A_control.csv')

belcap<- within(belcap, school <- relevel(school, ref = "Control"))

attach(belcap)

# Poisson model (null model) to baseline data to get a mean Poisson value

library(plyr)

library(nnet)

library(lme4)

model_base<-glm(dmfs_beg~1,family = "poisson")

summary(model_base) # mean poisson value = exp(1.8417) = 6.34

# Multinomial model to demonstrate correlation between baseline value and

model_multinomial<-multinom(school~dmfs_beg)

summary(model_multinomial)

continuous<-glm(dmfs_end~school+log(dmfs_beg),family = "poisson")

summary(continuous)
##########################################################################

# start loop to control number of simulations

overview_all<-read.csv("750 overview all belcap.csv")
```

```
overview_esd<-read.csv("750␣overview␣esd␣belcap.csv")

overview_mw<-read.csv("750␣overview␣mw␣belcap.csv")

overview_ohe<-read.csv("750␣overview␣ohe␣belcap.csv")

overview_ohy<-read.csv("750␣overview␣ohy␣belcap.csv")

sample_size<-750

for(a in 1:1000){

  tryCatch({ #stops loop ending if error occurs. the error that can occur

    ##########################################################################

    # Simulate baseline data

    baseline<-rpois(sample_size,6.34)

    prob_in_all<-exp(summary(model_multinomial)$coefficients[1,1]+summary
      (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(model_m
       +exp(summary(model_multinomial)$coefficients[2,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[3,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[4,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[5,1]+summary(model_mu

    prob_in_esd<-exp(summary(model_multinomial)$coefficients[2,1]+summary
      (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(model_m
       +exp(summary(model_multinomial)$coefficients[2,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[3,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[4,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[5,1]+summary(model_mu

    prob_in_mw<-exp(summary(model_multinomial)$coefficients[3,1]+summary(
      (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(model_m
       +exp(summary(model_multinomial)$coefficients[2,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[3,1]+summary(model_mu
       +exp(summary(model_multinomial)$coefficients[4,1]+summary(model_mu
```

```
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(model_mu

prob_in_ohe<-exp(summary(model_multinomial)$coefficients[4,1]+summary
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(model_m
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(model_mu

prob_in_ohy<-exp(summary(model_multinomial)$coefficients[5,1]+summary
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(model_m
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(model_mu

prob_in_con<-1/
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(model_m
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(model_mu
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(model_mu

treat_sim<-NULL

for(i in 1:length(prob_in_all)){

  t_sim<-rmultinom(1,1,c(prob_in_all[i],prob_in_esd[i],prob_in_mw[i],

  if(t_sim[1,1]==1){

    treat_sim_new<-"ALL"

  } else if(t_sim[2,1]==1){

    treat_sim_new<-"ESD"

  } else if(t_sim[3,1]==1){

    treat_sim_new<-"MW"
```

```
  } else if(t_sim[4,1]==1){

    treat_sim_new<-"OHE"

  } else if(t_sim[5,1]==1){

    treat_sim_new<-"OHY"

  } else{

    treat_sim_new<-"A_Control"

  }

  treat_sim<-rbind(treat_sim,treat_sim_new)

}

sim_data_base_treat<-data.frame(cbind(baseline,treat_sim))

names(sim_data_base_treat)<-c("baseline","treat")

sim_data_base_treat$baseline<-as.numeric(sim_data_base_treat$baseline
#####################################################################

# Simulate Outcome data

for(j in 1:length(prob_in_all)){

  if(sim_data_base_treat$treat[j]=="ALL"){

    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.1146-0.5224+0.4602*

  }else if(sim_data_base_treat$treat[j]=="ESD"){

    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.1146-0.1909+0.4602*

  }else if(sim_data_base_treat$treat[j]=="MW"){
```

```
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.1146-0.3523+0.4602*
  }else if(sim_data_base_treat$treat[j]=="OHE"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.1146-0.3447+0.4602*
  }else if(sim_data_base_treat$treat[j]=="OHY"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.1146-0.2710+0.4602*
  }else{
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.1146+0.4602*log(bas
}
#######################################################################
# Create baseline value categories
cate<-NULL
for (i in 1:length(baseline)){
  if(sim_data_base_treat$baseline[i]<=quantile(sim_data_base_treat$ba
    category<-"A"}else if (sim_data_base_treat$baseline[i]<=quantile(
      category<-"B"}else if (sim_data_base_treat$baseline[i]<=quantil
        category<-"C"}else if (sim_data_base_treat$baseline[i]<=quant
          category<-"D"}else {
            category<-"E"
          }
  cate<-rbind(cate,category)
}
```

```
final_data<-cbind(sim_data_base_treat,cate)


model_ran<- glmer(outcome~treat+(1|cate),data=final_data,
                  family=poisson,nAGQ=20,
                  glmerControl(optimizer = "bobyqa"))

sum_ran<-summary(model_ran)









if(isSingular(model_ran)==TRUE){

  x<-NULL

  repeat {

    print(paste("error␣on␣iteration",a,sep="␣"))



    baseline<-rpois(sample_size,6.34)

    prob_in_all<-exp(summary(model_multinomial)$coefficients[1,1]+sum
      (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(mod
       +exp(summary(model_multinomial)$coefficients[2,1]+summary(mode
       +exp(summary(model_multinomial)$coefficients[3,1]+summary(mode
       +exp(summary(model_multinomial)$coefficients[4,1]+summary(mode
       +exp(summary(model_multinomial)$coefficients[5,1]+summary(mode

    prob_in_esd<-exp(summary(model_multinomial)$coefficients[2,1]+sum
      (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(mod
```

```
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(mode

prob_in_mw<-exp(summary(model_multinomial)$coefficients[3,1]+summ
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(mod
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(mode

prob_in_ohe<-exp(summary(model_multinomial)$coefficients[4,1]+sum
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(mod
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(mode

prob_in_ohy<-exp(summary(model_multinomial)$coefficients[5,1]+sum
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(mod
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(mode

prob_in_con<-1/
  (1+exp(summary(model_multinomial)$coefficients[1,1]+summary(mod
   +exp(summary(model_multinomial)$coefficients[2,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[3,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[4,1]+summary(mode
   +exp(summary(model_multinomial)$coefficients[5,1]+summary(mode

treat_sim<-NULL

for(i in 1:length(prob_in_all)){

  t_sim<-rmultinom(1,1,c(prob_in_all[i],prob_in_esd[i],prob_in_mw

  if(t_sim[1,1]==1){
```

```
    treat_sim_new<-"ALL"
  } else if(t_sim[2,1]==1){
    treat_sim_new<-"ESD"
  } else if(t_sim[3,1]==1){
    treat_sim_new<-"MW"
  } else if(t_sim[4,1]==1){
    treat_sim_new<-"OHE"
  } else if(t_sim[5,1]==1){
    treat_sim_new<-"OHY"
  } else{
    treat_sim_new<-"A_Control"
  }
  treat_sim<-rbind(treat_sim,treat_sim_new)
}
sim_data_base_treat<-data.frame(cbind(baseline,treat_sim))
names(sim_data_base_treat)<-c("baseline","treat")
sim_data_base_treat$baseline<-as.numeric(sim_data_base_treat$base
#################################################################
# Simulate Outcome data
for(j in 1:length(prob_in_all)){
```

```
  if(sim_data_base_treat$treat[j]=="ALL"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.2309-0.5580+0.0
  }else if(sim_data_base_treat$treat[j]=="ESD"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.2309-0.1953+0.0
  }else if(sim_data_base_treat$treat[j]=="MW"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.2309-0.4132+0.0
  }else if(sim_data_base_treat$treat[j]=="OHE"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.2309-0.3506+0.0
  }else if(sim_data_base_treat$treat[j]=="OHY"){
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.2309-0.3082+0.0
  }else{
    sim_data_base_treat$outcome[j]<-rpois(1,exp(1.2309+0.0789*log
}
#####################################################################
# Create baseline value categories
cate<-NULL
for (i in 1:length(baseline)){
  if(sim_data_base_treat$baseline[i]<=quantile(sim_data_base_trea
    category<-"A"}else if (sim_data_base_treat$baseline[i]<=quant
      category<-"B"}else if (sim_data_base_treat$baseline[i]<=qua
```

```
            category<-"C"}else if (sim_data_base_treat$baseline[i]<=q

              category<-"D"}else {

                category<-"E"
              }

      cate<-rbind(cate,category)

    }

    final_data<-cbind(sim_data_base_treat,cate)


    model_ran<- glmer(outcome~treat+(1|cate),data=final_data,
                      family=poisson,nAGQ=20,
                      glmerControl(optimizer = "bobyqa"))

    sum_ran<-summary(model_ran)




  if(isSingular(model_ran)==FALSE) {
    print(paste("fixed␣iteration",a,sep="␣"))
    break

  }

  }
}

#########################################################################
# Poisson model for simulated dataset, ignoring baseline heterogeneit

model_no<-glm(outcome~treat,data=final_data,family = "poisson")

sum_no<-summary(model_no)
```

```
    # Apply the four parametric methods to the simulated dataset

    model_cont<-glm(outcome~treat+log(baseline+1),data=final_data,family

    sum_cont<-summary(model_cont)

    model_off<-glm(outcome~treat+offset(log(baseline+1)),data=final_data,

    sum_off<-summary(model_off)

    model_cat<-glm(outcome~treat+cate,data=final_data,family = "poisson")

    sum_cat<-summary(model_cat)

    ##########################################################################
    #verbeke method

    suppressMessages(attach(sim_data_base_treat))

    longdata<-data.frame(out=c(baseline,outcome),
                         parti=factor(rep(paste('p', 1:sample_size, sep='
                         time=as.factor(rep(0:1, each=sample_size)),
                         treat_group=rep(treat, 2))

    suppressMessages(library(glmmADMB))

    model_ver <- glmmadmb(out ~ treat_group*time+
                            (1+time|parti),
                          family = "poisson",
                          data = longdata)

    sum_ver<-summary(model_ver)


    if (ver_coef_all==sum_ver$coefficients[8,1]) stop(paste("error␣on␣ite
}, error=function(e){cat("ERROR␣:",conditionMessage(e), "\n")})



##############################################################################
```

```
#mantel haneszel RR for ALL

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$ALL

experiment_cat<-split(experiment,experiment$cate)

exp_a<-experiment_cat$A

exp_b<-experiment_cat$B

exp_c<-experiment_cat$C

exp_d<-experiment_cat$D

exp_e<-experiment_cat$E

control<-split_treat$A_Control

control_cat<-split(control,control$cate)

con_a<-control_cat$A

con_b<-control_cat$B

con_c<-control_cat$C

con_d<-control_cat$D

con_e<-control_cat$E

numerator_a<-sum(exp_a$outcome)/length(exp_a$outcome)

numerator_b<-sum(exp_b$outcome)/length(exp_b$outcome)

numerator_c<-sum(exp_c$outcome)/length(exp_c$outcome)

numerator_d<-sum(exp_d$outcome)/length(exp_d$outcome)
```

```
numerator_e<-sum(exp_e$outcome)/length(exp_e$outcome)

denominator_a<-sum(con_a$outcome)/length(con_a$outcome)

denominator_b<-sum(con_b$outcome)/length(con_b$outcome)

denominator_c<-sum(con_c$outcome)/length(con_c$outcome)

denominator_d<-sum(con_d$outcome)/length(con_d$outcome)

denominator_e<-sum(con_e$outcome)/length(con_e$outcome)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e

weight_a<-(sum(con_a$outcome)*length(exp_a$outcome))/(length(con_a$outcom
weight_b<-(sum(con_b$outcome)*length(exp_b$outcome))/(length(con_b$outcom
weight_c<-(sum(con_c$outcome)*length(exp_c$outcome))/(length(con_c$outcom
weight_d<-(sum(con_d$outcome)*length(exp_d$outcome))/(length(con_d$outcom
weight_e<-(sum(con_e$outcome)*length(exp_e$outcome))/(length(con_e$outcom


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e

mant_hans_ALL<-final_numerator/final_denominator

#########################################################################
#mantel haneszel RR for ESD

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$ESD

experiment_cat<-split(experiment,experiment$cate)

exp_a<-experiment_cat$A
```

```
exp_b<-experiment_cat$B

exp_c<-experiment_cat$C

exp_d<-experiment_cat$D

exp_e<-experiment_cat$E

control<-split_treat$A_Control

control_cat<-split(control,control$cate)

con_a<-control_cat$A

con_b<-control_cat$B

con_c<-control_cat$C

con_d<-control_cat$D

con_e<-control_cat$E

numerator_a<-sum(exp_a$outcome)/length(exp_a$outcome)

numerator_b<-sum(exp_b$outcome)/length(exp_b$outcome)

numerator_c<-sum(exp_c$outcome)/length(exp_c$outcome)

numerator_d<-sum(exp_d$outcome)/length(exp_d$outcome)

numerator_e<-sum(exp_e$outcome)/length(exp_e$outcome)

denominator_a<-sum(con_a$outcome)/length(con_a$outcome)

denominator_b<-sum(con_b$outcome)/length(con_b$outcome)

denominator_c<-sum(con_c$outcome)/length(con_c$outcome)

denominator_d<-sum(con_d$outcome)/length(con_d$outcome)
```

```
denominator_e<-sum(con_e$outcome)/length(con_e$outcome)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e

weight_a<-(sum(con_a$outcome)*length(exp_a$outcome))/(length(con_a$outcom
weight_b<-(sum(con_b$outcome)*length(exp_b$outcome))/(length(con_b$outcom
weight_c<-(sum(con_c$outcome)*length(exp_c$outcome))/(length(con_c$outcom
weight_d<-(sum(con_d$outcome)*length(exp_d$outcome))/(length(con_d$outcom
weight_e<-(sum(con_e$outcome)*length(exp_e$outcome))/(length(con_e$outcom


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e

mant_hans_ESD<-final_numerator/final_denominator

#########################################################################
#mantel haneszel RR for MW

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$MW

experiment_cat<-split(experiment,experiment$cate)

exp_a<-experiment_cat$A

exp_b<-experiment_cat$B

exp_c<-experiment_cat$C

exp_d<-experiment_cat$D

exp_e<-experiment_cat$E
```

```
control<-split_treat$A_Control

control_cat<-split(control,control$cate)

con_a<-control_cat$A

con_b<-control_cat$B

con_c<-control_cat$C

con_d<-control_cat$D

con_e<-control_cat$E

numerator_a<-sum(exp_a$outcome)/length(exp_a$outcome)

numerator_b<-sum(exp_b$outcome)/length(exp_b$outcome)

numerator_c<-sum(exp_c$outcome)/length(exp_c$outcome)

numerator_d<-sum(exp_d$outcome)/length(exp_d$outcome)

numerator_e<-sum(exp_e$outcome)/length(exp_e$outcome)

denominator_a<-sum(con_a$outcome)/length(con_a$outcome)

denominator_b<-sum(con_b$outcome)/length(con_b$outcome)

denominator_c<-sum(con_c$outcome)/length(con_c$outcome)

denominator_d<-sum(con_d$outcome)/length(con_d$outcome)

denominator_e<-sum(con_e$outcome)/length(con_e$outcome)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e
```

```
weight_a<-(sum(con_a$outcome)*length(exp_a$outcome))/(length(con_a$outcom
weight_b<-(sum(con_b$outcome)*length(exp_b$outcome))/(length(con_b$outcom
weight_c<-(sum(con_c$outcome)*length(exp_c$outcome))/(length(con_c$outcom
weight_d<-(sum(con_d$outcome)*length(exp_d$outcome))/(length(con_d$outcom
weight_e<-(sum(con_e$outcome)*length(exp_e$outcome))/(length(con_e$outcom


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e

mant_hans_MW<-final_numerator/final_denominator

#########################################################################
#mantel haneszel RR for OHE

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$OHE

experiment_cat<-split(experiment,experiment$cate)

exp_a<-experiment_cat$A

exp_b<-experiment_cat$B

exp_c<-experiment_cat$C

exp_d<-experiment_cat$D

exp_e<-experiment_cat$E

control<-split_treat$A_Control

control_cat<-split(control,control$cate)

con_a<-control_cat$A

con_b<-control_cat$B

con_c<-control_cat$C
```

```
con_d<-control_cat$D

con_e<-control_cat$E

numerator_a<-sum(exp_a$outcome)/length(exp_a$outcome)

numerator_b<-sum(exp_b$outcome)/length(exp_b$outcome)

numerator_c<-sum(exp_c$outcome)/length(exp_c$outcome)

numerator_d<-sum(exp_d$outcome)/length(exp_d$outcome)

numerator_e<-sum(exp_e$outcome)/length(exp_e$outcome)

denominator_a<-sum(con_a$outcome)/length(con_a$outcome)

denominator_b<-sum(con_b$outcome)/length(con_b$outcome)

denominator_c<-sum(con_c$outcome)/length(con_c$outcome)

denominator_d<-sum(con_d$outcome)/length(con_d$outcome)

denominator_e<-sum(con_e$outcome)/length(con_e$outcome)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e

weight_a<-(sum(con_a$outcome)*length(exp_a$outcome))/(length(con_a$outcom
weight_b<-(sum(con_b$outcome)*length(exp_b$outcome))/(length(con_b$outcom
weight_c<-(sum(con_c$outcome)*length(exp_c$outcome))/(length(con_c$outcom
weight_d<-(sum(con_d$outcome)*length(exp_d$outcome))/(length(con_d$outcom
weight_e<-(sum(con_e$outcome)*length(exp_e$outcome))/(length(con_e$outcom


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e
```

```
mant_hans_OHE<-final_numerator/final_denominator

#########################################################################
#mantel haneszel RR for OHY

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$OHY

experiment_cat<-split(experiment,experiment$cate)

exp_a<-experiment_cat$A

exp_b<-experiment_cat$B

exp_c<-experiment_cat$C

exp_d<-experiment_cat$D

exp_e<-experiment_cat$E

control<-split_treat$A_Control

control_cat<-split(control,control$cate)

con_a<-control_cat$A

con_b<-control_cat$B

con_c<-control_cat$C

con_d<-control_cat$D

con_e<-control_cat$E

numerator_a<-sum(exp_a$outcome)/length(exp_a$outcome)

numerator_b<-sum(exp_b$outcome)/length(exp_b$outcome)
```

```
numerator_c<-sum(exp_c$outcome)/length(exp_c$outcome)

numerator_d<-sum(exp_d$outcome)/length(exp_d$outcome)

numerator_e<-sum(exp_e$outcome)/length(exp_e$outcome)

denominator_a<-sum(con_a$outcome)/length(con_a$outcome)

denominator_b<-sum(con_b$outcome)/length(con_b$outcome)

denominator_c<-sum(con_c$outcome)/length(con_c$outcome)

denominator_d<-sum(con_d$outcome)/length(con_d$outcome)

denominator_e<-sum(con_e$outcome)/length(con_e$outcome)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e

weight_a<-(sum(con_a$outcome)*length(exp_a$outcome))/(length(con_a$outcom
weight_b<-(sum(con_b$outcome)*length(exp_b$outcome))/(length(con_b$outcom
weight_c<-(sum(con_c$outcome)*length(exp_c$outcome))/(length(con_c$outcom
weight_d<-(sum(con_d$outcome)*length(exp_d$outcome))/(length(con_d$outcom
weight_e<-(sum(con_e$outcome)*length(exp_e$outcome))/(length(con_e$outcom


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e

mant_hans_OHY<-final_numerator/final_denominator

#########################################################################

# results from 5 methods

#sample_size_experiment<-length(experiment$baseline)
```

```
#sample_size_control<-length(control$baseline)

no_adj_coef_all<-sum_no$coefficients[2,1]

no_adj_coef_esd<-sum_no$coefficients[3,1]

no_adj_coef_mw<-sum_no$coefficients[4,1]

no_adj_coef_ohe<-sum_no$coefficients[5,1]

no_adj_coef_ohy<-sum_no$coefficients[6,1]

no_adj_se_all<-sum_no$coefficients[2,2]

no_adj_se_esd<-sum_no$coefficients[3,2]

no_adj_se_mw<-sum_no$coefficients[4,2]

no_adj_se_ohe<-sum_no$coefficients[5,2]

no_adj_se_ohy<-sum_no$coefficients[6,2]

off_coef_all<-sum_off$coefficients[2,1]

off_coef_esd<-sum_off$coefficients[3,1]

off_coef_mw<-sum_off$coefficients[4,1]

off_coef_ohe<-sum_off$coefficients[5,1]

off_coef_ohy<-sum_off$coefficients[6,1]

off_se_all<-sum_off$coefficients[2,2]

off_se_esd<-sum_off$coefficients[3,2]

off_se_mw<-sum_off$coefficients[4,2]

off_se_ohe<-sum_off$coefficients[5,2]
```

```
off_se_ohy<-sum_off$coefficients[6,2]

cont_coef_all<-sum_cont$coefficients[2,1]

cont_coef_esd<-sum_cont$coefficients[3,1]

cont_coef_mw<-sum_cont$coefficients[4,1]

cont_coef_ohe<-sum_cont$coefficients[5,1]

cont_coef_ohy<-sum_cont$coefficients[6,1]

cont_se_all<-sum_cont$coefficients[2,2]

cont_se_esd<-sum_cont$coefficients[3,2]

cont_se_mw<-sum_cont$coefficients[4,2]

cont_se_ohe<-sum_cont$coefficients[5,2]

cont_se_ohy<-sum_cont$coefficients[6,2]

cat_coef_all<-sum_cat$coefficients[2,1]

cat_coef_esd<-sum_cat$coefficients[3,1]

cat_coef_mw<-sum_cat$coefficients[4,1]

cat_coef_ohe<-sum_cat$coefficients[5,1]

cat_coef_ohy<-sum_cat$coefficients[6,1]

cat_se_all<-sum_cat$coefficients[2,2]

cat_se_esd<-sum_cat$coefficients[3,2]

cat_se_mw<-sum_cat$coefficients[4,2]

cat_se_ohe<-sum_cat$coefficients[5,2]
```

```
cat_se_ohy<-sum_cat$coefficients[6,2]

ran_coef_all<-sum_ran$coefficients[2,1]

ran_coef_esd<-sum_ran$coefficients[3,1]

ran_coef_mw<-sum_ran$coefficients[4,1]

ran_coef_ohe<-sum_ran$coefficients[5,1]

ran_coef_ohy<-sum_ran$coefficients[6,1]

ran_se_all<-sum_ran$coefficients[2,2]

ran_se_esd<-sum_ran$coefficients[3,2]

ran_se_mw<-sum_ran$coefficients[4,2]

ran_se_ohe<-sum_ran$coefficients[5,2]

ran_se_ohy<-sum_ran$coefficients[6,2]


ver_coef_all<-sum_ver$coefficients[8,1]

ver_coef_esd<-sum_ver$coefficients[9,1]

ver_coef_mw<-sum_ver$coefficients[10,1]

ver_coef_ohe<-sum_ver$coefficients[11,1]

ver_coef_ohy<-sum_ver$coefficients[12,1]

ver_se_all<-sum_ver$coefficients[8,2]

ver_se_esd<-sum_ver$coefficients[9,2]

ver_se_mw<-sum_ver$coefficients[10,2]
```

```
ver_se_ohe<-sum_ver$coefficients[11,2]

ver_se_ohy<-sum_ver$coefficients[12,2]


iteration<-cbind(#sample_size_experiment,sample_size_control,
  no_adj_coef_all,off_coef_all,cont_coef_all,cat_coef_all,ran_coef_all,ma
  no_adj_coef_esd,off_coef_esd,cont_coef_esd,cat_coef_esd,ran_coef_esd,ma
  no_adj_coef_mw,off_coef_mw,cont_coef_mw,cat_coef_mw,ran_coef_mw,mant_ha
  no_adj_coef_ohe,off_coef_ohe,cont_coef_ohe,cat_coef_ohe,ran_coef_ohe,ma
  no_adj_coef_ohy,off_coef_ohy,cont_coef_ohy,cat_coef_ohy,ran_coef_ohy,ma
  no_adj_se_all,off_se_all,cont_se_all,cat_se_all,ran_se_all,ver_se_all,
  no_adj_se_esd,off_se_esd,cont_se_esd,cat_se_esd,ran_se_esd,ver_se_esd,
  no_adj_se_mw,off_se_mw,cont_se_mw,cat_se_mw,ran_se_mw,ver_se_mw,
  no_adj_se_ohe,off_se_ohe,cont_se_ohe,cat_se_ohe,ran_se_ohe,ver_se_ohe,
  no_adj_se_ohy,off_se_ohy,cont_se_ohy,cat_se_ohy,ran_se_ohy,ver_se_ohy)

iteration_all<-cbind(#sample_size_experiment,sample_size_control,
  no_adj_coef_all,off_coef_all,cont_coef_all,cat_coef_all,ran_coef_all,ma
  no_adj_se_all,off_se_all,cont_se_all,cat_se_all,ran_se_all,ver_se_all)

iteration_esd<-cbind(#sample_size_experiment,sample_size_control,
  no_adj_coef_esd,off_coef_esd,cont_coef_esd,cat_coef_esd,ran_coef_esd,ma
  no_adj_se_esd,off_se_esd,cont_se_esd,cat_se_esd,ran_se_esd,ver_se_esd)

iteration_mw<-cbind(#sample_size_experiment,sample_size_control,
  no_adj_coef_mw,off_coef_mw,cont_coef_mw,cat_coef_mw,ran_coef_mw,mant_ha
  no_adj_se_mw,off_se_mw,cont_se_mw,cat_se_mw,ran_se_mw,ver_se_mw)

iteration_ohe<-cbind(#sample_size_experiment,sample_size_control,
  no_adj_coef_ohe,off_coef_ohe,cont_coef_ohe,cat_coef_ohe,ran_coef_ohe,ma
  no_adj_se_ohe,off_se_ohe,cont_se_ohe,cat_se_ohe,ran_se_ohe,ver_se_ohe)

iteration_ohy<-cbind(#sample_size_experiment,sample_size_control,
  no_adj_coef_ohy,off_coef_ohy,cont_coef_ohy,cat_coef_ohy,ran_coef_ohy,ma
  no_adj_se_ohy,off_se_ohy,cont_se_ohy,cat_se_ohy,ran_se_ohy,ver_se_ohy)

overview_all<-rbind(overview_all,iteration_all)

overview_esd<-rbind(overview_esd,iteration_esd)
```

```
overview_mw<-rbind(overview_mw,iteration_mw)

overview_ohe<-rbind(overview_ohe,iteration_ohe)

overview_ohy<-rbind(overview_ohy,iteration_ohy)


overview_all<-as.data.frame(overview_all)

overview_esd<-as.data.frame(overview_esd)

overview_mw<-data.frame(overview_mw)

overview_ohe<-data.frame(overview_ohe)

overview_ohy<-data.frame(overview_ohy)

setwd("C:/PhD/belcap␣sim")

write.csv(overview_all,file = "750␣overview␣all␣belcap.csv",row.names =FA

write.csv(overview_esd,file = "750␣overview␣esd␣belcap.csv",row.names =FA

write.csv(overview_mw,file = "750␣overview␣mw␣belcap.csv",row.names =FALS

write.csv(overview_ohe,file = "750␣overview␣ohe␣belcap.csv",row.names =FA

write.csv(overview_ohy,file = "750␣overview␣ohy␣belcap.csv",row.names =FA

print(a)


}
```

## 8.3 Appendix 3: Third time point

```
# Load original polys data

setwd("C:/PhD")

polyps_edited<-read.csv('polyps edited.csv')

attach(polyps_edited)

# Poisson model (null model) to baseline data to get a mean Poisson value

library(plyr)

model_base<-glm(count_b~1,family = "poisson")

summary(model_base) # mean poisson value = exp(3.05164) = 21.15

# Logistic model to demonstrate correlation between baseline value and tr
# This is used for treatment allocation in the simulation

model_logistic<-glm(treatment~count_b,family = "binomial")

summary(model_logistic)

##########################################################################

# start loop to control number of simulations

overview<-NULL

number_iterations<-100

sample_size_each_iteration <-100

RR_1<-c(0.5)

RR_2<-c(0.25)
```

```
for(b in 1:length(RR_1)){

for(a in 1:number_iterations){

  tryCatch({ #stops loop ending if error occurs. the error that can occur
#########################################################################

# Simulate baseline data

baseline<-rpois(sample_size_each_iteration,21.15)

prob_in_treat<-exp(summary(model_logistic)$coefficients[1,1]+summary(mode
                (1+exp(summary(model_logistic)$coefficients[1,1]+summary(

treat_sim<-NULL

for(i in 1:length(prob_in_treat)){

  t_sim<-rbinom(1,1,prob_in_treat[i])

  treat_sim<-rbind(treat_sim,t_sim)

}

sim_data_base_treat<-data.frame(cbind(baseline,treat_sim))

#########################################################################

# Simulate 2nd time point data

second_data<-NULL

for(k in 1:length(prob_in_treat)){

  if(sim_data_base_treat$V2[k]==0){

    second<-rpois(1,exp(log(baseline[k])))

  }else{
```

```
    second<-rpois(1,exp(log(RR_1[b])+log(baseline[k])))
  }
  second_data<-rbind(second_data,second)
}
sim_data_base_treat_mid<-cbind(sim_data_base_treat,second_data)
names(sim_data_base_treat_mid)<-c("baseline","treat","mid_adj")
##########################################################################
# Simulate Outcome data
outcome_data<-NULL
for(k in 1:length(prob_in_treat)){
  if(sim_data_base_treat_mid$treat[k]==0){
  outcome<-rpois(1,exp(log(sim_data_base_treat_mid$baseline[k])))
  }else{
  outcome<-rpois(1,exp(log(RR_2[b])+log(sim_data_base_treat_mid$baseline[
  }
  outcome_data<-rbind(outcome_data,outcome)
}
sim_data_base_treat_out<-cbind(sim_data_base_treat_mid,outcome_data)
names(sim_data_base_treat_out)<-c("baseline","treat","mid_adj","outcome_a
##########################################################################
```

```
# Create baseline value categories

cate<-NULL

for (i in 1:length(baseline)){

  if(sim_data_base_treat_out$baseline[i]<=quantile(baseline,probs = 0.2))

    category<-"A"}else if (sim_data_base_treat_out$baseline[i]<=quantile(

      category<-"B"}else if (sim_data_base_treat_out$baseline[i]<=quantil

        category<-"C"}else if (sim_data_base_treat_out$baseline[i]<=quant

          category<-"D"} else{

            category<-"E"
          }

  cate<-rbind(cate,category)

}

final_data<-cbind(sim_data_base_treat_out,cate)

#########################################################################
# Poisson model for simulated dataset, ignoring baseline heterogeneity

model_no<-glm(outcome_adj~treat,data=final_data,family = "poisson")

sum_no<-summary(model_no)

# Apply the four parametric methods to the simulated dataset

model_cont<-glm(outcome_adj~treat+log(baseline),data=final_data,family =

sum_cont<-summary(model_cont)
```

```
model_off<-glm(outcome_adj~treat+offset(log(baseline)),data=final_data,fa

sum_off<-summary(model_off)

model_cat<-glm(outcome_adj~treat+cate,data=final_data,family = "poisson")

sum_cat<-summary(model_cat)

suppressMessages(library(lme4))

model_ran<- glmer(outcome_adj~treat+(1|cate),data=final_data,family=poiss

sum_ran<-summary(model_ran)

#########################################################################
#verbeke method

suppressMessages(attach(sim_data_base_treat_out))

longdata<-data.frame(outcome=c(baseline,mid_adj,outcome_adj),
                     parti=factor(rep(paste('p', 1:sample_size_each_itera
                     time=as.factor(rep(0:2, each=sample_size_each_iterat
                     treat_group=rep(treat, 3))

suppressMessages(library(glmmADMB))

model_ver <- glmmadmb(outcome ~ treat_group*time+(1+time|parti), family =

sum_ver<-summary(model_ver)


if (ver_coef==sum_ver$coefficients[4,1]) stop(paste("error␣on␣iteration",
  }, error=function(e){cat("ERROR␣:",conditionMessage(e), "\n")})

#################################################################
# Mantel-Haneszel

split_treat<-split(final_data,final_data$treat)

experiment<-split_treat$`1`
```

```
experiment_cat<-split(experiment,experiment$cate)
exp_a<-experiment_cat$A
exp_b<-experiment_cat$B
exp_c<-experiment_cat$C
exp_d<-experiment_cat$D
exp_e<-experiment_cat$E
control<-split_treat$`0`
control_cat<-split(control,control$cate)
con_a<-control_cat$A
con_b<-control_cat$B
con_c<-control_cat$C
con_d<-control_cat$D
con_e<-control_cat$E
numerator_a<-sum(exp_a$outcome_adj)/length(exp_a$outcome_adj)
numerator_b<-sum(exp_b$outcome_adj)/length(exp_b$outcome_adj)
numerator_c<-sum(exp_c$outcome_adj)/length(exp_c$outcome_adj)
numerator_d<-sum(exp_d$outcome_adj)/length(exp_d$outcome_adj)
numerator_e<-sum(exp_e$outcome_adj)/length(exp_e$outcome_adj)
denominator_a<-sum(con_a$outcome_adj)/length(con_a$outcome_adj)
denominator_b<-sum(con_b$outcome_adj)/length(con_b$outcome_adj)
```

```
denominator_c<-sum(con_c$outcome_adj)/length(con_c$outcome_adj)

denominator_d<-sum(con_d$outcome_adj)/length(con_d$outcome_adj)

denominator_e<-sum(con_e$outcome_adj)/length(con_e$outcome_adj)

risk_a<-numerator_a/denominator_a
risk_b<-numerator_b/denominator_b
risk_c<-numerator_c/denominator_c
risk_d<-numerator_d/denominator_d
risk_e<-numerator_e/denominator_e

weight_a<-(sum(con_a$outcome_adj)*length(exp_a$outcome_adj))/(length(con_
weight_b<-(sum(con_b$outcome_adj)*length(exp_b$outcome_adj))/(length(con_
weight_c<-(sum(con_c$outcome_adj)*length(exp_c$outcome_adj))/(length(con_
weight_d<-(sum(con_d$outcome_adj)*length(exp_d$outcome_adj))/(length(con_
weight_e<-(sum(con_e$outcome_adj)*length(exp_e$outcome_adj))/(length(con_


final_numerator<-(weight_a*risk_a)+(weight_b*risk_b)+(weight_c*risk_c)+(w
final_denominator<-weight_a+weight_b+weight_c+weight_d+weight_e

mant_hans<-final_numerator/final_denominator

#########################################################################

# results from 5 methods

sample_size_experiment<-length(experiment$baseline)

sample_size_control<-length(control$baseline)

no_adj_coef<-sum_no$coefficients[2,1]

no_adj_se<-sum_no$coefficients[2,2]

off_coef<-sum_off$coefficients[2,1]

off_se<-sum_off$coefficients[2,2]
```

```
cont_coef<-sum_cont$coefficients[2,1]

cont_se<-sum_cont$coefficients[2,2]

cat_coef<-sum_cat$coefficients[2,1]

cat_se<-sum_cat$coefficients[2,2]

ran_coef<-sum_ran$coefficients[2,1]

ran_se<-sum_ran$coefficients[2,2]

ver_coef<-sum_ver$coefficients[6,1]

ver_se<-sum_ver$coefficients[6,2]

test_coef<-sum_ver$coefficients[2,1]+sum_ver$coefficients[6,1]

alt_coef<-sum_ver$coefficients[2,1]

full_coef<-sum_ver$coefficients[2,1]+sum_ver$coefficients[3,1]+sum_ver$co

iteration<-cbind(RR_1[b],RR_2[b],sample_size_experiment,sample_size_contr
             cont_coef,cat_coef,ran_coef,ver_coef,test_coef,alt_coef,full
             log(mant_hans),no_adj_se,off_se,cont_se,cat_se,ran_se,ver_se

overview<-rbind(overview,iteration)

print(a)

}

print(paste("completed␣RR␣=",RR_1[b],sep="␣"))

}

overview<-as.data.frame(overview)
```

# References


[1] Agresti, A. (2002). Categorical Data Analysis. 2nd ed. Wiley, p.251.

[2] Agresti, A. (2019). An Introduction to Categorical Data Analysis. 3rd ed. Wiley, p.65.

[3] Altman, D. and Bland, J. (1999). Statistics notes: How to randomise. BMJ, 319(7211), pp.703-704.

[4] Ashburn, A., Fazakarley, L., Ballinger, C., Pickering, R., McLellan, L.D. and Fitton, C. (2007). A randomised controlled trial of a home based exercise programme to reduce the risk of falling among people with Parkinson's disease. J. Neurol. Neurosurg. Psychiatry 78, 678–684.

[5] Blanca, M., Arlacon, R., Arnau, J., Bono, R. and Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option?. Psicothema, 29(4), pp.552-557.

[6] Böhning, D., Dietz, E., Schlattmann, P., Mendona, L. and Kirchner, U. (1999). The zero-inflated Poisson model and the DMF-Index in dental epidemiology. Journal of the Royal Statistical Society, Ser. A 160, 195-209.

[7] Chao, W.H., Palta, M. and Young, T. (1997). Effect of omitted confounders on the analysis of correlated binary data. Biometrics 53, 678–689.

[8] Collett, D. (2015). Modelling survival data in medical research. Boca Raton, FL: CRC Press,pp.67.

[9] Cummings, P. (2019). Analysis of Incidence Rates. Milton: Chapman and Hall/CRC, pp.175

[10] Dobson, A. and Barnett, A. (2002). An introduction to generalized linear models. 2nd ed. Chapman & Hall / CRC, pp.95-105.

[11] Dobson, A. and Barnett, A. (2002). An introduction to generalized linear models. 2nd ed. Chapman & Hall / CRC, pp.152.

[12] Ema.europa.eu. 2021. [online] Available at: ¡https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials.pdf¿ [Accessed 7 September 2021].

[13] Forbes, C. (2011). Statistical Distributions. 4th ed. New Jersey: Wiley, p.93.

[14] Fox, J. (2016). Applied regression analysis and generalized linear models. Los Angeles: SAGE, pp.418-419.

[15] Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized exper- iments with nonlinear regressions and omitted covariates. Biometrika 71, 431–444.

[16] Goodwin, V.A., Richards, S.H., Henley, W., Ewings, P., Taylor, A.H. and Campbell, J.L. (2011). An exercise intervention to prevent falls in people with Parkinson's disease: a pragmatic randomised controlled trial. J. Neurol. Neurosurg. Psychiatry 82, 1232–1238.

[17] Gray, P. and Hildebrand, K. (2000). Fall risk factors in Parkinson's disease. J. Neurosci. Nurs. 32, 222–228.

[18] Hernández, A., Steyerberg, E. and Habbema, J. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. Journal of Clinical Epidemiology, 57(5), pp.454-460.

[19] Hosmer Jr., D., Lemeshow, S. and Sturdivant, R. (2013). Applied Logistic Regression. 3rd ed. Wiley, pp.12-15.

[20] Jørstad, E.C., Hauer, K., Becker, C. and Lamb, S.E. (2005). Measuring the psychological outcomes of falling: A systematic review. J. Am. Geriatr. Soc. 53, 501–510.

[21] Kent, D., Rothwell, P., Ioannidis, J., Altman, D. and Hayward, R. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials, vol. 11, no. 1.

[22] Kahan, B., Jairath, V., Doré, C. and Morris, T. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials, 15(1).

[23] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retropsective studies. J. Natl. Cancer Inst. 22 719–748.

[24] Martin, T., Weatherall, M., Anderson, T.J. and MacAskill, M.R. (2015). A Randomized Controlled Feasibility Trial of a Specific Cueing Program for Falls Management in Persons With Parkinson Disease and Freezing of Gait. J. Neurol. Phys. Ther. 39, 179–184.

[25] Matthews, J. (2000). An Introduction to Randomized Controlled Clinical Trials. London: Arnold, p.1.

[26] Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), p.370.

[27] Neuhaus, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. Journal of the American Statistical Association 93, 1124–1129.

[28] Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between and Within Cluster Covariate Effects in the Analysis of Clustered Data, Biometrics. 54, 638-645.

[29] nhs.uk. (2019). Bowel polyps. [online] Available at https://www.nhs.uk/conditions/bowel-polyps [Accessed 23 Jul. 2019].

[30] nhs.uk. (2020). Parkinson's Disease. [online] Available at: ¡https://www.nhs.uk/conditions/parkinsons-disease/¿ [Accessed 27 December 2020].

[31] Nyström, H., Nordström, A. and Nordström, P. (2016). Risk of Injurious Fall and Hip Fracture up to 26 y before the Diagnosis of Parkinson Disease: Nested Case–Control Studies in a Nationwide Cohort. PLOS Med. 13.

[32] Palta, M. and Yao, T.J. (1991). Analysis of longitudinal data with unmeasured confounders. Biometrics 47, 1355–1369.

[33] Piantadosi, S. (1997). Clinical trials. New York: Wiley, p.10.

[34] Pickering, R.M., Grimbergen, Y.A.M., Rigney, U., Ashburn, A., Mazibrada, G., Wood, B., Gray, P., Kerr, G. and Bloem, B.R. (2007). A

meta-analysis of six prospective studies of falling in Parkinson's disease. Mov. Disord. 22, 1892–1900.

[35] Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models. Biometrics, 42(2), p.311.

[36] Greenland, S. and Robins, J. (1985). Estimation of a Common Effect Parameter from Sparse Follow-Up Data. Biometrics, 41(1), p.55.

[37] Sedgwick, P. (2014). Explanatory trials versus pragmatic trials. BMJ, 349(nov13 3), pp.g6694-g6694.

[38] Senn, S. (2021). Stephen Senn: Being Just about Adjustment (Guest Post). [online] Error Statistics Philosophy. Available at: https://errorstatistics.com/2020/03/16/stephen-senn-being-just-about-adjustment-guest-post [Accessed 7 September 2021].

[39] Shamley, D. and Wright, B. (n.d.). A comprehensive and practical guide to clinical trials. pp.12-14.

[40] Song, J. and Chung, K. (2010). Observational Studies: Cohort and Case-Control Studies. Plastic and Reconstructive Surgery, 126(6), pp.2234-2242.

[41] Suresh, K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. Journal of Human Reproductive Sciences, 4(1), p.8.

[42] Timothy J. Legg, C. (2019). Cohort study: Finding causes, examples, and limitations. [online] Medical News Today. Available at: https://www.medicalnewstoday.com/articles/281703.php [Accessed 27 Jun. 2019].

[43] Verbeke, G., Spiessens, B. and Lesaffre, E. (2001). Conditional Linear Mixed Models. The American Statistician, 55(1), pp.25-34.

[44] Verbeke, G., Fieuws, S., Lesaffre, E., Kato, B., Foreman, M., Broos, P. and Milisen, K. (2006). A comparison of procedures to correct for base-line differences in the analysis of continuous longitudinal data: a case-study. Journal of the Royal Statistical Society: Series C (Applied Statistics), 55(1), pp.93-101.

[45] Verbeke, G. and Fieuws, S. (2006). The effect of miss-specified baseline characteristics on inference for longitudinal trends in linear mixed models. Biostatistics, 8(4), pp.772-783.

[46] Wang, H.C., Lin, C.C., Lau, C.I., Chang, A., Sung, F.C. and Kao, C.H. (2014). Risk of accidental injuries amongst Parkinson disease patients. Eur. J. Neurol. 21, 907–913.

[47] Winer, B. J. (1970). Statistical principles in experimental design. England: McGraw-Hill.

[48] Weir, C. and Lees, K. (2003). Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. Statistics in Medicine, 22(5), pp.705-726.

[49] Woodward, M. (1999). Epidemiology. Study Design and Analysis. Chapman & Hall /CRC, BocaRaton.

[50] Yardley, L. and Smith, H. (2002). A prospective study of the relationship between feared consequences of falling and avoidance of activity in community-living older people. Gerontologist 42, 17–23.