University of Southampton

Faculty of Medicine

Primary Care and Population Sciences

The analysis of falls counts from falls prevention trials in people with Parkinson's

by

Han Zheng

ORCID ID 0000-0003-3134-6947

Thesis for the degree of Doctor of Philosophy

May 2019

University of Southampton

Abstract

Faculty of Medicine

Primary Care and Population Sciences

Thesis for the degree of Doctor of Philosophy

The analysis of falls counts from falls prevention trials in people with Parkinson's

by Han Zheng

Falls are a common recurrent event for People with Parkinson's (PwP) and may result in injuries and loss of independence in daily activities. Falls prevention trials evaluate whether an intervention is effective in reducing falls. The traditional analysis is the logistic regression, but Negative Binomial (NB) models have become widely used recently. The distribution of the falls count is usually heavily skewed, with a relatively small mean and a few outlying large numbers. These large counts are a challenge in modelling falls count because they may have great influence in model estimation, especially when there is imbalance between groups.

This thesis focuses on examining the statistical methods used in analysing falls counts, especially the NB model. Diagnostic plots specifically designed to assessing the influence of outliers on NB modelling are developed in this context, so that the outliers can be easily identified.

The falls counts during a pre-randomisation baseline period is usually strongly correlated with the falls counts during an outcome period. Approaches to incorporating the baseline count in modelling outcome falls counts are examined in three motivating datasets and simulations carried out generating data resembling the characteristics of real data with respect to the methods used to collect the falls count. Data from trials with prospectively collected outcome counts and retrospectively collected baseline counts are examined using an actual dataset and simulations to check whether this design impacts on model estimation. Overall, including the logged baseline count as a covariate in NB regression was shown to have satisfying power and to be robust when the underlying assumption does not hold.

Some alternative count response models to the standard NB model are also considered: Poisson Inverse Gaussian models for heavily skewed data; zero-inflated NB to check for potential zero-inflation in falls counts; right-censored/right-truncated NB to reduce the influence of large falls counts; finite mixture Poisson model to accommodate the frequent fallers as a subpopulation; and random-effects NB models to explore the possibility of modelling longitudinal falls counts. They all show potential in dealing with specific issues in analysing falls data.

Contents

Figures

Tables

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Ruth Pickering, who is the kindest and most caring person I know. She is not only knowledgeable but also has a great sense of humor. I will always remember the fun time of having coffee with her (usually in Costa). I also want to thank my co-supervisor Dr Alan Kimber for all the help during the last four years.

Many thanks go to Brian Yuen, Scott Harris, and Karlien Paas in my office for their encouragement while I was writing up the thesis.

Finally, I would like to offer my special thanks to my parents. They have been very supportive of me throughout my PhD. I always enjoyed our weekly video calls.

Nomenclature

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BOE plot	Baseline/Outcome Event plot
CI	Confidence Interval
CMF	Cumulative Mass Function
CNB	Conditional Negative Binomial
FMM	Finite Mixture Model
FOG	Freezing of Gait
FRR	Falls Rate Ratio
GLM	Generalized Linear Model
HP	Heterogeneity Parameter
IRLS	Iteratively Reweighted Least Squares
IRR	Incidence Rate Ratio
LoFE	Line of Falls Equality
MLE	Maximum Likelihood Estimate
NB	Negative Binomial
NB-basic	The basic NB model that includes only one covariate: the group allocation
NB-full	The full NB model that includes additional covariates that are commonly collected in a falls prevention trial
NB-null / Poi-null	The NB/Poisson model that does not include the baseline count
NB-unlogged / Poi-unlogged	The NB/Poisson model including the untransformed baseline count as a covariate
NB-logged / Poi-logged	The NB/Poisson model including the logged baseline count as a covariate
NB-offset / Poi-offset	The NB/Poisson model including the logged baseline count as offset

NB1 The NB model with a variance function of $\mathrm{Var}(y) = \mu + \delta\mu$

NB2 The NB model with a variance function of $\mathrm{Var}(y) = \mu + \alpha\mu^2$ or $\mathrm{Var}(y) = \mu + \mu^2/\theta$

NB-rc The right-censored NB model

NB-rt The right-truncated NB model

NB-H Heterogeneous Negative Binomial

OR Odds Ratio

PDF Probability Density Function

PMF Probability Mass Function

PIG Poisson Inverse-Gaussian

PIG-null The PIG model that does not include the baseline count

PIG-logged The PIG model including the logged baseline count as a covariate

PwP People with Parkinson's

Random-effects NB model

LT Linear Time model: the random-effects NB model that includes time as a continuous covariate; this model assumes a linear time effect.

LTI Linear Time Interaction model: the random-effects NB model that includes time as a continuous variable for the main effect as well as in the interaction between time and intervention; this model assumes that the intervention effect changes linearly over time.

FT Factorial Time model: the random-effects NB model that includes time as a factor; this model is fitted to examine the assumption of the linear time effect in the LT model.

FTI Factorial Time Interaction model: the random-effects NB model that includes time as a factor for the main effect as well as in the interaction between time and intervention; this model is fitted to examine the assumption in the LTI model that the intervention effect changes linearly overtime.

UPDRS Unified Parkinson's Disease Rating Scale

ZINB Zero-Inflated Negative Binomial

The thesis contains a review of software functionality, and to make sure the software/program names, software module names, and command/function names are distinguishable, different fonts are used:

- **Statistical package:** the “Courier” font is used to denote a statistical package (for example, the SAS software) or a programming language (for example, the R language).
- **Package module:** the bold font is used to denote a module in a statistical package. The term module refers to a command (such as **nbreg**) in Stata, a procedure in SAS (such as **GENMOD**), a command in SPSS (such as **GENLIN**), a package (such as **MASS**) in R, and a module in Python (such as **statsmodel**).
- **Command, function, and code:** the “Courier New” font is used to denote: function in an R package (for example, the `glm.nb()` function from the **MASS** package in R) ; options in any statistical package; and other code.

UNIVERSITY OF Southampton

Research Thesis: Declaration of Authorship

Print name:	Han Zheng

Title of thesis:	The analysis of falls counts from falls prevention trials in people with Parkinson’s

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as: [please list references below]:

Chapter 6 of the thesis was published in the Biometrical Journal (Zheng et al. 2018) and rewritten slightly to fit the format of the thesis: .

Zheng, H., Kimber, A., Goodwin, V.A., Pickering, R.M., 2018. A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials. .

Biometrical J. 60, 66–78. .

Signature:		Date:	

Chapter 1

Introduction

Falls are common events for people with Parkinson's (PwP) and often result in injuries (Gray and Hildebrand, 2000; Nyström et al., 2016; Pickering et al., 2007; Wang et al., 2014). Frequent fallers may develop psychological difficulties, such as loss of confidence in daily activities due to fear of falling, which greatly affects their quality of life (Jørstad et al., 2005; Yardley and Smith, 2002).

Randomised Control Trials (RCTs) have been conducted to evaluate the potential of treatments for preventing falls, and they are usually referred to as falls prevention trials. Typically, after randomisation, participants in an intervention group receive an intervention while participants in a control group receive usual care. The number of falls experienced by each person during a follow-up period is recorded as the outcome, and then compared between groups to evaluate whether the intervention reduces the occurrence of falls. The falls count may be collected prospectively via a falls dairy, or retrospectively by asking participants to recall the number of falls that occurred during a period in the past. In addition to the outcome falls count, it is common to also obtain the number of falls during a pre-randomisation baseline period, which is referred to as the baseline falls count in this thesis.

In general, the distribution of a falls count is positively skewed and heavy tailed, with a relatively small mean and a few large outliers. Various methods are available for analysing data from falls prevention trials. A common approach is to dichotomise falls counts into a binary variable and fit a logistic model, which yields an Odds Ratio (OR) to estimate the size of the intervention effect. However, valuable information is lost in the process, resulting in low statistical power. In contrast to the logistic model, a count response model yields an Incidence Rate Ratio (IRR) to estimate the risk of falling based on all falls. The IRR is called a Falls Rate Ratio (FRR) in the context of a falls prevention trial, and the 95% Confidence

Interval (CI) of the FRR is often reported. Poisson regression is the standard and most commonly used count response model. It has an underlying assumption of equidispersion, which means that the response variance is equal to the mean. However, this assumption rarely holds for falls data due to heterogeneity, which arises when important dependent variables are not included in models or not observed in trials. This was described by Winkelmann (2008) as “the explanatory variables do not account for the full amount of individual heterogeneity in the conditional mean of the dependent variable.” Failing to account for heterogeneity in a model results in overdispersion, defined as the variance being greater than the mean (Hilbe, 2011). This is a major challenge in the analysis of falls counts, as overdispersion leads to inflated type I error rates in model-based hypothesis tests.

In recent years, Negative binomial (NB) regression has grown popular and become the recommended statistical model for falls data (Gillespie et al., 2012). It can be regarded as an extension of Poisson regression—fundamentally, NB regression is a Poisson model with a gamma-distributed random subject effect in the model to accommodate overdispersion. Compared to Poisson regression, NB is more robust to heterogeneity: the model-based standard errors (SEs) of regression coefficients are not underestimated to the same extent, and the type I error rates of the model-based tests are closer to the nominal level.

1.1 Research objectives

Despite the growing popularity of NB models for falls data, there are a few challenges in practice. The aim of this thesis is to address the issues in analysing falls counts from falls prevention trials in PwP, especially for NB and NB-related models. In particular, the thesis seeks to address the following topics:

- **Utilising the baseline count in statistical analysis:** Incorporating the baseline falls count in an NB model is expected to improve the statistical power in the testing of intervention effect, because the falls count during a baseline period is usually strongly correlated to the outcome count. Some trials collect outcome counts prospectively but baseline counts retrospectively. In such trials there is a discrepancy in the collection methods between the two counts, which would be anticipated to affect the relationship between them. The thesis seeks to examine how best to incorporate the baseline count in NB modelling, and how a discrepancy in methods of collecting the counts impacts on modelling.

- **Large outcome falls counts:** One of the challenge in modelling falls counts is the presence of outliers. Most people record only a few falls during trial follow-up, but occasionally some record massive numbers. Although NB models are based on a long-tailed distribution, large counts may still be influential in model estimation, but it is not straightforward to identify whether a large count is indeed influential and how it impacts on the estimation of the intervention effect. The study aims to develop diagnostics plots in the context of a falls prevention trial where a baseline count has been collected, develop software to automate the production of the plots, and explore statistical approaches to reduce the influence of the large outcome counts.
- **NB functionality in statistical packages:** An aim of the thesis is to review statistical packages regarding their functionality in NB modelling to facilitate researchers in selecting a package that best meet their requirements.
- **Alternative count response modes:** In addition to NB models, alternative count response models are studied in the context of modelling falls count.

1.2 Motivating datasets from falls prevention trials

Three motivating datasets from falls prevention trials for PwP were made available to this project. Each of the three trials is comprised of a baseline period, and one, two, or more post-randomisation follow-up periods of falls collection.

1.2.1 Goodwin et al. dataset

The Goodwin et al. (2011) trial is an RCT carried out in the South West of England. One hundred and thirty PwP meeting the following eligibility criteria were recruited: with a diagnosis of Parkinson's, with a history of at least two falls in the year prior to enrolment (the number was obtained via a retrospective question at the screening interview but not recorded other than for checking eligibility), with mobilising ability, and resident in or registered with a general practitioner in Devon.

The recruited PwP were randomised to either an intervention (n=64) or a control group (n=66), but not informed of group allocation until they had finished a 10-week baseline prospective falls collection period. During 20 weeks of follow-up, each participant prospectively recorded the number of falls they experienced in diaries; the number recorded during the baseline period is referred to as the baseline falls count. Note that this number is different to the retrospective falls count, which was only used to check eligibility.

After the baseline period (weeks 1-10), the participants in the intervention group received one strength and balance group exercise session and two home exercises in each week. The intervention sessions lasted for 10 weeks (weeks 11-20), and the participants were followed up for another 10 weeks (weeks 21-30). Throughout the trial period, both the intervention and control groups received usual care that was delivered by a clinical team blind to group allocation. The number of falls recorded during the two outcome periods, weeks 11-20 and 21-30, are referred to as the intervention and outcome falls counts respectively.

The following baseline characteristics were made available to this project: sex, age, the number of years since diagnosis of Parkinson's, the Hoehn and Yahr stage, and living status.

1.2.2 Martin et al. dataset

The Martin et al. (2015) trial is a parallel delayed-start RCT carried out in New Zealand. The aim of the study was to investigate whether cueing could reduce the risk of falling. Cueing improves gait for PwP (Nieuwboer, 2008; Spaulding et al., 2013) and may alleviate Freezing of Gait (FOG), which is a symptom associated with PwP, but there is currently no evidence that a cueing program reduces the risk of falling (Rocha et al., 2014).

Twenty-one participants with diagnosis of Parkinson's, over 65 years old, with FOG, independently mobile, and with stable Parkinson's medication were recruited and randomized to an Immediate-Start (IS, n=12) or a 6-month Delayed-Start (DS, n=9) group. The delayed-start design is an alternative to the standard parallel-group RCT, aiming to increase the recruitment rate (Spineli et al., 2017). All participants received the same intervention four weeks (IS) or six months (DS) after entering the trial. The intervention was a home-based exercise and education program, which provided instruction on cued exercises using a metronome. Participants were instructed to record falls in a daily diary.

The dataset made available to this project included the count of falls in each of the 30 weeks, but no baseline characteristics were available in the dataset.

To make the dataset comparable with the other two, only the first 24 weeks are included in the analysis (see Table 1-1). Within these weeks, the IS group has records of falls counts after the intervention starts (at week 5), while the DS group is regarded as a control group. The reduced dataset thus follows the form of a standard RCT with four weeks of baseline and twenty weeks of follow-up falls collection. In the following text, IS and DS are referred to as the intervention and control groups respectively.

Table 1-1 Structure and included weeks in the Martin et al. dataset.

Group	**Included data set**		**Excluded data set**
	Weeks 1-4	Weeks 5-24	Weeks 24-30
IS (intervention group)	Waiting	Receiving intervention	Receiving intervention
DS (control group)	Waiting	Waiting	Receiving intervention

The Martin et al. trial has a limitation compared with the other two trials discussed in this section: its sample size of 21 is too small to detect even a large intervention effect.

1.2.3 EXSart dataset

The EXSart trial (Ashburn et al., 2007) is an RCT carried out in the UK between October 2002 and April 2005. The eligibility criteria were: with a diagnosis of Parkinson's, independently mobile, community dwelling, with a history of at least two falls in the previous year (the number was obtained via a retrospective question and referred to as the baseline falls count), and having passed a gross cognitive impairment test. One hundred and forty-two people were enrolled and randomised to an intervention (n=70) or control (n=72) group. The participants in the control group received usual care while the intervention group received a 6-week home based exercise programme. After the programme the participants in the intervention group were telephoned monthly to encourage them to continue the exercises. Baseline characteristics that were collected included sex, age, the number of years since diagnosis of Parkinson's, the Hoehn and Yahr stage, UPDRS (defined in section 2.1), and living status (alone, with partner, and with family/friends/other).

The design of the EXSart trial is similar to the other two, but with one major difference—the baseline and outcome falls were recorded using different methods. The baseline falls count was obtained retrospectively, by asking the participants how many falls they had experienced in the previous year in the screening interview. After the randomisation, the participants were instructed to prospectively record the number of falls they experienced in diaries with telephone reminders to do so. The falls count was analysed for the first 8 weeks and from 8 weeks to 6 months periods, they are referred to as the intervention and follow-up counts, respectively.

1.3 Structure of the thesis

The rest of the thesis is presented in the following structure:

- Chapter 2 is a review of the literature, background to falls prevention trials for PwP, and the statistical methods used in the thesis. In section 2.1, the diagnosis, treatment, and measures of severity for Parkinson's are introduced, and the intervention on falls in PwP are summarised. The designs for falls prevention trials are described in section 2.2, especially relating to the baseline count of falls and the two methods commonly used to collect the counts in falls prevention trials. In section 2.3, we introduce the statistical analyses that can be used for modelling falls, focussing on the NB model and NB-related models. Diagnostics for assessing and validating NB models are also described in this section.
- In Chapter 3, NB models are fitted to the three datasets, and compared to the Poisson model. The goal of this chapter is to 1) help understand the characteristic of the data, and 2) highlight the limitation of the NB modelling when the baseline count is not included in the model.
- Chapter 4 and Chapter 5 aim to provide tools to facilitate applied statisticians in analysing falls counts using NB models.
    - In Chapter 4, five statistical packages (`Stata`, `SAS`, `SPSS`, `R`, and `Python`) are reviewed regarding their functionality for fitting the NB and NB-related models. Some models can be fitted using several different modules in a package, each supporting different post-estimation commands for

  producing diagnostics, and the parameterisation of the models may differ across modules and packages. The difference between these options are discussed to assist in selecting a statistical package that best meets the specific goals of modelling.

  - Chapter 5 covers the application of diagnostic plots for NB models and a new plot is introduced in section 5.2. It is designed specifically for falls prevention trials where an outcome and baseline count are available but can also be used for other trials with similar features. An existing diagnostic plot, the covariate-adjusted probability plot, is described and examined in section 5.3. The possibility of using this plot to provide a visual inspection of overdispersion is discussed. In section 5.4, these plots are produced to show diagnostic statistics from the Poisson and NB models fitted to the three datasets.

- In Chapter 6 and Chapter 7, a number of approaches to incorporating the baseline count in NB or Conditional NB (CNB) models are compared based on the analysis of data from the motivating trials, and simulation studies.

  - Chapter 6 is motivated by the Goodwin et al. dataset and focuses on scenarios where the outcome and baseline counts are highly correlated, and heterogeneity in the outcome count is controlled by incorporating the baseline count in modelling.

  - Chapter 7 is motivated by the EXSart dataset, where the outcome was collected prospectively but the baseline count retrospectively, which introduces greater discrepancy between the two counts, violating the assumption underlying the CNB model. The goal of this chapter is to examine the robustness of the models to this discrepancy.

- Chapter 8 describes some other count response models. In section 8.1, the Poisson Inverse Gaussian model, which can model heavily skewed count data, were fitted to the EXSart dataset. The issue of zero-inflation was examined in the Goodwin et al. dataset in section 8.2. The possibility of using the right-censored or right-truncated NB model to reduce the influence of large counts is considered in section 8.3. Section 8.4 examines whether a finite mixture model could outperform an NB model in accommodating heterogeneity and large counts. Section 8.5 explores the benefits of fitting the random-effects NB model to a longitudinal count dataset.

- In Chapter 9, the contribution and the limitations of the study are summarised and discussed. Potential future research is suggested.

Chapter 2

Background

2.1 Falls among people with Parkinson’s

In 1817, an English doctor named James Parkinson published an essay (Parkinson, 1817) describing a neurological condition that led to “involuntary tremulous motion, with lessened muscular power, in parts not in action and even when supported; with a propensity to bend the trunk forwards, and to pass from a walking to a running pace: the senses and intellects being uninjured.” Parkinson named this condition “Shaking Palsy”, but today it is better known as Parkinson’s disease, or now just Parkinson’s. In his essay, Parkinson reported some cases who were afraid of falling forwards, and chose to either walk “on toes and forepart of the feet” or “take much quicker and shorter steps” to avoid falling. Since then, the condition is recognised as a risk factor of falling world-wide. While one third of elderly people experience at least one fall every year (Tinetti et al., 1988), for People with Parkinson’s (PwP), the proportion is doubled (Wood et al., 2002).

Guidance from the *National Collaborating Centre for Chronic Conditions* (National Collaborating Centre for Chronic Conditions, 2006) defined Parkinson’s as “a progressive neurodegenerative condition resulting from the death of the dopamine containing cells of the substantial nigra”. It is the second most common neurodegenerative disorder, with a prevalence of about 0.3% in the general population of industrialised countries (Goetz and Pal, 2014). As the average age of the global population is increasing, and Parkinson’s is becoming more recognized, the number of PwP is anticipated to further increase in the future (Rubenis, 2007).

To this date, the physiological mechanism for the high risk of falling among PwP is unclear. Freezing of gait (FOG) was found to be related to falling (Latt et al., 2009). Another possible reason causing PwP to fall is activities involving switching from one movement to another.

Foongsathaporn et al. (2016) found that switching movement, especially in the vertical direction (such as getting out of a car), is correlated to falling.

Pickering et al. (2007) conducted a meta-analysis of six prospective studies of falling in Parkinson's. The authors found 213 of 461 participants (46%, 95% CI: 38 to 54%) fell within 3 months, and even amongst those with no history of falling in the previous year the rate was 21%. A systematic review conducted by Deandrea et al. (2010) showed that PwP had a significantly higher risk of falling (OR: 2.71, 95% CI: 1.08 to 6.84) and recurrent falling (OR: 2.84, 95% CI: 1.77 to 4.58) than community-dwelling older people without Parkinson's.

A prospective cohort (Paul et al., 2017) conducted in Australia studied Fall-Related Hospital Admissions (FRHA) and injuries in the general elderly (≥ 60 years). The authors found that 2.5% of FRHAs were for people with a diagnosis of Parkinson's, while PwP only comprised 1.7% of the population in the same age group. PwP had a higher rate ratio for FRHAs (1.63, 95% CI: 1.59-1.67) than people without Parkinson's, as well as having longer hospital length of stay (median: 9 days versus 6 days). Another finding of the study was that PwP had a higher risk of injury (rate ratio: 1.47, 95% CI: 1.43-1.51). Around 67% and 35% of fall-related Parkinson's admissions in the study were related to injury and fracture, respectively.

Injury and fracture are common consequences of falls (Cumming et al., 1990). Genever et al. (2005) reported in a retrospective cohort study that the risk of injury for PwP is about twice that of a control group. A study conducted by Melton et al. (2006) showed a similar result: the risk of sustaining a fracture is 2.2 times higher in PwP than in non-Parkinson's community dwellers, and the risk of hip fracture specifically is 3.2 times higher for PwP. A number of studies indicated that the higher risk of injury for PwP is associated with falling. Allcock et al. (2009) reported from a prospective study that 32% and 1.2% of falls experienced by PwP resulted in injury and fracture respectively. Nyström et al. (2016) found an increased risk of fall-related injury (OR: 1.19) for PwP up to ten years before Parkinson's diagnosis, and an increased risk of fall-related hip fracture (OR: 1.36) more than 15 years before diagnosis, which suggests that injurious falls are likely to be related to the progress of Parkinson's from an early stage.

There may be psychological consequences of falls, including fear of falling, avoidance of daily activities, as well as loss of confidence and independence (Foongsathaporn et al., 2016;

Jørstad et al., 2005). The quality of life for PwP is considerably affected. In a falls study for elderly community dwellers, Tinetti et al. (1988) reported that 48% people who fell were afraid of falling, and 26% restrained their daily activities (e.g. shopping) to avoid falling.

The standard medication for Parkinson's is levodopa (Rascol et al., 2002), which is used as a dopamine-replacement therapy (Fahn et al., 2004). Although dopaminergic medicine is effective in decreasing bradykinesia and rigidity, it has less effect on falling (Keus et al., 2004). Despite using medicine or neurosurgery, PwP typically experience worsening body function and deteriorated daily activities as their condition progresses (Nijkrake et al., 2007).

A potential intervention to prevent falling is physiotherapy, which provides exercises, aids, education, and advice to PwP (Deane et al., 2002). It has been shown to improve the strength, postural balance, and motor co-ordination (Keus et al., 2004).

In clinical trials, the severity of Parkinson's is commonly measured by the Unified Parkinson's Disease Rating Scale (UPDRS) (Tomlinson et al., 2014), which comprises six sections (Wade, 1992): I. Mentation, behaviour, and mood; II. Activities of daily living; III. the Motor examination; IV. Complications of therapy; and two stand-alone scales V. the Modified Hoehn and Yahr staging; and VI. the Schwab and England Activities of Daily Living Scale. In 2008, the Movement Disorder Society (MDS) published a revision of this scale known as the MDS-UPDRS (Goetz et al., 2008), consisting of revisions to the first four parts of the original scale, and excluding Parts V and VI. The Section III Motor Examination is similar in both versions of the UPDRS and includes questions regarding: speech, facial expression, rigidity, finger tapping, hand movements, pronation/supination of hands, toe tapping, leg agility, arising from chair, gait, freezing of gait, postural stability, posture, global spontaneity of movement, postural tremor of hands, kinetic tremor of hands, rest tremor amplitude, and constancy of rest tremor. Because the questions of motor examination are most relevant to falling, Section III is often the only section of the UPDRS included in falls prevention trials in Parkinson's. Although participants should, ideally, rate their own disability (Goetz et al., 2008), the motor examination has to be assessed by a qualified clinician such as a physiotherapist.

The Hoehn and Yahr scale (Hoehn and Yahr, 1967) is a commonly used assessment of severity of Parkinson's for clinical use and research in its own right. The scale classifies Parkinson's into five stages: "1. Unilateral involvement only usually with minimal or no functional disability; 2. Bilateral or midline involvement without impairment of balance; 3. Bilateral disease: mild to moderate disability with impaired postural reflexes; 4. Severely disabling disease; the patient is still able to walk and stand unassisted but is markedly incapacitated; 5. Confinement to bed or wheelchair unless aided". Goetz et al. (2004) questioned the reliability of this scale, and proposed a modified Hoehn and Yahr scale, which is more specific for intermediate disease stages. The modified version consists of: "1.0 - Unilateral involvement only; 1.5 - Unilateral and axial involvement; 2.0 - Bilateral involvement without impairment of balance; 2.5 - Mild bilateral disease with recovery on pull test; 3.0 - Mild to moderate bilateral disease; some postural instability; physically independent; 4.0 -Severe disability; still able to walk or stand unassisted; 5.0 - Wheelchair bound or bedridden unless aided."

One may anticipate that PwP with poor UPDRS motor examination or more severe Hoehn and Yahr stage would experience more falls than those with medium scores since maintaining postural balance is more challenging for them. However, Pickering et al. (2007) showed an inverse U-shaped curve in a plot of the falls count against the UPDRS motor examination score: PwP fell more frequently as the UPDRS rating increased to begin with, but the falling rate decreased as UPDRS further increased. PwP with serious balancing difficulties may restrict their daily activities to avoid falls and fall related injuries.

2.2 Falls prevention trials

In falls prevention trials, outcomes assessing falling may include the rate of falling (or near falling), number of fallers, time to first fall, or fall related fractures (Gillespie et al., 2012).

In this study, we focus on the rate of falls, as it is both practical and has become increasingly popular in falls prevention trials. The rate of falls is the number of falls experienced by a participant in a certain period of time. The rate is interchangeable with the count of falls if the length of observation is the same for all trial participants. The length of observation is usually planned to be equal, but in practice participants may drop out of follow-up, usually assumed to be missing at random.

In addition to the outcome falls count collected during a post-randomisation period, it is not uncommon for falls prevention trials to collect the falls count during a pre-randomisation baseline period, which is typically planned to have equal duration for all participants. A meta-analysis (Pickering et al., 2007), which included six prospective studies of falling in Parkinson's from 2000 to 2004, found that the strongest explanatory variable for falls is the number of falls in the previous year.

2.2.1 Methods for collecting the count of falls

In a falls prevention trial, there are two main approaches to collecting information on the occurrence of falls: the retrospective and prospective methods.

For both methods the falls count is reported by the trial participants, so the reported number of falls depends on each participant's subjective understanding of falling. It is essential to give a clear definition in order to exclude falls related to external reasons. Clark et al. (1993) defined a fall as "an event that resulted in a person coming to rest unintentionally on the ground or other lower level, not as a result of a major intrinsic event or overwhelming hazard". This definition is strict and precise for falls prevention trials in PwP, but other versions have also been used, some of which are not clear and thus may be problematic. There are debates for categorizing falls in ambiguous settings. For example, if a person learns to collapse on chair or bed intentionally to avoid injury, does it count as a fall?

The retrospective method is a one-off question, asking participants to recall the number of falls they experienced over a specified period of time in the past. This approach is easy to implement, and thus has become a standard question in trial screening interviews to obtain a history of falling, which is often used as an eligibility criterion—for example, a trial may only enrol PwP who have experienced at least two falls in the previous year. The reason for choosing a threshold on a retrospective falls count is to limit the study population to people with a higher risk of falling. If most of the participants experienced no falls during the trial, the statistical power for the analysis is anticipated to be low. People who have already experienced falls are more likely to have subsequent falls and thus setting the selection criterion on the history of falls should increase the power of a trial (Cook and Wei, 2003).

When the number of falls is collected using the prospective method, participants are instructed to record each occurrence of falling in diaries during a period of follow-up. Lamb et al. (2005) recommended recording falls in monthly or shorter diaries, and participants should be reminded to complete their diaries by telephone. The prospective approach is more time-consuming and expensive, but the advantage of better accuracy compensates its higher cost.

It is not uncommon that participants of falls prevention trials record a large number of falls during a study period. This has been found in trials that collect the falls count prospectively or retrospectively. The mechanism under frequent falling is not clear for PwP. A possible reason is that the frequent fallers have a much higher risk of falling than average PwP; Another possibility is that when the falls count is collected retrospectively, participants may recall a large number because they had experienced frequent falling during a few weeks before the interview. The true reason that large falls counts occurred in falls prevention trials in PwP remains an open question.

Over the years, concerns have arisen regarding the objectivity of the retrospective method (Tinetti et al., 1988). Several studies have been conducted to examine discrepancies between counts obtained from the retrospective and prospective approaches. Cummings et al. (1988) compared both methods in a prospective study. People (not necessarily PwP) over the age of 60 were instructed to record falls and were followed up weekly for 12 months. At the end the study, they were asked to recall whether a fall had occurred within the trial period. There were 13% participants who failed to recall falls which were reported on their prospective diaries. The correlation between the number of falls recorded prospectively and retrospectively was only between 0.28 and 0.59 at 3, 6 and 12 months. The authors concluded that the number of falls recalled by elderly people has limited accuracy as they tend to forget falls. Because the memory of falling might be reinforced by the process of prospective recording, the true bias may be even larger than reported in the study.

Peel (2000) conducted a similar study with a duration of 12 months and showed results in line with Cummings et al., but with falling categorised as having “fallen at least once” or not. A considerable disagreement was found between the two methods: kappa agreement coefficient of 0.7, sensitivity of 79.5%, and specificity of 91.4% for reports of falling. The study further examined the accuracy of recalling the number of falls: over a third of the participants did not recall the number correctly, and the proportion of recalled falls became smaller as the number of falls increased.

Mackenzie et al. (2006) found 4% of participants gave false positive self-reports of falling and 13% gave false-negatives in a 12-month study (kappa = 0.84). The sensitivity of the retrospective recall over six months was 56% (95% CI: 44.1 to 67.5).

Although these three studies were targeted towards the general elderly rather than PwP, the conclusion is anticipated to hold for PwP for two reasons: 1) PwP are predominantly elderly people—the onset age of Parkinson's is rarely before 50 (Tysnes and Storstein, 2017); and 2) evidence suggests that Parkinson's is related to memory loss (National Collaborating Centre for Chronic Conditions, 2006). Therefore, forgetting falls is likely to be an even more serious problem among PwP than among the general elderly population.

Mackenzie et al. also suggested that the retrospective method may introduce bias into group comparisons. The participants in their intervention group were found to recall falls more accurately (sensitivity of 71.0%) than those in their control group (sensitivity of 40.5%), suggesting that intervention may improve the recall of falling, which would make the estimate of effect smaller than, or in the opposite direction to, its true value. A Cochrane review of falls prevention trials (Gillespie et al., 2012) described the methods of collecting outcome falls counts used in the trials and found that 55% adopted the prospective method while the rest either adopted the retrospective method or not clearly stated. The latter method was considered to be at high risk of introducing bias. The authors recommended that falls "should be recorded daily and monitored monthly."

McLennan et al. (1972) conducted a trial to investigate whether Parkinson's affects the hand-writing, which is usually referred to as micrographia. They found micrographia occurred in 5% of participants as the first symptom recorded. Difficulty in hand-writing may discourage PwP from recording each fall in their diaries. Although they could ask a carer to record falls for them, this provides a reason that a prospectively reported number of falls might be lower than the true value.

2.3 Statistical methods for the analysis of falls counts data

In falls prevention trials, the number of falls experienced by PwP is typically zero, one, or two, but in some cases very large numbers are reported. The distribution has the following pattern (an example is shown in Figure 2-1):

- Positive skew: the mean number of falls in the example dataset is 5.93 and the median is 1.
- With a few outliers: the maximum value in the example dataset is 499.

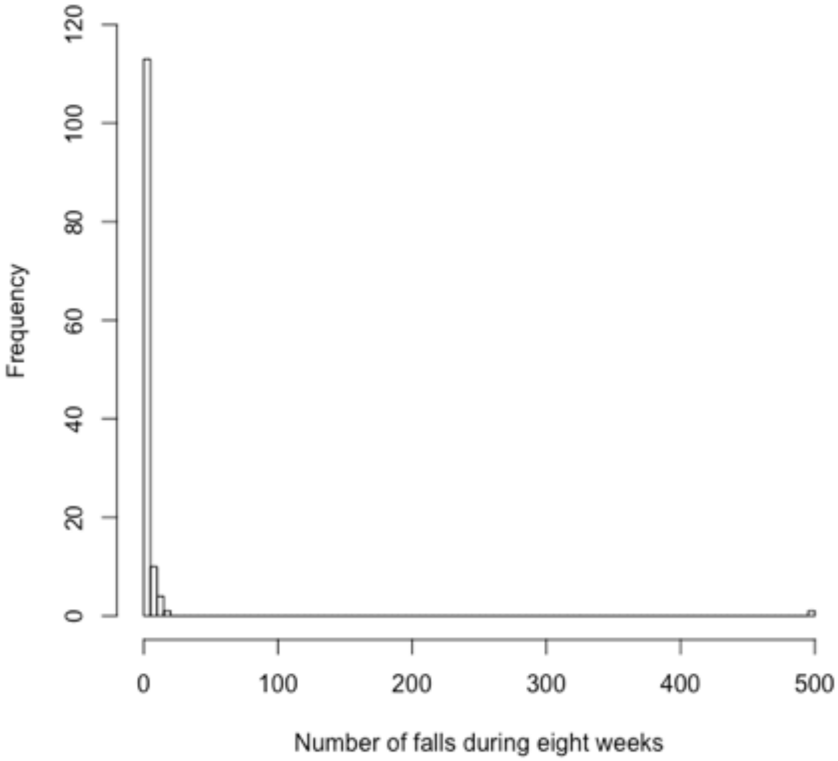


Figure 2-1 Distribution of an outcome falls count from the EXSart (Ashburn et al., 2007) trial (n=125).

Because of the skew, procedures that assume the normal distribution, such as the t-test, ANOVA, or linear regression, should not be used. The Mann-Whitney U test (also called Wilcoxon rank sum test) is a nonparametric approach and is widely used for skewed data. Aban et al. (2009) conducted a simulation study to compare the performance of models and tests in the analysis of NB-distributed count data. The study showed that the Mann-Whitney U test had lower statistical power than model-based tests, especially when the simulated data were closer to Poisson—that is, less overdispersed (see sections 2.3.1

and 2.3.3). The study also suggests that Mann-Whitney U test was conservative: its type I error rates were found to be lower than the nominal level 0.05 (minimum around 0.01) when the simulated data were closer to Poisson. Another drawback of the standard Mann-Whitney U test is that it cannot control for covariates, which is often needed in clinical trials. Two other nonparametric tests, the van Elteren and the $\mathrm{T}_{\mathrm{adap2}}$ tests, are capable of controlling for just one stratification variable, and they are limited to the comparison of only two groups (Jakobsen et al., 2015). The probabilistic index model (Thas et al., 2012; Vermeulen et al., 2015) can control for covariates and may be considered as a generalisation of the Mann-Whitney U test, it is a semiparametric model and its robustness to model misspecification has not yet been fully investigated for count data.

A traditional method of analysing counts of falls is to set a threshold to dichotomise the outcome count and then fit a logistic model. In practice, people not falling or falling at least once are categorized as “non-fallers” and “fallers” respectively, and people falling less than twice and at least twice as “seldom fallers” and “frequent fallers”. Donaldson et al. (2009) conducted a systematic review of fall prevention RCTs for community-dwelling older people. The review found that the most popular method of analysing falls data in trials reported during the period 1994 to November 2006 was reporting percentages and the Odds Ratios (OR) of people falling once or more (47 of 83 trials, 57%). Castañeda and Gerritse (2010) pointed out that ignoring the subsequent events results in a great loss of information. Cumming et al. (1990) argued that the proportion of participants with at least one fall within a time interval only focuses on the first fall. Since each fall has a risk of resulting in injury, subsequent falls should not be ignored. They further argued that the risk of the first fall is correlated to the length of the study. If a trial lasts a year, the proportion of people falling at least once would be much larger than a trial lasting ten weeks, assuming the average falls rate is the same in the two trials.

The advantage of fitting a logistic model is that outliers do not have a large impact on the estimation, but information is lost during this process, resulting in low statistical power. The power loss may be negligible when analysing a rare event but falling is a common recurrent event for many PwP. The dichotomising approach not only ignores the higher risk of injuries for frequent fallers, but also discards valuable information that would otherwise have been incorporated in statistical analysis.

A similar approach is to categorise the outcome count into an ordinal outcome (for example, “not falling,” “1-5 falls,” “6-20 falls,” and “more than 20 falls”) and fit an ordinal response model. Similar to the dichotomisation approach, modelling the ordinal outcome is robust to outliers, and it is anticipated to have higher statistical power than the standard logistic model. However, the categorisation of count variable still involves the subjective decision of choosing cut-points, and the issue of wasting information persists, although less information is lost compared to the standard logistic models.

A more appropriate approach for analysing falls data is to fit a count response regression model, which is based on the mean rate and results in an estimate of the effect size as an Incidence Rate Ratio (IRR), referred to as a Fall Rate Ratio (FRR) in this context. Because the FRR is based on all falls, it is easier to extrapolate to a wider population. An alternative model for a count outcome is based on the median of counts using quantile regression (Koenker and Bassett, 1978; Machado and Santos Silva, 2005). Although this model is robust to outliers, this study focuses on modelling the mean falls rate because 1) it usually requires a smaller sample size; 2) the estimated intervention effect on the mean rate may be easier to interpret by practitioners than the effect on the median; and 3) the effects of skewness/outliers on the estimation of a mean rate based model can be mitigated by using a suitable distribution for the outcome.

In some cases, logistic models were used because the cut-points bear clinical meaning; however, Sroka and Nagaraja (2018) argued that this should not be invoked as a justification for dichotomising count data because count response models can also produce ORs. They proposed an approach of fitting geometric, Poisson, or NB models with a log-odds link function, where the odds are based on the probability of the outcome count being greater or equal to the cut-point divided by the probability of the complement. Because these models share the same log-odds link function, their estimates are comparable to those of logistic models. The authors provided a mathematical proof showing that ORs estimated from count response models were more efficient (with higher Fisher information and smaller variance) than those from logistic models, and that the power improvement increases exponentially (especially for NB models) as the mean of the count increases; these were confirmed by their simulation study: the 95% CIs of OR from

NB models were only 61-69% as wide as those from logistic models, and the MSEs of $\log(\mathrm{OR})$ in NB were only 37.4-50.9% as high as those from logistic models.

In addition to examining the performance of each model when they were correctly specified, Sroka and Nagaraja (2018) included simulations to check their performance under model-specification. They simulated data from NB model (with various level of extra-Poisson variance) and compared the estimates for a binary covariate between Poisson and logistic models. The results bore out the robustness of logistic models to outliers: the Poisson models yielded heavily biased estimates when they are overdispersed, while the logistic models consistently gave estimates with low bias.

Though logistic models are robust to outliers, their power is generally low in the context of falls prevention trials and thus researchers are moving away from their use. Sroka and Nagaraja's (2018) approach of calculating ORs from count response model will not be further investigated in this thesis, because preventing the first or second fall, which are commonly-used cut-points for the dichotomisation approach, is not the goal of falls prevention trials for PwP in most cases; instead, the thesis focuses on count response models, with the aim of attenuating the effects of outliers and overdispersion.

2.3.1 Poisson regression model

Poisson regression is the most widely used statistical model for count data (Cameron and Trivedi, 2013). It assumes that a discrete random variable Y follows a Poisson distribution with parameter μ ($\mu > 0$). The Probability Mass Function (PMF) of Y is given by

$$f(y;\mu) = \frac{\exp(-\mu)\,\mu^{y}}{y!}, \qquad (2\text{-}1)$$

where $y \in \{0, 1, \dots\}$.

One possible derivation of the Poisson distribution is as the limit of the binomial distribution: $f(y;n,p) = \binom{n}{y} p^{y}(1-p)^{n-y}$, as the number of trials n approaches infinity and the probability of success p approaches zero, such that np equals a constant μ (McCullagh and Nelder, 1989).

As the Poisson model is a member of the Generalized Linear Models (GLM) family (McCullagh and Nelder, 1989), covariates can be introduced into the model via a linear predictor, linked to μ by the log function, in the form

$$g(\mu)= \log(\mu) =\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{2-2}$$

where $\boldsymbol{x}$ is a vector of m covariates $x_1, x_2, \dots, x_m$ and with coefficients vector $\boldsymbol{\beta} =(\beta_1, \beta_2, \dots, \beta_m)^{\mathrm{T}}$. Let β_k be the parameter for a group allocation variable x_k ($x_k = 1$ for the intervention group and $x_k = 0$ for the control group), then $\exp(\beta_k)$ is the FRR for the intervention effect.

Essentially, in Poisson regression the rate of an event is analysed, but when all subjects have the same exposure, say, one year, a falls count is effectively the falls rate per year. The number of falls occurring during the same length of observation is referred to as a falls count in the thesis for the sake of simplification. If participants are lost to follow up due to reasons independent of the risk of falling and group allocation (such as administrative reasons), the counts of the falls they experience after dropping out are assumed to be missing at random and not informative to model estimation (Balakrishnan, 2014). In this case, the length of follow-up periods, termed exposure, is different across subjects because they are shorter for those who drop-out, and this needs to be accommodated in the model by including the exposure as an offset.

An underlying property of the Poisson distribution is that the variance equals the mean; therefore, for Poisson regression:

$$\mathrm{E}(Y|\boldsymbol{x}) = \mathrm{Var}(Y|\boldsymbol{x}). \tag{2-3}$$

which is termed *equidispersion*.

In practice, when Poisson models are fitted to count data (including counts of falls), the response variance is often greater than the mean, and this problem is referred to as *Poisson overdispersion*, or *overdispersion* for short (Hilbe, 2011). Overdispersion results in inflated type I error rates in statistical tests based on Poisson models, leading to false positive test results (Breslow, 1990).

As overdispersion is common in falls data, the basic Poisson model is generally not suitable for analysing datasets from falls prevention trials. The reasons causing overdispersion and the alternative models are discussed in the following sections.

2.3.2 Heterogeneity and mixed Poisson models

Overdispersion may be caused for a number of reasons. In Poisson regression, each of the recurrent events is assumed to be independent of the others. If there are unobserved (latent) prognostic variables, the variance will exceed its mean. Overdispersion may also arise when there is positive correlation between events, which is termed *positive occurrence dependence*. If the explanatory variables in a model do not incorporate all the heterogeneity across individuals in a study, the problem is called *unobserved heterogeneity*, or *heterogeneity* for short (Winkelmann, 2008).

Let Y denote a count variable that follows a Poisson distribution with a conditional expectation of

$$\mathrm{E}(Y|\boldsymbol{x}, s) = \exp\,(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})s, \tag{2-4}$$

where $Y \in \{0, 1, \dots\}$, $\boldsymbol{x}$ is a vector denoting the m observed covariates $x_1, x_2, \dots, x_m$, $\boldsymbol{\beta}$ is a vector of the m coefficients, and s is a random variable representing the effects of the unobserved heterogeneity. Suppose that s follows a distribution with density $g(s)$, then

$$f(y|\boldsymbol{x}) = \int_0^\infty f(y|\boldsymbol{x}, s)g(s)ds, \tag{2-5}$$

where $s > 0$. Based on (2.5), we can construct a mixed Poisson model with a random subject effect s, in which the heterogeneity is accommodated (Lawless, 1987). To ensure identifiability of the regression parameters, without loss of generality, we choose that s satisfies that $\mathrm{E}(s) = 1$. There are three distributions that are commonly employed: the gamma, the inverse-Gaussian, and the log-normal distribution. They result in the following three mixed Poisson models: NB, Poisson Inverse Gaussian (PIG) (Dean et al., 1989), and Poisson log-normal (Winkelmann, 2008). The NB model is the only one with a closed form solution for likelihood function, while estimation of the other two models is based on simulation or quadrature (Hilbe, 2011).

There are alternative models based on the same underlying mixed Poisson distribution. Cook and Wei (2003) proposed the Conditional Negative Binomial (CNB) model, which is derived from the joint distribution of a baseline and outcome count, with a shared gamma-distributed random subject effect, by conditioning on the baseline count. The NB model can also be derived from the joint distribution by marginalising over s, as shown in equation (2-5). The CNB model is discussed in detail in section 6.2.1.

The main focus of the thesis is the NB model, but the PIG model is investigated in section 8.1.

2.3.3 Negative binomial regression model

The NB regression was first described by Glynn and Buring (1996) as a statistical model for analysing the rate of a recurrent event in medical studies.

The NB regression model is a generalization of the standard Poisson model, allowing extra variation in the outcome count of falls by including a random subject effect that follows a gamma distribution with mean 1 and variance α. The variance function of NB model is derived as

$$\mathrm{Var}(Y) = \mu + \alpha\mu^2. \tag{2-6}$$

It is sometimes referred to as the NB2 model (Hilbe, 2011), as the extra variance over the mean is provided by the product of α and the quadratic form of the mean (μ^2). Another parameterization for the variance function is

$$\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta,} \tag{2-7}$$

where $\theta = 1/\alpha$. The variance function in the (2-7) parameterisation is less straightforward but is the default in some statistical packages.

The parameter α reflects the amount of Poisson overdispersion: as α approaches zero, the NB model tends to a Poisson model; a larger α indicates greater overdispersion. Hilbe (2011) refers to α as the Heterogeneity Parameter (HP), because it indicates how much heterogeneity has been accounted for in the gamma component.

The PMF of the NB distribution (Hilbe, 2014) is:

$$f(y;\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\,\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}, \tag{2-8}$$

where $y \in \{0, 1, \dots\}$.

A linear predictor for the parameter μ in the NB model can be set up in an identical way to that for a Poisson model. An FRR is again calculated by exponentiating the coefficient of the covariate in question. The difference between NB and Poisson regression is that, because overdispersion is accommodated in the subject effect, NB produces more accurate model-based SEs, and it also results in a type I error rate that is closer to the nominal level in significance testing of intervention effects (Lawless, 1987).

Because NB is a GLM if α is constrained to be constant (Lawless, 1987), it can be estimated using an amended version of the Iteratively Reweighted Least Squares (IRLS) algorithm, which is used in the estimation of GLM models (Nelder and Wedderburn, 1972). The procedure for fitting the NB model as a GLM (Hilbe, 2011; Hilbe and Robinson, 2016) is to: 1) estimate α with a working $\hat{\mu}$ extracted from an initial iteration of Poisson or the previous NB GLM fit, and 2) estimate the regression coefficients with α fixed to $\hat{\alpha}$ from the previous step. These steps are repeated iteratively until convergence is achieved.

A factor for trial allocation may be included in an NB model, and then tested using the Wald, score, or Likelihood Ratio (LR) test. Aban et al. (2009) recommended using the score test when the sample size is small (less than 50) because its type I error is closer to the nominal level than other tests, and using the Wald and LR tests for trials with reasonably large sample sizes because of their higher power. They found that the power of the Wald and LR tests were almost identical. Because the Wald test is the default option for NB modelling in most statistical packages, it will be the focus of the analysis in the following chapters.

The goodness of fit of NB models can be examined using the Akaike Information Criterion (AIC) or Bayesian information criterion (BIC) statistics. AIC is defined as (Akaike, 1974):

$$\mathrm{AIC} = -2\hat{\mathcal{L}} + 2m, \tag{2-9}$$

where $\hat{\mathcal{L}}$ is the maximised log likelihood of the model and m is the number of estimated parameters; BIC is defined as (Schwarz, 1978):

$$BIC = -2\hat{\mathcal{L}} + \log(n)\, m. \tag{2-10}$$

where n is the number of subjects.

Although the NB model is recommended for analysing falls data (Gillespie et al., 2012; Robertson et al., 2005), it was found to be under-used (20/83, 24%) in a systematic review of falls prevention trials from 1994 to November 2006 (Donaldson et al., 2009).

There are other types of NB models, but NB2 is the standard and most commonly used NB model (Hilbe, 2011). It is conventional to refer to NB2 as the NB model for simplification. We follow this terminology unless a different type of NB model is used, in which case we switch to the full name NB2. A review of statistical packages regarding their functionality for fitting other types of NB models is included in Chapter 4. These models are described below.

Linear negative binomial model (NB1)

In linear NB models, the variance function is parameterised as $\mathrm{Var}(Y) = \mu + \delta\mu$ where $\delta > 0$ (Cameron and Trivedi, 1986). This model is usually referred to NB1 because, compared with the quadratic form of μ in the NB2 variance function, the linear NB model has a variance function with a linear form of μ.

The NB1 distribution is derived from $Y \sim \mathrm{Poisson}(\lambda_i)$, where $\lambda_i \sim \mathrm{gamma}(1/\delta, \mu)$. The PMF of NB1 is

$$f(y; \mu, \omega) = \frac{\Gamma(y+\mu)}{\Gamma(y+1)\,\Gamma(\mu)} \left(\frac{1}{1+\delta}\right)^{\mu} \left(\frac{\delta}{1+\delta}\right)^{y}. \tag{2-11}$$

Truncated and censored negative binomial models

A truncated distribution arises when the range of the outcome Y is a subset of the range of the original distribution (Rigby et al., 2017). For example, if in a falls trial all the participants reporting over 100 falls are excluded, the resultant distribution of the falls count is right-truncated, and if the participants who report less than 2 falls are excluded, the distribution is left-truncated. A censored distribution occurs when the exact value of the observation y_i is unknown given 1) $y_i \leq c_1$ (left censoring), 2) $y_i \geq c_2$ (right censoring), or 3) $c_3 \leq y_i \leq c_4$ (interval censoring) where c_1 to c_4 are known positive integers (Cameron and Trivedi, 2013). When the outcome count is truncated or censored, it is necessary to adjust the PMF of the NB model to accommodate the censoring or truncation structure.

Terza (1985) proposed the right-censored model as a solution to fit count response models to survey data, in which a common way to collect a count of an event is via a survey question, and one answer is commonly specified as "x times or more" where x stands for a positive integer. Cameron and Trivedi (2013) considered this model as a solution to attenuate the great influence of outliers.

The estimation of right-censored NB models is similar to the survival models. Let d_i be a censoring indicator such that $d_i = 1$ if $y_i \leq c$ and $d_i = 0$ if $y_i > c$, where c is a cut-point. For n independent observations, the log-likelihood function of a right-censored NB model is given by:

$$\mathcal{L}(\mu, \alpha) = \sum_{i=1}^{n} [d_i \log\left(f(y_i, \mu, \alpha)\right) + (1 - d_i) \log\left(1 - F(c, \mu, \alpha)\right)], \tag{2-12}$$

where $f(y_i; \mu, \alpha)$ are $F(y_i; \mu, \alpha)$ is the PMF and Cumulative Mass Function (CMF) of NB, respectively (Brännäs, 1992; Cameron and Trivedi, 2013).

Grogger and Carson (1988a, 1988b) were the first to study truncated count response models: they proposed the zero-truncated Poisson and NB models to analyse count data with no zeros — a special case of left-truncation. Gurmu and Trivedi (1992) generalized the truncated models for left- and right-truncated distribution. The PMF of right-truncated NB model (Gurmu and Trivedi, 1992) can be written as:

$$\Pr(Y = y \mid Y \leq c) = \frac{f(y; \mu, \alpha)}{F(c; \mu, \alpha)}. \tag{2-13}$$

Cameron and Trivedi (2013) pointed out that the right-truncated and right-censored models can both be used to solve the issue of outliers, but less information is lost in the right-censored model because large counts are revalued and labelled as right-censored, instead of dropped.

Zero-Inflated Negative Binomial (ZINB)

When there are excessive zero counts, it is possible that some zeros are generated from a process that is different to the count process. For example, if a participant of a falls prevention trials has fallen at least once during an observational period but has recorded no falls in the falls dairy, this zero count would be a different process to the zero count reported by another participant who has not experienced any falls during the period. It would be impossible to distinguish these two source of zeros, so the natural approach is to account for the excessive zeros in a model.

To solve the issue of excessive zeros, Lambert (1992) proposed the Zero-Inflated Poisson (ZIP) model, which assumes the zeros are generated from a binary component and a count component: the binary component generates excessive zeros with probability π. The count component accounts for the remaining $1-\pi$ probability and is assumed to follow a Poisson distribution (including zeros).

The Zero-Inflated NB (ZINB) model is an extension to the ZIP model by allowing both zero-inflation and overdispersion (Yau et al., 2003). In ZINB models, the distribution of the response variable Y can be written as (Ridout et al., 2001)

$$\Pr(Y=0)=\pi+(1-\pi)\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}} \tag{2-14}$$

for zero counts, and

$$\Pr(Y=y)=(1-\pi)\frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\,\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \tag{2-15}$$

for $y>0$; π is usually parametrised with a logit link such that

$$\log\left(\frac{\pi}{1-\pi}\right)=\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}, \tag{2-16}$$

where $\boldsymbol{z}$ is the vector of *inflation covariates* and $\boldsymbol{\gamma}$ is the vector for the corresponding coefficients.

The Vuong test (Vuong, 1989) compares two non-nested models fitted to the same data. It is regularly used to test zero-inflation by comparing the ZINB to the standard NB model, or ZIP to the standard Poisson model. The test statistic is asymptotically normally distributed, with positive test statistic values in favour of zero-inflated models while negative values in favour of standard NB or Poisson models. Desmarais and Harden (2013) showed that the Vuong test is biased in favour of choosing the ZINB model instead of the standard NB model. The authors proposed correcting the Vuong test statistic using the AIC or BIC statistic. They conducted a simulation study to compare the test results with corrections to that for the original test. The results confirmed that the Vuong test without corrections rarely rejects ZINB when the true model was NB but performed the best when the true model was ZINB. The simulations also showed that the AIC-based correction moderately favours ZINB, while the BIC-based correction favours NB.

Heterogeneous NB (NB-H) model

In the heterogeneous NB (NB-H) model, α in equation (2.6) is modelled with a linear predictor (Hilbe, 2011; Venkataraman et al., 2016), taking the form of $\log(\alpha_i) = \exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma})$.

2.3.4 Diagnostic statistics

Anscombe Residual

Residuals are informative for examining the fit of a model and the variability that remains unexplained by a model. McCullagh and Nelder (1989) described that an ideal residual “can be used to explore the adequacy of fit of a model, in respect of choice of variance function, link function and terms in the linear predictor.” Among the numerous types of standardised residuals available for NB regression, the Anscombe residuals (Anscombe, 1972) are reasonably normally distributed, and heterogeneity and outliers are easily identified (Hilbe, 2011; McCullagh and Nelder, 1989).

The Anscombe residual for the NB model (Hilbe, 2011) is defined as

$$r_i^A = \frac{\frac{3}{\hat{\alpha}}\left\{(1+\ \hat{\alpha}y_i)^{\frac{2}{3}} - (1+\ \hat{\alpha}\hat{\mu}_i)^{\frac{2}{3}}\right\} + 3\left(y_i -\ \hat{\mu}_i^{\frac{2}{3}}\right)}{2(\hat{\alpha}\hat{\mu}_i^2 +\ \hat{\mu}_i)^{\frac{1}{6}}}. \tag{2-17}$$

For models in the exponential family, Anscombe residuals approximately follow the standard normal distribution (Cameron and Trivedi, 2013). However, the NB model is not a member of the exponential family and I have not found literature that discusses whether the Anscombe residuals for NB models approximate the standard normal distribution. Hilbe (2011) recommended the Anscombe residual for NB models because it generally achieves better normality than the other residuals. Because of this, the Anscombe residual is reported for model diagnostics throughout the thesis.

Leverage

The diagonal of the *hat* matrix, a vector, is called *leverage*, and is a measure of the overall extremeness of the values in the explanatory variables for each subject (Madsen and Thyregod, 2010).

In NB regression, the hat matrix (Atkinson and Riani, 2012; Hilbe, 2011) is defined as:

$$\boldsymbol{h} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{\frac{1}{2}}, \tag{2-18}$$

where $\boldsymbol{X}$ is an $n \times m$ matrix denoting m explanatory variables for n subjects, $\boldsymbol{\beta}$ is a vector of the m coefficients, $\boldsymbol{W}$ is a diagonal matrix where the element in row i and column i is:

$$w_{i,i} = \frac{1}{V(\hat{\mu}_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \tag{2-19}$$

$V(.)$ is the variance function for NB regression in equation (2-6), and η_i is the i^{th} element of $\boldsymbol{\eta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$.

A high leverage suggests that the explanatory variables of this subject show low agreement with the other subjects, and this subject may potentially be influential in model estimation (Cook and Weisberg, 1982; Davison, 2003).

Cook's distance

In linear regression, the influence of each subject can be measured by Cook's distance (Cook, 1977). The measure of influence of subject i in a GLM (Williams, 1987) equivalent to Cook's distance for linear regression is given by

$$\mathrm{COOKD}_i = \frac{h_i}{m(1-h_i)}(r_i^p)^2, \tag{2-20}$$

where h_i is the i^{th} element of the leverage vector, and r_i^p is the standardised Pearson residual for subject i:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1-h_i)}}. \tag{2-21}$$

This measure is an approximation of $2m^{-1}[\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\widehat{\boldsymbol{\beta}}_{(i)})]$, where the $m \times 1$ vector $\widehat{\boldsymbol{\beta}}_{(i)}$ denotes the Maximum Likelihood Estimates (MLE) of $\boldsymbol{\beta}$ after subject i is deleted, and $\mathcal{L}(\widehat{\boldsymbol{\beta}}_{(i)})$ is the log likelihood evaluated at $\widehat{\boldsymbol{\beta}}_{(i)}$. Statistics based on assessing the effect of deletion are usually referred to as *deletion diagnostics* (Atkinson and Riani, 2012).

Cook's distance is a useful tool for detecting outliers. Because it is based on both the standardised Pearson residuals and the leverage, the Cook's distance measures the overall influence of each subject on the goodness of fit of a model. It is especially useful if the leverage of a subject is large but the residual is small, or vice versa.

DFBETA

The DFBETA statistic (Belsley et al., 1980; Williams, 1987) is a deletion diagnostic that approximates the influence of removing subject i on the estimation of a regression coefficient. The DFBETA of β_j ($j = 1, \dots, m$) for subject i is defined as the j^{th} element of

$$\mathbf{DFBETA}_i = w_{i,i}^2 (1-h_i)^{-\frac{1}{2}}\, r_i^p (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x}_i \approx \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}, \tag{2-22}$$

where vector $\boldsymbol{x}_i^{\mathrm{T}}$ is the i^{th} row of $\boldsymbol{X}$.

2.3.5 Assessing overdispersion in Poisson and NB models

Pearson dispersion statistic

The Pearson dispersion statistic (Hilbe and Robinson, 2016) can be used to assess both Poisson and NB overdispersion. The statistic is defined as (Wood, 2017):

$$\text{Pearson dispersion statistic} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)(n-m)}\ , \qquad \text{(2-23)}$$

where $V(\hat{\mu}_i)$ is the variance function of the model, that is, $\hat{\mu}_i$ for Poisson and $\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i$ for NB regression.

If the dispersion statistic is close to 1, it suggests that the model is equidispersed; if the statistic is greater than 1, it suggests overdispersion. However, this statistic is associated with sample size, so that a model with dispersion statistic only slightly greater than 1 may indicate overdispersion if the sample size is large. Two formal tests for Poisson and NB models are introduced as follows.

Testing Poisson overdispersion using the boundary likelihood ratio test

Because Poisson can be regarded as a special case of NB regression as α approaches 0, the two models are nested. A boundary likelihood ratio test with a boundary at $\alpha = 0$ can be conducted to assess whether there is overdispersion in the Poisson model (Cameron and Trivedi, 2013), with the deviance given by:

$$\text{LR} = -2(\mathcal{L}_{Poi} - \mathcal{L}_{NB}) \qquad \text{(2-24)}$$

where $\mathcal{L}_{Poi}$ and $\mathcal{L}_{NB}$ denote the log-likelihood of a Poisson model and the NB model fitted to the same data and with the same linear predictor. The test statistic LR in equation (2-24) asymptotically follows one half chi-square distribution, so the P value from the chi-square statistic is divided by 2 (Schlattmann, 2009).

Testing NB overdispersion using the Kim and Lee score test

The NB model accommodates overdispersion in a gamma distributed subject effect. However, if overdispersion exceeds that from the NB variance, the NB model itself can also be overdispersed.

Xue and Deddens (1992) included the extra-variation in NB regression as a multiplicative random effect v_i

$$Y_i|v_i \sim NB(\mu_i v_i, \alpha), \tag{2-25}$$

such that in addition to the heterogeneity accommodated by the gamma component in the NB2 model, there is hidden heterogeneity v_i with mean 1 and variance σ^2 that is not controlled for in the NB model. The authors proposed a score test for σ^2 as a means for testing for NB overdispersion.

Kim and Lee (2018) pointed out that the Xue and Dedden test treated β and α as fixed, and showed it to be conservative in a simulation study. The authors proposed a score test that adjusts for the uncertainty of $\hat{\beta}$ and $\hat{\alpha}$, and the test statistic is given by:

$$T = \frac{S(\hat{\beta}, \hat{\alpha})}{\sqrt{\hat{I}(\hat{\beta}, \hat{\alpha})}} \sim N(0,1), \tag{2-26}$$

where $S(.)$ is the score function for $\sigma^2 = 0$:

$$S(\beta, \alpha) = \frac{\partial \log L(\beta, \alpha, \sigma^2)}{\partial \sigma^2}\bigg|_{\sigma^2=0} = \frac{1}{2}\sum_i \frac{(y_i - \mu)^2 - (\mu_i + \alpha\mu_i y_i)}{(1 + \alpha\mu_i)^2}. \tag{2-27}$$

The formula for $\hat{I}(\hat{\beta}, \hat{\alpha})$ is not given here due to its length; see Kim and Lee (2018) for details.

Kim and Lee showed that their score test had higher power than Xue and Deddens's test and its type I error is closer to the nominal level 0.05.

2.3.6 Robust standard errors

Assume a model is fitted to a sample of n observations and $\widehat{\boldsymbol{\beta}}$ is the vector of MLEs of the model parameters, the robust variance estimator (also referred to as the sandwich estimator) for $\widehat{\boldsymbol{\beta}}$ (Hardin, 2003; Huber, 1967; White, 1980) is defined as

$$\mathbf{V}_{\text{rob}}(\widehat{\boldsymbol{\beta}}) = \mathbf{H}^{-1}\mathbf{M}\mathbf{H}^{-\text{T}}, \tag{2-28}$$

where $\mathbf{H}$ is the Hessian of the log likelihood, and

$$\mathbf{M} = \frac{n}{n-1}\sum_{i=1}^{n} \mathbf{u}_i^{\mathrm{T}} \mathbf{u}_i \,, \tag{2-29}$$

where $\mathbf{u}_i$ is the score at the MLE $\widehat{\boldsymbol{\beta}}$ for subject i.

The robust standard errors for $\widehat{\boldsymbol{\beta}}$ are given as

$$\mathbf{SE}_{\mathrm{rob}}\big(\widehat{\boldsymbol{\beta}}\big) = \sqrt{\mathrm{diag}(\mathbf{M})}. \tag{2-30}$$

The robust standard error is a consistent estimator of the standard deviation of the sampling distribution of $\hat{\beta}$ even when the variance function is misspecified (Hardin and Hilbe, 2007), and they are robust when there is overdispersion and model misspecification (Hardin, 2003).

2.3.7 Finite mixture models

If a sample is drawn from a homogeneous population, the count outcome y can be described using a PMF $f(y|\lambda)$ where λ is a parameter of population. However, if there are subpopulations such that for each sub-population j there is a corresponding λ_j, this is referred to as unobserved population heterogeneity (Böhning, 1999; Böhning and Seidel, 2003).

A Finite Mixture Model (FMM) accommodates the population heterogeneity using a mixture model such that

$$f(y,P) = \sum_{j}^{k} p_j f(y,\lambda_j)\,, \tag{2-31}$$

where p_j is the proportion of the subpopulation (also known as component) j; the number of components k may or may not be pre-specified (Schlattmann, 2009).

2.3.8 Longitudinal negative binomial models

Negative binomial models assume that observations are independent of each other (Breslow, 1996). This assumption is violated when observations are correlated within an observed data structure, which is usually termed *panels* (Cameron and Trivedi, 2013). Such data are called *panel data*, and often referred to as *longitudinal data* when "each of a number of subjects or patients give rise to a vector of measurements representing the same variable observed at a number of different time points" (Everitt, 1995).

When falls counts are recorded prospectively, they may be made available in the dataset as the number of falls recorded during, say, each week of an observation window. The datasets usually include an *id* variable to denote the subject and a *time* variable to indicate the weeks.

In general, there are four approaches for analysing count panel data: Generalized Estimation Equations (GEE), unconditional/conditional fixed-effects, and random-effects models (Hilbe, 2011). In the thesis, only the random-effects NB model is used in the analysis (section 8.5), the GEE and fixed-effects NB models are described here because Chapter 4 contains a review of software functionality for fitting NB related models.

Generalized Estimation Equations NB models

GEE models are an extension to GLMs that modify the variance function with a specified correlation matrix structure that parameterizes within-subject correlation (Liang and Zeger, 1986). The goodness of fit of GEE models can be assessed using the AIC, QIC (Quasi-likelihood Information Criteria), and CIC (Correlation Information Criteria) statistics; see Hin and Wang (2009) and Pan (2001) for details. Hin and Wang (2009) conducted a simulation study and showed that CIC had higher sensitivity and specificity in choosing the correct correlation structures than both QIC and AIC. The command to produce the QIC and CIC statistics in `Stata` is described in section 4.2.

Fixed-effects NB models

Unconditional fixed-effects NB models treat the subject effect as fixed, including a categorical subject indicator in the model linear predictor to estimate a different intercept for each participant of the trial, but it is only applicable when the number of participants (n) is small: a large n results in too many parameters in the linear predictor (Cameron and Trivedi, 2013). Hilbe (2011) suggested that n should be "less than 20" as a guide. When there are a large number of subjects, it is preferable to use a conditional fixed-effects NB model, which conditions on the subject effect through a sufficient statistic $\sum_t y_{it}$, where y_{it} denotes the number of counts for subject i during the t^{th} observation period (Hausman et al., 1984).

Although the fixed-effects model accounts for the subject effect, they are of little practical value for falls prevention trials. The reason is that the main goal of these trials is to study the effect of an intervention, estimated from the time-invariant regressor of group allocation, which is collinear with the fixed subject effect because it is also time-invariant (Hardin and Hilbe, 2002).

Random-effects NB models

Hausman et al. (1984) proposed the random-effects NB model for analysing longitudinal counts data.

Let y_{it} be the falls count in the t^{th} observation for subject i. Assume $y_{it}|\gamma_{it} \sim \text{Poisson}(\gamma_{it})$, where $\gamma_{it}|\delta_i \sim \text{gamma}(\mu_{it}, \delta_i)$ and δ_i is the heterogeneity parameter for subject i. The PMF of Y_{it} conditional on μ_{i1} and δ_i is given by

$$\Pr(Y_{it} = y_{it}|\, \mu_{i1}, \delta_i) = \frac{\Gamma(y_{it} + \mu_{it})}{\Gamma(y_{i1} + 1)\Gamma(\mu_{it})} \left(\frac{1}{1 + \delta_i}\right)^{\mu_{it}} \left(\frac{\delta_i}{1 + \delta_i}\right)^{y_{i1}}, \tag{2-32}$$

so that $\text{Var}(Y_{it}) = \mu_{it} + \delta_i \mu_{it}$. The PMF in (2-32) is similar to the PMF of an NB1 model as shown in equation (2-11), but δ_i now varies across subjects such that

$$\frac{1}{1 + \delta_i} \sim \text{Beta}(r, s), \tag{2-33}$$

which yields a PMF (Hausman et al., 1984; Hilbe, 2011) given by

$$\begin{aligned} &\Pr(Y_{it} = y_{it}; \mu, r, s) \\ &\quad = \frac{\Gamma(r + s) + \Gamma\left(r + \sum_{t=1}^{n_i} \mu_{it}\right) + \Gamma\left(s + \sum_{t=1}^{n_i} y_{it}\right)}{\Gamma(r)\Gamma(s)\Gamma\left(r + s + \sum_{t=1}^{n_i} \mu_{it} + \sum_{t=1}^{n_i} y_{it}\right)} \\ &\quad \times \prod_{t=1}^{n_i} \frac{\Gamma(\mu_{it} + y_{it})}{\Gamma(\mu_{it})\Gamma(y_{it} + 1)}. \end{aligned} \tag{2-34}$$

An advantage of the random-effects NB model over the fixed-effects NB model is that the time-invariant variables (including the intervention effect and baseline characteristics) can be included as covariates.

As shown in equations (2-11) and (2-32), the random-effects NB model is essentially an extension of the NB1 model: δ_i is constant across subjects in NB1 but varies in the random-effects NB model. Therefore, an NB1 model fitted to the panel data is sometimes compared to the random-effects NB model in an LR test.

Chapter 3

The characteristics of the three motivating datasets: exploratory analysis

3.1 Introduction

This chapter includes an exploratory analysis of the three datasets made available to this project. Each dataset is from a falls prevention trial aiming to investigate whether a physiotherapy or exercise intervention is effective in preventing falls among PwP.

In general, participants in the three trials recorded or recalled the number of falls they had experienced during a baseline period. After the randomisation, they prospectively recorded falls during an intervention period, in which they received physiotherapy programmes. In two trials, they were followed up for a period of time after the programme had ended. Although all three datasets share a common structure, they have unique characteristics that are important to data analysis.

Summary statistics for all the variables made available for each dataset are reported in section 3.2. Poisson and NB models are fitted to the falls counts during the outcome period(s) of each dataset. The models are compared and evaluated to provide an outline of how much Poisson overdispersion can be explained by observed baseline variables (excluding the baseline count).

This chapter aims to answer the following questions:

- If the baseline count is not collected in a falls prevention trial, would an NB model suffice to accommodate overdispersion for a small to medium sample size?
- Do the baseline characteristics reduce heterogeneity, and which of these variables are most important in this respect?
- Are outliers influential in the estimation and testing of the intervention effect?

3.2 Exploratory analysis for the three datasets

To better understand the three datasets, especially for Poisson overdispersion and other issues in NB modelling, both Poisson and NB regression were fitted to the three datasets and referred to as:

- ***basic* models**: including a Poisson (*Poi-basic*) and an NB (*NB-basic*) models that include only one covariate — the group allocation;
- ***full* models**: including a Poisson (*Poi-full*) and an NB (*NB-full*) models that include the group allocation and the baseline characteristics that are commonly collected in falls prevention trials — such as age, sex, severity of Parkinson's, and social status.

The *basic* models yield an FRR to estimate the intervention effect, without controlling for other subject-specific variables; whilst the *full* models incorporate the baseline characteristics, which are anticipated to reduce the heterogeneity and improve the statistical power in testing intervention effects.

The models are compared regarding the estimation of the intervention effect and goodness of fit. For each covariate included in a Poisson/NB model, the P value is reported from the likelihood ratio (LR) test to examine the explanatory power of the variable in modelling falls counts. For the categorical variables included in the *full* models, the largest category is chosen as the reference category.

To examine overdispersion in each model, the Pearson dispersion statistic is produced. We test Poisson and NB overdispersion using the boundary overdispersion test and NB overdispersion score test, respectively.

The statistical analysis was conducted in `R` (version 3.5.0). The Poisson and NB models are fitted using the `glm()` function in the **stats** package and the `glm.nb()` function in the **MASS** package, respectively. The P values from the NB overdispersion score test was produced using the code made available by the authors (Kim and Lee, 2018).

Note that though the baseline falls count is by far the most important regression covariate for predicting future falls (Pickering et al., 2007), it is not included in the *full* models in this chapter. How to incorporate a baseline falls count in NB models will be discussed in Chapter 6 and Chapter 7.

3.2.1 The Goodwin et al. dataset

Figure 3-1 shows the distribution of the falls count during the baseline (weeks 1-10), intervention (weeks 11-20), and follow-up periods (weeks 21-30) in the Goodwin et al. (2011) dataset. Overall, the distributions during the three periods are similar: most participants reported only a few falls, but a small number of people recorded outlying large numbers. Figure 3-2 shows that the falls counts during the three periods are strongly correlated; also, the largest counts shown in Figure 3-1 were reported by the same participants. The trial participants, including the most frequent fallers, tended to report consistent falls rates.

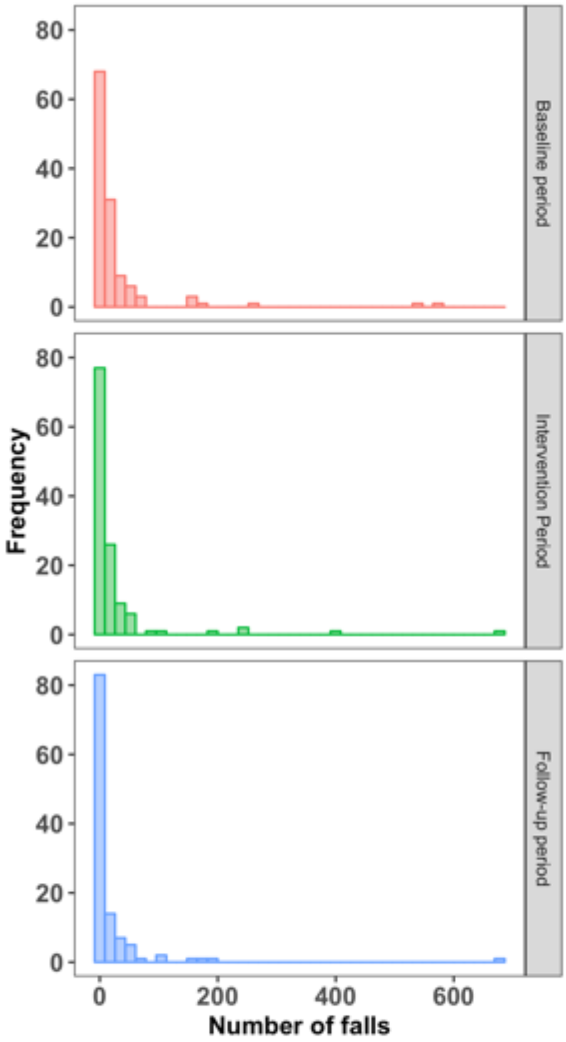


Figure 3-1 Goodwin et al. dataset: distribution of the falls count during baseline (weeks 1-10, n=124), intervention (weeks 11-20, n=125), and follow-up (weeks 21-30, n=126) periods.

As shown in Table 3-1, the average falls rate of the control group is around thirty falls per ten weeks across the baseline, intervention, and follow-up periods. Although the risk of falling may increase as the Parkinson's progresses, the average falls rate of the control group is relatively stable during the whole trial, which may be explained by the short length of the study (30 weeks in total). The average count in the intervention group, in comparison, decreases from 26.48 per 10 weeks during the baseline period to 17.93 per 10 weeks during the intervention period, and is further reduced to 7.36 per 10 weeks during the follow-up period.

The ranges of the reported falls counts are wide: the maximum is over five hundred falls within 10 weeks (this participant reported the largest counts for all three periods; see Table 3-3). In both groups, the falls counts have much greater variances than the means, suggesting the presence of overdispersion if the heterogeneity is not sufficiently accounted for by covariates in a statistical model.

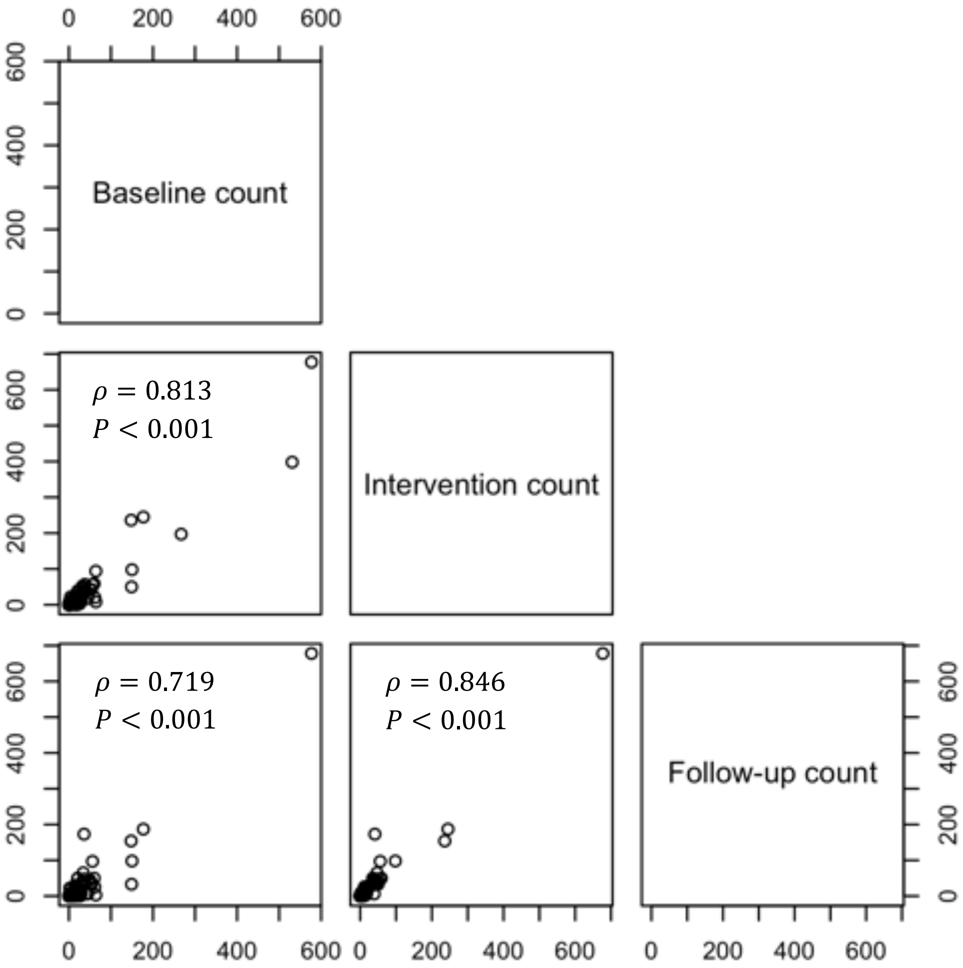


Figure 3-2 Scatter matrix plots of the baseline, intervention, and follow-up falls count in the Goodwin et al. dataset. The Spearman's correlation coefficient ρ and P value is shown in each subplot.

Table 3-1 Summary of the complete falls count during the baseline, intervention, and follow-up periods in the Goodwin et al. trial

Period	Group	N	Missing	Mean	Median	SD (variance)	Range
Baseline period (weeks 1-10)	Intervention	60	4 [a]	26.48	6.5	77.18 (5956.65)	0-531
	Control	64	2	29.08	6.5	78.13 (6103.53)	0-577
	Total	124	6	27.82	6.5	77.36 (5985.15)	0-577
Intervention period (weeks 11-20)	Intervention	61	3	17.93	3.0	56.44 (3185.60)	0-398
	Control	64	2	32.25	6.0	93.41 (8724.89)	0-677
	Total	125	5	25.26	5.0	77.63 (6025.84)	0-677
Follow-up period (weeks 21-30)	Intervention	56	8 [b]	7.36	2.5	11.17 (124.85)	0-49
	Control	60	6 [c]	31.88	4.0	94.32 (124.85)	0-678
	Total	116	14	20.04	3.0	69.11 (4775.87)	0-678

[a] ID 1 has missing values during weeks 4-10 (baseline periods);

[b] ID 18 and 97 from the intervention group dropped out at week 22;

[c] ID 51 and 101 from the control group dropped out at weeks 23 and 21, respectively

The extreme skew of the distribution suggests that Poisson models are likely to be overdispersed, but overdispersion cannot be verified unless a model is actually estimated. The Poisson and NB models are fitted to the intervention and follow-up falls counts. Following the nomenclature defined in section 3.1, the Poisson and NB models are referred to as: *Poi-basic* and *NB-basic* when the group allocation is the sole covariate; *Poi-full* and *NB-full* when group allocation, gender, age, years since diagnosed with Parkinson's, Hoehn and Yahr stage, living status are included in the models. The baseline characteristics that are included as covariates in the *full* models are summarized in Table 3-2.

Table 3-2 Goodwin et al. trial: characteristics of the participants at baseline

		Intervention group (n=64)	**Control group** (n=66)	**Total** (n=130)
Sex				
	Male	39 (61%)	35 (53%)	74 (57%)
	Female	25 (29%)	31 (47%)	56 (43%)
Age (years)				
	Mean (SD)	72.0 (8.6)	70.1 (8.3)	71.0 (8.5)
	Range	50-87	53-89	50-89
Years since diagnosis				
	Mean (SD)	9.1 (6.4)	8.2 (6.4)	8.7 (6.4)
	Range	1-26	1-30	1-30
Hoehn and Yahr				
	Stage 1	4 (6%)	9 (14%)	13 (10%)
	Stage 2	31 (48%)	28 (42%)	59 (45%)
	Stage 3	16 (25%)	21 (32%)	37 (28%)
	Stage 4	13 (20%)	8 (12%)	21 (16%)
Living status				
	Alone	14 (22%)	19 (29%)	33 (25%)
	With partner	48 (75%)	44 (67%)	92 (71%)
	With family/friends	1 (2%)	2 (3%)	3 (2%)
	Other	1 (2%)	1 (2%)	2 (2%)

Table 3-3 lists the ten most frequently falling participants during each of the baseline, intervention, and follow-up periods. The two groups are reasonably balanced for large baseline counts. During the intervention period, eight out of ten most frequently falling participants were in the control group, possibly because the intervention has reduced the falls rate. Because two frequently falling participants in the intervention group dropped out before the follow-up period started, only one of the ten most frequently falling participants during the follow-up periods was in the intervention group. The baseline characteristics of the participants shown in Table 3-3 are listed in Table 3-4, and no clear pattern could be seen.

Table 3-3 Ten largest falls counts during each of the baseline, intervention, and follow-up periods of the Goodwin et al. trial

Period	ID	Group	Baseline count	Intervention count	Follow-up count
Baseline	75	Control	577.0	677	678
	97	Intervention	531.0	398	-
	18	Intervention	267.0	197	-
	95	Control	177.0	245	187
	9	Control	150.0	98	98
	108	Intervention	149.0	50	33
	45	Control	148.0	236	154
	7	Intervention	64.0	8	3
	101	Control	64.0	94	-
	112	Control	61.5	20	25
Intervention	75	Control	577	677	678
	97	Intervention	531	398	-
	95	Control	177	245	187
	45	Control	148	236	154
	18	Intervention	267	197	-
	9	Control	150	98	98
	101	Control	64	94	-
	12	Control	60	59	50
	44	Control	38	57	48
	22	Control	56	56	97
Follow-up	75	Control	577	677	678
	95	Control	177	245	187
	116	Control	36	41	173
	45	Control	148	236	154
	9	Control	150	98	98
	22	Control	56	56	97
	99	Control	34	47	65
	11	Control	21	35	50
	12	Control	60	59	50
	96	Intervention	21	39	49

For each period, the rows are sorted by the grey column

Table 3-4 Baseline characteristics of the frequent fallers listed in Table 3-3.

ID	Group	Sex	Age	Years since diagnosis	Hoehn & Yahr	Living status
7	Intervention	Male	73	10	Stage 3	With partner
9	Control	Male	71	30	Stage 4	With partner
11	Control	Female	69	20	Stage 3	With partner
12	Control	Male	81	5	Stage 3	With partner
18	Intervention	Male	79	13	Stage 3	With partner
22	Control	Female	72	12	Stage 2	With partner
44	Control	Male	64	6	Stage 2	Alone
45	Control	Male	55	21	Stage 3	With partner
75	Control	Male	71	2	Stage 3	Alone
95	Control	Male	62	7	Stage 4	With partner
96	Intervention	Female	78	5	Stage 1	Alone
97	Intervention	Male	74	10	Stage 4	With partner
99	Control	Male	78	8	Stage 2	With partner
101	Control	Male	70	14	Stage 3	With partner
108	Intervention	Male	55	2	Stage 2	With partner
112	Control	Female	75	5	Stage 3	With partner
116	Control	Male	67	1	Stage 1	With partner

We shall first look at the models fitted to the falls count during the intervention period.

Intervention period

As shown in Table 3-5, the *Poi-basic* model yields an FRR of 0.556 for the intervention effect, which indicates that during weeks 11-20 the fall rate was 44% lower for people received the intervention than those received the usual care.

There is evidence of overdispersion in the *Poi-basic* model. Firstly, the AIC of *Poi-basic* is more than ten times higher than the AIC of *NB-basic*. Secondly, *Poi-basic* results in an enormous dispersion statistic (225.215). In addition, the boundary LR overdispersion test yields a significant result (P<0.001). Although the LR test of the intervention effect is significant (P<0.001) in *Poi-basic*, it has little meaning since overdispersion often leads to false positives in model-based hypothesis testing, especially when *NB-basic* yields a non-significant (P=0.072) test result.

The estimates of the intervention effect in the *NB-basic* and *Poi-basic* models are almost identical, but the SE of the estimate is much larger in *NB-basic*. This is because SEs in overdispersed Poisson regression are typically underestimated.

The large dispersion statistics and significant NB overdispersion score test result (P<0.001) suggests that the *NB-basic* may also be overdispersed. Consequently, the test of the intervention effect based on *NB-basic* is likely to be liberal.

Table 3-5 Goodwin et al. dataset: *Poi-basic* and *NB-basic* models fitted to the intervention count (n=125)

	Poi-basic				***NB-basic***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-0.587	0.037	0.556 (0.517, 0.599)	< 0.001	-0.587	0.322	0.556 (0.294, 1.051)	0.072
HP					3.187			
Dispersion	225.215				2.825			
AIC	10026.4				936.7			
Overdispersion test	P < 0.001[a]				P < 0.001[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

A solution to remedying overdispersion is to include more regression covariates, because they may explain heterogeneity at the subject level. This approach also increases the statistical power of the model.

As shown in Table 3-6, the FRRs of the intervention effect from *Poi-full* (0.513) and *NB-full* (0.538) are close to the FRRs from the *basic* models. During weeks 11-20, the risk of falling for people who received the intervention was expected to be halved compared to those who received the usual care, with the remaining covariates held constant. The *Poi-full* model is less overdispersed than *Poi-basic*, but the dispersion statistic is still large (90.624). This is also borne out by the significant boundary LR overdispersion test (P<0.001), and the SEs that are an order of magnitude smaller than the SEs from *NB-full*.

The LR test of the intervention effect in the *NB-full* model is significant (P=0.044), but this may be caused by the imbalance of large falls count between the two trial arms. Eight out of the ten most frequently falling participants during the intervention period, including ID

75 who recorded the largest falls count in all three periods, were in the control group (see Table 3-3), which may lead to an overestimated intervention effect in preventing falls.

Table 3-6 Goodwin et al. dataset: *Poi-full* and *NB-full* models fitted to the intervention count (n=125)

	Poi-full				***NB-full***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	IRR (95% CI)	P
Intervention	-0.667	0.038	0.513 (0.476, 0.553)	< 0.001	-0.619	0.289	0.538 (0.303, 0.955)	0.044
Female	-1.557	0.055	0.211 (0.189, 0.235)	< 0.001	-1.194	0.300	0.303 (0.167, 0.550)	0.001
Age	-0.020	0.002	0.980 (0.976, 0.985)	< 0.001	-0.015	0.018	0.985 (0.950, 1.022)	0.469
Years since diagnosis	-0.017	0.003	0.984 (0.977, 0.990)	< 0.001	0.011	0.027	1.011 (0.958, 1.066)	0.710
Hoehn & Yahr				< 0.001				0.003
Stage 1	-0.148	0.102	0.863 (0.706, 1.055)		0.543	0.513	1.722 (0.623, 4.755)	
Stage 2			1				1	
Stage 3	1.534	0.051	4.635 (4.190, 5.126)		1.104	0.353	3.015 (1.497, 6.072)	
Stage 4	1.752	0.058	5.767 (5.139, 6.473)		1.485	0.469	4.413 (1.742, 11.180)	
Living status				< 0.001				0.458
With partner			1				1	
Alone	0.373	0.043	1.452 (1.332, 1.583)		0.263	0.354	1.301 (0.645, 2.624)	
With family/friends	0.882	0.253	2.416 (1.463, 3.989)		0.264	1.194	1.302 (0.122, 13.871)	
Residential home	-2.904	1.002	0.055 (0.008, 0.399)		-2.410	1.434	0.090 (0.005, 1.538)	
HP					2.365			
Dispersion	90.624				1.459			
AIC	6918.6				914.5			
Overdispersion test	P < 0.001[a]				P = 0.038[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

Among the baseline characteristics that are included in *NB-full* as covariates, sex and Hoehn and Yahr scale were significant (P = 0.001 and 0.003 respectively). A female PwP has 69.7% lower falls rate than a male PwP with the same baseline characteristics and in the same group. The CIs of FRR for Stage 3 and 4 Hoehn and Yahr scale both do not contain 1, which indicate significant higher risk of falling for PwP with Hoehn and Yahr Stage 3 and 4 than with Stage 2 (the reference level).

Although *NB-full* has a smaller dispersion statistic (1.459) than *NB-basic* (2.825), the NB overdispersion score test indicates that it is also overdispersed (P=0.038).

Follow-up period

As shown in Table 3-1, four participants dropped out within the first four weeks of the intervention period. With the assumption that this was due to missing at random, the number of weeks available for the follow-up counts was considered as the exposure of the follow-up count, and the logged exposure was included in the models as an offset.

As shown in Table 3-7 and Table 3-8, the *Poi-basic* and *Poi-full* models fitted to the follow-up count are both significantly overdispersed (P<0.001). Similar to the Poisson models fitted to the intervention count, they yield large dispersion statistics and AICs.

Table 3-7 Goodwin et al. dataset: *Poi-basic* and *NB-basic* models fitted to the follow-up count (n=120)

	Poi-basic				***NB-basic***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-1.378	0.052	0.252 (0.227, 0.279)	<0.001	-1.247	0.347	0.287 (0.144, 0.572)	0.001
HP					3.511			
Dispersion	148.8				1.5			
AIC	7342.1				817.3			
Overdisperison test	P < 0.001[a]				P = 0.018[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

The AIC of the *NB-basic* model is considerably smaller than *Poi-basic*. The FRR of the intervention effect is estimated by *NB-basic* to be 0.287 (P<0.001) without controlling for the baseline characteristics, which indicates an astonishing 71% lower falls rate during the follow-up period experienced by people who received the intervention, compared to those receiving the usual care.

Table 3-8 shows the estimation of the *Poi-full* and *NB-full* models. The estimated intervention effect from *NB-full* (FRR=0.361, P=0.004) is relatively close to that from *NB-basic* (FRR=0.287, P<0.001), suggesting the intervention reduced falls rate by more than 60%. The large estimated effect was possibly due to the group imbalance regarding frequent fallers. As shown in Table 3-3, the nine most frequently falling participants during the follow-up period were all in the control group. Although the *NB-basic* and *NB-full* models yielded remarkably large intervention effect during the follow-up period, the estimates may be influenced by the large counts in the control group (see Figure 5-9 in Chapter 5).

The NB overdispersion score test suggests that *NB-basic* is significantly overdispersed (P=0.018) but *NB-full* is not (P=0.191). Sex is the only baseline characteristic with a significant LR test result (P=0.010). The LR test of Hoehn and Yahr scale was significant when modelling the intervention falls count but not the follow-up count.

Table 3-8 Goodwin et al. dataset: *Poi-full* and *NB-full* models fitted to the follow-up count (n=120)

	Poi-full				***NB-full***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-1.361	0.053	0.256 (0.231, 0.285)	< 0.001	-1.020	0.330	0.361 (0.188, 0.693)	0.004
Female	-1.245	0.055	0.288 (0.258, 0.321)	< 0.001	-0.987	0.344	0.373 (0.188, 0.737)	0.010
Age	-0.028	0.003	0.972 (0.967, 0.977)	< 0.001	-0.027	0.021	0.973 (0.933, 1.014)	0.273
Years since diagnosis	-0.015	0.004	0.985 (0.977, 0.992)	< 0.001	0.023	0.030	1.023 (0.963, 1.087)	0.499
Hoehn & Yahr				< 0.001				0.053
Stage 1	0.376	0.082	1.457 (1.239, 1.713)		1.098	0.569	2.997 (0.970, 9.262)	
Stage 2			1				1	
Stage 3	1.433	0.055	4.192 (3.760, 4.673)		0.956	0.406	2.601 (1.163, 5.815)	
Stage 4	1.172	0.070	3.228 (2.810, 3.708)		0.997	0.527	2.710 (0.953, 7.709)	
Living status				< 0.001				0.586
With partner			1				1	
Alone	0.588	0.047	1.800 (1.639, 1.976)		0.412	0.404	1.510 (0.678, 3.360)	
With family/friends	0.378	0.267	1.460 (0.859, 2.480)		-0.343	1.336	0.709 (0.050, 10.026)	
Residential home	-1.502	0.583	0.223 (0.070, 0.707)		-1.449	1.430	0.235 (0.014, 3.994)	
HP					2.963			
Dispersion	76.2				1.3			
AIC	5529.2				814.4			
Overdispersion test	P < 0.001[a]				P = 0.191[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

3.2.2 The Martin et al. dataset

Figure 3-3 shows the distributions of the falls rates (per week) during the 4-week baseline and 20-week follow-up periods in Martin et al. (2015) dataset. Note that the falls rates are reported instead of the counts, because a large proportion of subjects missed at least one week of records in falls diaries (assumed to occur at random and not informative), and thus have varying lengths of observation time. No baseline covariates were made available to the project, so only *Poi-basic* and *NB-basic* are examined in this section.

The sample size of the Martin et al. dataset is only 21, but the distributions of the falls counts are nevertheless heavily skewed (Figure 3-3). Similar to the Goodwin et al. dataset, the follow-up falls rate is correlated to the baseline rate (see Figure 3-4).

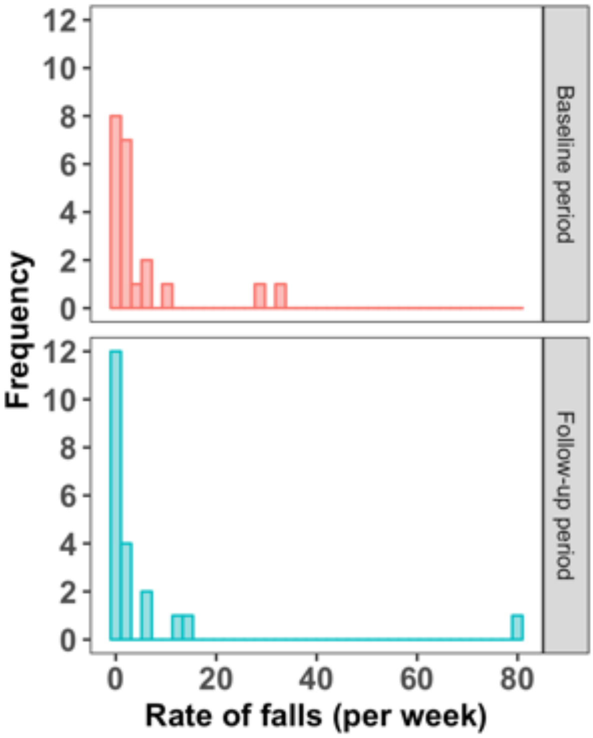


Figure 3-3 Martin et al. dataset: distribution of the falls counts during baseline (4 weeks, n=21) and follow-up (20 weeks, n=21) periods

.

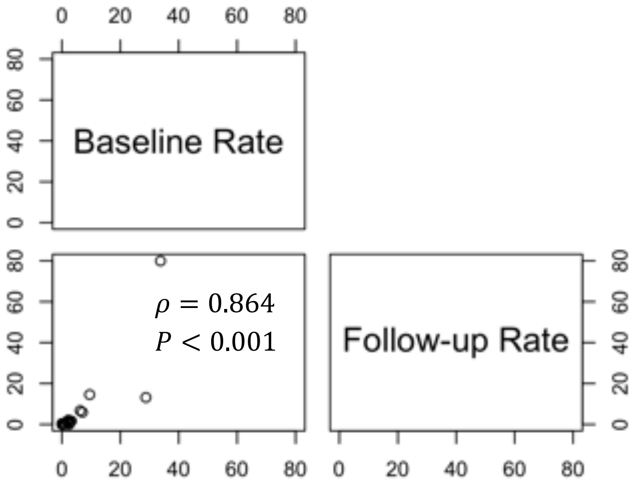


Figure 3-4 Martin et al. dataset: scatter matrix plots of the baseline and follow-up falls rate (per week) The Spearman's correlation coefficient ρ and P value is shown in each subplot.

Table 3-9 shows the summary statistics of the baseline and follow-up fall rates. It is interesting that for a trial with such a small sample size, one participant (ID CU21) reported eighty falls per week (see Table 3-10) and 1599 falls in total during the follow-up period. Because ID CU21 was randomised to the intervention group, the average falls rates of the intervention group are higher than those of the control group. Outliers in a small dataset are anticipated to be influential in the model estimation.

Table 3-9 Martin et al. dataset: summary of the falls rate (per week) during the baseline and intervention periods

		N	Mean	Median	SD (variance)	Range
Baseline period (weeks 1-4)	Intervention	12	6.69	1.5	11.68 (136.46)	0-33.75
	Control	9	2.83	2.3	3.17 (10.06)	0-9.50
	Total	21	5.04	1.75	9.11 (82.90)	0-33.75
Follow-up period (weeks 5-24)	Intervention	12	8.60	0.3	22.80 (519.82)	0-79.95
	Control	9	3.04	1.0	4.78 (22.84)	0-14.55
	Total	21	6.22	9.0	17.41 (303.00)	0-79.95

The most frequently falling participants during each of the baseline and follow-up period are listed in Table 3-10. The falls rate reported by ID CU21 is outlying large compared to the other participants.

Table 3-10 Ten largest falls rates during the baseline and intervention periods of the Martin et al. trial.

Period	ID	Group	Baseline rate	Intervention rate
Baseline	CU21	Intervention	33.8	80.0
	CU11	Intervention	28.8	13.2
	CU02	Control	9.5	14.6
	CU01	Intervention	7.0	5.9
	CU09	Control	6.2	6.7
	CU06	Intervention	3.2	1.6
	CU20	Control	3.0	1.6
	CU14	Control	2.5	0.9
	CU18	Intervention	2.5	0.0
	CU03	Control	2.2	2.1
Intervention	CU21	Intervention	33.8	80.0
	CU02	Control	9.5	14.6
	CU11	Intervention	28.8	13.2
	CU09	Control	6.2	6.7
	CU01	Intervention	7.0	5.9
	CU03	Control	2.2	2.1
	CU06	Intervention	3.2	1.6
	CU20	Control	3.0	1.6
	CU10	Intervention	1.8	1.5
	CU12	Control	1.8	1.0

For each period, the rows are sorted by the grey column

Table 3-11 presents the results of the *Poi-basic* and *NB-basic* model. The boundary LR overdispersion test indicates significant Poisson overdispersion (P<0.001), which is confirmed by the large dispersion statistic and AIC of the *Poi-basic* model, compared with those of *NB-basic*.

Table 3-11 Martin et al. dataset: *Poi-basic* and *NB-basic* models (n=21)

	Poi-basic				***NB-basic***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	1.156	0.050	3.178 (2.864, 3.526)	< 0.001	1.041	0.856	2.833 (0.472, 16.993)	0.250
HP					3.753			
Dispersion.	660.3				1.4			
AIC	7482.9				204.2			
Overdisperison test	P < 0.001[a]				P = 0.353[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

The *NB-basic* model yields an FRR of 2.833 for the intervention effect, which indicates that the risk of falling for people who received the intervention increased by 183% (P=0.250). The reason the FRR is greater than 1 is that participant CU21 recorded 1599 falls during the follow-up period, which is more than five times of the second largest number, and this participant was in the intervention group (see Table 3-10). This result indicates that, when the sample size is small, the estimation of the intervention effect is extremely susceptible to the influence of outliers, which are unlikely to be balanced between groups.

The NB overdispersion score test suggests that the *NB-basic* model is not significantly overdispersed (P=0.353).

3.2.3 The EXSart dataset

Figure 3-5 shows the distributions of the falls counts in the EXSart dataset (Ashburn et al., 2007) during the baseline (1 year prior to the screening interview), intervention (first 6 weeks), and follow-up (between 8 weeks and 6 months) periods. As shown in Figure 3-5, the distribution of the baseline count is more skewed than the intervention and follow-up counts, because 1) the baseline period lasts for one year, longer than the other two periods; and 2) the falls count during the baseline period was obtained retrospectively, which has lower precision than the prospective method, especially when a person falls frequently.

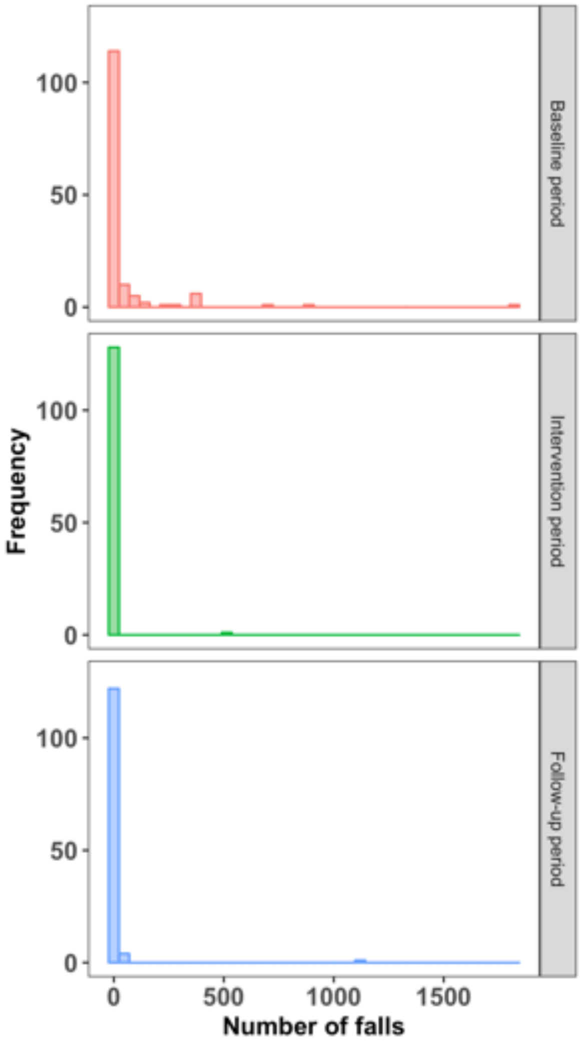


Figure 3-5 EXSart dataset: distribution of the count of falls during the baseline (1 year, n=142), intervention (first 8 weeks, n=129), and follow-up (between 8 weeks and 6 months, n=127) periods.

Compared to the other two datasets, the baseline falls rates in the EXSart dataset are less consistent with the rates during the intervention and follow-up periods (see Figure 3-6). The baseline rates for the frequent fallers in the Goodwin et al. dataset are close to the intervention and follow-up rates; even the frequent fallers in the Martin et al. dataset, which has a very small sample size, have consistent baseline and follow-up falls rates. The participants in the EXSart trial reporting large baseline falls counts did not record as many falls during the intervention and follow-up periods.

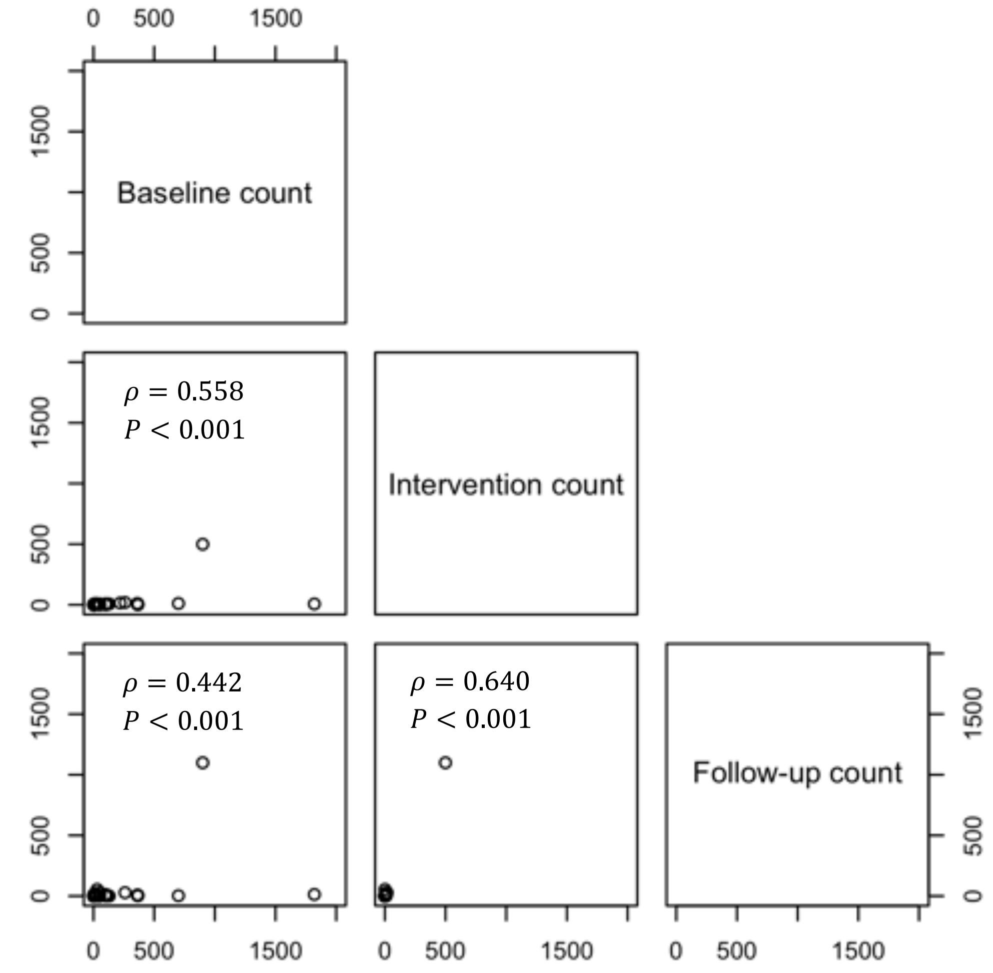


Figure 3-6 EXSart dataset: scatter matrix plots of the baseline, intervention, and follow-up falls count The Spearman's correlation coefficient ρ and P value is shown in each subplot.

As shown in Table 3-12, outlying large counts were reported during the baseline, intervention, and follow-up periods, with the maximums of 1820, 499, and 1099 respectively. A difference to the other two datasets is that there were no zero or single falls during the baseline period, because only PwP who had fallen twice or more were recruited to the trial. Therefore, the distribution of the baseline count is left-truncated.

Table 3-12 EXSart dataset: summary of falls counts during the baseline, intervention, and follow-up periods

Period	Group	N	Missing	Mean	Median	SD (variance)	Range
Baseline period (12 months)	Intervention	70	0	49.90	6.0	225.30 (50761.51)	2-1820
	Control	72	0	61.21	5.0	154.17 (23768.17)	2-900
	Total	142	0	55.63	5.5	191.94 (36841.30)	2-1820
Intervention period (first 8 weeks)	Intervention	65	5	1.83	1.0	3.18 (10.11)	0-19
	Control	64	8	10.12	1.0	62.16 (3863.76)	1-499
	Total	129	13	5.93	1.0	44 (1969.38)	0-499
Follow-up period (8 weeks to 6 months)	Intervention	64	6	3.14	1.0	5.39 (29.01)	0-29
	Control	63	9	21.33	1.0	138.19 (19095.55)	0-1099
	Total	127	15	12.17	1.0	97.44 (9494.12)	0-1099

Similar to section 3.2.1 (the Goodwin et al. dataset), the *Poi-basic* and *NB-basic* models are fitted to the falls count during the intervention (first 8 weeks), and follow-up (between 8 weeks and 6 months) periods. Also, *Poi-full* and *NB-full* are fitted, and they include the group allocation and the following baseline characteristics as covariates: gender, age, years since diagnosed with Parkinson's, Hoehn and Yahr stage, UPDRS rating, and living status (the baseline characteristics are summarised in Table 3-13).

Table 3-13 EXSart trial: characteristics of the participants at baseline

	Intervention group (n=70)	**Control group** (n=72)	**Total** (n=142)
Sex			
Male	38 (54%)	48 (67%)	86 (61%)
Female	32 (46%)	24 (33%)	56 (39%)
Age			
Mean (SD)	72.7 (9.6)	71.6 (8.8)	72.2 (9.2)
Range	44-91	52-90	44-91
Years since diagnosis			
Mean (SD)	7.7 (5.8)	9.0 (5.8)	8.3 (5.8)
Range	1-31	1-30	1-31
Hoehn and Yahr			
Stage 2	8 (11%)	8 (11%)	16 (11%)
Stage 3	44 (63%)	48 (67%)	92 (65%)
Stage 4	18 (18%)	16 (22%)	34 (24%)
UPDRS			
Mean (SD)	19.77 (8.82)	22.17 (11.90)	20.98 (10.32)
Range	3-41	4-74	3-74
No. of missing	1	2	3
Living status			
Alone	18 (26%)	16 (22%)	34 (24%)
With partner	43 (61%)	52 (72%)	95 (67%)
With family/friends / other	9 (13%)	4 (6%)	13 (9%)

Table 3-14 show the most frequently falling participants during the baseline, intervention, and follow-up periods. An interesting finding is that some participants who had recalled a large baseline falls count only reported a few falls during the intervention and follow-up periods. A typical example is ID 71, who recalled 1820 falls during the baseline period and only 7 and 13 falls during the intervention and follow-up periods.

Table 3-14 Ten largest falls rates during the baseline, intervention, and follow-up periods of the EXSart trial.

Period	ID	Group	Baseline count	Intervention count	Follow-up count
Baseline	71	Intervention	1820	7	13
	28	Control	900	499	1099
	23	Control	700	11	3
	99	Control	366	1	3
	68	Control	365	9	7
	106	Intervention	365	1	2
	113	Intervention	365	2	8
	126	Control	365	6	7
	30	Control	360	11	8
	109	Intervention	260	19	28
Intervention	28	Control	900	499	1099
	109	Intervention	260	19	28
	114	Control	220	15	Na
	23	Control	700	11	3
	30	Control	360	11	8
	84	Intervention	20	11	8
	63	Intervention	100	9	0
	68	Control	365	9	7
	118	Intervention	52	8	29
	69	Control	120	7	5
Follow-up	28	Control	900	499	1099
	48	Control	30	1	55
	118	Intervention	52	8	29
	109	Intervention	260	19	28
	1	Control	15	6	25
	102	Control	100	5	14
	71	Intervention	1820	7	13
	100	Control	3	6	13
	86	Control	8	4	11
	10	Control	6	1	9

For each period, the rows are sorted by the grey column

Table 3-15 Baseline characteristics of the frequent fallers listed in Table 3-14.

ID	Group	Sex	Age	Years since diagnosis	Hoehn & Yahr	UPDRS	Living status
1	Control	Male	90	15	Stage 4	31	With partner
10	Control	Male	80	17	Stage 3	22	With partner
23	Control	Female	78	30	Stage 4	46	With partner
28	Control	Male	60	11	Stage 4	-	With partner
30	Control	Male	53	20	Stage 4	74	With partner
48	Control	Male	56	12	Stage 3	6	With partner
63	Intervention	Female	69	5	Stage 3	16	With partner
68	Control	Male	67	5	Stage 4	30	With partner
69	Control	Female	86	6	Stage 4	26	With partner
71	Intervention	Male	67	16	Stage 4	23	With partner
84	Intervention	Male	59	10	Stage 3	16	With partner
86	Control	Female	75	8	Stage 3	27	With partner
99	Control	Male	76	5	Stage 3	12	With partner
100	Control	Male	71	10	Stage 3	16	With partner
102	Control	Male	84	4	Stage 4	36	With partner
106	Intervention	Male	74	5	Stage 4	27	With partner
109	Intervention	Male	76	31	Stage 4	18	With partner
113	Intervention	Female	76	7	Stage 4	41	Alone
114	Control	Female	60	15	Stage 3	22	Alone
118	Intervention	Female	57	10	Stage 4	30	With partner
126	Control	Female	63	1	Stage 4	33	With partner

The baseline characteristics of the participants listed in Table 3-14 are shown in Table 3-15, which indicates that the frequent falling participants were all rated Stage 3-4 on the Hoehn and Yahr scale (more severe Parkinson's), and no other pattern can be seen.

First, we compare the models on the falls count during the first 6 weeks, during which the participants received the home-based exercise and strategy programme.

Intervention period

Table 3-16 and Table 3-17 present the *basic* and *full* models, respectively. Note that the number of participants included in the *full* models is smaller than the number in the *basic* models, because three participants have missing values for the covariate UPDRS. Both the *Poi-basic* and *Poi-full* models are significantly (P<0.001) overdispersed. The dispersion statistic in *Poi-full* model (3.100) is much smaller than that in *Poi-basic* (660.328), which bears out that incorporating baseline characteristics can be effective in adjusting for heterogeneity when modelling falls counts. Besides, there is a much larger difference in AIC between *Poi-basic* and *NB-basic*, than between *Poi-full* and *NB-full*.

Table 3-16 EXSart dataset: *Poi-basic* and *NB-basic* models fitted to the intervention count (n=129)

	Poi-basic				***NB-basic***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-1.710	0.100	0.181 (0.148, 0.220)	< 0.001	-1.710	0.348	0.181 (0.091, 0.360)	< 0.001
HP					3.753			
Dispersion.	660.3				1.4			
AIC	7482.9				204.2			
Overdisperison	P < 0.001[a]				P < 0.001[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

The FRR of the intervention estimated from *NB-basic* is 0.181 (P<0.001), which indicates a more than eighty percent reduction in the falls rate for people who received the intervention. The estimated effect has a marked difference to that from the *NB-full* model (FRR=0.763, P=0.253). The reason for this great discrepancy is that one participant (ID 28) in the control group reported an outlying large falls count during the intervention period (see Table 3-14). Because the UPDRS rating of this participant was missing, this outlier is not included in *NB-full*. The huge impact of one subject on the model estimation highlights the danger of outliers.

The NB overdispersion score test indicates significant NB overdispersion in *NB-basic* (P<0.001) but not in *NB-full* (P=0.396). The dispersion statistic of *NB-full* (1.1) is closer to one than that of *NB-basic* (1.4), and it also has a smaller HP (1.002) compared to *NB-basic* (3.753). The *NB-full* model accommodates overdispersion better than *NB-basic*, not only because heterogeneity is to some extent controlled by the baseline characteristics, but also because ID 28 is not included in *NB-full*.

The number of years since Parkinson's diagnosis and Hoehn and Yahr scale are significant in the *NB-full* model (P=0.025 and 0.019 respectively). Participants who were rated Hoehn and Yahr Stage 4 had significantly higher falls rate (FRR: 2.040, 95% CI: 1.103 to 3.772) than those rated Stage 3. The trend that people with more severe Hoehn and Yahr rating have higher risk of falling is in line with the finding in the Goodwin et al. trial (see Table 3-6)

Table 3-17 EXSart dataset: *Poi-full* and *NB-full* models fitted to the intervention count (n=126)

	Poi-full				***NB-full***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-0.180	0.130	0.835 (0.645, 1.081)	0.166	-0.270	0.234	0.763 (0.480, 1.213)	0.253
Female	-0.022	0.133	0.979 (0.752, 1.273)	0.870	-0.098	0.246	0.907 (0.557, 1.475)	0.694
Age	-0.015	0.008	0.985 (0.970, 1.000)	0.049	-0.011	0.014	0.989 (0.962, 1.018)	0.442
Years since diagnosis	0.051	0.009	1.052 (1.034, 1.071)	< 0.001	0.041	0.020	1.042 (1.001, 1.084)	0.025
Hoehn & Yahr				< 0.001				0.019
Stage 2	-0.847	0.325	0.429 (0.225, 0.816)		-0.829	0.453	0.437 (0.178, 1.071)	
Stage 3			1				1	
Stage 4	0.698	0.160	2.010 (1.464, 2.759)		0.713	0.310	2.040 (1.103, 3.772)	
UPDRS	-0.002	0.006	0.998 (0.986, 1.010)	0.752	0.000	0.013	1.000 (0.975, 1.027)	0.973
Living status				0.015				0.274
With partner			1				1	
Alone	-0.323	0.188	0.724 (0.499, 1.050)		-0.396	0.306	0.673 (0.367, 1.234)	
With family/friends /others	-0.817	0.366	0.442 (0.214, 0.913)		-0.635	0.535	0.530 (0.184, 1.529)	
HP					1.002			
Dispersion	3.1				1.1			
AIC	570.4				478.7			
Overdispersion test	P < 0.001[a]				P = 0.396[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

Follow-up period

The *basic* and *full* models are presented in Table 3-18 and Table 3-19, respectively. The Poisson models are significantly overdispersed (P<0.001), and again with much larger dispersion statistics and AIC than the NB models.

The FRR from the *NB-basic* model is 0.147, close to the FRR from *NB-basic* during the intervention period. The model estimation is again influenced by the outlier ID 28, who reported 1099 falls between 6 weeks and 8 months. The *NB-full* model, which does not include this subject, yields an FRR of 0.724 (P=0.189), which is also close to that from *NB-full* fitted to the intervention count.

Table 3-18 EXSart dataset: *Poi-basic* and *NB-basic* models fitted to the follow-up count (n=127)

	Poi-basic				***NB-basic***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-1.916	0.076	0.147 (0.127, 0.171)	< 0.001	-1.916	0.357	0.147 (0.073, 0.298)	< 0.001
HP					3.865			
Dispersion.	448.6				5.7			
AIC	8975.6				674.1			
Overdispersion test	P < 0.001[a]				P < 0.001[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

The NB overdispersion score test suggests that the *NB-basic* model is significantly overdispersed (P<0.001) but *NB-full* is not (P=0.296).

The Hoehn and Yahr scale is the only significant covariate in the *NB-full* model. During the follow-up period, the falls rate was higher for people with a more severe Hoehn and Yahr rating at the baseline — the same pattern also shows in the intervention period. Compared with the participants with a Hoehn and Yahr rating of Stage 3, the falls rate for those with a rating of Stage 2 was eighty-five percent lower (FRR: 0.149, CI: 0.052 to 0.426), whilst for the participants with a rating of Stage 4 the rate was more than three times higher (FRR: 3.049, 95% CI: 1.618 to 5.748).

Table 3-19 EXSart dataset: *Poi-full* and *NB-full* models fitted to the follow-up count (n=124)

	Poi-full				***NB-full***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-0.305	0.102	0.737 (0.603, 0.901)	0.003	-0.323	0.241	0.724 (0.450, 1.167)	0.189
Female	-0.250	0.108	0.779 (0.629, 0.965)	0.019	-0.330	0.251	0.719 (0.437, 1.182)	0.200
Age	-0.032	0.006	0.968 (0.956, 0.980)	< 0.001	-0.016	0.015	0.984 (0.955, 1.014)	0.234
Years since diagnosis	0.045	0.007	1.046 (1.032, 1.061)	< 0.001	0.035	0.021	1.036 (0.994, 1.079)	0.087
Hoehn & Yahr				< 0.001				< 0.001
Stage 2	-2.155	0.388	0.116 (0.054, 0.250)		-1.901	0.529	0.149 (0.052, 0.426)	
Stage 3			1				1	
Stage 4	1.044	0.125	2.842 (2.218, 3.640)		1.115	0.320	3.049 (1.618, 5.748)	
UPDRS	-0.031	0.005	0.970 (0.960, 0.980)	< 0.001	-0.017	0.014	0.983 (0.956, 1.010)	0.199
Living status				< 0.001				0.325
With partner								
Alone	-0.461	0.157	0.630 (0.462, 0.860)		-0.421	0.312	0.656 (0.354, 1.217)	
With family/friends /others	-0.728	0.298	0.483 (0.268, 0.871)		-0.522	0.554	0.594 (0.198, 1.778)	
HP					1.219			
Dispersion	6.0				1.1			
AIC	857.2				554.5			
Overdispersion test	P < 0.001[a]				P = 0.296[b]			

[a] Boundary LR overdispersion test; [b] NB overdispersion score test.

3.3 Summary

In this chapter, the three datasets made available to this project were explored and described regarding the distribution of falls counts and other variables.

The Poisson and NB models were compared using the three datasets, with two scenarios considered to resemble the procedure of analysis in practice: the *basic* models that compare the falls rates between the two trial arms without controlling for other variables, and the *full* models that in addition include the baseline characteristics as covariates. The baseline falls count, which is the main topic of Chapter 6 and Chapter 7, has not been considered as a covariate for the NB models in this chapter.

The three trials share a similar trial design (after the transformation of the Martin et al. dataset) — falls are collected during a pre-randomisation baseline period and one or more follow-up periods after the randomisation. In general, the falls counts have heavily skewed distributions. Typically, a small proportion of subjects recorded considerably more falls than the others, and they are influential in model estimation.

All the fitted Poisson models were significantly overdispersed. For this reason, the covariates in all the fitted Poisson models were significant. The AICs of Poisson models were an order of magnitude higher than those of the NB models with the same covariates.

The dispersion statistics in *Poi-full* models were smaller than the *Poi-basic* models, and the AIC disparities between *Poi-full* and *NB-full* were not as large as between *Poi-basic* and *NB-basic*. This suggests that incorporating baseline characteristics in models explains the heterogeneity to some extent and reduces overdispersion.

The NB models accommodate the overdispersion, and thus fitted the three datasets better than the Poisson models: they not only had lower AIC values, but also resulted in smaller dispersion statistics than the Poisson models. In addition, the boundary LR overdispersion tests, which directly compares the goodness of fit of NB and Poisson models, produce small P values (<0.001).

The *NB-basic* models showed lower statistical power than the *NB-full* models, as indicated by their dispersion statistics. The Hoehn and Yahr scale was a significant covariate in the *NB-full* models fitted to the EXSart dataset and the intervention counts in the Goodwin et al. dataset, suggesting the risk of falling is associated with the severity of Parkinson's. Sex was found to be significant in the Goodwin et al. dataset, but not in the EXSart dataset. The number of years since Parkinson's diagnosis was significant only when modelling the intervention falls in the EXSart dataset.

Large outcome falls counts were found to be a crucial issue in modelling falls. They appear in all three datasets, including the Martin et al. dataset (with the smallest sample of only 21 people): one participant recorded 1599 falls within 20 weeks. How to cope with the large counts has been a great challenge in statistical modelling. The estimation of the fitted NB models is influenced by outliers, especially when the sample size is small, and when large falls counts are not balanced between two groups.

NB overdispersion, which is rarely addressed in falls prevention trials, was examined in the fitted NB models. Most of the *NB-basic* models reported in this chapter are significantly overdispersed, including covariates reduces the dispersion statistic and may control for overdispersion in *NB-full* (compare Table 3-7 and Table 3-8 for example).

In conclusion, falls data often result in Poisson overdispersion, which leads to false positives if it is not controlled in a model. Including the baseline characteristics as model covariates is effective in reducing heterogeneity. The NB models fit the falls data better than the Poisson models, but the model estimation is not robust to outliers.

Chapter 4

A review of NB functionality in five statistical packages

4.1 Introduction

Over the years, NB regression has become the standard statistical model for analysing falls counts from falls prevention trials. Because of their increasing popularity, NB models are supported in various statistical packages.

Five packages—Stata, SAS, SPSS, R, and Python—are reviewed regarding their functionality for NB modelling. Among the five packages, Stata, SAS, and SPSS are commercial software, while R and Python are open-source statistical/general-purpose programming languages that support NB modelling.

In each package, the NB functionality is generally provided in a module, which is named differently across packages. For example, a module is called a "procedure" in SAS and a "command" in Stata. Note that a module in the R language is referred to as an "R package," which may be confused with a statistical package. For clarity, the name of a module is shown in bold font. We discuss the pros and cons of the five statistical packages in general and review the NB modules within each package.

The NB2 model is considered as the standard and most commonly-used version of NB models. It is referred to as the NB model throughout the thesis, but the full name NB2 is used in this chapter to distinguish from the other types of NB models.

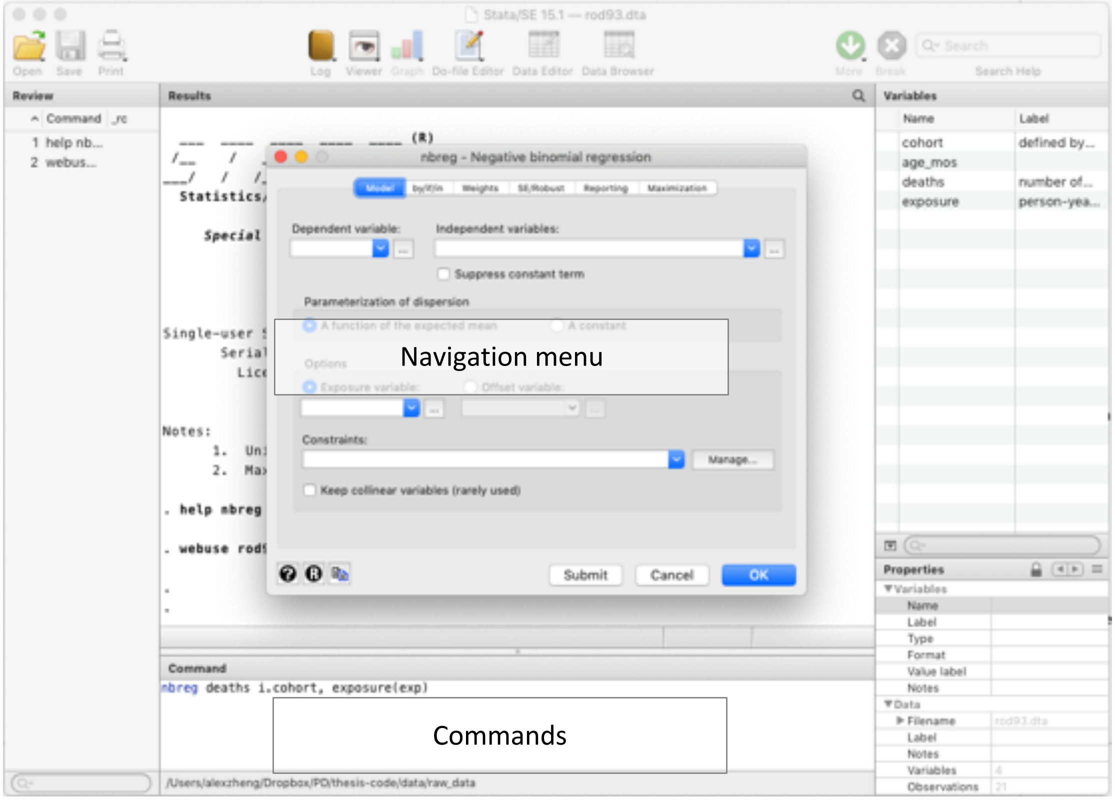
Table 4-1 summarizes the modules that provide functionality for fitting NB2 and other types of NB models in each of the five packages. The NB2 model can be fitted using all the five packages, but the options and post-estimation commands from each module provide different model-based statistics. These are also reviewed in this chapter.

Table 4-1 Functionality of NB models in Stata, SAS, SPSS, R, and Python

Stats. packages	**Version**	**NB2 Modules**	**Other NB models and modules**
Stata	v. 15	**nbreg** **glm** **countfit**	▪ NB1: **nbreg** ▪ ZINB: **zinb, countfit** ▪ NB-H: **gnbreg** ▪ NB GEE: **xtnbreg**, **xtgee**, **qic** ▪ Random-/fixed-effects NB: **xtnbreg**
SAS	9.4 x64	**GENMOD**	▪ ZINB: **GENMOD** ▪ NB GEE: **GENMOD**
SPSS	v.22	**GENLIN**	▪ NB GEE: **GENLIN**
R	3.5.0	**MASS** **msme** **COUNT** **aod** **mgcv** **gamlss**	▪ NB1: **gamlss**, **COUNT** ▪ ZINB: **gamlss**, **pscl** ▪ NB-H: **msme**, **aod** ▪ Zero-hurdle NB: **gamlss**, **pscl** ▪ Censored/truncated NB: **gamlss.cens**, **gamlss.tr**
Python	3.6.4	**statsmodels**	▪ NB GEE: **statsmodels**

4.2 NB modelling in **Stata**

Stata is a commercial statistical package that is widely used in medical research, public health in particular (Dembe et al., 2011). Stata not only supports a wide range of statistical models, but also provides a navigation menu making a user-friendly software interface (see Figure 4-1). In addition, Statas supports *community-contributed commands* so that professional users can implement functions that are missing in the original package, and upload the commands to the *Stata Journal*, the *Statistical Software Component* (Boston College Department of Economics, n.d.) archive, or other websites to make them easily accessible by other users (StataCorp, 2015).



```
nbreg deaths i.cohort, exposure(exp)
```

Navigation menu

Commands

Figure 4-1 User interface of `Stata/SE 15.1` on macOS 10.14

Two commonly used `Stata` modules for fitting NB2 models are the **nbreg** and **glm** commands. They produce almost identical regression coefficients but different diagnostic statistics. `Stata` also provides commands to fit other types of NB models. These commands are introduced and compared in this section.

The **nbreg** and **glm** commands

The most straightforward way to fit an NB2 model is using the **nbreg** command (which stands for NB regression). An alternative to **nbreg** is to use the GLM command (**glm**) and specify the distribution family as negative binomial, that is, `family(nbinomial ml)` in the `Stata` syntax.

In practice, these two commands result in close but not identical estimates. A more substantial difference lies in the range of post-estimation statistics. For example, only the **glm** command provides the Pearson dispersion statistic, which is useful to assess whether there is model overdispersion (see section 2.3.5). Another difference is that the **glm** command reports α in the form of a variance function, while it is listed along with other estimates in the output of **nbreg**, with SEs and CIs reported. There are also minor differences in the syntax between the two commands. The option for reporting the IRR for each covariate is `irr` in **nbreg** and `eform` in **glm**.

A dilemma is that only **nbreg** provides the boundary likelihood ratio test to test overdispersion while only **glm** produces diagnostic statistics for examining the influence of outliers, which is especially needed in the case of falls prevention trials as subjects reporting large falls counts are suspected to be overly influential in model estimation, so both need to be run.

The **zinb** command

The **zinb** command fits an ZINB model with the inflation covariates specified using the option `inflate()`, and the **zip** command for ZIP models uses the same syntax.

In the previous versions of `Stata`, the **zinb** and **zip** commands reported the result of the Vuong test for zero-inflation when the option `vuong` is specified, but this option has been removed from `Stata 15`. If the option is specified in `Stata 15`, a warning message is shown to remind users that the standard Vuong test may result in biased result for testing zero-inflation. Desmarais (2013) provided two community-contributed commands **zinbcv** and **zipcv** to support AIC- and BIC-based corrections of the Vuong tests for zero-inflation in ZINB and ZIP models (see section 2.3.3).

The **xtnbreg** and **xtgee** commands

The **xtnbreg** command supports conditional fixed-effects NB models in the `fe` option, random-effects NB models in the `re` option, and NB GEE models in the `pa` option.

The **xtgee** command fits an NB GEE model when `family(nbinomial)` is specified. The syntax of **xtgee** is similar to **xtnbreg** but with a few minor differences:

- Before running the **xtnbreg** command, the dataset must be specified as a panel data using the **xtset** command; in **xtgee** the subject and time variables of a longitudinal dataset can be specified in the options `i()` and `t()`, respectively.
- **xtnbreg** uses the `irr` option (same as **nbreg**) to report IRRs from the fitted GEE models, while **xtgee** uses the `eform` option (same as **glm**).

`Stata` does not support estimating HPs from GEE models. Instead, the common approach is to 1) fit a standard NB2 model using **nbreg** or **glm** and store the estimate of HP, and 2) fit a GEE model with HP fixed to the stored value. In **xtgee** this is done by specifying HP in the `family(nbinomial[HP])` option. For example, if the estimate of HP from **nbreg** is 1.5, the option should be specified as `family(nbinomial 1.5)` in **xtgee**. Note that `xtnbreg, pa` estimates the model with HP fixed to 1 and no option is available to change this value, thus **xtgee** is preferable than **xtnbreg** for fitting NB GEE models.

The **qic** command

When GEE models are used in practice, it is often desired to examine the goodness of fit statistics to facilitate decision making in choosing correlation structures, but these statistics are not reported in either **xtnbreg** or **xtgee**.

The **qic** command (Cui, 2007) is a community-contributed command that fits GEE models using the same syntax as **xtgee** and reports the QIC and CIC statistics (see section 2.3.8). The **qic** command does not support specifying a covariate as a factor—users have to create dummy variables manually.

The **countfit** command

The **countfit** command is included in a community-contributed commands combo **spost13_ado** (Long and Freese, 2006). It compares four different count models (Poisson, NB2, ZIP, and ZINB), showing the corresponding regression coefficients, BIC/AIC, and the difference in the predicted and observed probabilities. This command is very useful in model selection.

Others

Some other varieties of NB models described in section 2.3.3 are also supported in `Stata`. NB1 models can be fitted by specifying the option `dispersion(constant)` in the **nbreg** command; the default option for `dispersion()` is `mean`, which is the NB2 model. NB-H models can be fitted using the **gnbreg** command, where the predictor for estimating α is specified in the `lnalpha()` option.

4.3 NB modelling in SAS

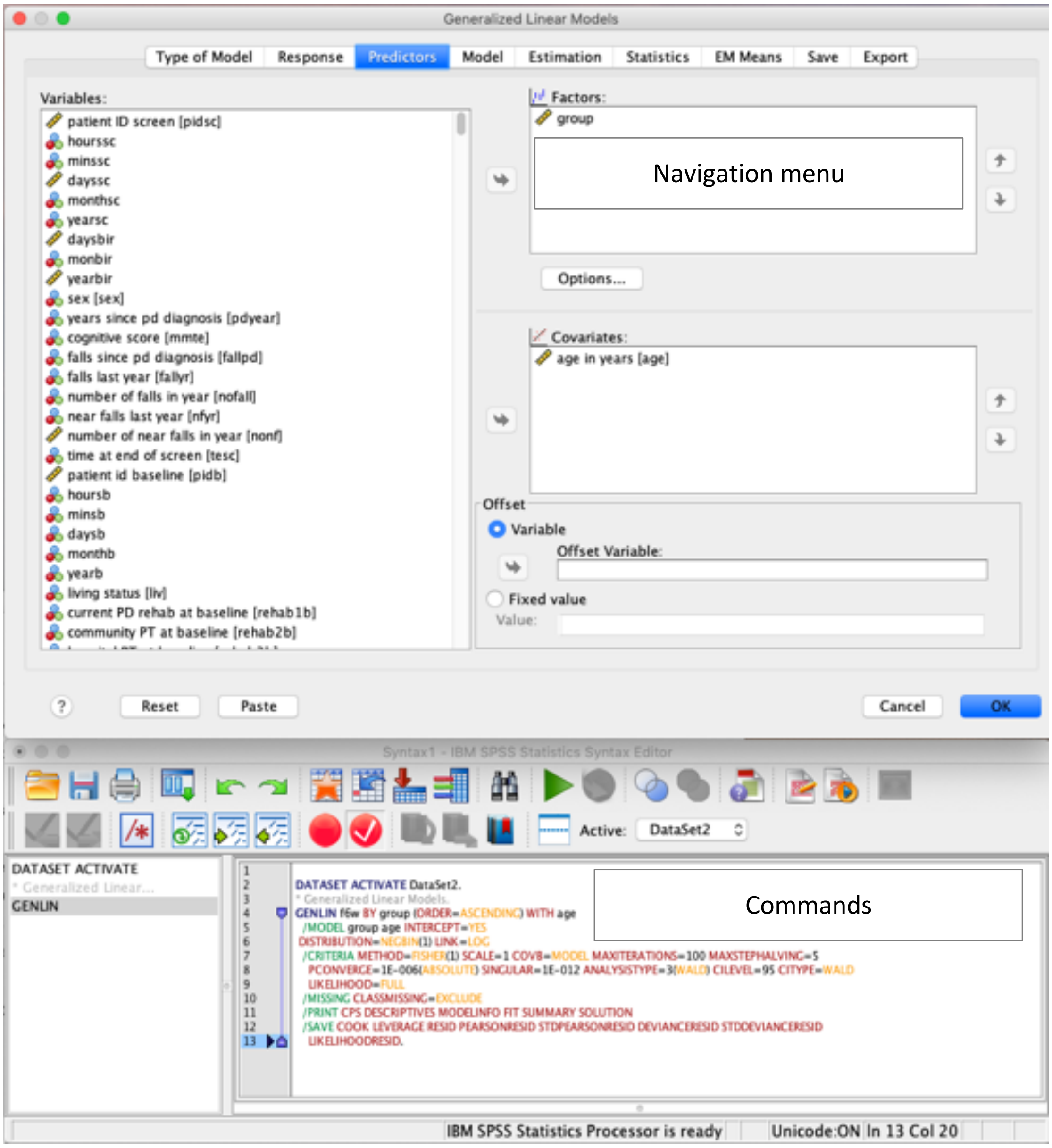
SAS (SAS Institute Inc., 2013) supports NB2 and ZINB in its generalized linear modelling procedure (**GENMOD**) by specifying DIST=NEGBIN and DIST=ZINB, respectively. The **GENMOD** procedure produces the following diagnostic statistics: Cook's distance, leverage, and DFBETA. In addition, the **GENMOD** procedure has an option PLOTS=ALL for producing various residual and diagnostic plots.

It is straightforward to fit an NB GEE model in SAS—by specifying the option REPEATED in the **GENMOD** procedure. The subject ID and the correlation structure are specified using the subject and TYPE options, respectively. Unlike Stata, fitting NB GEE models in SAS does not require providing a value for HP.

4.4 NB modelling in SPSS

Similar to Stata, statistical analysis can be conducted in SPSS (IBM Corp., 2017) using the syntax, or the menus and graphical user interface (see Figure 4-2).

NB2 modelling is supported in the GLM command (**GENLIN**) by specifying the options DISTRIBUTION=NEGBIN(1) and LINK=LOG. If the /REPEATED option is specified, the **GENLIN** command fits an NB GEE model.



```
DATASET ACTIVATE DataSet2.
* Generalized Linear Models.
GENLIN f6w BY group (ORDER=ASCENDING) WITH age
  /MODEL group age INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
   PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
   LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
  /SAVE COOK LEVERAGE RESID PEARSONRESID STDPEARSONRESID DEVIANCERESID STDDEVIANCERESID
   LIKELIHOODRESID.
```

Figure 4-2 User interface of SPSS 24 on macOS 10.14

4.5 NB modelling in R

R (R Core Team, 2016) is an open-source programming language for statistical analysis. The R language is free, powerful, and widely used in academia. It is distributed as a set of R core packages, which provide support for standard statistical analyses, numerical computations, and constitute the foundation of the R language. The R core packages were written and maintained by the R Foundation (R-Foundation, 2015). In addition, statistical models and methods are implemented as user-programmed packages, which are written and maintained by R users and made accessible to others.

For commercial packages such as Stata, SAS, and SPSS, the company that owns the intellectual property of a package is responsible for validating and maintaining the package to make sure it conducts the analyses correctly (not for user-submitted commands), and it also provides technical support to consumers who purchased a license. The source code is compiled and thus the models are essentially fitted in verified "black boxes". By contrary, R as well as the other open-source languages such as Python, distribute their source code freely so any programming bug or error could be found by users. The online community for R (such as the Stack Overflow website and the R-help mailing list) is blooming and friendly. When facing difficulties in using R, users can raise questions on the online forums and usually get prompt replies.

Another advantage of the R language is that making a package is easy and straightforward. When an author publishes a paper to propose a new statistical method or model, it has become standard to also publish an R package, so that readers can try it on their own datasets. This approach has greatly accelerated the promotion of new statistical methods. Although it is also common to publish papers with community-contributed commands in Stata, the R language is more popular due to its modern language features such as object-oriented programming and functional programming. While Stata is a powerful scriptable software, it is not a programming language and its customisability is limited compared to R. As a result, algorithms to fit cutting-edge models are mostly produced in R exclusively.

In the following sections, the available R packages that can be used to fit NB models are introduced.

The **MASS** package

A popular R function for NB2 modelling is glm.nb() in the **MASS** package (Venables and Ripley, 2002). As glm.nb() was developed as an extension of the glm() function in the **stats** package (the NB2 model can be considered as a GLM model when α or θ is fixed, as discussed in section 2.3.3; **stats** is a core package), most of the post-estimation generic functions in glm can be directly applied to a glm.nb() model fit.

The glm.nb() function yields an estimate of θ in equation (2-7), while most of the other software parameterises via α in equation (2-6). As discussed in 2.3.3, the parameterisation using θ may be counter intuitive, because a smaller θ indicates greater overdispersion, and the NB2 model tends to a Poisson model as θ approaches infinity.

Note that if the **MASS** package is used for NB2 modelling, the generic function `confit()` is linked to the `confit.glm()` function in the **MASS** package, which reports the profile CI instead of the standard model-based CI (which assumes that each parameter estimator is approximately normally distributed). If standard model-based CIs are preferred to profile CIs, the `confit.lm()` function should be used.

The **msme** package

A less well-known `R` function for NB2 modelling is `nbinomial()` in the **msme** package (Hilbe and Robinson, 2014). This package was written to include the functions and datasets used in Hilbe and Robinson's (2016) book *Methods of Statistical Model Estimation*.

The output of `nbinomial()` function is generally similar to the `glm.nb()` function except for a few enhancements:

- The parameterization of overdispersion can be specified either via α (when the option `family` is specified as the default `"nb2"`) or θ (when `family` is specified as `"negBinomial"`).
- The CIs for the estimates of parameters are reported in the output of the `summary()` function.
- HP can be parameterized as a linear predictor via a log link-function, that is the NB-H model described in section 2.3.3.
- The dispersion statistic is produced.
- The `alrt()` function in the **msme** package can be used for testing Poisson overdispersion using the boundary LR test. The support of the boundary LR test is missing in the **MASS** package.

The limitations of the **msme** package is that it does not support as many diagnostic options as **MASS**. While the leverage of each subject can be calculated with the `hatvalues()` function in **msme**, the Cook's distances cannot be produced.

A useful feature of the **msme** package is that the `P__disp()` function produces the dispersion statistic for a fitted `glm()` Poisson model or a `glm.nb()` model. This is a useful tool that is missing in the **stats** and **MASS** packages.

The **COUNT** package

The **COUNT** package was written for a book, *Negative Binomial Regression* (Hilbe, 2011). This package includes the data and code used in the book, including the `ml.nb2()` and `ml.nb1()` functions, which support the NB2 and NB1 models, respectively.

The **aod** package

The **aod** package (Lesnoff et al., 2012) includes various functions for overdispersed count data or proportions, including the `negbin()` function for NB modelling.

Similar to the `nbinomial()` function in the **msme** package, `negbin()` supports both the NB2 and NB-H models: a linear predictor can be specified in the `random` option to estimate α.

The **mgcv** package

The **mgcv** package (Wood, 2017) is designed for Generalized Additive Models (GAMs). As GAMs are a generalization of GLMs, this package can also be used for fitting the NB2 model. This is done by specifying the option `family=nb()` in the `gam()` function. Similar to `glm.nb()` in **MASS**, `gam()` parameterises NB2 with θ. Although another **mgcv** function named `negbin()` also supports NB2 modelling, a value of θ has to be given as an argument.

The **gamlss** package

The **gamlss** package is an `R` package for fitting the Generalised Additive Models for Location, Scale and Shape (GAMLSS) model, which is a very flexible framework that can be used to model more than a hundred discrete, continuous and mixed distributions, and it is also a generalization of the GLM family (Stasinopoulos and Rigby, 2007). The NB2 model is supported in the **gamlss** package via the `NBI()` function, which returns a `gamlss.family` object that is passed to the `gamlss()` function as an argument (`family`) to specify the distribution of the response variable.

It should be noted that there is a confusion in the nomenclature of the distributions in the **gamlss** package: `NBI` refers to the distribution that is conventionally named NB2, and `NBII` refers to NB1.

A feature of the **gamlss** package is that it supports right-, left-, and interval-censoring/truncation for any model that is a member of the gamlss family, in the **gamlss.cens** and **gamlss.tr** packages respectively (Hilbe, 2011; Stasinopoulos et al., 2017). These two packages provide functionality for fitting right-censored or right-truncated NB models.

The **pscl** package

Although the **pscl** package (Jackman et al., 2007) does not include functions for fitting the NB2 model, it includes a number of useful tools for a `glm.nb()` fit and other types of NB models. For example, the `odTest()` function can be used to examine overdispersion using the boundary likelihood rate test. In addition, the zero-inflated NB model can be fitted using the `zeroinfl()` function.

The **pscl** package supports the Vuong test in the `vuong()` function, as well as the corrections based on AIC and BIC (see section 2.3.3).

4.6 NB modelling in **`Python`**

`Python` (Python Core Team, 2015) is a general-purpose programming language. Over the past decade, `Python` has become increasingly popular in data science and statistics. Although `Python` is not as commonly used in medical statistics as in machine learning, a `Python` module **statsmodels** (Seabold and Perktold, 2010) is available for fitting statistical models. The **statsmodels** module is based on **NumPy** arrays (Oliphant, 2006), which is numeric computing package, and **pandas** data frames (McKinney, 2010), which is a counterpart of the `R` **dataframe** in the **base** package. The **NumPy** and **pandas** modules also have good performance in terms of speed due to `Python` incorporating very fast optimisation methods.

NB2 models can be fitted using the GLM function `GLM()` in **statsmodels**, by specifying the argument `family= families.NegativeBinomial()`. NB GEE models can be fitted using the `gee()` function with the same option for the `family` argument as in `GLM()`.

4.7 Discussion

In this chapter, five statistical packages were reviewed regarding the functionality for NB modelling. All the packages support the widely-used NB2 models and some other types of NB models. Different modules also result in different model-based statistics, even for modules within the same package. There are a few practical points worth mentioning here:

`Stata` provides a very complete set of NB commands, but only a few post-estimation diagnostic statistics are supported. As `Stata` is widely used in public health, learning to fit an NB model would be relatively straightforward for an applied researcher. In terms of the cost, the price of `Stata` is lower than `SPSS` or `SAS`.

`SAS` is a popular statistical package for medical studies, especially in the pharmaceutical industry. Although `SAS` only supports three types of NB models (NB2, ZINB, and NB GEE) in **PROC GENMOD**, it produces a number of diagnostic statistics and plots. However, `SAS` is more expensive than the other four packages.

Among the five reviewed packages, `SPSS` is most user-friendly. Its graphical user-face is well designed and straightforward, especially for people with no prior knowledge of programming. Data input in `SPSS` is convenient, but the functionality for NB modelling is limited.

`R` is a free and powerful programming language for statistics and computation. Although rarely used in medical research a few years ago, `R` is now a fast-growing language, and is used more and more by medical researchers. `R` is distributed freely, but there is a drawback: users often have to seek help from the online community. This is particularly tricky if a package is rarely used or no longer maintained. Although the authors of `R` packages are usually responsible for maintaining the projects and answering technical questions, they are not obliged to. Therefore, the questions or bugs may not be dealt with promptly—most authors only contribute to the projects in their free time. Conducting a statistical analysis in `R` requires more advanced programming skills than the commercial packages, but it provides greater flexibility.

`Python` is a programming language commonly used in data science, and it has become increasingly popular for data analysis. Given `Python`'s rising popularity, more `Python` modules are expected to support NB models in the future.

Chapter 5

Diagnostic plots for NB modelling of falls data

5.1 Introduction

The distribution of a falls count reported by PwP is usually positively skewed, with most participants reporting relatively small counts whilst a few participants report large counts, and when a count response model is fitted, these large counts are generally highly influential in model estimation. Because NB regression can fit a more heavily skewed distribution to count data than Poisson regression, it copes with large falls counts better than Poisson models. However, if a falls count is extremely large, it may exceed the capacity of the NB model to accommodate outliers. The results presented in Chapter 3 suggest that large outcome counts greatly influence model estimation, and the estimation of the intervention effect was sensitive to outliers. Even a single large count may substantially change the estimated intervention effect.

Large counts are a great challenge in NB modelling. Although they may have a major impact on model estimation, large counts are not always influential: if the covariates have good predictive power, they may be fitted very well. Hence, pinpointing outliers, quantifying their influence on model estimation, and understanding how individual counts impact on model estimation is essential to statistical analysis of falls counts.

The diagnostic statistics introduced in section 2.3.4 are useful in model checking, as they assess the influence of each subject from a quantitative perspective, but they may not be straightforward to use in practice. Sifting through the diagnostics for each participant is time-consuming and error-prone, especially for a large dataset. A visual inspection of a diagnostic plot, in comparison, should be intuitive and easy to interpret.

In this chapter, a new diagnostic plot specifically designed for modelling falls counts from falls prevention trials is introduced, and it is produced for the NB models presented in

Chapter 3, that is, without the baseline count included. An existing diagnostic plot is examined in the context of NB modelling. These diagnostic plots are discussed regarding the traits of each model and dataset. An R package (see Appendix A) was written to automate the production of NB diagnostic plots.

This chapter provides a tool for examining NB model diagnostics graphically in the context of falls prevention trials, and other trials with recurrent events as the outcome and also collected at baseline.

5.2 Baseline/Outcome Event plot

Standard diagnostic plots for NB regression are not generally helpful in relation to analysing data from a falls prevention trial, specifically, the core requirement of assessing diagnostic statistics relating to the participants reporting high falls rates. Therefore, a diagnostic plot for falls data should present both the diagnostic statistic and the corresponding falls rate at the subject level. Ideally, the plots should not only be useful in identifying the influential subjects, but also help in inspecting patterns of diagnostic statistics in the context of their baseline/outcome fall rates. They should facilitate the examination of whether the estimated intervention effect, which is usually the main research question, is substantially influenced by a few outliers.

Four new diagnostic plots are proposed to present the following statistics: Cook's distance, leverage, Anscombe residual, and DFBETA. The collective set of plots are referred to as Baseline/Outcome Event (BOE) plots in the thesis, and each plot is referred to by the name of the model diagnostic presented.

For the sake of demonstration, two examples — a Cook's distance plot and a DFBETA plot — are shown in Figure 5-1 and Figure 5-2, respectively. Both plots are based on an NB model fitted to a simulated two-arm trial dataset (n=200). Assume the falls count is collected during a baseline and a follow-up period, both lasting for one month. Let the NB distribution be denoted by $\mathrm{NB}(\mu, \alpha)$. The baseline count in both groups follows the distribution $\mathrm{NB}(30,1)$, while the outcome count follows the distribution $\mathrm{NB}(20,1)$ in the intervention group and $\mathrm{NB}(30,1)$ in the control group, so that on average the outcome falls count is 33% lower in the intervention group. The same gamma-distributed subject effect is used in generating the outcome and baseline counts for the same subject (more details of the simulation are given in section 6.2.1). The fitted NB model includes only one covariate: the group allocation.

As shown in Figure 5-1 and Figure 5-2, the BOE plots are based on a scatter plot with the outcome rate plotted against the baseline rate, both axes on a logarithmic scale. The diagnostic statistic for each subject is indicated by the size of the plotting symbol. In order to include zero falls, 0.5 is added to both rates before log-transformation. A vertical and a horizontal dashed line are plotted at the location of $\log(0.5)$ to indicate zero counts. In order to compare the two groups, the plotting symbols from different groups are shown in different colours. Because Anscombe residuals and DFBETAS may be negative, they are plotted with triangular symbols, with upside-down triangles indicating negative values (see Figure 5-2).

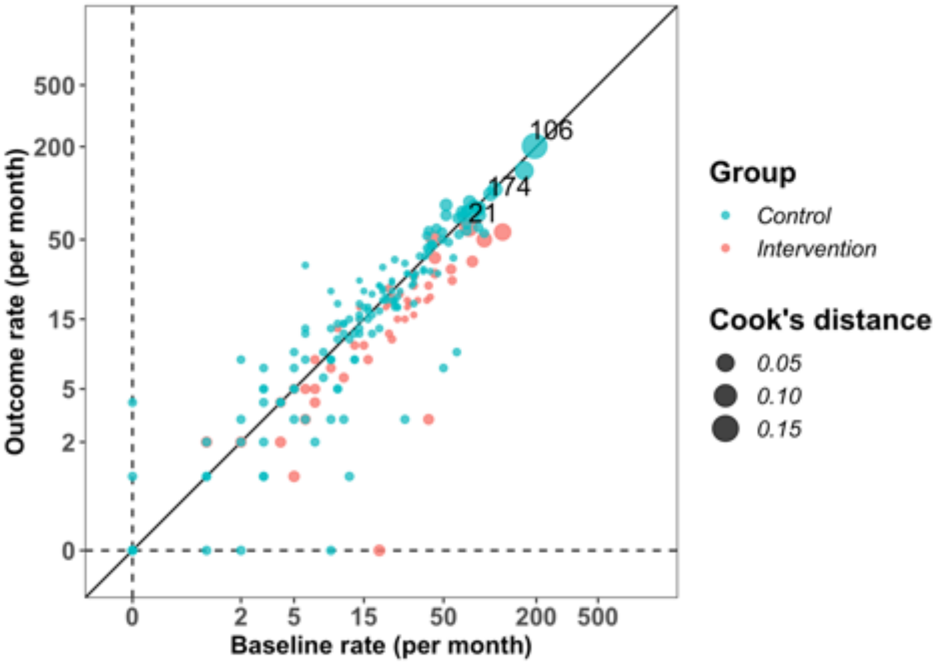


Figure 5-1 Demonstration of the Cook’s distance plot (n=200). The IDs of the subjects with the largest Cook’s distance are shown as labels.

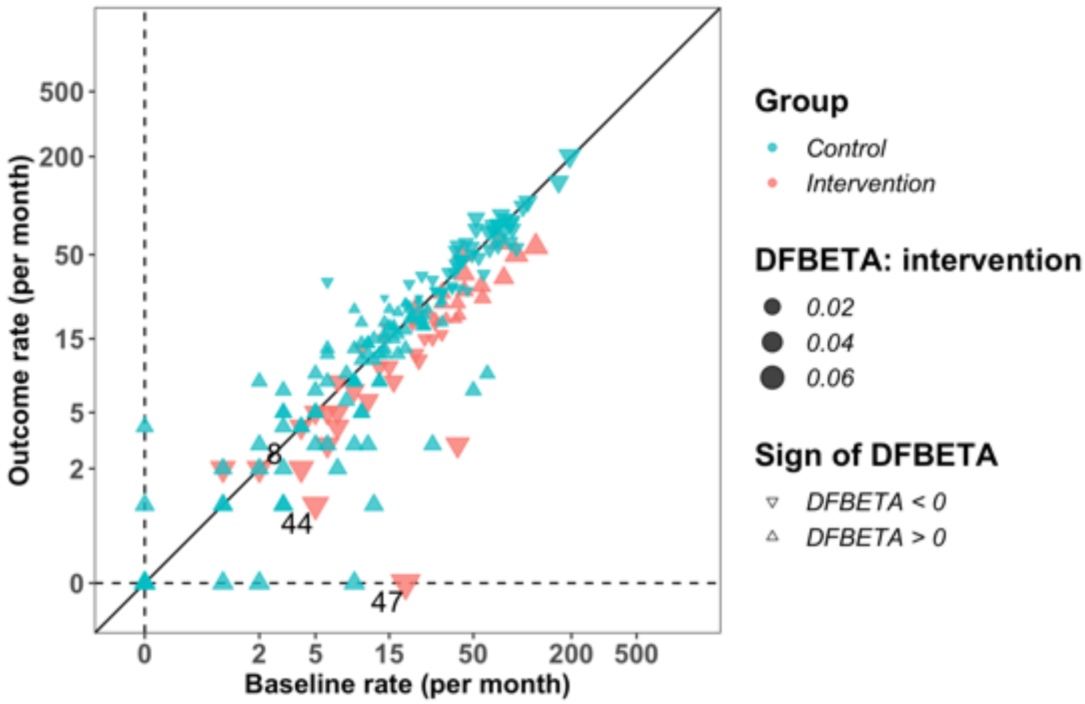


Figure 5-2 Demonstration of the DFBETA plot (n=200). The IDs of the subjects with the largest absolute values in DFBETA are shown as labels.

The reason that the rates are plotted on a logarithmic scale is to cope with the skewed distribution of the falls rate. The x- and y-axes are labelled with the untransformed falls rates to improve readability (example Figure 5-1 and Figure 5-2). Compared to the original scale, the plotting symbols under a logarithmic scale are more evenly distributed, so that it is easier to compare different diagnostic statistics for the same subject across plots based on the location of the plotting symbol on different plots. For example, one may be interested in examining the Cook's distance and residual for the participant with the highest falls rate during the outcome period. It is straightforward to compare the two statistics in the corresponding BOE plots, because the plotting symbols lie at the top of the body of points.

To provide a reference line for comparing the outcome and baseline rate, the BOE plots include a Line of Falls Equity (LoFE), defined as a line with the slope 1 and intercept 0. The x- and y-axes have the same range so that the LoFE is the diagonal of the plot. If an outcome rate is exactly the same as the baseline rate, the plotting symbol would perfectly lie on the LoFE, which is shown as the diagonal in the plots (for example see Figure 5-1 and Figure 5-2).

If the intervention effect is effective in preventing falls, the red symbols for the intervention group are anticipated to be below the blue symbols for the control group (this pattern is clear in Figure 5-1 and Figure 5-2). This is especially useful for examining the estimated intervention effect: if an intervention has no effect in reducing falls rate, the estimated intervention effect may still be significant because of a few outliers, but this would be apparent from the BOE plot.

Another function of the LoFE is to show the period effect. If the plotting symbols in the control group are symmetric around the line (for example see Figure 5-1 and Figure 5-2), it shows that the falls rate in the control group is relatively constant across periods, suggesting at most a mild period effect. If the symbols are generally above the LoFE, it indicates that the outcome falls rate is higher than the baseline level, possibly due to the worsening of body balance or progress of Parkinson's. Although most plotting symbols from the control group may also be below the LoFE, this should be relatively rare if the outcome and baseline falls counts are obtained using the same collection method, because the falls rate is not anticipated to decrease if a participant is not given an intervention. However, the pattern is possible when the outcome count is collected prospectively but the baseline count is collected retrospectively, because trial participants may overestimate how many falls they experienced when asked to recall the number at baseline. For example, a participant who falls twice per week on average but by chance falls more frequently in the few weeks prior to the screening interview, say, 5 times per week, this participant may give an approximate baseline count by multiplying 5 by 52 (to arrive at the falls count in the previous year), so that the number would be greater than the actual falls count. Another possibility is that frequent fallers may stop recording falls because of the continued effort of recording every fall event in diaries, and if only the outcome count was collected prospectively, this would result in a lower than anticipated outcome rate. The average falls rate in the control group may also be lower during the follow-up period due to regression to the mean, especially when the eligibility criteria include a threshold for the baseline count (for example, "falling at least twice in the previous year").

The four diagnostic statistics are chosen for the following reasons. The Cook's distance is a deletion diagnostic that approximates the effect of deleting a subject on the goodness of fit of the model. The Cook's distance shows the overall influence of each subject on model estimation, but it does not indicate why a subject is influential. The leverage and Anscombe residual, however, measure respectively how extreme are the covariates from the typical values in the sample and whether the subject conforms to the fitted model. Comparing the leverage and residual for a subject shows whether a large Cook's distance is due to outlying covariates, or poor model fit for the subject. The DFBETA shows the influence of each subject on the estimate of the intervention effect, the focus of a falls prevention trial.

The four statistics, when compared with each other, provide valuable diagnostic information on the outliers. For instance, the combination of a large Cook's distance and a large residual indicates a poor agreement between an outcome count and the model fit. If this pattern is found for all frequent fallers, it suggests that the fitted model cannot accommodate the outliers. It would then be of interest to know to what extent they have altered the estimate of the intervention effect using the DFBETA plot. Another example is when the Cook's distance and leverage are large, but the residual is small. This suggests that the subject is so influential that reducing the residual of this subject becomes a priority of model estimation, and the case would not be identified in a residual analysis alone (Hilbe, 2011).

The BOE plots are supported in the R package **NBDiagnostics**. The package includes a function `nbdiagnostic()` to fit an NB model, and the fitted model is then passed into the `boeplot()` function as an argument to produce the BOE plots (for details see Appendix A).

5.3 The covariate-adjusted probability plot for NB models

Holling et al. (2016) proposed a covariate-adjusted probability plot as a diagnostic plot for a fitted count response model. Suppose there are n subjects in a sample. Let the outcome count for each subject i be y_i $(i = 1, \dots, n)$ and the vector of covariates be denoted by $\boldsymbol{x}_i$. Let f_y denote the frequency of counts in the sample $y_1, \dots, y_n$ with a range from 0 to the maximum of y_i. Assume that the variable Y_i follows a distribution $p_y(\lambda(\boldsymbol{\theta}, \boldsymbol{\eta}_i))$, where $\boldsymbol{\theta}$ is a vector of unknown parameters, $\boldsymbol{\eta}_i$ is a vector of known parameters, and $\lambda(.,.)$ is a known function that links $\boldsymbol{\theta}$ and $\boldsymbol{\eta}_i$ to p_y. In the case of NB model:

$$\begin{aligned} p_y\big(\lambda(\boldsymbol{\theta},\boldsymbol{\eta}_i)\big) &= \mathrm{NB}(y|\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}),\alpha) \\ &= \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\,\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)^{\alpha^{-1}}\left(\frac{\alpha\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\alpha\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)^{y}. \end{aligned} \quad (5\text{-}1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta},\alpha)^{\mathrm{T}}$ and $\boldsymbol{\eta}_i = \boldsymbol{x}_i$. The authors defined the covariate-adjusted probability as:

$$\hat{p}_y\big(\widehat{\boldsymbol{\theta}}_n\big) = \frac{1}{n}\sum_{i=1}^{n} p_y(\hat{\lambda}_i)\,, \quad (5\text{-}2)$$

where $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta}$ and $\hat{\lambda}_i = \lambda(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\eta}_i)$. For a fitted NB model $\widehat{\boldsymbol{\theta}}_n$=$\big(\widehat{\boldsymbol{\beta}},\hat{\alpha}\big)^{\mathrm{T}}$.

The authors proved that the covariate-adjusted probability $\hat{p}_y$ and f_y/n converges if the model is correctly specified. Therefore, the covariate-adjusted probability plot enables comparison of the estimated probability marginalising over the distribution of the covariates to the observed probability (that is, the relative frequency f_y/n).

If a Poisson and an NB model include identical covariates, their covariate-adjusted probabilities can be shown in one plot to compare their goodness of fit. This plot could be used to show 1) whether Poisson overdispersion is present, and, 2) is the NB model a good fit to the dataset?

Figure 5-3 show the covariate-adjusted probabilities of Poisson and NB models from six simulated datasets. Each set of data comprised of 500 subjects in group 1 with group mean of $\exp(1)$, and 500 subjects in group 2 with group mean of $\exp(1.2)$. When the dataset is equidispersed (Figure 5-3 a), the covariate-adjusted probabilities of the Poisson and NB models are indistinguishable from the observed probabilities. For this subplot, the Poisson model is anticipated to fit well because the dataset is simulated from a Poisson distribution, while the NB model yields a small HP and thus its estimation is close to Poisson. As datasets become more overdispersed, indicated by greater α, the Poisson models show much worse goodness of fit, while the NB models remain close to the observed probabilities.

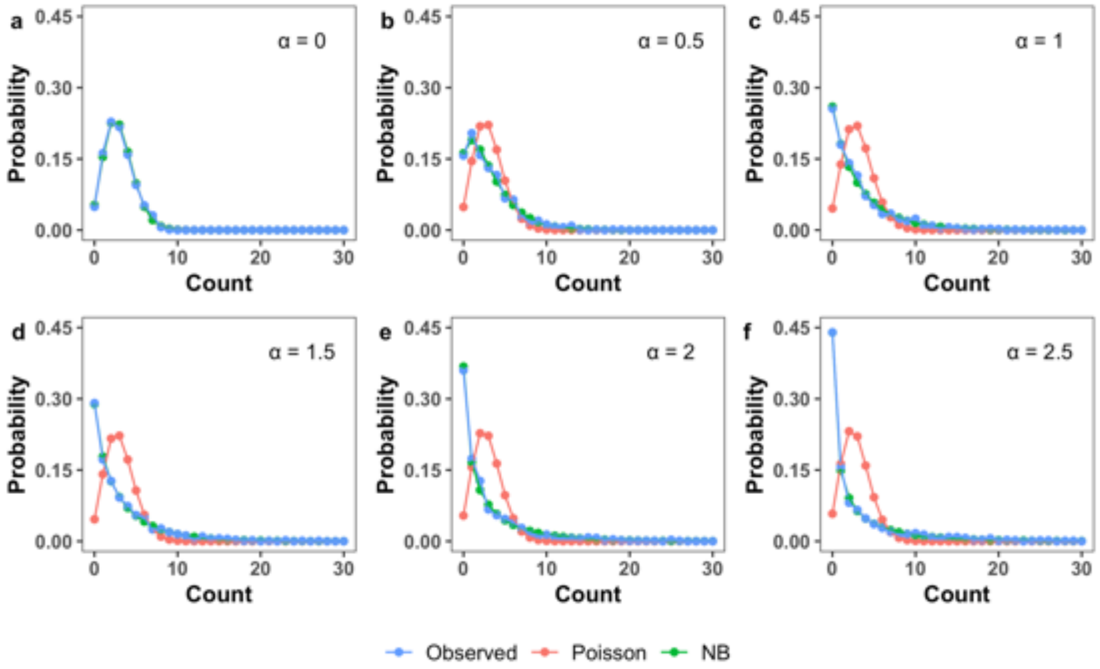


Figure 5-3 Covariate-adjusted probability of NB and Poisson models on simulated data with two groups (small difference in group means). For each subplot 500 counts (y_i) are generated from $\mathrm{NB}(\exp(0.2x_i + 1)\,, \alpha)$, where $i = 1, \ldots, 1000$; the binary group indicator $x_i = 0$ for $i \leq 500$ and $x_i = 1$ from the rest. The NB and Poisson models include the same covariate x_i.

Another example is shown in Figure 5-4. The simulation settings are same as in Figure 5-3, except for the group means, which are $\exp(1)$ in in group 1 and $\exp(2.2)$ in group 2. Because the difference between group means is bigger than in Figure 5-3, the observed probabilities show a bimodal pattern, especially for small α. Again, the covariate-adjusted probabilities from the NB models are indistinguishable from the observed probabilities. The Poisson models yield much worse goodness of fit than the NB models, except when $\alpha = 0$.

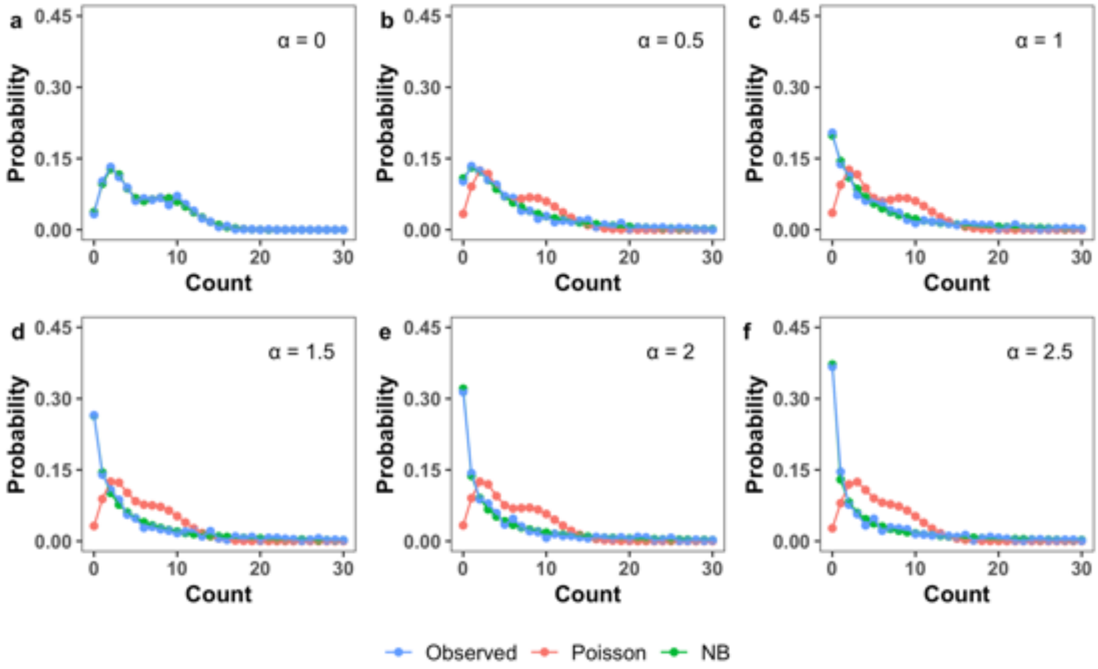


Figure 5-4 Covariate-adjusted probability of NB and Poisson models on simulated data with two groups (large difference in group means). For each subplot 1000 counts are generated from $\mathrm{NB}(\exp(1.2x_i + 1), \alpha)$, where $i = 1, \dots, 1000$; the binary group indicator $x_i = 0$ for $i \le 500$ and $x_i = 1$ from the rest. The NB and Poisson models include the same covariate x_i.

The two examples demonstrate that the covariate-adjusted probability plot is a practical tool for graphically illustrating whether NB should be used instead of Poisson regression. This plot is also produced by the R package **NBDiagnostics**.

5.4 Application of NB diagnostics plots to three falls datasets

The diagnostic plots described in this chapter are produced for the NB models included in Chapter 3, using the Goodwin et al. (2011), Martin et al. (2015), and EXSart (Ashburn et al., 2001) datasets.

The NB models included in Chapter 3 are: 1) *NB-basic*, which includes only one covariate—the group allocation; and 2) *NB-full*, which includes both the group allocation and baseline characteristics as covariates (the baseline count is not included in both models). The covariate-adjusted probabilities from *Poi-basic* and *Poi-full* are compared with those from *NB-basic* and *NB-full* in plots for visualisation of Poisson overdispersion.

5.4.1 Goodwin et al. dataset

Figure 5-5 presents the comparison of the Poisson and NB models in section 3.2.1 regarding goodness of fit, by plotting their covariate-adjusted probabilities and the observed probabilities. In each subplot, the covariate-adjusted probabilities from NB model are closer to the observed probabilities than those from Poisson model, indicating that the NB models fit the data better than the Poisson models.

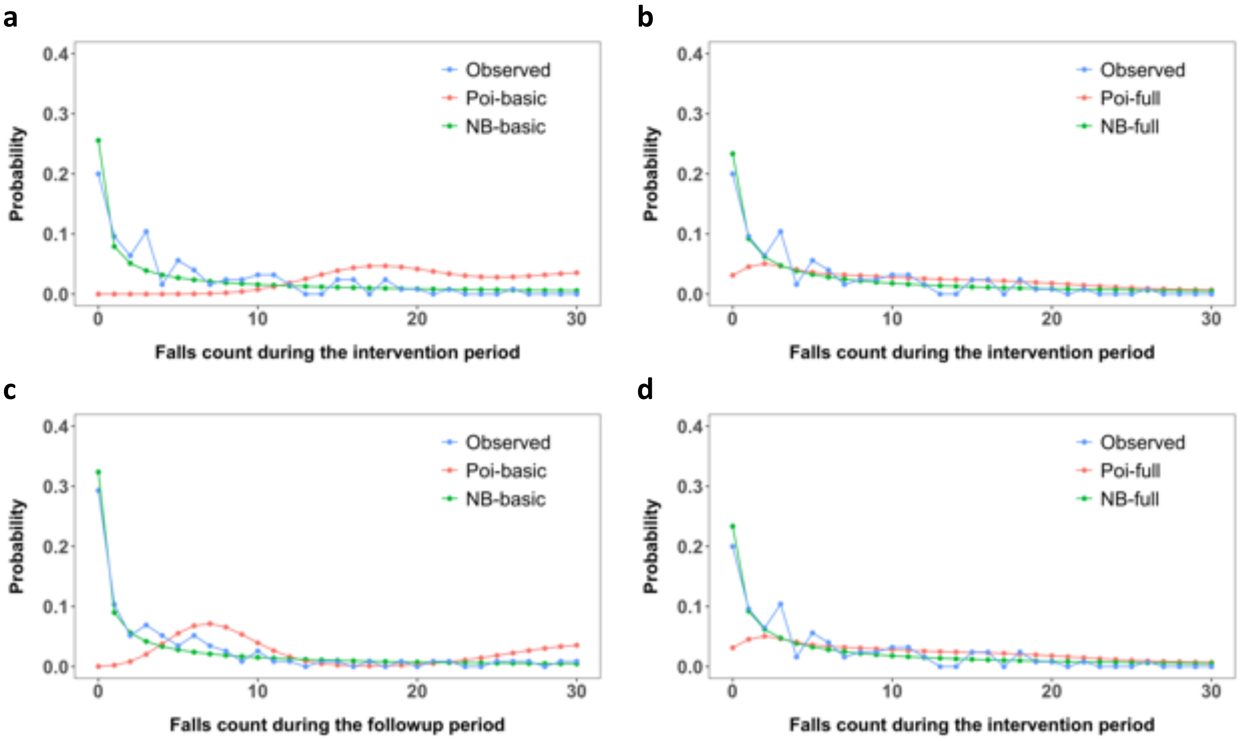


Figure 5-5 Covariate-adjusted probability plots for model comparisons in the Goodwin et al. dataset. **(a)** intervention falls count: *Poi-basic* versus *NB-basic* (n=116); **(b)** intervention falls count: *Poi-full* versus *NB-full* (n=116); **(c)** follow-up falls count: *Poi-basic* versus *NB-basic* (n=130); **(d)** follow-up falls count: *Poi-full* versus *NB-full* (n=130).

Figure 5-6 shows the Cook's distance plots of the *NB-basic* and *NB-full* models for counts from the intervention and follow-up periods. Overall, most plotting symbols are close to the LoFEs. The blue plotting symbols (control group) are symmetric with respect to the LoFE, indicating that the period effect is small, and the falls rates during both the intervention and follow-up periods are consistent with the baseline rate. There is a tendency for the red symbols to fall mostly below the LoFE, suggesting that the participants in the intervention group had lower outcome falls rates.

The plotting symbols at the top of Figure 5-6 a) and c) are bigger than those in b) and d), suggesting that the frequent fallers are more influential in the *NB-basic* models than in the *NB-full* models. Participant ID 18 (from the intervention group) reporting around 30 falls/week during both the baseline and intervention period, yields the third largest Cook's distance in the *NB-basic* model on intervention falls (Figure 5-6 a). The great influence of this subject does not persist in the *NB-full* model (Figure 5-6 b). The participant ID 75 had reported the largest falls count during all three periods. Even though the plotting symbol of ID 75 lies on the LoFE of all four subplots, the corresponding Cook's distance from *NB-basic* is large (around 0.6). In comparison, the Cook's distance of ID 75 from *NB-full* is around 0.2, though this value is still higher than the Cook's distances for most participants. The pattern shown in the Cook's distance plots implies that *NB-full* better accounts for large counts than *NB-basic*, but it is still limited when an outcome count is very large and the baseline count is not included in the model.

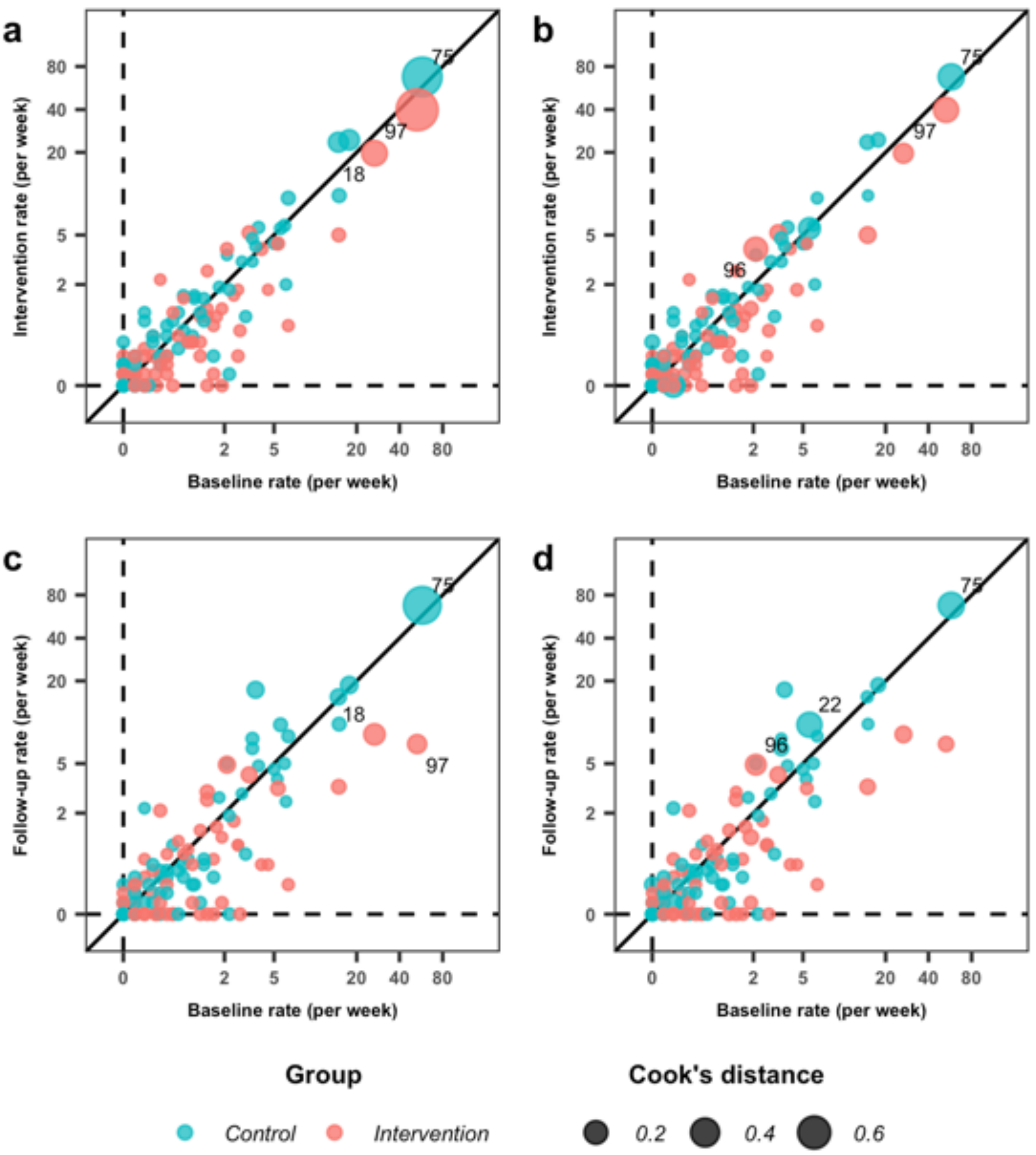


Figure 5-6 Cook's distance plots for the models fitted to the Goodwin et al. dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=116); **(b)** *NB-full* fitted to the intervention falls count (n=116); **(c)** *NB-basic* fitted to the follow-up falls count (n=130); **(d)** *NB-full* fitted to the follow-up falls count (n=130). In each subplot, the three subjects with the largest Cook's distances are labelled with their ID.

Having identified the influential subjects in each model, we examine the leverage and Anscombe residuals (shown in Figure 5-7 and Figure 5-8 respectively) of the subjects with large Cook's distances.

The leverage plots have little diagnostic value for the *NB-basic* model, because it includes only one binary covariate — group allocation. The leverage plots for the *NB-full* models does not indicate that the participants with high Cook's distance have unusual covariate values.

The Anscombe residuals show a similar pattern in all four models: the residuals are negative when the outcome rate is small and positive when the outcome rate is large, that is, large outcome counts are underestimated by the model and small counts overestimated. This indicates that the models do not sufficiently accommodate the variance of the data, and they fit poorly for the large numbers (indicated by the large sizes of the plotting symbols at the top-right corners in Figure 5-8). This result shows that the large Cook's distances for the frequently falling participants are not because of peculiar values in the covariates, but because the residuals are large.

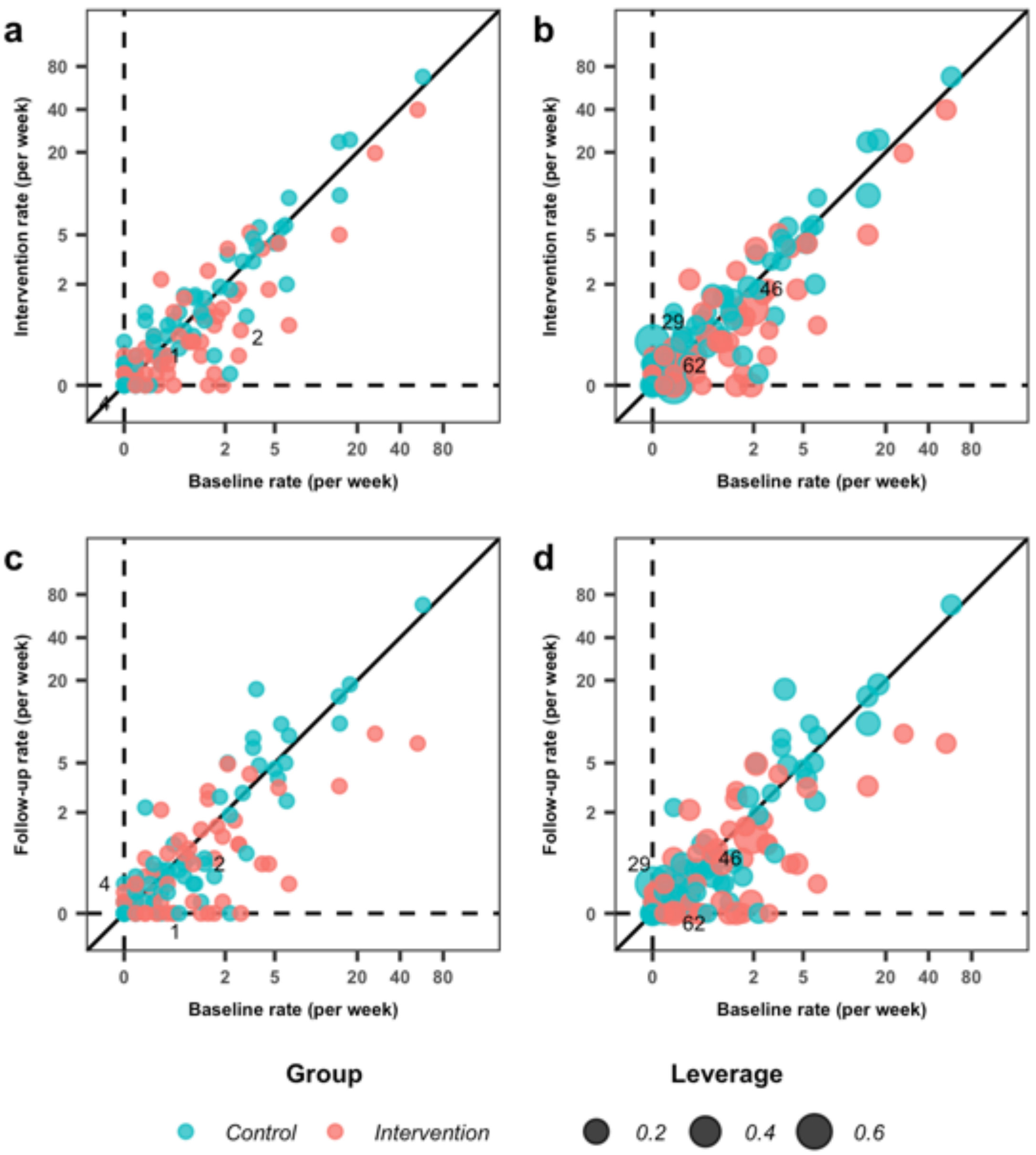


Figure 5-7 Leverage plots for the models fitted to the Goodwin et al. dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=116); **(b)** *NB-full* fitted to the intervention falls count (n=116); **(c)** *NB-basic* fitted to the follow-up falls count (n=130); **(d)** *NB-full* fitted to the follow-up falls count (n=130). In each subplot, the three subjects with the largest leverage are labelled with their ID.

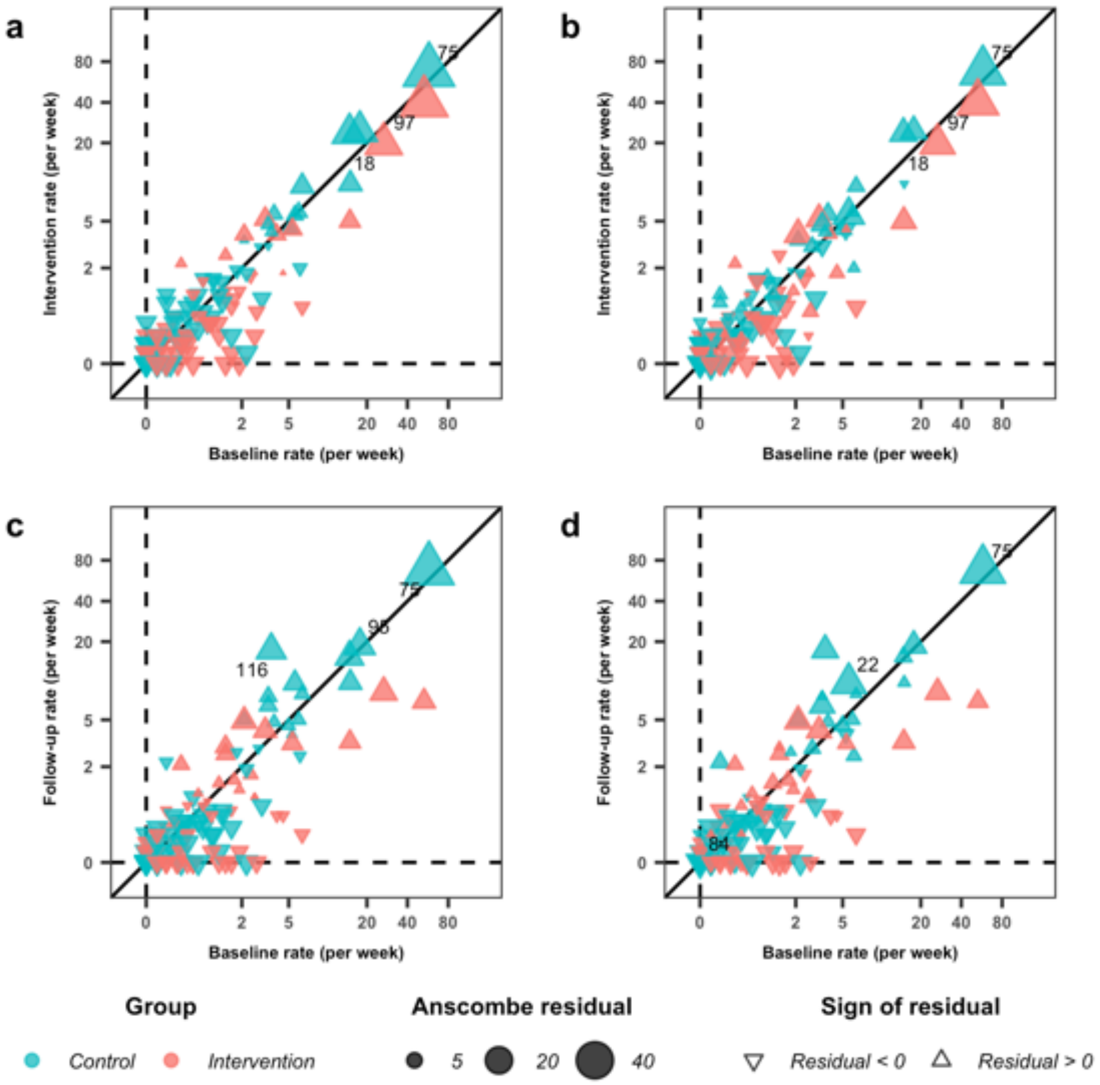


Figure 5-8 Anscombe residuals plots for the models fitted to the Goodwin et al. dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=116); **(b)** *NB-full* fitted to the intervention falls count (n=116); **(c)** *NB-basic* fitted to the follow-up falls count (n=130); **(d)** *NB-full* fitted to the follow-up falls count (n=130). In each subplot, the three subjects with the largest absolute values of Anscombe residuals are labelled with their ID.

The DFBETA of the intervention effect is shown in Figure 5-9. As discussed in section 3.2.1, the nine most frequently falling participants during the follow-up period were all in the control group. They are shown at the top-right corner of subplots c) and d), and they all have negative DFBETA, which indicate that excluding these subjects from the model would result in a larger regression coefficient, that is, the FRR for the intervention effect would be closer to 1. This is in line with the extreme intervention effects estimated from *NB-basic* (FRR: 0.287; Table 3-7) and *NB-full* (FRR: 0.361; Table 3-8). The plotting symbols of these participants are generally close to the LoFE, indicating that they had a consistent falls rate during the baseline, intervention, and follow-up periods. Thus, it is anticipated that they will be less influential if the baseline count is incorporated in the models.

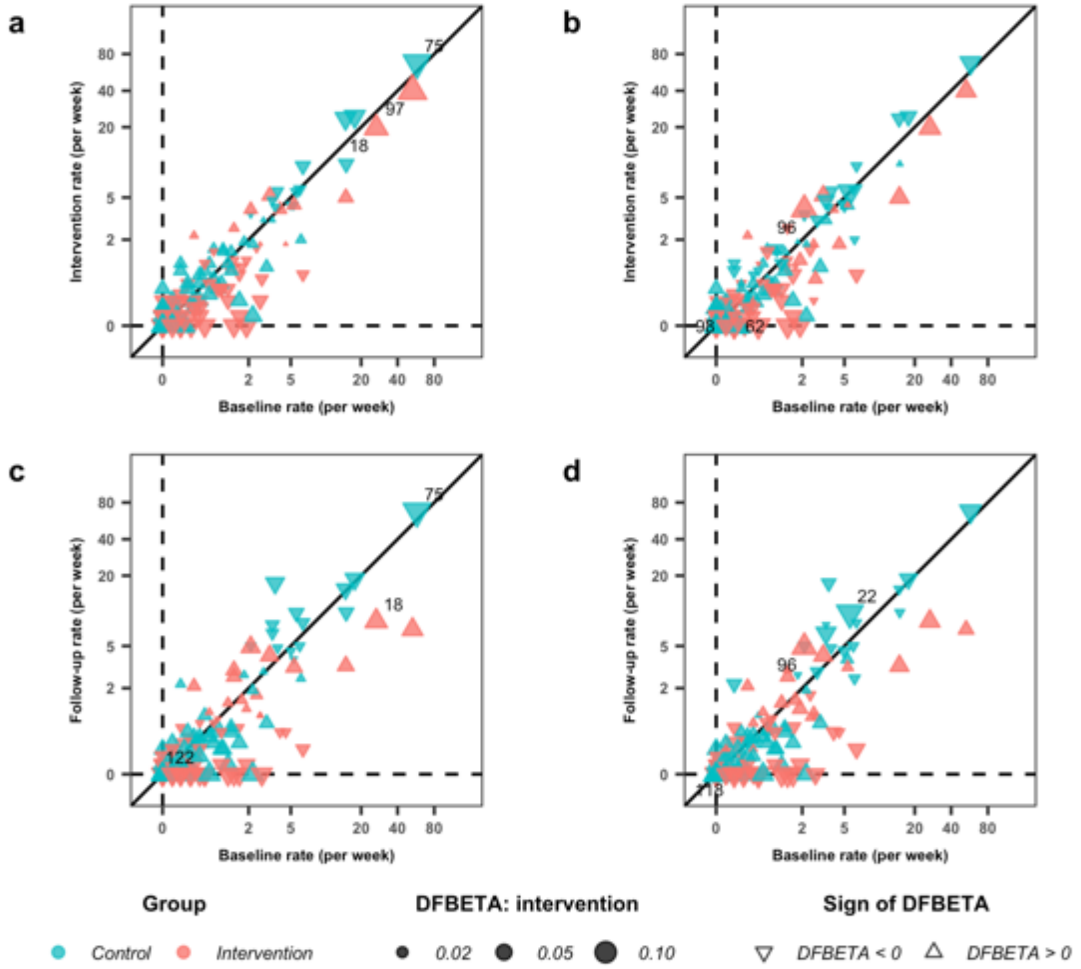


Figure 5-9 Intervention DFBETA plots for the models fitted to the Goodwin et al. dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=116); **(b)** *NB-full* fitted to the intervention falls count (n=116); **(c)** *NB-basic* fitted to the follow-up falls count (n=130); **(d)** *NB-full* fitted to the follow-up falls count (n=130). In each subplot, the three subjects with the largest absolute values of DFBETA are labelled with their ID.

5.4.2 Martin et al. dataset

Figure 5-10 compares the covariate-adjusted probabilities from the *Poi-basic* and *NB-basic* for the Martin et al. dataset, from which it is clear that the *NB-basic* model has a better fit than *Poi-basic*. The covariate-adjusted probabilities from the NB model (the green curve in the figure) are reasonably close to the observed probabilities (the blue curve), considering the sample size is only 21.

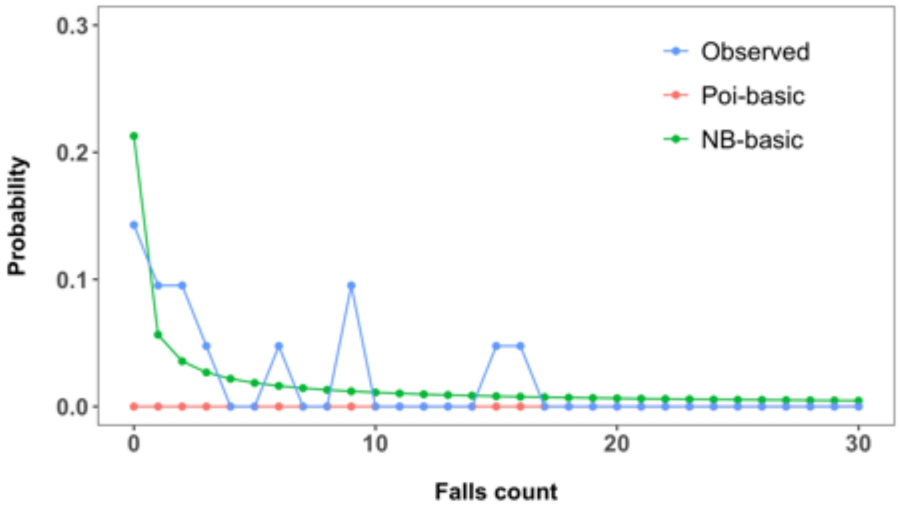


Figure 5-10 Covariate-adjusted probability plots for model comparisons in the Martin et al. dataset: *Poi-basic* versus *NB-basic* (n=21).

BOE plots from *NB-basic* fitted to the Martin et al. dataset are shown in Figure 5-11 to Figure 5-14. Similar to the plots for the Goodwin et al. dataset, the plotting symbols are close to the LoFE, which implies that the falls rate is stable across the baseline and follow-up periods. There are two reasons for the strong correlation between the baseline and intervention rates: 1) falls counts were collected prospectively during both periods; 2) the baseline and follow-up periods were relatively short (4 and 20 weeks respectively) and there is no gap in between, so the risk of falling during the intervention period was not considerably different from the baseline risk.

Figure 5-11 shows that participant CU21 recorded the highest falls rate during both the baseline and intervention periods, and this participant also showed the greatest Cook’s distance in the NB model. The three participants with the largest Cook’s distance had fallen more frequently than the others.

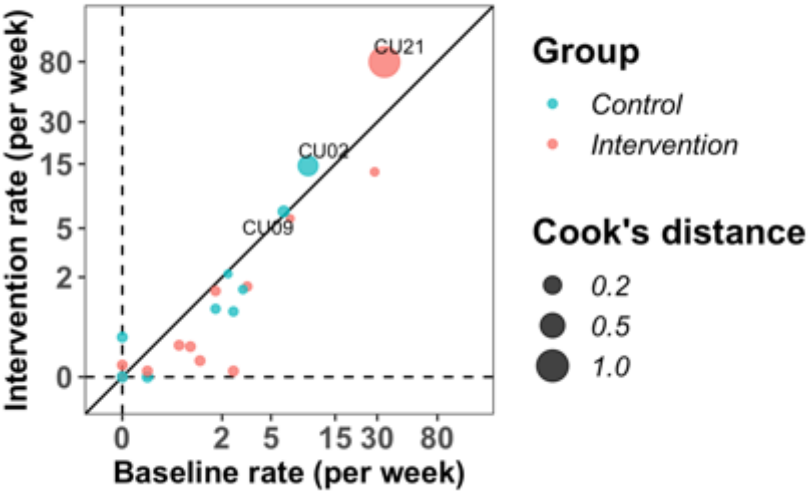


Figure 5-11 Cook's distance plot for *NB-basic* fitted to the Martin et al. dataset (n=21). The three subjects with the largest Cook's distances are labelled with their ID.

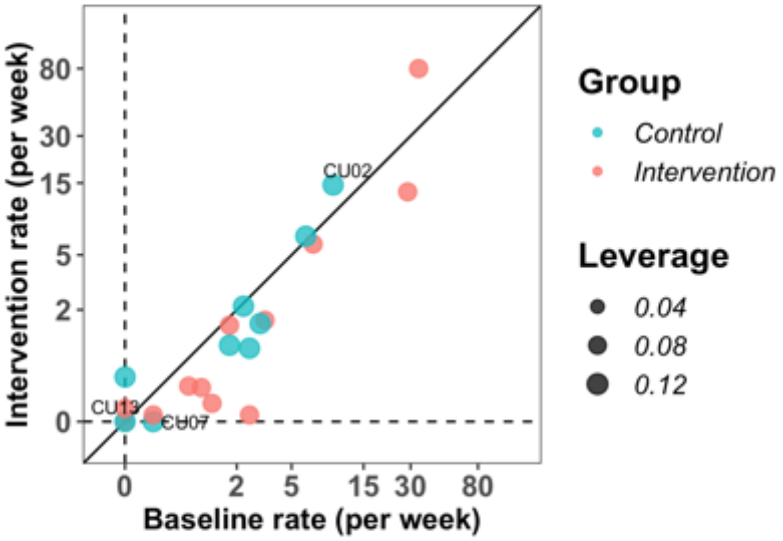


Figure 5-12 Leverage plots for *NB-basic* fitted to the Martin et al. dataset (n=21). The three subjects with the largest leverage are labelled with their ID.

As shown in Figure 5-12, all subjects have small leverage, again because the model has only one covariate—group.

In addition to the large Cook's distances, CU21 and CU02 show large Anscombe residuals (Figure 5-13). The small outcome rates typically yielded negative residuals, while the large outcome rates yielded positive residuals, suggesting the model has not fully accommodated the skewness of the data.

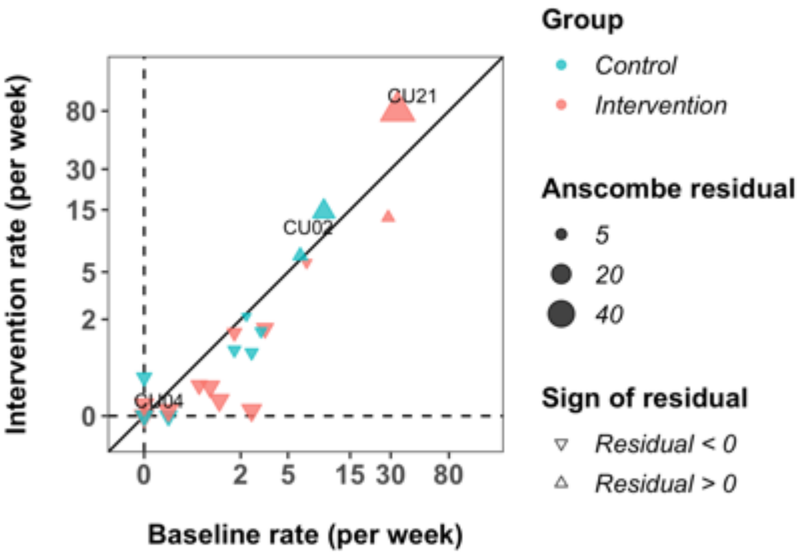


Figure 5-13 Anscombe residual plots for *NB-basic* fitted to the Martin et al. dataset (n=21). The three subjects with the largest absolute values of Anscombe residuals are labelled with their ID.

Figure 5-14 shows the DFBETA for the intervention effect from *NB-basic*. Because of the small sample size, the subjects with the largest or smallest outcome counts have remarkably large DFBETA, suggesting that these subjects have large impacts on the estimation of the intervention effect. *NB-basic* yields an FRR of 2.833 (see Table 3-11) for intervention effect, which is not in line with the pattern shown in the BOE plots: the blue symbols (control group) do not show a trend of falling under the red dots. This suggest that the extreme FRR is likely to be influenced by the outliers.

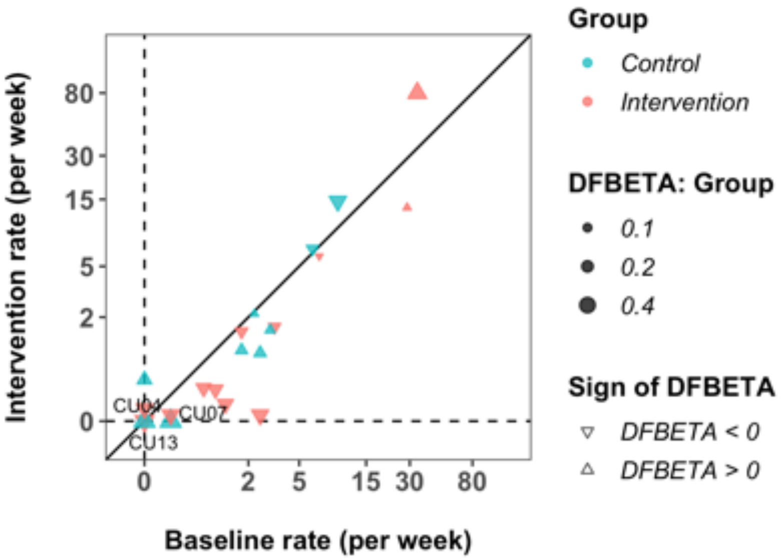


Figure 5-14 Intervention DFBETA plots for *NB-basic* fitted to the Martin et al. dataset (n=21). The three subjects with the largest absolute values of DFBETA are labelled with their ID.

5.4.3 EXSart dataset

Figure 5-15 compares the covariate-adjusted probabilities from Poisson and NB models for the EXSart dataset. The NB models again fit the dataset better than the Poisson models, and the covariate-adjusted probabilities from the *NB-full* models are closer to the observed probabilities than those from *NB-basic*.

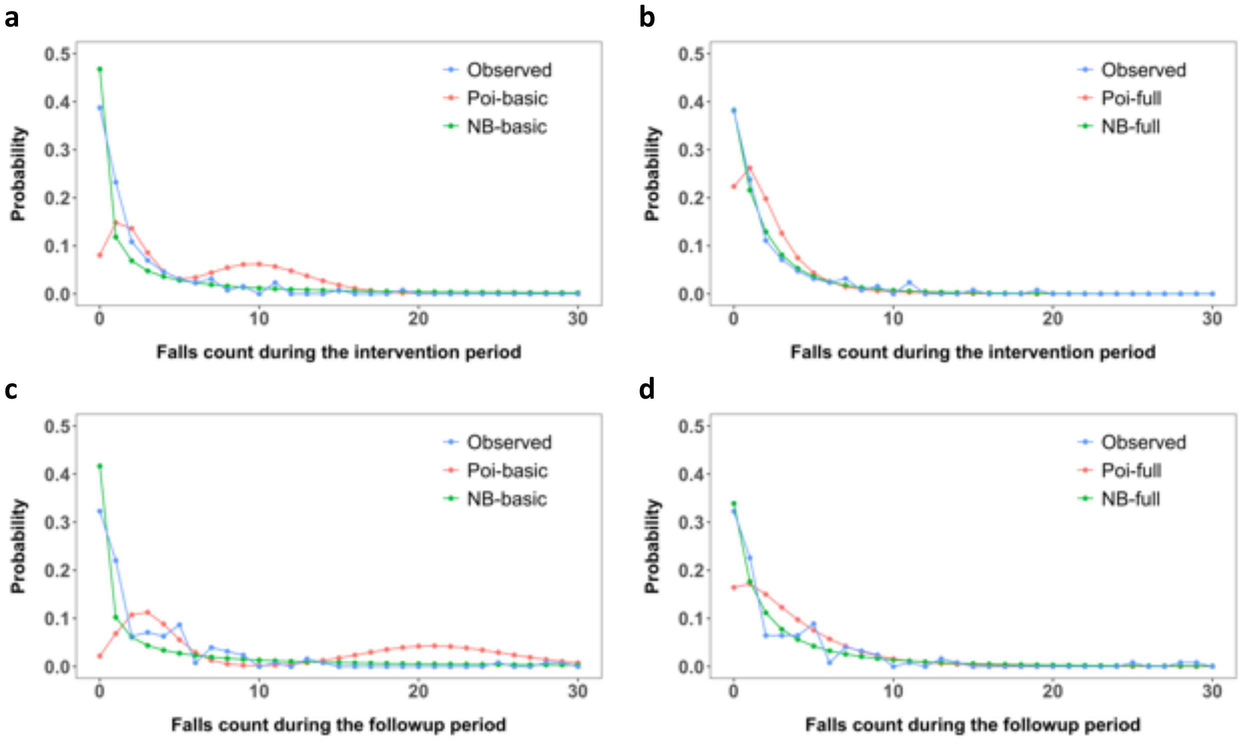


Figure 5-15 Covariate-adjusted probability plots for model comparisons in the EXSart dataset. **(a)** intervention falls count: *Poi-basic* versus *NB-basic* (n=129); **(b)** intervention falls count: *Poi-full* versus *NB-full* (n=126); **(c)** follow-up falls count: *Poi-basic* versus *NB-basic* (n=127); **(d)** follow-up falls count: *Poi-full* versus *NB-full* (n=124).

The EXSart trial is different to the other two in that the baseline falls count was obtained by asking the participants to retrospectively recall how many falls they had experienced during the year prior to the screening interview. The correlation between a retrospective and a prospective falls count would be expected to be weaker than that between two prospectively collected counts.

As shown in Figure 5-16, the plotting symbols deviate from the LoFE to a greater extent than those in the other two datasets, confirming the weaker correlation between retrospective baseline and prospective follow-up counts. The most frequently falling participants during the intervention and follow-up periods, ID 28, has the largest Cook's distances (around 5) from *NB-basic* for both the intervention and follow-up counts. ID 28 does not appear in Figure 5-16 b) and d) because this participant was excluded from the *NB-full* models due to missing data in UPDRS.

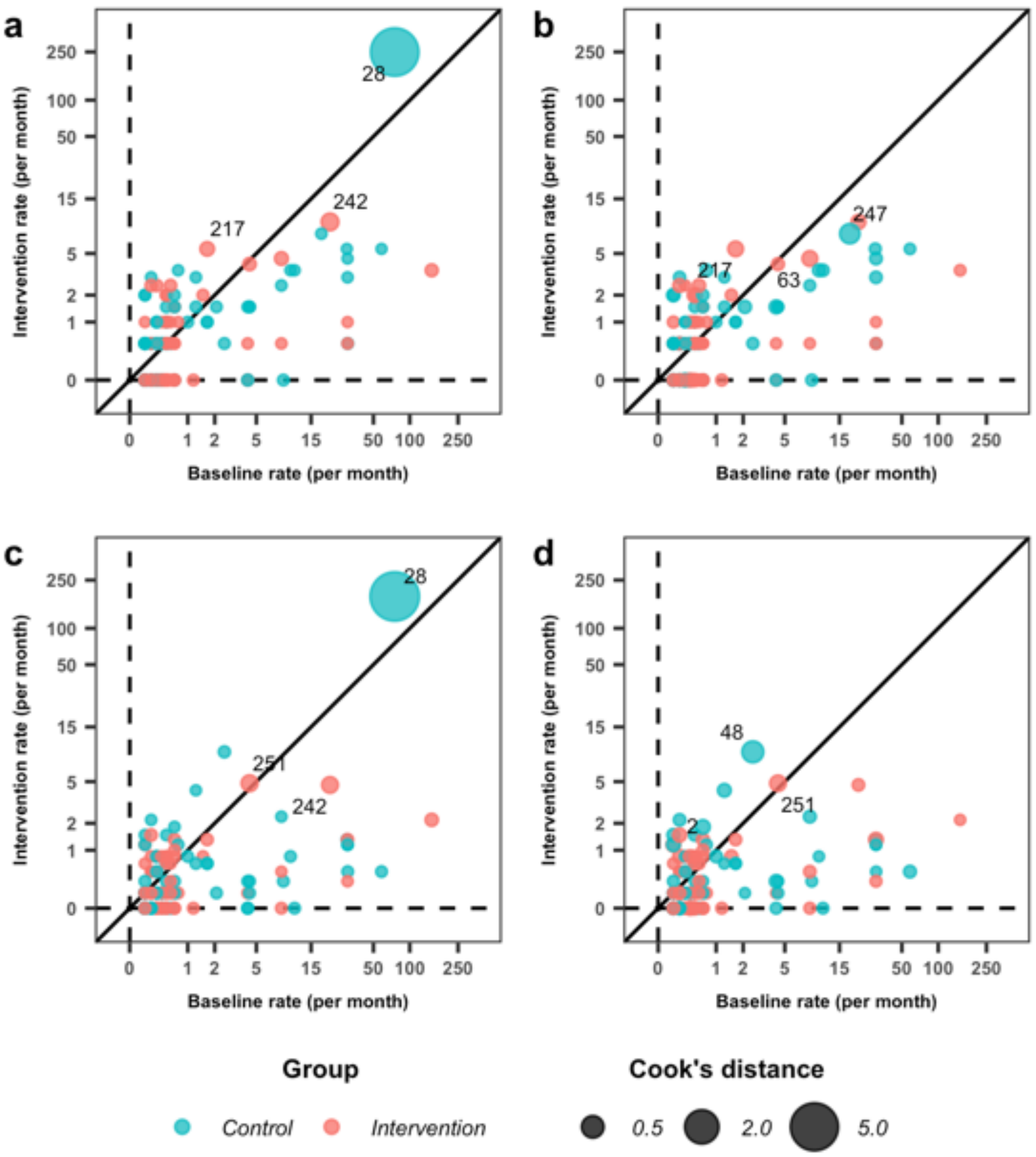


Figure 5-16 Cook's distance plots for the models fitted to the EXSart dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=129); **(b)** *NB-full* fitted to the intervention falls count (n=126); **(c)** *NB-basic* fitted to the follow-up falls count (n=127); **(d)** *NB-full* fitted to the follow-up falls count (n=124). In each subplot, the three subjects with the largest Cook's distances are labelled with their ID.

Figure 5-17 shows the leverage plots. The participant with the largest Cook's distance did not have large leverage, indicating that the large influence is not due to the particular values in covariates.

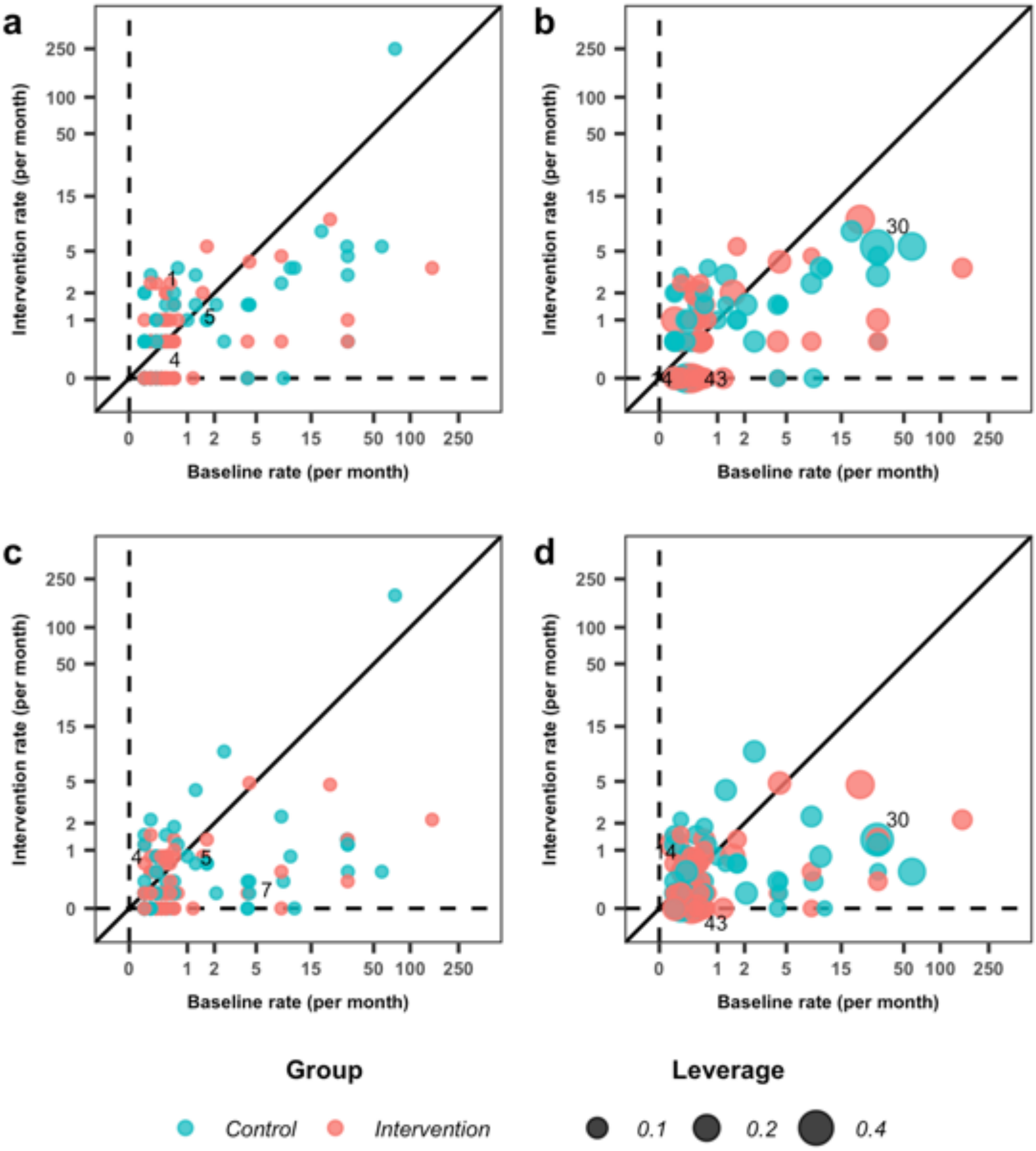


Figure 5-17 Leverage plots for the models fitted to the EXSart dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=129); **(b)** *NB-full* fitted to the intervention falls count (n=126); **(c)** *NB-basic* fitted to the follow-up falls count (n=127); **(d)** *NB-full* fitted to the follow-up falls count (n=124). In each subplot, the three subjects with the largest leverage are labelled with their ID.

Figure 5-18 shows the Anscombe residuals from each NB model. Similar to the other two datasets, the small outcome rates have negative residuals while the large outcome rates have positive residuals. ID 28 has a massive positive Anscombe residual of around 70. This suggests that the large counts have not been fully accommodated in any of the models.

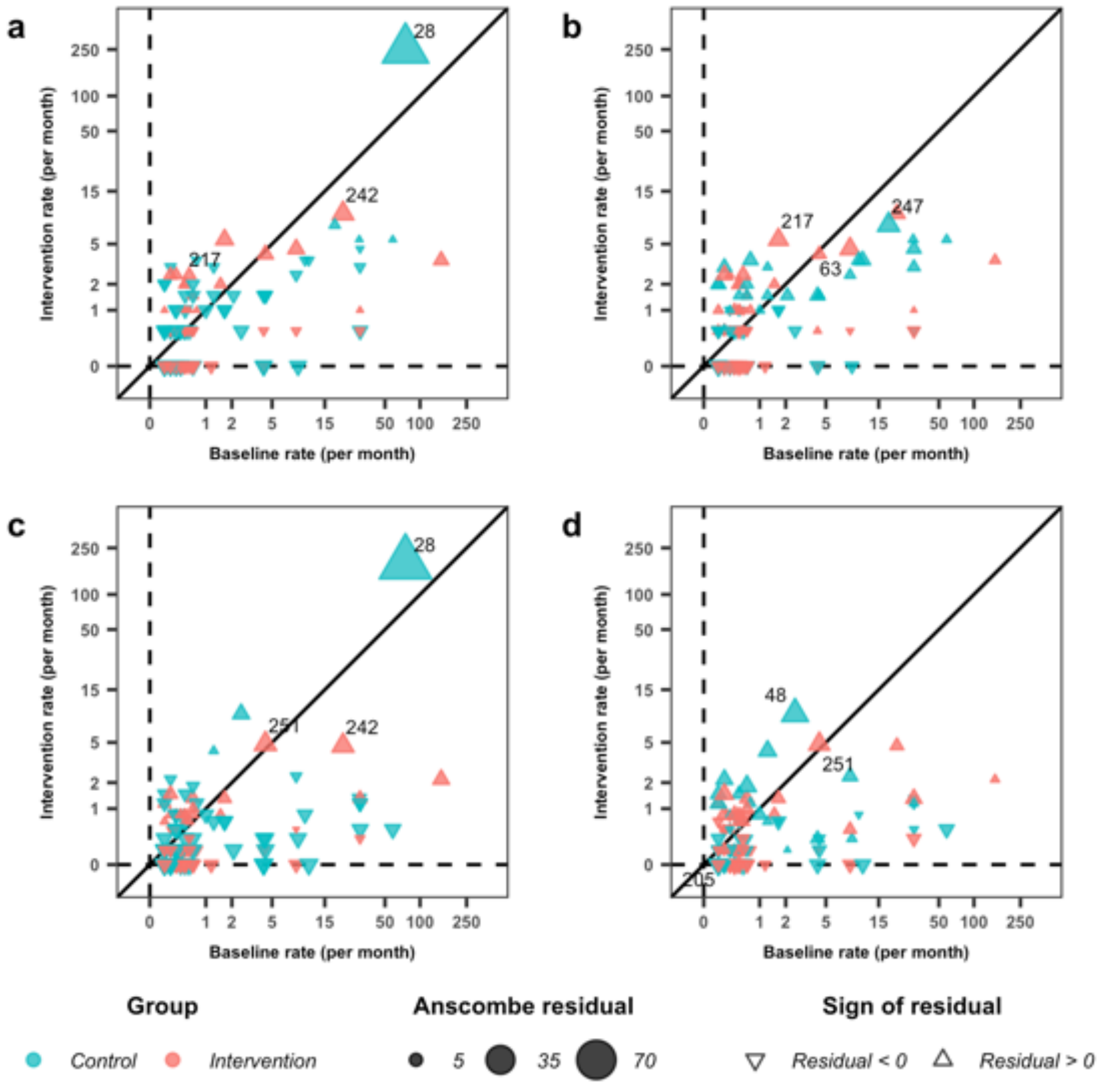


Figure 5-18 Anscombe residual plots for the models fitted to the EXSart dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=129); **(b)** *NB-full* fitted to the intervention falls count (n=126); **(c)** *NB-basic* fitted to the follow-up falls count (n=127); **(d)** *NB-full* fitted to the follow-up falls count (n=124). In each subplot, the three subjects with the largest absolute values of Anscombe residuals are labelled with their ID.

Figure 5-19 shows that ID 28 is influential in the estimation of the intervention effect, shown by the large negative DFBETA values (around -0.15 for both the intervention counts and follow-up counts).

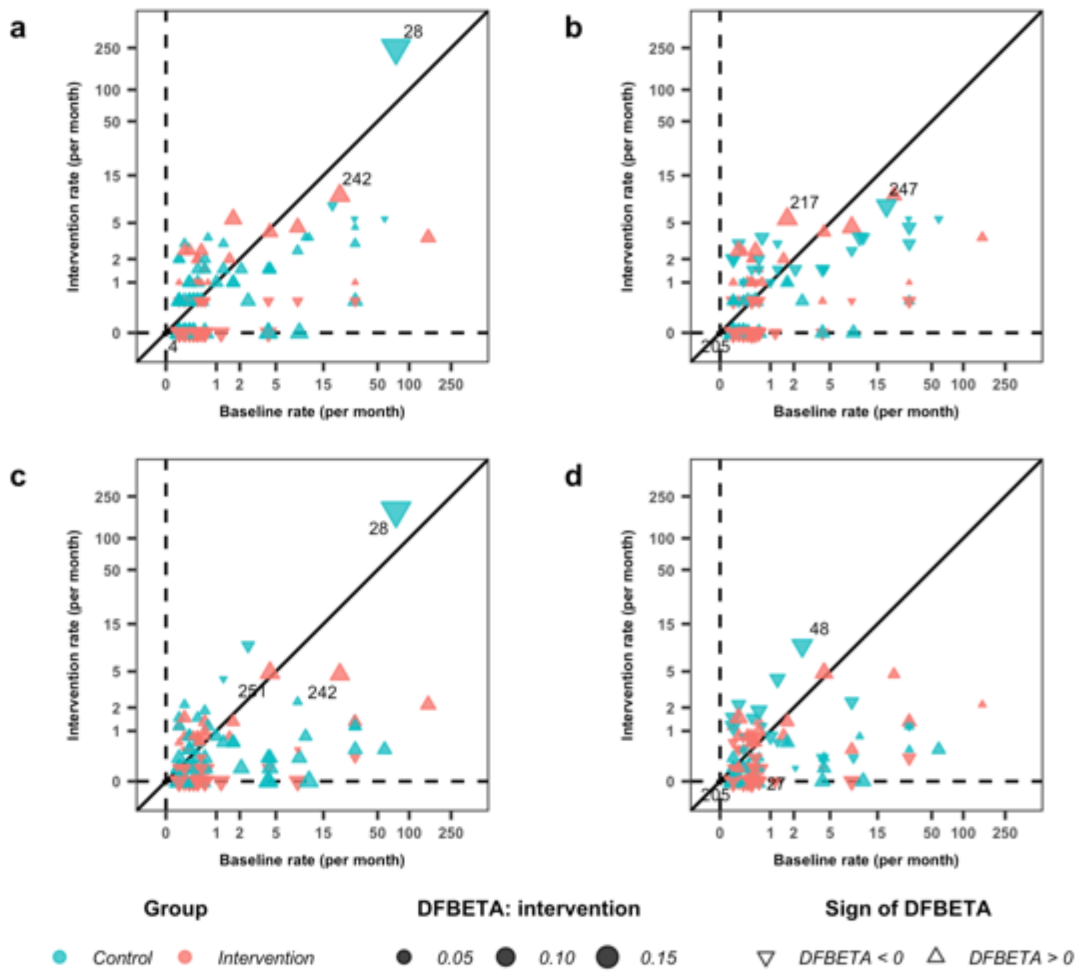


Figure 5-19 Intervention DFBETA plots for the models fitted to the EXSart dataset. **(a)** *NB-basic* fitted to the intervention falls count (n=129); **(b)** *NB-full* fitted to the intervention falls count (n=126); **(c)** *NB-basic* fitted to the follow-up falls count (n=127); **(d)** *NB-full* fitted to the follow-up falls count (n=124). In each subplot, the three subjects with the largest absolute values of DFBETA are labelled with their ID.

5.5 Discussion

Outlying large outcome counts are a major challenge for modelling falls data, as they often result in model overdispersion and are influential in the estimation of the intervention effect. Diagnostic plots of the model diagnostic statistics assist in pinpointing influential subjects, but existing plots are limited for analysing data from falls prevention trials. Four BOE plots were described in this chapter to present the following diagnostic statistics: Cook’s distance, leverage, Anscombe residual, and DFBETA for intervention effect. These plots provide useful diagnostic information and are straightforward to interpret.

BOE plots show whether the outcome falls rate is consistent with the baseline rate. An LoFE is included in the BOE plot to provide a reference line of constant falls rate across periods. Because the falls rate may change over time due to disease progression, it would be desirable to assess the possibility of a period effect, and this can be done by examining whether the plotting symbols from the control group are symmetrical around the LoFE. The BOE plots also provide a visualisation of the intervention effect, and the estimate of the effect from an NB model can be examined in the DFBETA plot regarding whether it could be due to just a few influential outliers.

Plotting covariate-adjusted probabilities from NB models is also discussed in this chapter. By overlaying the observed probabilities with the covariate-adjusted probabilities from the NB and Poisson models with the same covariates, this plot proves to be an effective diagnostic tool for examining Poisson overdispersion.

The diagnostic plots described above were produced for the NB models in Chapter 3. The covariate-adjusted probability plots show that all the NB models resulted in a much better fit to the falls data than the Poisson models, which conforms to the LR overdispersion tests in Chapter 3. Overall, the large outcome counts are highly influential when baseline counts are not included in the model, as shown in the Cook's distance plots. The Anscombe residuals are mostly negative for small outcome counts and positive for large counts. This indicates that the variance of the outcome count exceeds the NB variances ($\mu + \alpha\mu^2$).

The three datasets each has its own characteristics that result in different patterns in the BOE plots. During the follow-up period of the Goodwin et al. trial, most of the participants who recorded the largest falls counts were in the control group. The BOE plots showed that the frequently falling participants during the follow-up period are influential in model estimation, especially for the estimation of the intervention effect. The Martin et al. dataset has a small sample size, and therefore each participant has great influence on the estimation of the intervention effect, as indicated by the DFBETA plot. EXSart is the only dataset in which the baseline falls were collected retrospectively, while the method of collecting the outcome falls rate was consistent with that for the baseline rate in the Goodwin et al. and Martin et al. datasets. As a result, the outcome and baseline rates have much weaker correlation in the EXSart dataset (the plotting symbols deviate from the LoFEs to a greater extent in the BOE plots).

By comparing the four BOE plots, a reader gains a full picture of the influential subjects. An `R` package was written to produce the diagnostic plots for NB models for dataset from falls prevention trials, as well as trials for other recurrent events with a baseline count.

Chapter 6

Comparison of approaches to incorporating the baseline count in NB-related models

Most of this chapter has been published in *Biometrical Journal* (Zheng et al., 2018). An exception is section 6.6, which is presented for the first time here. The author of this thesis (HZ) is the first author and the main contributor to the publication and conducted the analysis as well as the simulation study. The formulae and tables in section 6.2 to 6.5 are the same as used in the published paper. The chapter has been rewritten slightly to fit the formatting of the thesis and examines models for the falls count in the Goodwin et al. (2011) dataset during the intervention period, including the intervention effect and baseline count. Other baseline characteristics are not considered.

6.1 Introduction

A common design for falls prevention trials is to collect the number of falls experienced by each participant during a baseline period (prior to randomisation) as the baseline falls count, and during a follow-up period (after randomisation and the onset of the intervention) as the outcome count. Outcome falls counts are often analysed using count response models to calculate an FRR as the estimate of the intervention effect, but how best to incorporate baseline counts in modelling remains a question.

Vickers and Altman (2001) discussed methods for analysing RCTs with a continuous variable measured at the baseline and as an outcome at follow up. They commented that the most straightforward method, basing analysis solely on the outcomes in each trial group, does not cope with the potential imbalance of the baseline measurements between groups. The authors recommended including the baseline measurement as a regressor instead, because it copes with the baseline measurement regardless of whether they are balanced between groups. Although Vickers and Altman demonstrated the issue using normally distributed data, their argument is applicable to falls counts as well.

Including a baseline falls count in NB models is expected to increase the statistical power, and in addition, control for overdispersion. If the heterogeneity of a model is primarily due to unobserved latent subject-specific prognostic variables, the outcome and baseline counts are anticipated to be correlated as they measure the same person. Because the baseline period of a trial is prior to the randomisation, the number of falls during the period is not confounded with the intervention. Therefore, including a baseline count in a model accounts for the latent variables.

Despite the great benefit of including baseline falls counts in models, few Parkinson's researchers have recognized the importance of utilising this information in statistical analysis. Whether the baseline count was incorporated in modelling and how it was done is often not explicitly described in papers. In a Cochrane review of falls prevention trials (Gillespie et al., 2012), the authors recommend using NB regression for analysing falls data, but did not provide guidance on incorporating baseline counts in the model, nor did they review how this was done in practice: the baseline count may be ignored, categorised into a discrete covariate, or included after transformation. The lack of description implies that the baseline count has been largely overlooked.

Cook and Wei (2003) proposed the Conditional Negative Binomial (CNB) model to incorporate the baseline count. Similar to the NB model, the heterogeneity is modelled in CNB as a gamma distributed random subject effect. The difference is that the CNB model is based on a mixed Poisson distribution in which a baseline count shares the same random subject effect as the outcome. This enables modelling the outcome count conditioning on the baseline count; while an NB model ignoring the baseline count can be deemed a model marginalising over the random effect.

To ensure that only PwP with high risk of falling could enter the study, some trials restricted participation to those with baseline counts greater than a threshold. This design increases statistical power, but it also results in a truncated distribution for the baseline count, which violates the mixed Poisson distribution underlying the CNB model. For a trial with this design, the threshold value must be specified in the CNB model to accommodate the truncation. A problem of this design is that it is expensive, because it requires recruiting more PwP during the baseline period, many of whom may not fall. An alternative approach is to, 1) ask the interviewees at the screening interview to recall the number of falls they had experienced during a period of time in the past, 2) only recruit people who recalled more falls than the eligibility criterion, and 3) obtain a baseline count during a baseline period using the prospective method. This approach is more cost-effective, as only the PwP who are likely to fall enter the prospective study. Because this approach does not result in a truncated baseline count, the CNB model does not need to be adjusted based on the eligibility criterion.

The motivating dataset of this chapter was that reported by Goodwin et al. (2011). An eligibility criterion was that participants had to report having fallen at least twice in the previous year, obtained by a retrospective question asked at a screening interview prior to enrolment and baseline. As discussed above, this did not result in a truncated count during the prospective baseline diary collection period, and thus the CNB model does not need to account for the truncation in the baseline count.

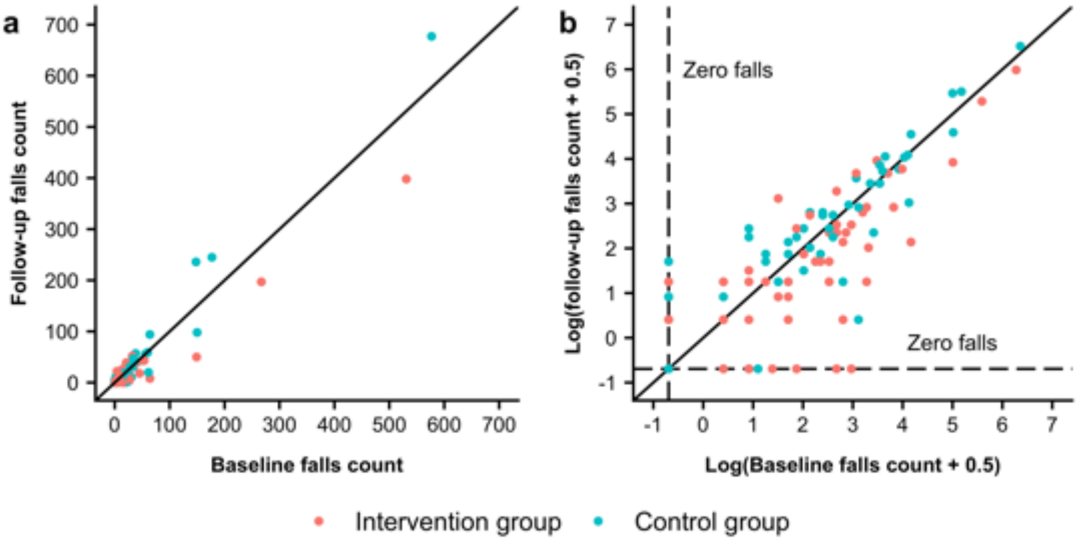


Figure 6-1 Goodwin et al. dataset: follow-up falls counts against baseline falls counts (n=124, Spearman ρ=0.813, P<0.001). The diagonal line is the LoFE described in section 5.2. In subplot b, 0.5 is added to both counts before log-transformation to include zero counts.

Figure 6-1 shows the falls counts in the Goodwin et al. dataset. The outcome counts are plotted against the baseline counts, on the linear (Figure 6-1a) and logarithmic (Figure 6-1b) scale. The falls counts have a relatively small mean and a few outlying large numbers. The figure indicates that the outcome counts, in a broad sense, follow a linear relationship with the baseline counts. In isolation, the large outcome counts may be classified as outliers because they are far from the body of cases in Figure 6-1, but overall they show a strong agreement with the corresponding baseline counts in the scatter plots.

Although CNB can be used to model the relationship between the outcome and baseline counts, it is not widely supported in statistical packages, which raises a question—how to incorporate the baseline count in an NB model so that the model reflects the underlying mixed Poisson distribution?

In this chapter, different approaches are compared to incorporating the baseline falls count in NB and Poisson models. Their performance is compared to that of CNB, which is considered as the benchmark model. The models are fitted to the Goodwin et al. dataset, and model diagnostic statistics are further examined. A simulation study is conducted to compare the models regarding bias, power, type I error rate, and the standard error of the intervention effect, under scenarios reflecting our motivating dataset. Statistical significance of the intervention effect was assessed using the Wald test, because it is typically the default model-based test in statistical packages. P values from score tests were also calculated and compared to those from Wald tests.

6.2 Models incorporating the baseline count

6.2.1 Mixed Poisson distribution with subject-specific heterogeneity

Suppose m subjects are enrolled in a trial, which is comprised of a baseline period (indicated by $j = 0$; prior to randomisation) and an outcome period (indicated by $j = 1$; post randomisation). Let t_0 denote the duration of the baseline period, which is assumed to be the same for all subjects (common in falls prevention trials), and t_{i1} the duration of the outcome period for subject i, where $i = 1, \dots, m$. The subjects may have varying length of the outcome period due to dropout, assumed to occur at random. At randomisation, subject i is allocated either to an intervention (denoted by $x_i = 1$) or a control group ($x_i = 0$). Let y_{i0} and y_{i1} denote the number of falls experienced by subject i during the baseline and outcome periods respectively. If variables Y_{i0} and Y_{i1} both follow Poisson distribution,

$$\Pr(Y_{i0} = y_{i0}; \lambda_0, t_0) = \frac{(\lambda_0 t_0)^{y_{i0}} \exp(-\lambda_0 t_0)}{y_{i0}!} \tag{6-1}$$

and

$$\Pr(Y_{i1} = y_{i1}; \lambda_1, t_{i1}) = \frac{(\lambda_1 \exp(\beta x_i)\, t_{i1})^{y_{i1}} \exp(-\lambda_1 \exp(\beta x_i)\, t_{i1})}{y_{i1}!}, \tag{6-2}$$

where λ_0 is the average falls rate during the baseline period, λ_1 is the average rate for the control group during the outcome period, and β is the logarithm of the FRR for the intervention effect.

Let $\mu_0 = \lambda_0 t_0$ and $\mu_{i1}{=}\lambda_1 \exp(\beta x_i)\, t_{i1}$, the expectation and variance of Y_{i0} and Y_{i1} are

$$\mathrm{E}(Y_{i0}) = \mathrm{Var}(Y_{i0}) = \mu_{i0} \tag{6-3}$$

$$\mathrm{E}(Y_{i1}) = \mathrm{Var}(Y_{i1}) = \mu_{i1}. \tag{6-4}$$

If there is heterogeneity in the baseline and outcome counts, they will both be overdispersed, that is, the variance of Y_{ij} will be greater than the expectation μ_{ij}, which violates the assumption of equidispersion in Poisson regression.

We further assume that the heterogeneity is brought about by a gamma distributed random subject effect s_i with mean 1 and variance α. The conditional probability distributions of Y_{i0} and Y_{i1} given s_i is a mixed Poisson distribution (Cook and Wei, 2003) given by

$$Y_{i0}|s_i \sim \mathrm{Poisson}(s_i \mu_{i0}) \tag{6-5}$$

$$Y_{i1}|s_i \sim \mathrm{Poisson}(s_i \mu_{i1}). \tag{6-6}$$

The counting process underlying Y_{i0} and Y_{i1} was described by Cook et al. (2005) as a time-homogeneous Poisson process, because heterogeneity is introduced by a latent subject-specific effect, such that Y_{i0} and Y_{i1} are conditionally independent given s_i.

Marginalising over s_i yields the PMF of NB regression:

$$\Pr(Y_{i1} = y_{i1}; \mu_{i1}, \alpha) = \frac{\Gamma(y_{i1} + \alpha^{-1})}{\Gamma(y_{i1} + 1)\Gamma(\alpha^{-1})}\left(\frac{1}{1 + \alpha\mu_{i1}}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_{i1}}{1 + \alpha\mu_{i1}}\right)^{y_{i1}}. \tag{6-7}$$

The variance of Y_{i1} in (6-7) is

$$\mathrm{Var}(Y_{i1}) = \mu_{i1} + \alpha\mu_{i1}^2, \tag{6-8}$$

where $\alpha\mu_{i1}^2$ accommodates the extra variance exceeding that in Poisson regression, which can be regarded as an NB model with α approaching zero. Conversely, NB regression is a generalisation of the Poisson model, with the same log link function $\mathrm{g}(\mathrm{E}(Y_{i1}))$ = $\log(\mu_{i1})$ = η_{i1}, where η_{i1} is the linear predictor of the model.

From the mixed Poisson distribution described in equations (6-5) and (6-6), Cook and Wei (2003) derived the conditional distribution of Y_{i1} given the baseline y_{i0} as:

$$\Pr(Y_{i1} = y_{i1} | y_{i0}; \lambda_0, \lambda_1, \beta, \alpha) = \frac{\Gamma(y_{i0} + y_{i1} + \alpha^{-1})}{\Gamma(\alpha^{-1} + y_{i0})\Gamma(y_{i1} + 1)}\frac{(1 + \alpha\mu_{i0})^{\alpha^{-1}+y_{i0}}(\alpha\mu_{i1})^{y_{i1}}}{(1 + \alpha(\mu_{i0} + \mu_{i1}))^{\alpha^{-1}+y_{i0}+y_{i1}}}, \tag{6-9}$$

and their Conditional Negative Binomial (CNB) model fits this distribution to data.

As introduced in section 2.3.3, the estimate of α from NB regression is referred to as the Heterogeneous Parameter (HP) by Hilbe (2011). It shows the degree of heterogeneity remaining unaccounted for by the model covariates. Therefore, including more covariates in an NB model may, to some degree, explain heterogeneity, and would be expected to result in a smaller HP. In contrast, the $\hat{\alpha}$ from a CNB model is the variance of the underlying random subject effects and is estimated from both the baseline and the outcome counts. Larger $\hat{\alpha}$ indicates stronger association between the outcome and baseline counts. Because of the distinct interpretations of the two estimates, the estimate of α is referred to as HP for NB model and as $\hat{\alpha}$ for CNB model.

6.2.2 Including the baseline count as a covariate in NB models

Although CNB model is a direct derivation from the mixed Poisson distribution put forward in (6-5) and (6-6), a commonly seen alternative is to include the baseline count as a covariate in NB regression. In this section some alternative approaches to incorporating the baseline count in an NB regression are described so that it captures the correlation of Y_{i0} and Y_{i1}.

Given the random subject effect s_i, the conditional expectations of Y_{i0} and Y_{i1} given s_i in (6-5) and (6-6) are

$$\mathrm{E}(Y_{i0}|s_i) = \lambda_0 s_i t_0 \tag{6-10}$$

$$\mathrm{E}(Y_{i1}|s_i, x_i, t_{i1}) = \lambda_1 s_i \exp(\beta x_i)\, t_{i1}. \tag{6-11}$$

Hence

$$\mathrm{E}(Y_{i1}|s_i, x_i, t_{i1}) = \frac{\lambda_1}{\lambda_0 t_0} \exp(\beta x_i) \mathrm{E}(Y_{i0}|s_i) t_{i1}. \tag{6-12}$$

Marginalising over s_i in (6-10) and (6-11) gives

$$\mathrm{E}(Y_{i0}) = \int_0^\infty \mathrm{E}(Y_{i0}|s_i) f_S(s_i) ds_i \tag{6-13}$$

$$\mathrm{E}(Y_{i1}|x_i, t_{i1}) = \int_0^\infty \mathrm{E}(Y_{i1}|s_i, x_i, t_{i1}) g_S(s_i) ds_i\,, \tag{6-14}$$

where $g_S(.)$ is the PDF of s_i. Based on equation (6-11), equation (6-14) can be written as

$$\begin{aligned}\mathrm{E}(Y_{i1}|x_i, t_{i1}) &= \int_0^\infty \frac{\lambda_1}{\lambda_0 t_0} \exp(\beta x_i)\, \mathrm{E}(Y_{i0}|s_i) t_{i1} g_S(s_i) ds_i \\ &= \frac{\lambda_1}{\lambda_0 t_0} \exp(\beta x_i)\, t_{i1} \int_0^\infty \mathrm{E}(Y_{i0}|s_i) g_S(s_i) ds_i \\ &= \frac{\lambda_1}{\lambda_0 t_0} \exp(\beta x_i)\, \mathrm{E}(Y_{i0}) t_{i1}\end{aligned} \tag{6-15}$$

Taking the logarithm of both sides of (6-15) yields

$$\log\big(\mathrm{E}(Y_{i1}|x_i, t_{i1})\big) = \log\left(\frac{\lambda_1}{\lambda_0 t_0}\right) + \beta x_i + \log\big(\mathrm{E}(Y_{i0})\big) + \log(t_{i1}). \tag{6-16}$$

Putting the moment estimator of Y_{i0}, that is, $y_{i0}/1 = y_{i0}$, instead of $\mathrm{E}(Y_{i0})$ gives

$$\log\big(\mathrm{E}(Y_{i1}|\, x_i, t_{i1})\big) = \log\left(\frac{\lambda_1}{\lambda_0 t_0}\right) + \beta x_i + \log(y_{i0}) + \log(t_{i1}). \qquad \text{(6-17)}$$

Equation (6-17) suggests a Poisson/NB regression model for Y_{i1} of the following form:

$$\log\big(\mathrm{E}(Y_{i1}|y_{i0}, x_i, t_{i1})\big) = g(\mu_{i1}) = \zeta + \beta x_i + \log(y_{i0}) + \log(t_{i1}), \qquad \text{(6-18)}$$

where the constant term $\log(\lambda_1/(\lambda_0 t_0))$ in (6-17) is absorbed in the intercept ζ.

This suggests that, compared to including the baseline count y_{i0} as an untransformed regressor, it is more appropriate to include $\log\,(y_{i0})$ as an offset. If the exposure t_{i1} varies across subjects, the offset in the model is the combined term $\log(y_{i0}) + \log(t_{i1})$, which can be reduced to $\log\,(y_{i0})$ if t_{i1} is the same over i.

The performance of NB regression with the following four linear predictors are compared: 1) ignoring the baseline count y_{i0}; 2) including the untransformed y_{i0} as a covariate; 3) including $\log\,(y_{i0})$ as a covariate; and 4) including $\log\,(y_{i0})$ as an offset. The results of Poisson models with the same four linear predictors are produced for comparison, and the CNB model is included as the benchmark model.

The NB / Poisson Models with the four linear predictors described above are referred to as *NB-null* / *Poi-null*, *NB-unlogged* / *Poi-unlogged*, *NB-logged* / *Poi-logged*, and *NB-offset* / *Poi-offset* respectively in Zheng et al. (2018). They are introduced and described below.

Ignoring the baseline count (*NB-null/Poi-null*)

As discussed earlier, the HP in NB regression shows how much variability relative to Poisson has been introduced by s_i, and remains unexplained by any explanatory variables in the model. When the baseline count is ignored, the intervention indicator x_i and the exposure $\log\,(t_{i1})$ are the only explanatory variables in the linear predictor:

$$g(\mu_{i1}) = \zeta + \beta x_i + \log(t_{i1}). \qquad \text{(6-19)}$$

The HP from *NB-null* is estimated based on the outcome count only, thus its value is anticipated to be close to $\hat{\alpha}$ from CNB model.

Including the unlogged baseline count as a covariate (*NB-unlogged/Poi-unlogged*)

In the *NB-unlogged* and *Poi-unlogged* models, the baseline count y_{i0} is included as a covariate, with no transformation. The HP from an *NB-unlogged* model is anticipated to be smaller than that from *NB-null*, because the included covariate y_{i0} may partially reduce heterogeneity.

The linear predictor in either *NB-unlogged* or *Poi-unlogged* is:

$$g(\mu_{i1}) = \zeta + \beta x_i + \psi y_{i0} + \log(t_{i1}), \tag{6-20}$$

where ψ is the regression coefficient for the baseline count.

Including the logged baseline count as a covariate (*NB-logged/Poi-logged*)

In the *NB-logged* and *Poi-logged* models, the log-transformed baseline count is included as a covariate to conform to the scaling of y_{i0} in (6-18). The linear predictor including the logged baseline count is given by:

$$g(\mu_{i1}) = \zeta + \beta x_i + \phi \log(y_{i0}) + \log(t_{i1}), \tag{6-21}$$

where ϕ is the regression coefficient for the logged baseline count. It is anticipated to be close to one when the data are in accordance with the mixed Poisson distribution.

If including logged baseline counts better accounts for the correlation between y_{i0} and the outcome y_{i1} than including the untransformed baseline counts, the HP from *NB-logged* is anticipated to be smaller than that from *NB-unlogged*. To include the subjects with $y_{i0} = 0$ (which cannot be logged), 0.5 was added to all the baseline counts before log-transformation.

Including the logged baseline count as an offset (*NB-offset/Poi-offset*)

Now we consider the *NB-offset* and *Poi-offset* models. Compared with the *NB-logged* and *Poi-logged* models, they are a closer match to the form of (6-18)—the baseline count is included as an offset, so the regression coefficient is constrained to be one. The linear predictor with baseline count as an offset term is:

$$g(\mu_{i1}) = \zeta + \beta x_i + \log(y_{i0}) + \log(t_{i1}). \tag{6-22}$$

Again, 0.5 is added to the baseline count to ensure that zero baseline counts can be log-transformed. If the approach of including logged baseline counts is more appropriate than including the untransformed value, the HP from *NB-offset* is anticipated to be lower than that of *NB-unlogged*.

6.3 Methods

Simulations and analysis were conducted using `R` (version 3.5.0). NB models were fitted using the `negbin()` function from the **aod** package, and Poisson models were fitted using the `glm()` function. CNB models were fitted using the `nlm()` function for non-linear minimisation (using code from the authors) .

P values were obtained from Wald tests, and reported along with estimates, 95% CI, and the AIC. Cook's distances were obtained from the `glm.nb()` function in the **MASS** package (described in section 4.5).

The models were fitted to the falls counts collected during the intervention period of the Goodwin et al. trial. To ensure that *NB-null* and *Poi-null* are comparable to the other models, one participant (ID 1) was excluded from analysis due to missing value in the baseline count.

6.4 Poisson/NB/CNB models fitted to the Goodwin et al. dataset

The models described in section 6.2 are fitted to the Goodwin et al. dataset and the estimates are shown in Table 6-1. *Poi-null*, which ignores both the baseline count and overdispersion, yields the largest AIC among all the fitted models as expected. By accounting for overdispersion, *NB-null* achieved a marked reduction in AIC (931.8 versus 9996.1 in *Poi-null*). Although the baseline count is not incorporated in *NB-null*, the model results in a lower AIC than any of the fitted Poisson models.

NB-unlogged results in a smaller AIC (844.2) than *NB-null*, and its HP is smaller as well, with ψ estimated to be 0.019. By including the logged baseline count instead of the untransformed count, the resultant model, *NB-logged*, further deceases AIC to 744.3, with $\hat{\phi}$ estimated to be 0.911. *NB-offset* yields a marginally higher AIC (745.5) than *NB-logged*, and a similar estimate of the intervention effect (FRR=0.698 in *NB-logged* and 0.707 in *NB-offset*). Both *NB-logged* and *NB-offset* yield significant intervention effects, with P values of 0.021 and 0.032 respectively. Their HPs are also much smaller than that of *NB-unlogged*.

The Poisson models show a similar pattern to the NB models. *Poi-unlogged* yields a smaller AIC than *Poi-null*. The AIC is further decreased in *Poi-logged* and *Poi-offset*, and the two models also give similar estimates for the intervention effect. The estimate of ψ from *Poi-unlogged* is 7.02×10^{-3} (close to zero), while $\hat{\phi}$ from NB-logged is 1.030 (close to one).

The CNB model yields smaller SE for $\hat{\beta}$ than the four NB models, as well as smaller P value from the Wald test. The estimate of α is 2.873 in CNB.

Table 6-1 Poisson/NB/CNB models fitted to the Goodwin et al. dataset (n=124).

Model	**AIC**	$\hat{\boldsymbol{\beta}}$ **(SE)**	**FRR (95% CI)**	**P**	$\hat{\boldsymbol{\psi}}$ **(SE)**	$\hat{\boldsymbol{\phi}}$ **(SE)**	**HP**
Poi-null	9996.1	-0.571 (0.037)	0.565 (0.525, 0.608)	< 0.001			
Poi-unlogged	3247.6	-0.472 (0.038)	0.624 (0.580, 0.672)	< 0.001	7.02×10^{-3} (6.78×10^{-5})		
Poi-logged	1131.5	-0.480 (0.037)	0.619 (0.575, 0.666)	< 0.001		1.030 (0.012)	
Poi-offset	1135.6	-0.479 (0.037)	0.619 (0.577, 0.666)	< 0.001			
NB-null	931.8	-0.572 (0.323)	0.565 (0.300, 1.064)	0.077			3.189
NB-unlogged	844.2	-0.391 (0.236)	0.677 (0.426, 1.074)	0.098	0.019 (0.004)		1.541
NB-logged	744.3	-0.359 (0.156)	0.698 (0.514, 0.948)	0.021		0.911 (0.048)	0.511
NB-offset	745.5	-0.346 (0.161)	0.707 (0.516, 0.970)	0.032			0.519
							$\hat{\alpha}$
CNB		-0.479 (0.051)	0.619 (0.561, 0.684)	< 0.001			2.873

Figure 6-2 displays diagnostic plots for *NB-unlogged* and *NB-logged* as a means of examining graphically whether the baseline count should be logged. The Anscombe residuals of *NB-unlogged* show a curvilinear pattern in Figure 6-2a—the residuals of the subjects with the largest fitted values deviating remarkably downwards from $y = 0$, which indicates that they are considerably overestimated by *NB-unlogged*. The pattern is confirmed by the Q-Q normal plot of the Anscombe residuals (Figure 6-2c). Although the Anscombe residuals in Figure 6-2d show satisfying normality, in Figure 6-2b it is clearly not a standard normal distribution (this was addressed in section 2.3.4). Figure 6-2e shows the Cook's distances from *NB-unlogged* in a BOE plot. The subjects who reported the largest baseline and outcome counts, the plotting symbols at the top-right corner of the plot, have the largest Cook's distances.

In contrast, for *NB-logged* the points in the residual-versus-fitted plot (Figure 6-2b) are reasonably symmetric around zero, and the residuals show satisfying normality in the Q-Q normal plot (Figure 6-2d). Comparing Figure 6-2 f) to e), the large outcome counts in *NB-logged* are not as influential as in *NB-unlogged*. The two subjects with the largest baseline and outcome counts are highly influential in *NB-unlogged*, but not in *NB-logged*. The subjects with inconsistent falls rates between the baseline and outcome periods, that is, those whose plotting symbols deviate furthest from the LoFE in Figure 6-2f, had the largest Cook's distances in *NB-logged*.

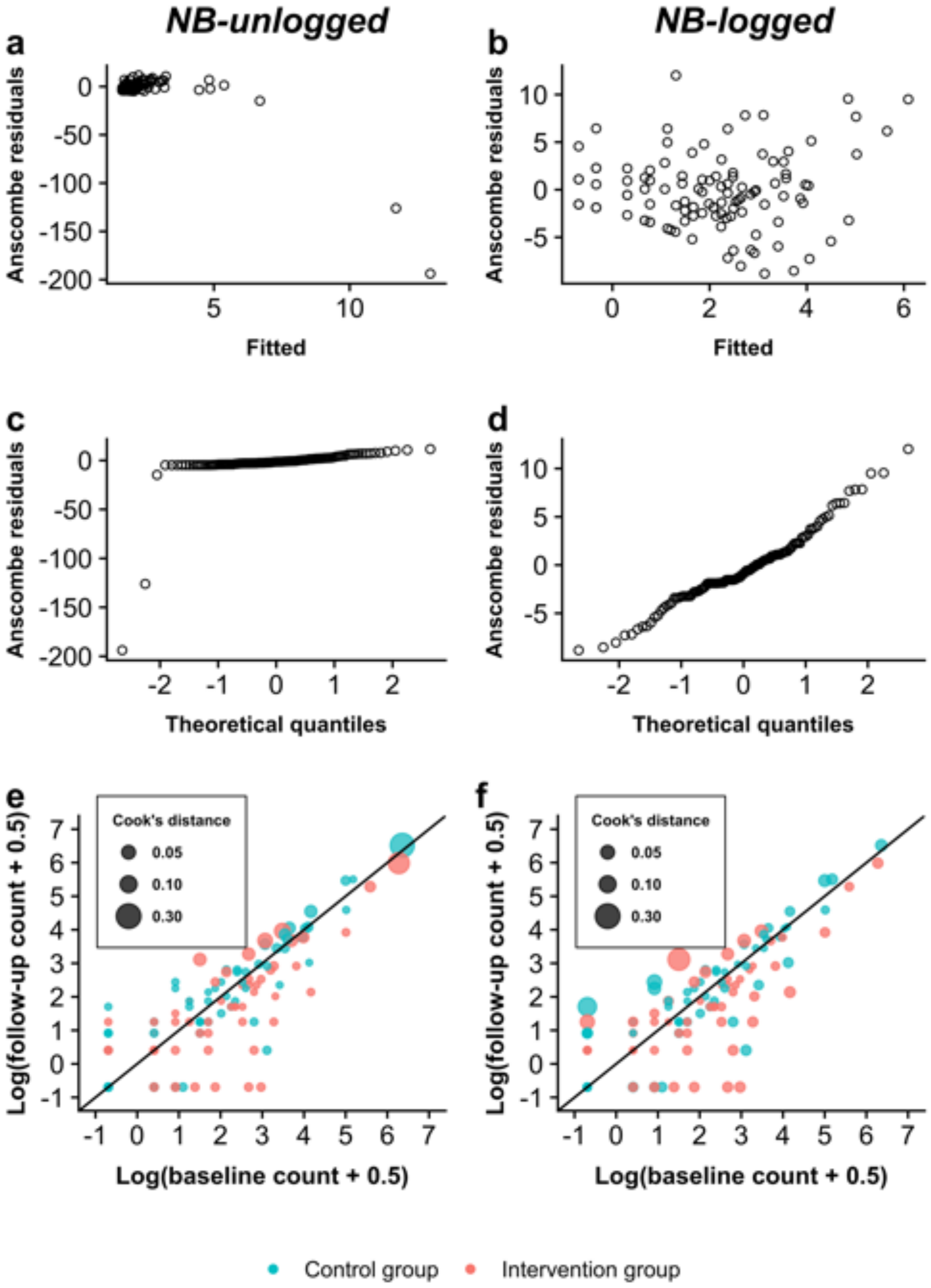


Figure 6-2 Goodwin et al. (2011) dataset: diagnostic plots from *NB-unlogged* versus *NB-logged* (n=124). **(a-b)** Anscombe residuals versus fitted values. **(c-d)** Normal Q-Q plot of Anscombe residuals. **(e-f)** The BOE plot presenting Cook's distance with x- and y- axes on a logarithmic scale (the diagonal line is the LoFE).

6.5 Simulation study and model comparison

6.5.1 Simulation datasets

The simulations were based on the Goodwin et al. dataset with some simplifications. For each simulation scenario 2000 trials were simulated from the mixed Poisson distribution (see section 6.3), with each trial comprised of m subjects (an example code is given in Appendix B). The first n subjects ($n = m/2$) were allocated to the control group and the rest to the intervention group. The length of baseline and outcome periods was assumed to be the same for all subjects ($t_0 = t_{i1} = 1$). The average baseline count (μ_{i0}) was set to 30, which is close to the number (around 28) reported in the Goodwin et al. dataset (see Table 3-1). We further assume that $\lambda_1 = \lambda_0$ so that for participants in the control group $\mu_{i1} = \mu_{i0} = 30$. A few sets of data were simulated and examined, and found to show a similar pattern to the Goodwin et al. data (not shown).

Twenty-four scenarios were considered expanding three factors: α, β, and sample sizes m:

- The variance of the gamma-distributed subject effect (α) was set at 3 to resemble $\hat{\alpha}$ from the CNB model fitted to the Goodwin et al. dataset (Table 6-1), and at 0.5 for less overdispersed data.
- An intervention effect close to that estimated from CNB in the Goodwin et al. dataset, $\beta = -0.4$ (FRR: 0.670), and a smaller intervention effect, $\beta = -0.2$ (FRR: 0.819), were considered for examining the power of the Wald test, while $\beta = 0$ (FRR: 1) was considered for examining the type I error rate.
- The datasets were simulated with total size (m) of 50, 100, 200, and 500, typical of small to medium sized falls prevention trials.

The NB, Poisson, and CNB models were fitted to each simulated dataset using R as described in section 6.3. From each fitted model, $\hat{\beta}$ and $\mathrm{SE}(\hat{\beta})$ were extracted, and the following statistics suggested by White (2010) to show the properties of the model estimators are reported:

$$\widehat{\mathrm{Bias}} = \mathrm{av}(\hat{\beta}) - \beta, \tag{6-23}$$

where $\mathrm{av}(\hat{\beta})$ is the average value (denoted as av) of the estimates of β from the 2000 datasets in a scenario. The Monte Carlo error (MCError) is reported to show certainty concerning $\widehat{\mathrm{Bias}}$:

$$\mathrm{MCError}\left(\widehat{\mathrm{Bias}}\right) = \frac{\mathrm{EmpSE}}{\sqrt{\mathrm{n_{sim}}}}, \tag{6-24}$$

where n_{sim} and EmpSE are the number of estimates and the empirical SE (that is, the standard deviation of $\hat{\beta}$), that is the standard deviation of $\hat{\beta}$ within the scenario. The ModSE is defined as the average of $\mathrm{SE}(\hat{\beta})$, that is, the average of the model-based SEs. White also suggested examining the model-based SEs from each model using the relative error, which is defined as:

$$\text{Relative Error} = \frac{\mathrm{ModSE}}{\mathrm{EmpSE}} - 1, \tag{6-25}$$

such that a positive relative error suggests that the model-based SE is overestimated, and *vice versa* for a negative relative error.

The following statistics were also computed across datasets within each scenario: $\mathrm{av}(\mathrm{HP})$, $\mathrm{av}(\hat{\alpha})$, $\mathrm{av}(\hat{\psi})$, and $\mathrm{av}(\hat{\phi})$. The datasets where the algorithm did not converge or yielded incorrect estimates (judged by $|\hat{\beta} - \beta| > 5$ or $\mathrm{SE}(\hat{\beta}) > 1$ were excluded, the selection criteria were chosen by inspecting the respective distributions of $\hat{\beta}$ and its SE). The proportion of simulated trials in which the null hypothesis of the Wald test of intervention effect was rejected was reported as the empirical power when $\beta \neq 0$, and the empirical type I error rate when $\beta = 0$. The empirical power and type I error rate of the score test for β were further examined. The P value of the score test was obtained from the `st.ml()` function in the **robNB** package (Aeberhard, 2016) in R.

To inspect the appropriateness of adding 0.5 in the log-transformation of the baseline count, separate simulations were carried out, with the same levels of α, β, and m as the main simulation, for *NB-logged* and *NB-offset* only, to compare their performance when different values (0.01, 0.1, and 1) are added.

6.5.2 Simulation results

In most cases, the algorithms of the models converged without raising errors (the number of successful convergences out of the 2000 repeats within each scenario are shown in Appendix C Table C-1).

As shown in Table 6-2, *NB-null* yielded the largest HPs in every scenario, with averages of HPs close to $\hat{\alpha}$ from the CNB model. The HPs from *NB-unlogged* are smaller than those from *NB-null*, but much larger than the HPs from *NB-logged* and *NB-offset* (the HPs of the latter two models are typically close). The estimates of ψ (the regression coefficient for the unlogged baseline count from *NB-unlogged*) and ϕ (the coefficient for the logged baseline count from *NB-logged*) are close to zero and one respectively.

Figure 6-3 shows the $\widehat{\text{Bias}}$ of $\hat{\beta}$ from the NB and CNB models, with the error bars showing the 95% CI calculated from MCError. Generally, $\hat{\beta}$ yielded by the NB and CNB models are close to the underlying value. The wide error bars for *NB-null* show that its estimates of the intervention effect have a higher variability than the other fitted models. Although the estimates of β from *NB-unlogged* have smaller variance than those from *NB-null*. The empirical SEs of $\hat{\beta}$ from *NB-unlogged* are larger than those from the *NB-logged*, *NB-offset*, and CNB models, which suggests that these three models are more efficient in estimating intervention effects. Also, the error bars from the *NB-logged*, *NB-offset*, and CNB models have similar widths when $\alpha = 3$ and $\alpha = 0.5$. In contrast, the error bars from *NB-null* were much wider when $\alpha = 3$, which suggests that $\hat{\beta}$ from *NB-null* has higher variability when the underlying distribution is more skewed (with more outliers), while the effect of outliers on the estimate of the intervention effect is mitigated by incorporating the baseline count.

The relative errors from each model are compared in Figure 6-4. Overall, the relative errors are small when $\alpha = 0.5$. The model-based SEs from *NB-null*, *NB-logged*, and *NB-offset* are typically lower than the empirical standard errors when the sample size is small, but the relative errors of the three models are generally low, especially for large sample sizes. When $\alpha = 3$, the model-based SEs from *NB-unlogged* are considerably larger than the corresponding empirical SEs, which agrees with the low type I error rate in the Wald test based on this model (Figure 6-5b).

The empirical power and type I error rates of the Wald test in NB and CNB models are presented in Figure 6-5. *NB-null* has the lowest empirical power, although its empirical type I error rates are relatively close to the nominal level (0.05). Because of the extra information from including the untransformed baseline count, *NB-unlogged* achieved greater power than *NB-null*, but the improvement when α=3 is not as large as when α=0.5. *NB-logged* and *NB-offset* result in almost identical power, which are substantially higher than the power for *NB-unlogged* in all scenarios, and they are only marginally less powerful than CNB. Similar to the CNB model, the power of the Wald test in *NB-logged* and *NB-offset* are less affected by change of α than that of *NB-unlogged*.

The CNB model not only has the greatest power, its type I error rates are also more stable than the other models—they are typically close to the nominal level regardless of the sample size. The type I error rates of *NB-null*, *NB-logged*, and *NB-offset* are moderately higher than the nominal level of 0.05 for small sample sizes (the maximum rate for *NB-logged* is 0.071 when m=50 and α=3), but approach 0.05 as the sample size increases; while the type I error rate for *NB-unlogged* is consistently deflated when α=3, without showing any trend of convergence towards 0.05, even when m=500. Type I error rates in *NB-offset* are closer to 0.05 than *NB-logged*, but the difference is small.

The simulations were repeated to assess the performance of the score test for the four NB models. For *NB-null*, *NB-logged*, and *NB-offset*, the empirical type I error rates of the score test are closer to the nominal level than the Wald test when the sample size is small (Figure 6-6b). The type I error rates of the score test in *NB-unlogged* deviate further from 0.05 than for the Wald test. For the scenarios with $\alpha = 3$, the test shifts from being liberal when the sample size is small (50 and 100) to being conservative when the sample size is large (200 and 500).

Different values (0.01, 0.1, 0.5, and 1) were added to the baseline count before log-transformation and the resultant *NB-logged*/*NB-offset* models were compared in simulations regarding the estimation and hypothesis testing of intervention effects. The results of the simulations show that the values examined do not have a large impact on the estimation of β (Table 6-3) or the Wald test (Table 6-4). The results show that adding different values (between 0.01 and 1) does not substantially change the model estimation.

As shown in Figure 6-7, $\hat{\beta}$ are generally close to the underlying values in Poisson models, except when $\beta = -0.4$ and $\alpha = 3$, in which *Poi-unlogged* underestimated $\hat{\beta}$. The relative error plot (Figure 6-8) illustrates that $\mathrm{SE}(\hat{\beta})$ are underestimated in every Poisson model, leading to the extremely inflated type I error rates (Figure 6-9b). Note that *Poi-logged* and *Poi-offset* have lower type I error rates than the other two Poisson models, and they also have higher power (Figure 6-9a).

Table 6-2 Estimates of HP from NB models, $\hat{\alpha}$ from CNB, $\hat{\psi}$ from *NB-unlogged*, and $\hat{\phi}$ from *NB-logged*

			$\mathrm{av(HP)}$				$\mathrm{av}(\hat{\alpha})$	$\mathrm{av}(\hat{\psi})$	$\mathrm{av}(\hat{\phi})$
α	β	m	*NB-null*	*NB-unlogged*	*NB-logged*	*NB-offset*	CNB	*NB-unlogged*	*NB-logged*
3	-0.4	50	2.931	1.152	0.018	0.019	2.988	0.033	1.017
		100	2.951	1.192	0.019	0.019	2.981	0.032	1.018
		200	2.982	1.223	0.018	0.019	2.994	0.031	1.018
		500	2.989	1.233	0.019	0.019	2.997	0.031	1.018
	-0.2	50	2.954	1.195	0.019	0.020	3.012	0.033	1.021
		100	2.969	1.225	0.019	0.020	3.000	0.032	1.019
		200	2.980	1.244	0.019	0.019	2.997	0.031	1.018
		500	2.989	1.257	0.020	0.020	2.996	0.031	1.018
	0	50	2.918	1.204	0.020	0.021	2.971	0.033	1.020
		100	2.972	1.250	0.020	0.021	2.999	0.032	1.019
		200	2.988	1.271	0.021	0.021	2.998	0.031	1.019
		500	2.994	1.284	0.021	0.021	3.000	0.031	1.018
0.5	-0.4	50	0.479	0.087	0.027	0.029	0.488	0.029	0.930
		100	0.491	0.093	0.027	0.029	0.496	0.028	0.929
		200	0.497	0.096	0.029	0.030	0.499	0.028	0.928
		500	0.499	0.098	0.029	0.030	0.500	0.028	0.929
	-0.2	50	0.484	0.089	0.026	0.029	0.493	0.029	0.930
		100	0.490	0.093	0.027	0.029	0.495	0.028	0.931
		200	0.496	0.097	0.029	0.030	0.498	0.028	0.929
		500	0.497	0.098	0.029	0.031	0.498	0.028	0.927
	0	50	0.483	0.090	0.027	0.030	0.491	0.029	0.928
		100	0.489	0.094	0.028	0.030	0.494	0.029	0.927
		200	0.494	0.098	0.030	0.031	0.497	0.028	0.926
		500	0.498	0.100	0.030	0.031	0.499	0.028	0.928

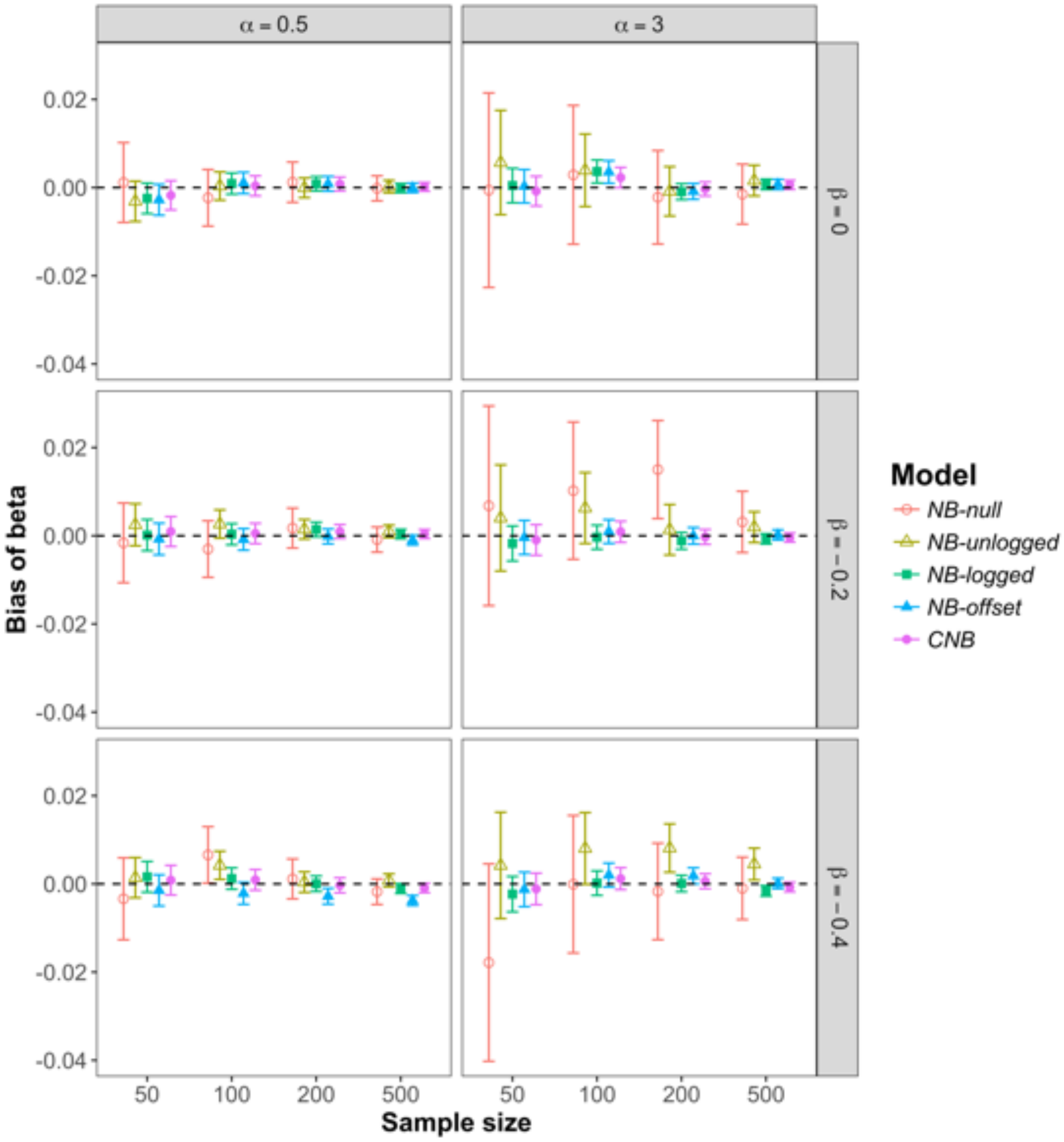


Figure 6-3 Bias plot of *NB-null*, *NB-unlogged*, *NB-logged*, *NB-offset*, and CNB. The $\widehat{\text{Bias}}$ of $\hat{\beta}$ are shown as the points with error bars (the 95% CI calculated from the MCError of $\widehat{\text{Bias}}$).

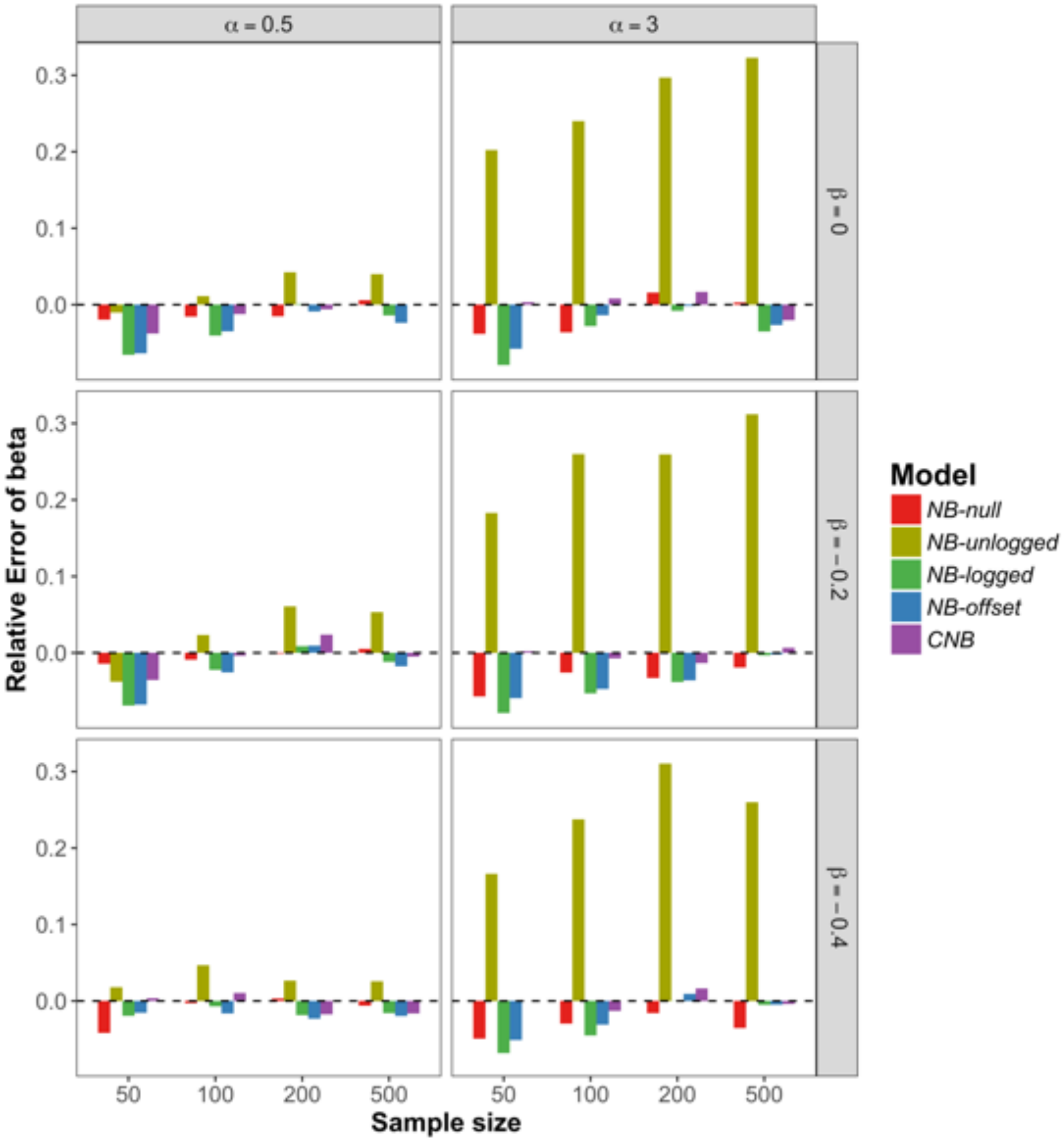


Figure 6-4 Relative error plot of *NB-null*, *NB-unlogged*, *NB-logged*, *NB-offset*, and CNB.

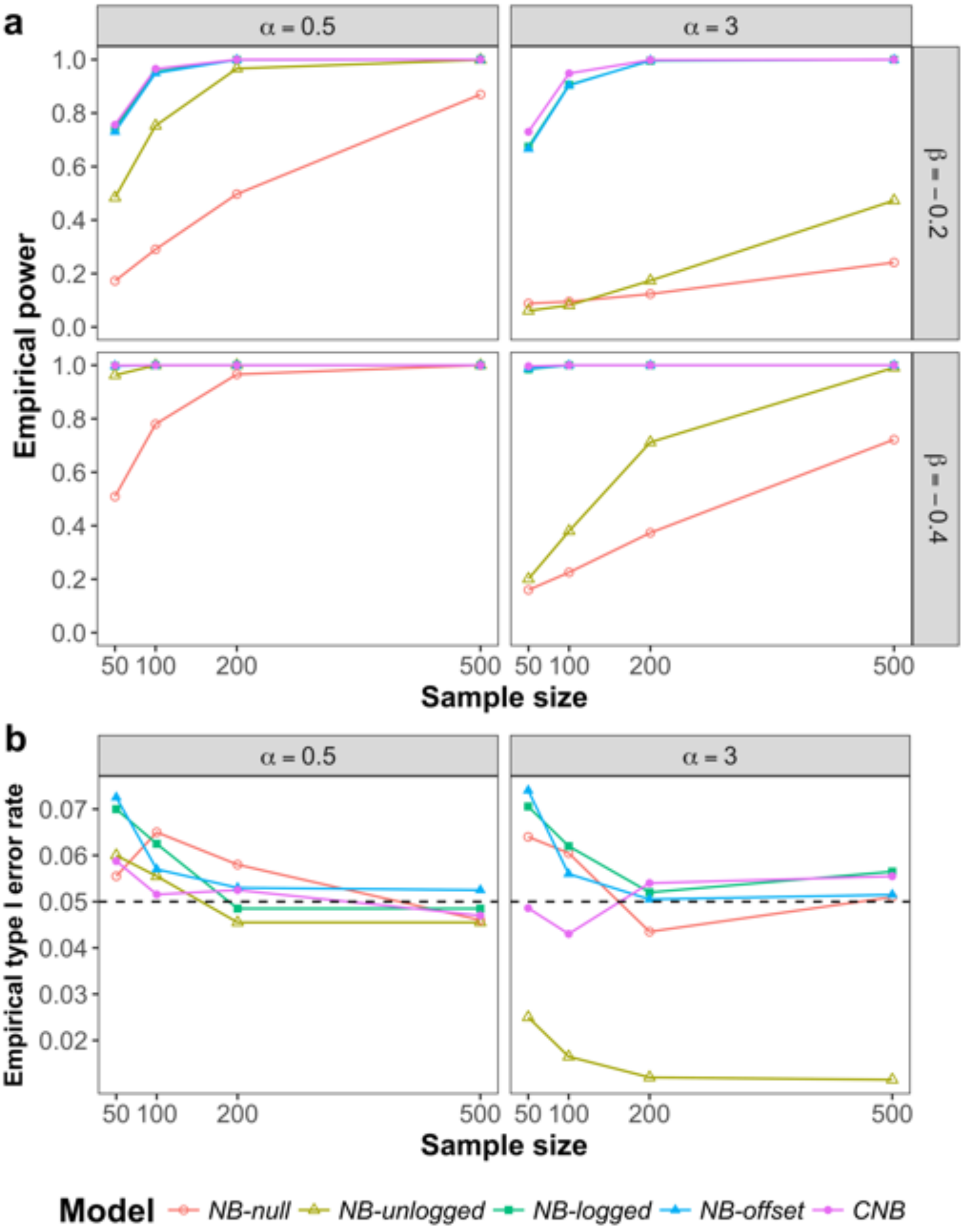


Figure 6-5 Performance of the Wald test from *NB-null*, *NB-unlogged*, *NB-logged*, *NB-offset*, and CNB in simulations. **(a)** Empirical Power; **(b)** Empirical type I error rates.

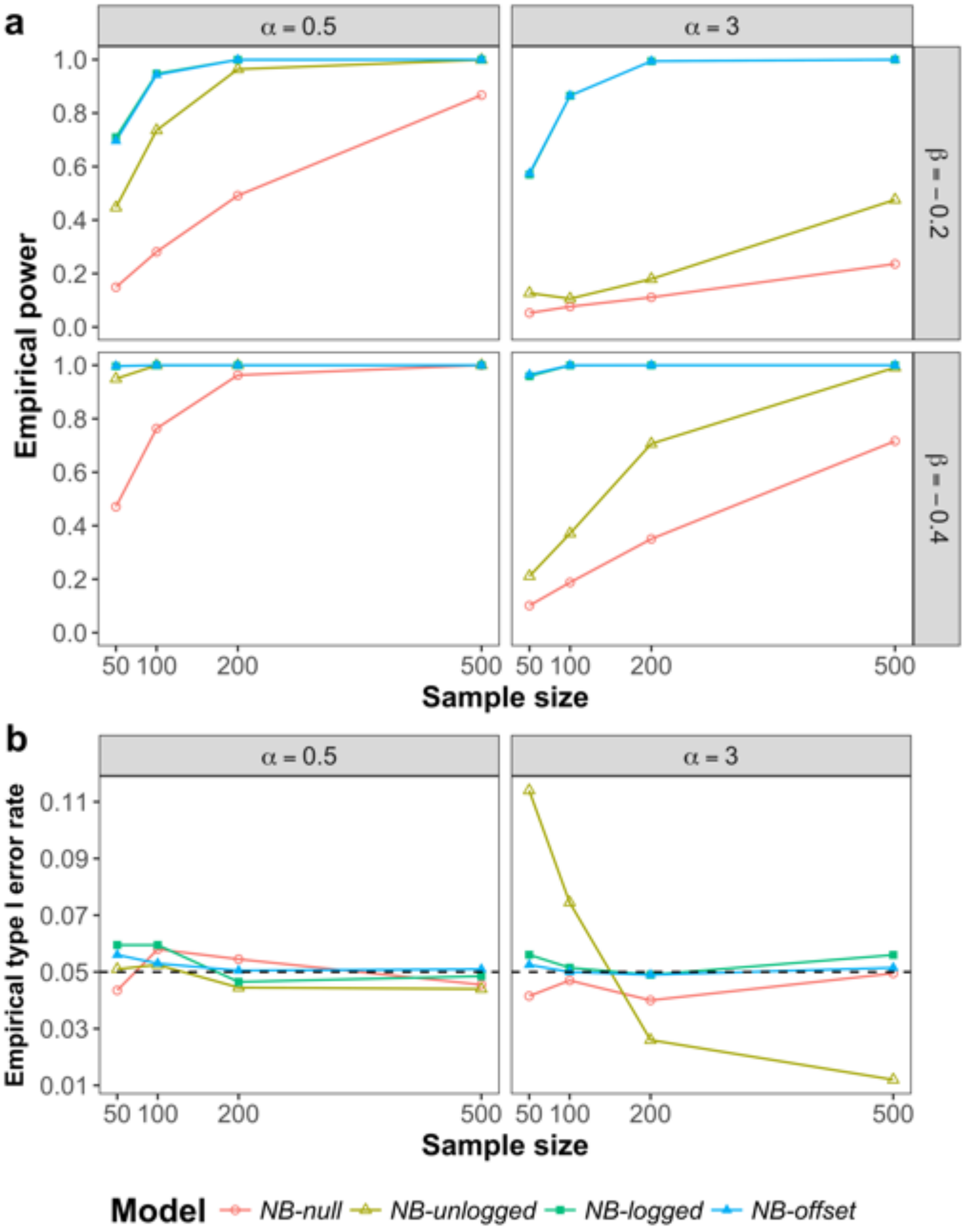


Figure 6-6 Performance of the score test from *NB-null*, *NB-unlogged*, *NB-logged*, and *NB-offset* in simulations. **(a)** Empirical power; **(b)** Empirical type I error rates. (Note that the scaling of y-axis is different to that in Figure 6-5b).

Table 6-3 The av($\hat{\beta}$) from *NB-logged* and *NB-offset* in each simulation scenario, with different values (0.001, 0.1, 0.5, and 1) added before log-transformation.

			NB-logged				***NB-offset***			
α	β	m	+0.01	+0.1	+0.5	+1	+0.01	+0.1	+0.5	+1
3	-0.4	50	-0.401	-0.401	-0.402	-0.403	-0.407	-0.404	-0.401	-0.398
		100	-0.397	-0.398	-0.400	-0.401	-0.402	-0.401	-0.398	-0.395
		200	-0.397	-0.398	-0.400	-0.401	-0.402	-0.401	-0.398	-0.395
		500	-0.398	-0.400	-0.402	-0.403	-0.403	-0.403	-0.400	-0.397
	-0.2	50	-0.201	-0.201	-0.202	-0.202	-0.203	-0.202	-0.200	-0.198
		100	-0.198	-0.199	-0.200	-0.201	-0.201	-0.201	-0.199	-0.197
		200	-0.199	-0.200	-0.201	-0.202	-0.202	-0.202	-0.200	-0.198
		500	-0.199	-0.200	-0.201	-0.201	-0.202	-0.201	-0.200	-0.198
	0	50	-0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000
		100	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
		200	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
		500	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.5	-0.4	50	-0.398	-0.398	-0.398	-0.399	-0.403	-0.402	-0.401	-0.401
		100	-0.398	-0.399	-0.399	-0.399	-0.403	-0.403	-0.402	-0.401
		200	-0.400	-0.400	-0.400	-0.400	-0.404	-0.404	-0.403	-0.402
		500	-0.401	-0.401	-0.401	-0.401	-0.405	-0.405	-0.404	-0.403
	-0.2	50	-0.200	-0.200	-0.200	-0.200	-0.201	-0.201	-0.201	-0.200
		100	-0.200	-0.200	-0.200	-0.200	-0.201	-0.201	-0.201	-0.200
		200	-0.198	-0.198	-0.199	-0.199	-0.201	-0.201	-0.200	-0.200
		500	-0.200	-0.200	-0.200	-0.200	-0.202	-0.201	-0.201	-0.201
	0	50	-0.002	-0.002	-0.002	-0.002	-0.003	-0.003	-0.003	-0.003
		100	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		200	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 6-4 Positive rate of the Wald test from *NB-logged* and *NB-offset* within each simulation scenario, with different values (0.001, 0.1, 0.5, and 1) added before log-transformation. The table presents the empirical Power when $\beta \neq 0$ and the empirical type I error rate when $\beta = 0$.

			NB-logged				***NB-offset***			
α	β	m	+0.01	+0.1	+0.5	+1	+0.01	+0.1	+0.5	+1
3	-0.4	50	0.963	0.982	0.985	0.981	0.985	0.986	0.988	0.980
		100	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	-0.2	50	0.627	0.660	0.674	0.658	0.668	0.668	0.666	0.654
		100	0.869	0.894	0.906	0.886	0.896	0.900	0.904	0.892
		200	0.990	0.994	0.996	0.995	0.997	0.996	0.997	0.994
		500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0	50	0.075	0.073	0.071	0.074	0.072	0.072	0.074	0.074
		100	0.060	0.059	0.062	0.060	0.058	0.057	0.056	0.061
		200	0.051	0.055	0.052	0.051	0.053	0.052	0.050	0.048
		500	0.054	0.053	0.056	0.060	0.061	0.057	0.052	0.054
0.5	-0.4	50	0.996	0.998	0.998	0.998	0.998	0.998	0.998	0.999
		100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	-0.2	50	0.738	0.742	0.746	0.746	0.722	0.722	0.730	0.738
		100	0.952	0.954	0.954	0.953	0.947	0.948	0.950	0.954
		200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0	50	0.068	0.068	0.070	0.074	0.072	0.071	0.073	0.072
		100	0.062	0.060	0.062	0.064	0.057	0.056	0.057	0.058
		200	0.049	0.048	0.048	0.050	0.052	0.052	0.053	0.053
		500	0.052	0.052	0.048	0.049	0.052	0.053	0.052	0.052

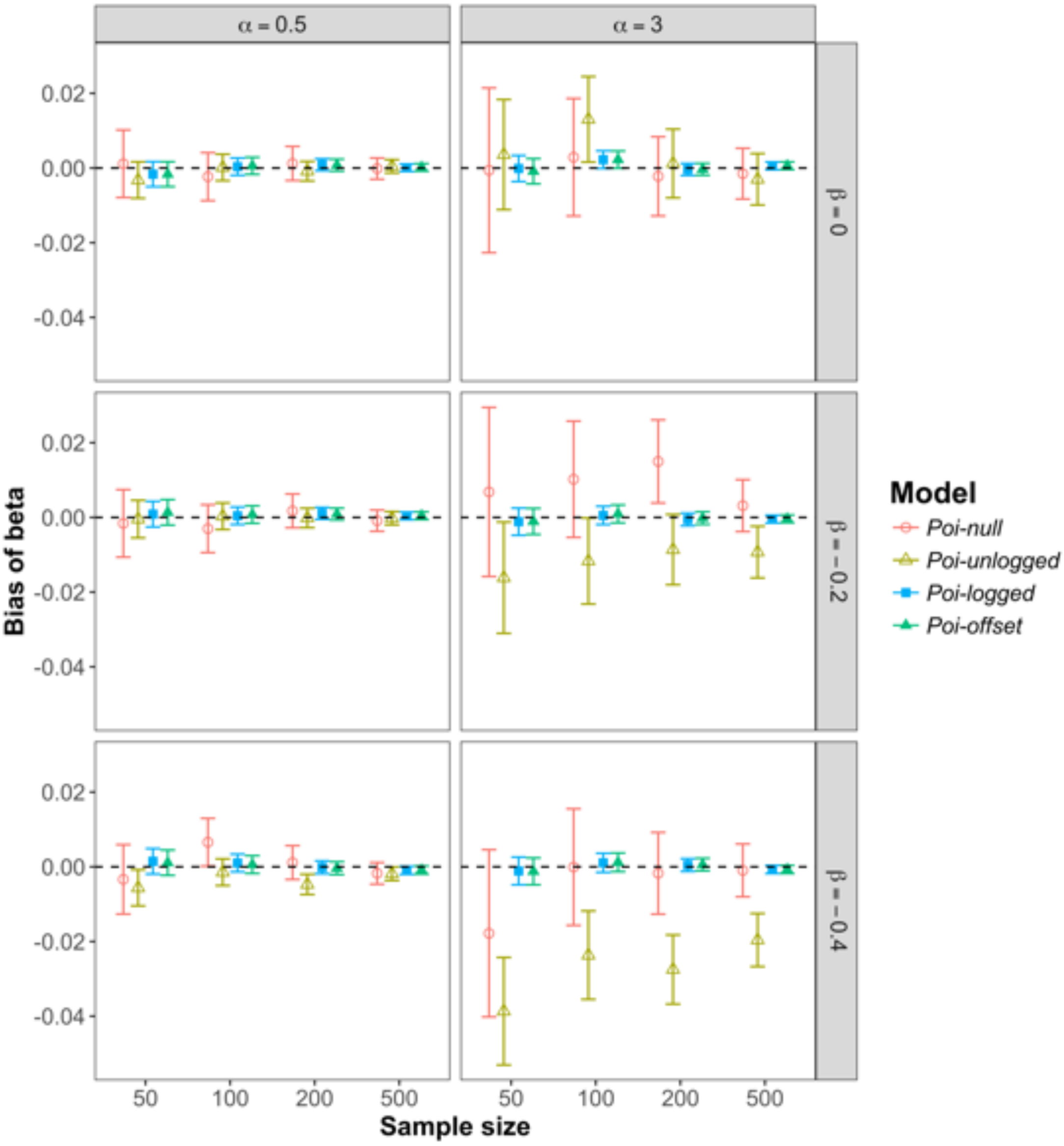


Figure 6-7 Bias plot of *Poi-null*, *Poi-unlogged*, *Poi-logged*, and *Poi-offset*. The $\widehat{\mathrm{Bias}}$ of $\hat{\beta}$ are shown as the points with error bars (the 95% CI calculated from the MCError of $\widehat{\mathrm{Bias}}$).

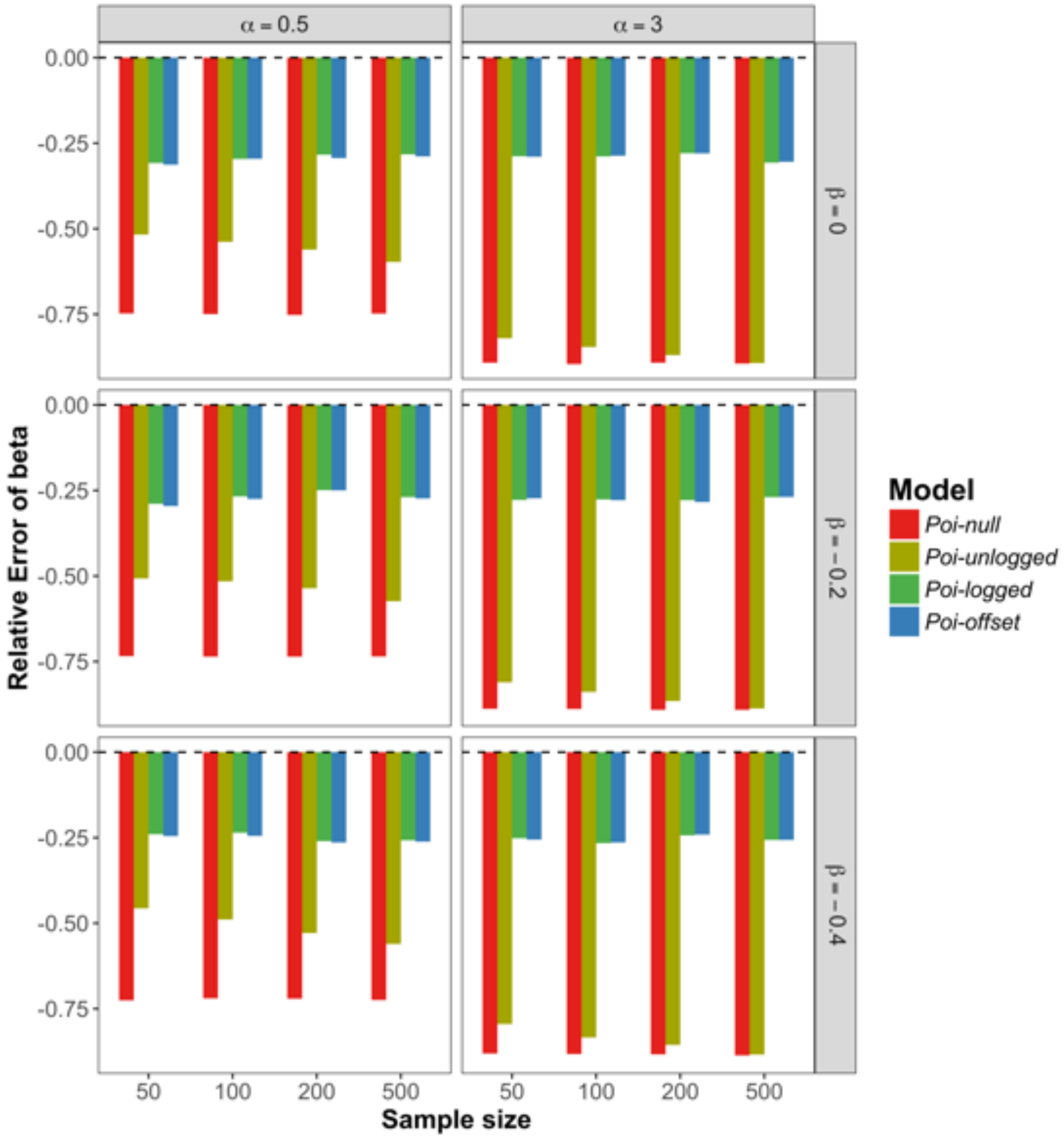


Figure 6-8 Relative error plot of *Poi-null*, *Poi-unlogged*, *Poi-logged*, and *Poi-offset*. (Note that the scaling of y-axis is different to that in Figure 6-4).

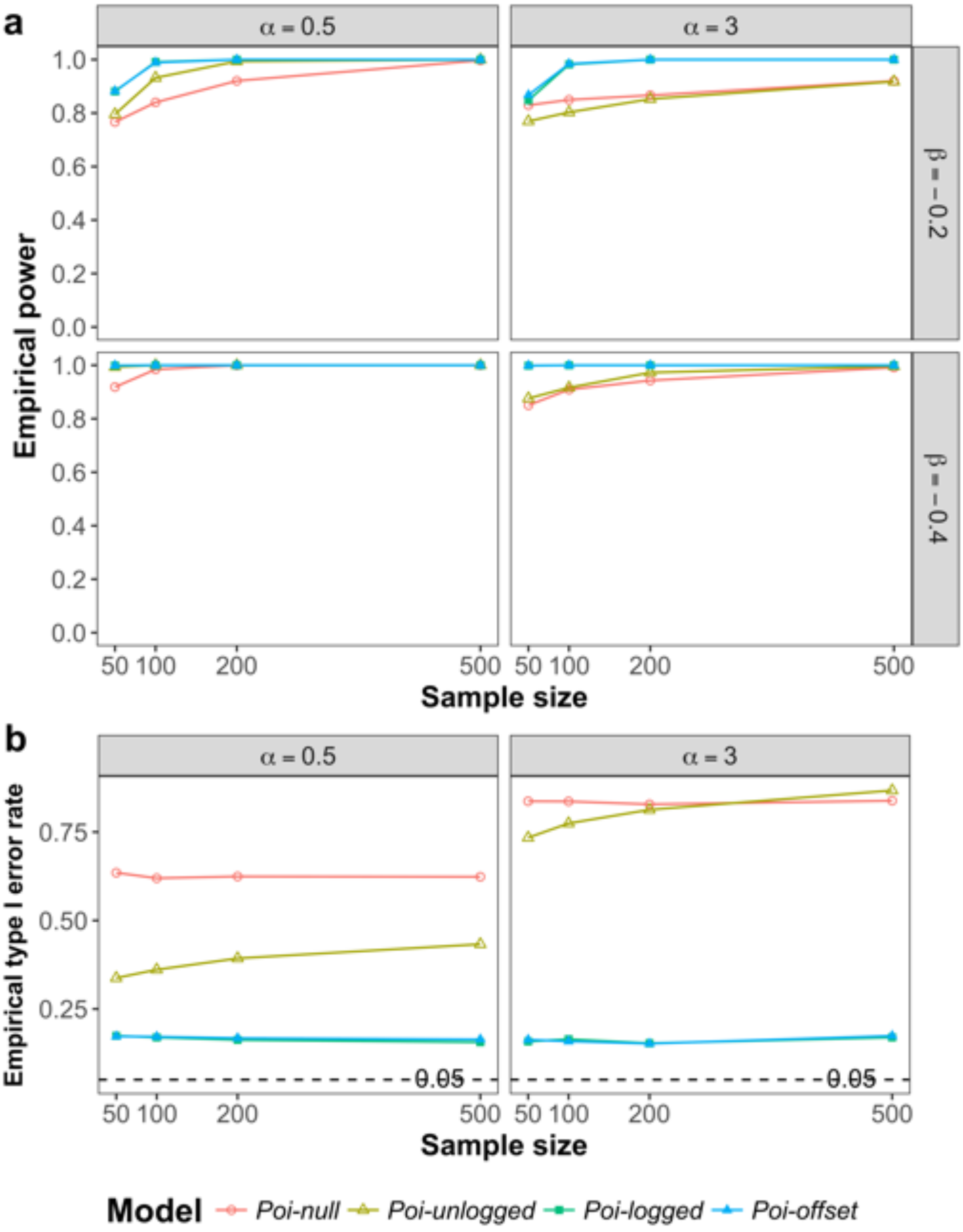


Figure 6-9 Performance of the Wald tests from *Poi-null*, *Poi-unlogged*, *Poi-logged*, and *Poi-offset* in simulations. **(a)** Empirical power; **(b)** Empirical type I error rates. (Note that the scaling in y-axis is different to those in Figure 6-5b and Figure 6-6b).

6.6 Sample size calculation for *NB-null, NB-logged,* and *NB-offset*

Sample size for *NB-null*

Zhu and Lakkis (2014) proposed a sample size formula for the LR test based on NB model, without a baseline count, that is, the *NB-null* model.

Let the ratio of the size of intervention group to that of control group be denoted by ρ. Let the average outcome falls rate for the control group be denoted by $\lambda_a = \lambda$ and let the corresponding falls rate for the intervention group be $\lambda_b = \lambda \exp(\beta)$. The null hypothesis for the test of $\beta = 0$ is:

$$H_0: \lambda_b/\lambda_a = 1$$

And the alternative:

$$H_1: \lambda_b/\lambda_a \neq 1$$

To distinguish it from α in the CNB model, the level of significance for the test is denoted by α^*. To achieve the power of $1 - \varphi$, Zhu and Lakkis show the number required in the control group to be:

$$m_a = \frac{\left(z_{\alpha^*/2}\sqrt{V_0} + z_\varphi\sqrt{V_1}\right)^2}{\beta^2}, \tag{6-26}$$

where $z_{\alpha^*/2} = \Phi^{-1}(\alpha^*/2)$ and $z_\varphi = \Phi^{-1}(\varphi)$; $\Phi(.)$ is the cumulative density function of the standard normal distribution; V_1 is the estimate of $m_a\mathrm{Var}(\hat{\beta})$ under the alternative hypothesis and is given by:

$$V_1 = \frac{1}{t_1}\left(\frac{1}{\lambda} + \frac{1}{\rho\lambda\exp(\beta)}\right) + \frac{(1+\rho)\alpha}{\rho}, \tag{6-27}$$

and V_0 is the estimate of $m_a\mathrm{Var}(\hat{\beta})$ under the null hypothesis. V_0 can be estimated using three approaches:

- Approach 1: because $\lambda_a = \lambda_b = \lambda$ under H_0, V_0 can be based on the rate in the control group, giving:

$$V_{01} = \frac{1+\rho}{t_1\rho\lambda} + \frac{(1+\rho)\alpha}{\rho}. \tag{6-28}$$

- Approach 2: V_0 can be based on the rates of both groups (λ_a and λ_b), so that:

$$V_{02} = V_1. \tag{6-29}$$

- Approach 3: maximizing the log-likelihood function underlying the LR test with λ_a/λ_b constrained to be 1 yields an MLE of the overall events rate, and based on this V_0 can be estimated by:

$$V_{03} = \frac{(1+\rho)^2}{t_1\rho(\lambda + \rho\lambda\exp(\beta))} + \frac{(1+\rho)\alpha}{\rho}. \tag{6-30}$$

The number required to achieve power of $1 - \varphi$ in the intervention group is:

$$m_b = \rho m_a, \tag{6-31}$$

and the total number required is $m_a + m_b$. The total number required by Approaches 1, 2, and 3 are referred to as m_1, m_2, and m_3, respectively.

The simulations in Zhu and Lakkis's paper showed that the sample sizes calculated using equation (6-26) generally achieved empirical power close to the nominal level 80%. The authors found that m_2 and m_3 reached the target 80% power in both the Wald and LR test in most scenarios, whilst m_1 underestimated the sample sizes in some scenarios.

Approximate sample size for *NB-logged* and *NB-offset*

Tango (2009) proposed a conditional score test for β given the baseline count y_{i0}. The test is derived from the same joint distribution, in equations (6-5) and (6-6), as used to derive the CNB (6-9). Unlike the CNB and NB models, the conditional score test does not requires specification of the distribution of the random subject effect s_i. A formula for sample size calculation of the two-tailed conditional score test is given by:

$$\begin{aligned} m = \frac{1}{\mu_0 t_0 k\theta(\exp(\beta) - 1)} \Bigg(& z_{\alpha^*/2}\sqrt{\frac{2(\exp(\beta)+1)(1+k\theta)(1+k\theta\exp(\beta))}{2+k\theta(\beta+1)}} \\ & + z_\varphi \sqrt{\frac{\beta(1+k\theta)^3 + (1+k\theta\exp(\beta))^3}{(1+k\theta)(1+k\theta\exp(\beta))}}\Bigg)^2 \end{aligned} \tag{6-32}$$

for power $1 - \varphi$, where $k = t_1/t_0$ is the ratio of duration and $\theta = \lambda_1/\lambda_0$ is the ratio of falls rates (that is, a period effect).

Tango showed that the conditional score test and CNB model resulted in almost identical estimates and CIs when fitted to a dataset from a trial of epileptic patients. As *NB-logged* and *NB-offset* had similar empirical power to the CNB model (see Figure 6-5a), Tango's sample size calculation may be useful as an approximation to the sample size required for *NB-logged* and *NB-offset*. A simulations study was conducted to check this.

We set $\lambda_1 = \lambda_0 = \lambda$ and $t_1 = t_0 = 1$. Twelve scenarios were considered spanning the combinations of β (-0.3, -0.2, and -0.1), α (0.5 and 3), and λ (15 and 30). The required m for each of the 12 scenarios was calculated by Tango's formula to achieve a power 80% in a $\alpha^* = 5\%$ level test. For each scenario 2000 datasets were simulated from the mixed Poisson distribution described in section 6.2.1, with balanced group size. The empirical power of the Wald test from each model was calculated to examine whether m from Tango's formula was sufficient to achieve 80% power.

Table 6-5 Sample size calculated from Tango's conditional score test and Zhu and Lakkis's formulae with $1 - \varphi = 80\%$ and α^*=5% ($t = 1$)

					Calculated sample size			
					Tango	Zhu and Lakkis*		
Scenario ID	α	λ	β	FRR: exp (β)	m	m_1	m_2	m_3
1	3	15	-0.1	0.905	430	9632	9640	9640
2	3	15	-0.2	0.819	112	2410	2414	2414
3	3	15	-0.3	0.741	52	1072	1074	1074
4	3	30	-0.1	0.905	216	9526	9530	9530
5	3	30	-0.2	0.819	56	2382	2384	2384
6	3	30	-0.3	0.741	26	1060	1062	1060
7	0.5	15	-0.1	0.905	430	1784	1792	1790
8	0.5	15	-0.2	0.819	112	448	452	452
9	0.5	15	-0.3	0.741	52	200	202	202
10	0.5	30	-0.1	0.905	216	1678	1680	1680
11	0.5	30	-0.2	0.819	56	420	422	422
12	0.5	30	-0.3	0.741	26	188	190	188

* The sample size was calculated for a trial with balanced group size ($\rho = 1$)

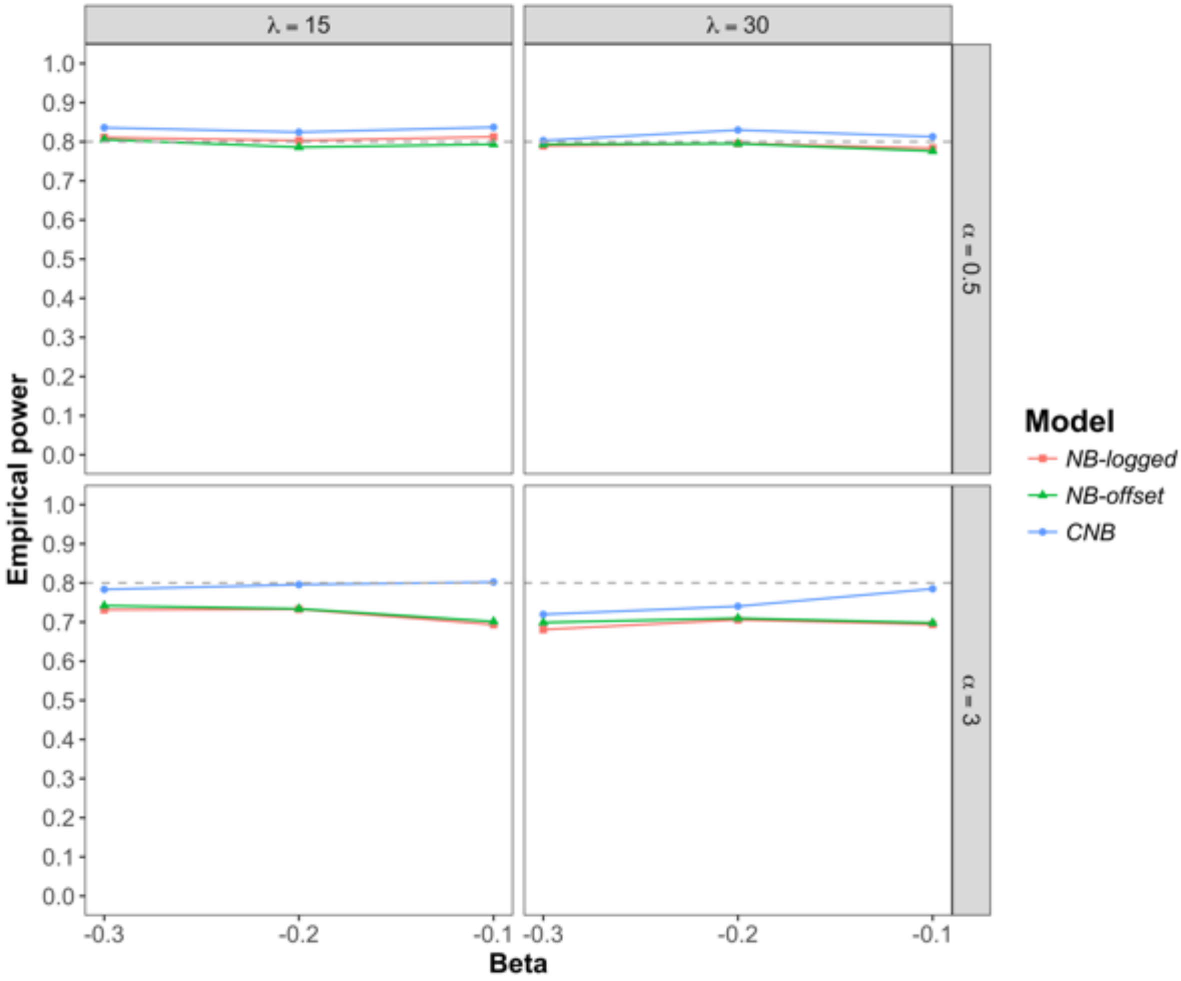


Figure 6-10 Empirical power of the Wald test based on *NB-logged*, *NB-offset*, and CNB models with the sample size calculated from the formula for Tango's score test with 80% power and a 5% significance level.

The sample sizes for the twelve simulation scenarios calculated from Tango's formula are displayed in Table 6-5. In the same table we also present the sample sizes m_1, m_2, and m_3 for *NB-null,* calculated by the Zhu and Lakkis formulae, for comparison. The three approaches from the Zhu and Lakkis formula result in similar sample sizes for *NB-null*, but the sample sizes are typically very large. Conditioning on the baseline count results in a remarkable reduction in the required sample size, especially when the outcome count (λt) is large, the intervention effect is small, or heterogeneity is great.

The empirical powers of the *NB-logged*, *NB-offset*, and CNB models obtained from simulations of size equal to m from Tango's formula to achieve power 80% are summarized in Figure 6-10. When $\alpha = 0.5$, the empirical power for all three models were relatively close to 80%. When $\alpha = 3$, the empirical powers for the CNB model was relatively close to

80% for $\lambda = 15$ but lower than 80% for $\lambda = 30$; while the empirical power for *NB-logged* and *NB-offset* was between 70% and 75%. This shows that Tango's equation (6-32) can be used to approximate the required sample size for smaller levels of heterogeneity, but the number could be inflated when a considerable heterogeneity is anticipated.

6.7 Discussion

NB regression has been widely used for analysing falls data. It is common to collect the falls count during a baseline period in a falls prevention trial, but it remains a question as how to incorporate a baseline count in statistical modelling. One approach is Cook and Wei's (2003) CNB model. The simulations in this chapter showed that CNB resulted in the highest power for the Wald test of the intervention effect among the compared models, and the tests had type I error rates closer to nominal level even for the smallest sample sizes considered in simulations. However, CNB is not currently supported in any statistical package. Another approach is to incorporate the baseline count using NB regression, which is supported in most popular statistical packages and commonly used in practice.

NB models ignoring the baseline count (*NB-null*) were examined. The empirical power from *NB-null* was noticeably lower than those from other NB and CNB models in simulations. *NB-unlogged* including the baseline count as a covariate without any transformation, was more powerful than *NB-null*, even though the scaling of the baseline count is not appropriate. However, *NB-unlogged* is conservative when $\alpha = 3$, even for the largest sample sizes ($m = 500$) examined in the simulations.

In the simulations in section 6.5, *NB-logged* and *NB-offset*, the NB models incorporating the log-transformed baseline count, had satisfying performance. They yielded $\hat{\beta}$ with smaller bias and variability than the estimates from *NB-unlogged*. The two models also produced more accurate SE estimates for $\hat{\beta}$. They were more powerful in testing the intervention effect than *NB-unlogged*, and they typically resulted in small HPs. Compared to the benchmark model (CNB) *NB-logged* and *NB-offset* were only slightly less powerful, and the disparity in power diminished as the sample size increased. Overall, *NB-logged* and *NB-offset* produced similar results: their estimates of β were similar, and $\hat{\phi}$ in *NB-logged* was generally close to one. This suggests that the logged baseline count, when included as an explanatory variable in NB regression, appropriately accounts for the relationship with the outcome.

The difference between HP and $\hat{\alpha}$ has been previously discussed in section 6.2.1. Loosely speaking, HP shows how much variability in the outcome count remains unexplained by the covariates in the model, while $\hat{\alpha}$ reflects how much heterogeneity in the outcome count can be accounted for by the baseline count. The HP from *NB-null* were similar to the $\hat{\alpha}$ from CNB in simulations, which suggests that the heterogeneity was accounted for in the model solely by the random subject effects. This conforms to the data generating mechanism underlying the simulations. *NB-unlogged*, which is essentially *NB-null* with an additional covariate (the untransformed baseline count) achieved a reduction in HP compared to *NB-null*. The HP were further reduced in *NB-logged* and *NB-offset*, which reflects the fact that the log-transformation yields a more appropriate scale for the baseline count as a covariate in NB regression. Poisson models do not accommodate overdispersion and hence are too liberal. Some might wonder: since the subject effect is shared in the baseline and outcome count, shouldn't including the logged baseline count as a covariate obviate overdispersion so that it is unnecessary to use NB models rather than a Poisson? The simulations showed this not to be the case: the type I errors of *Poi-logged* and *Poi-offset* were still too high (around 0.16); also, in the Goodwin et al. dataset, the AIC of these two Poisson models were higher than the AIC of *NB-null*, which does not even incorporate the baseline count.

For *NB-logged* and *NB-offset*, the empirical type I error rates of the Wald test were higher than the nominal level (0.05) for small sample sizes. As the sample size increases, the rates converged to 0.05. Aban et al. (2009) conducted a simulation study to compare the two types of hypothesis tests in two-group NB comparisons. They reported that the type I error rates for the Wald test were higher than 0.05 when the sample size was small (<200), as also shown in Figure 6-5b. Aeberhard et al. (2017) conducted a similar simulation study and reached the same conclusion. They recommended using the robust TETT (Tilted Exponential Tilting Test) for analysing small samples. To my knowledge, this test is only available in the R package **robNB** provided by the authors (Aeberhard, 2016). This chapter focused on the Wald test because it is the default hypothesis test in many statistical packages with functionality for NB modelling. When the sample size reached 200, the type

I error rates of the Wald test were only slightly higher than 0.05. Datasets were simulated with sample size 50, 100, 200, and 500, the range encompasses 80% of falls prevention trials included in a Cochrane interview (Gillespie et al., 2012). The simulations regarding the score test showed it to be more reliable than the Wald test when sample size is small: the type I error rates of the score test were closer to the nominal level than those of Wald test. However, NB-based score tests are not widely supported in statistical packages.

The results of the models fitted to the Goodwin et al. dataset substantiated the conclusions from the simulation study. The AIC of *NB-unlogged* was smaller than that of *NB-null*. *NB-logged* and *NB-offset* resulted in smaller AIC than *NB-unlogged*, and the Wald test did not indicate significance for the intervention effect (P=0.098) in *NB-unlogged*, while significance was shown in *NB-logged* (P=0.021) and *NB-offset* (P=0.032). This is in line with the simulation results that *NB-unlogged* has low power and is overly conservative. The model diagnostics showed that *NB-unlogged* cannot accommodate the large outcome counts in the Goodwin et al. dataset: they typically had large Cook's distances, but the issue does not appear to occur for *NB-logged* and *NB-offset*.

Because zero baseline counts cannot be logged, the pragmatic approach of adding 0.5 to the baseline counts before the log-transformation was used in this chapter. There is a trade-off in choosing the value for addition: if a smaller value, say 10^{-10}, is added to a baseline count n $(n > 0)$, the value of $\log{(n + 10^{-10})}$ would be closer to $\log{(n)}$ than $\log{(n + 0.5)}$, but $\log{(0 + 10^{-10})}$ would become -23, a large negative value. The choice of 0.05 is a standard continuity correction in practice and the simulations showed the results not to be sensitive to the choice between 0.01 and 1.

Tango (2009) proposed a formula for calculating the sample size for the conditional score test. The simulations in section 6.6 showed that this formula can also be used to calculate the required sample size of *NB-logged* and *NB-offset* when the degree of heterogeneity is low. For higher heterogeneity, the sample size calculated from the formula tend to be underestimated, so the number should be inflated to reach a required power. Alternatively, the sample size can be calculated using simulation-based methods, and the Tango formula can be used to provide a starting value for sample size. Compare the sample size calculated from the Tango formula and the Zhu and Lakkis (2014) formulae (which estimate the sample size required for *NB-null*), it is apparent that conditioning on the baseline count considerably reduce the required sample size for a falls prevention trial. The sample size calculated using the Zhu and Lakkis formulae are generally too large for falls prevention trials in PwP.

In conclusion, this chapter showed that NB models including logged baseline count as a covariate/offset is a viable alternative to the CNB model. A baseline falls count has great value when analysing a falls outcome count. It is generally recognized that a pre-randomisation baseline value of the outcome should be collected when designing an RCT (Assmann et al., 2000), and this also holds true when the outcome is a count in a falls prevention trial. *NB-logged* and *NB-offset* can be fitted in all statistical packages that support NB modelling. For medium to large sample sizes, *NB-logged* and *NB-offset* are almost as powerful as CNB, and the type I error rates of the Wald and score test are close to the nominal level. They have great practicality and are easily accessible for applied statisticians.

Chapter 7

Comparison of NB-based and CNB-based methods when the underlying assumption does not hold

7.1 Introduction

The previous chapter showed that when the CNB model is correctly specified, including the log-transformed baseline count in NB regression as a covariate (*NB-logged*) or offset (*NB-offset*) achieves comparable power to CNB in many circumstances. In this chapter, the *NB-logged* and *NB-offset* models are compared to CNB when the assumption underlying the CNB model does not hold.

A core assumption of CNB in modelling the outcome count is that the heterogeneity is introduced by a gamma-distributed subject effect that is shared with the baseline count for the same subject, so that heterogeneity is fully accounted for by incorporating the baseline count in the CNB model. The underlying counting process was referred to by Cook et al. (2005) as the time-homogeneous Poisson process, in which that heterogeneity is subject-specific and constant over time, but this may not be the case for falls prevention trials. For example, if a latent variable, such as progression in disease severity, increases the risk of falling, the variable may introduce different heterogeneity into the falls counts during the baseline and outcome periods, so that the heterogeneity cannot be treated as fixed if there is a relatively long interval between the baseline and outcome periods.

Another possibility is that the outcome and baseline falls counts may be collected using different methods. It is not uncommon to encounter a trial in which the outcome count is recorded prospectively in a falls diary, while the baseline count is obtained via a single retrospective question. As discussed in section 2.2.1, the logic behind this trial design is that although the prospective method is thought to be more accurate than the retrospective method, following up participants prospectively is more expensive and time-consuming. When designing falls prevention trials, researchers may choose to use the prospective method to collect the falls count during the outcome period, but use the less expensive retrospective method to collect a baseline count. However, for trials with this design, the falls counts collected during the baseline and outcome periods are subject to different measurement error processes, resulting in a discrepancy between the two variables, which is not accommodated in the subject effect underlying the CNB model.

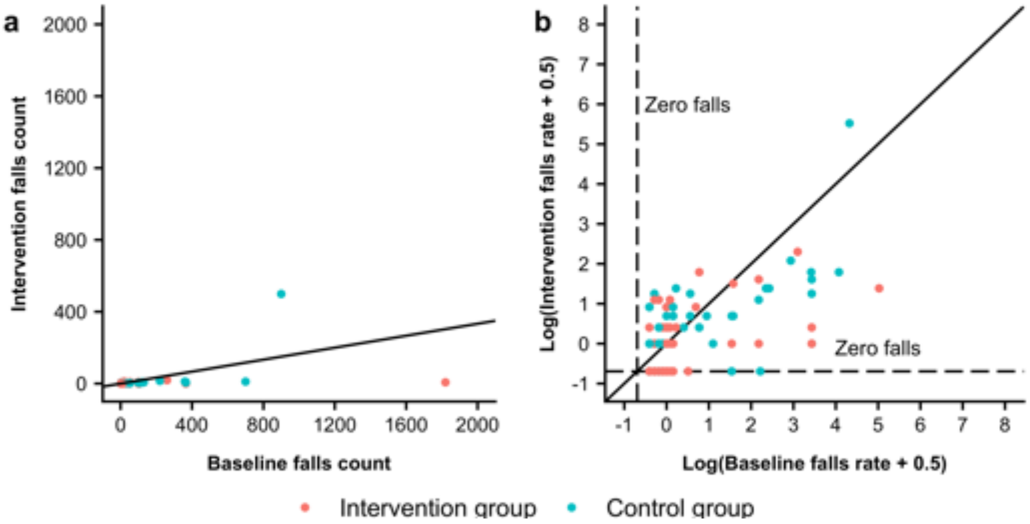


Figure 7-1 Distributions of falls counts in in the EXSart dataset (n=129, Spearman ρ=0.558, P<0.001). **(a)** Intervention falls count versus baseline falls count on the linear scale; **(b)** Intervention falls rate (per month) versus baseline falls rate (per month) on a logarithmic scale. The line is the LoFE that indicates the falls count/rate if the outcome rate is the same as the baseline rate.

The motivating dataset for this chapter, from the EXSart (Ashburn et al., 2007) trial, is an example of a falls prevention trial with a prospectively collected outcome count and retrospectively collected baseline count. The baseline and outcome counts show weaker correlation (Figure 7-1), in contrast to the high consistency between the corresponding variables in the Goodwin et al. dataset (see Figure 6-1), in which both the baseline and outcome counts were collected prospectively.

For the sake of simplification, the assumption that the outcome and baseline counts are generated from a time-homogeneous Poisson process is referred to as the assumption of subject-specific heterogeneity in the thesis. The question addressed in this chapter is whether the *NB-logged*, *NB-offset*, and CNB models are robust when the assumption of subject-specific heterogeneity is violated, such as when different methodologies for collecting falls counts are used in the baseline and outcome periods.

7.2 *NB-logged* and *NB-offset* models when the assumption of subject-specific heterogeneity is violated

Following the notation of Chapter 6, suppose a trial comprises of a pre-randomisation baseline period of duration t_0, and a post-randomisation outcome period of duration t_{i1}. The participants are randomised either to an intervention group ($x_i = 1$) or a control group ($x_i = 0$). For each participant the duration t_{i1} may vary due to drop out, assumed to occur at random.

Let y_{i0} and y_{i1} denote the baseline and outcome falls counts for trial participant i. As shown in equation (6-18), when y_{i0} is included in the linear predictor of NB regression, it should be log transformed to account for the log link-function, and the parameter for the logged baseline count is 1 if the assumption underlying CNB holds true. This is the basis of *NB-offset*, which fixes the coefficient of the logged baseline count to be 1 and the linear predictor is given by:

$$g(\mu_{i1}) = \zeta + \beta x_i + \log(y_{i0}) + \log(t_{i1}). \qquad (7\text{-}1)$$

where $\mu_{i1} = E(Y_{i1})$, β is the regression coefficient of the intervention effect, and ζ is the intercept.

Another approach considered in Section 6.2.2 is the *NB-logged* model, which relaxes the restriction and allows the coefficient of logged baseline count (denoted as ϕ) to vary. The linear predictor of *NB-logged* is given by:

$$g(\mu_{i1}) = \zeta + \beta x_i + \phi\log(y_{i0}) + \log(t_{i1}), \qquad (7\text{-}2)$$

As a result, the linear predictor of *NB-logged* has one more parameter ϕ (and degree-of-freedom) than the linear predictor of *NB-offset*.

The simulations in section 6.5 compared *NB-logged* and *NB-offset* in various scenarios, and the results showed that $\hat{\phi}$ was generally close to one, and the estimates from the two models were generally similar. This indicates that when data are generated from the time-homogeneous Poisson process underlying the CNB model, the results of *NB-logged* and *NB-offset* are almost indistinguishable as expected, which prompts the question: will the performance of the two models differ when the assumption of subject-specific heterogeneity is violated by using different methods to collect the baseline and outcome falls counts. By letting ϕ vary, this parameter in *NB-logged* may partially accommodate violations of the assumption, and for this reason, *NB-logged* is anticipated to perform better than *NB-offset* and CNB.

In the following sections, the *NB-logged* model is compared to *NB-null*, *NB-offset* and CNB using the EXSart dataset, and in datasets simulated to resemble prospectively collected outcome falls counts and retrospectively collected baseline counts. The *NB-null* model does not incorporate the baseline count and thus is not affected by the discrepancy in collection methods, but is included for comparison.

7.3 Methods

NB-null, *NB-logged*, *NB-offset*, and CNB were first fitted to the actual dataset of falls counts during the intervention (first 8 weeks) and outcome (week 9 to month 6) periods from the EXSart trial.

Because an eligibility criterion of the EXSart trial was that the participants must have fallen at least twice during the baseline period, there are no zero baseline counts in this dataset. Therefore, the logged baseline count was included in *NB-logged* and *NB-offset* without adding 0.5, which is different to the approach in section 6.4. The CNB model conditional on $y_i \geq 2$ was fitted to account for the eligibility criterion; see Cook and Wei (2003) for the detail of the implementation.

The analysis was conducted in R (version 3.3.0) using the same packages and functions described in section 6.3. From each model, the estimate of the intervention effect β, the model-based and robust SE (using the **sandwich** package) of $\hat{\beta}$, and the corresponding FRR with 95% CIs are reported. For the three NB models specifically, HP and AIC are reported. Furthermore, the estimate of ϕ is reported for *NB-logged*, and $\hat{\alpha}$ is reported for CNB.

The BOE plots described in Chapter 5 are presented for *NB-null* and *NB-logged* models for the four diagnostic statistics: Cook's distance, leverage, Anscombe residual, and DFBETA of $\hat{\beta}$ (see section 6.5.1).

7.4 NB/CNB models fitted to the EXSart dataset

Results from the *NB-null*, *NB-logged*, and *NB-offset*, and CNB models fitted to the EXSart dataset are shown in Table 7-1. Although the CNB model performed the best in simulations in section 6.5.2 and in only a few scenarios were there any simulated datasets where it did not achieve convergence, it did not converge when fitted to either the intervention or follow-up periods.

NB-null has the largest AIC among the three fitted NB models for both the intervention and follow-up counts, and it shows similar HPs when fitted to the intervention (3.593) and follow-up (3.865) counts. Compared with *NB-null*, the *NB-logged* models have much lower AICs (the intervention period: 491.5 versus 577.3; the follow-up period: 604.2 versus 674.1). *NB-null* and *NB-logged* also result in very different estimates for the intervention effects: the FRRs from *NB-null* are 0.181 and 0.147 for the intervention and follow-up periods respectively, whilst in *NB-logged* the respective FRRs are 0.780 and 0.686. The FRR estimated from *NB-null* suggests that the intervention reduced the falls rates by more than 80%, which contradicts the pattern shown in the corresponding BOE plots: the red symbols (intervention) in both Figure 7-2 and Figure 7-6 do not show an apparent trend of falling under the blue symbols, as suggested by the FRRs from *NB-null*. As the estimated intervention effect is probably overestimated, the significant test result (P<0.001) is likely to be a false positive.

In section 6.4, *NB-logged* and *NB-offset* were fitted to the Goodwin et al. dataset, they resulted in almost identical AICs and very similar estimates for the intervention effect. However, the two models show obvious differences when fitted to the EXSart dataset: 1) for both the intervention and follow-up periods, *NB-logged* and *NB-offset* yield noticeably different estimates of the intervention effect; 2) the AICs of the *NB-logged* models are smaller than that of *NB-offset* by a sizeable margin (the intervention period: 491.5 versus 504.0; the follow-up period: 604.2 versus 614.1); 3) *NB-logged* results in smaller HP than *NB-offset* (the intervention period: 1.310 versus 1.597; the follow-up period: 1.958 versus

2.220); and, 4) the estimate of ϕ from *NB-logged* was 0.724 for the intervention counts and 0.710 for the follow-up counts, whilst the estimates were close to 1 for *NB-logged* fitted the Goodwin et al. dataset.

As shown in Table 7-1, the robust SEs of $\hat{\beta}$ are bigger than the model-based SEs in *NB-null* by a considerable margin, but the former is only slightly larger in *NB-logged* and *NB-offset*.

Table 7-1 NB and CNB models fitted to the EXSart dataset.

Period	Model	AIC	$\widehat{\beta}$ (SE)	SE (Robust SE)	FRR (95% CI)	P	$\widehat{\phi}$ (SE)	HP
Intervention (n=129)	*NB-null*	577.3	-1.710	0.348 (0.803)	0.181 (0.091, 0.358)	< 0.001		3.593
	NB-logged	491.5	-0.248	0.254 (0.271)	0.780 (0.475, 1.283)	0.328	0.724 (0.074)	1.310
	NB-offset	504.0	-0.039	0.277 (0.273)	0.962 (0.559, 1.656)	0.888		1.597
								$\hat{\alpha}$
	CNB	Did not converge						-
Follow-up (n=127)	*NB-null*	674.1	-1.916	0.357 (0.851)	0.147 (0.073, 0.296)	< 0.001		3.865
	NB-logged	604.2	-0.377	0.275 (0.279)	0.686 (0.400, 1.175)	0.170	0.710 (0.085)	1.958
	NB-offset	614.1	-0.162	0.291 (0.292)	0.851 (0.480, 1.506)	0.578		2.220
								$\hat{\alpha}$
	CNB	Did not converge						-

The BOE plots for *NB-null* and *NB-logged* models are presented in Figure 7-2 to Figure 7-5 for the intervention period and Figure 7-6 to Figure 7-9 for the follow-up period. A participant (ID 28) reported the highest fall rate during both outcome periods (see also Table 3-14), and has the largest Cook's distance, Anscombe residual, and DFBETA in all *NB-null* and *NB-logged* models, but this subject is less influential in *NB-logged* than in *NB-null* as indicated by the smaller plotting symbols in the corresponding plots for the three diagnostic statistics. In particular, the ID 28 has large negative DFBETA from *NB-null* (see Figure 7-5 and Figure 7-9), which indicates that excluding the subject from *NB-null* would increase the regression coefficient, that is, if ID 28 is omitted the FRR would be closer to 1. This is in line with the extreme FRR (Table 7-1) from *NB-null*. Although the DFBETA of ID 28 is also negative in *NB-logged*, this participant's influence on the estimation of intervention effect is smaller in *NB-logged* than in *NB-null*. This suggests that both models may not sufficiently accommodate the largest outcome counts, but including the baseline count reduces their influence.

Figure 7-4 and Figure 7-8 compare the Anscombe residuals from *NB-null* and *NB-logged*. Apart from ID 28, the largest residuals from *NB-null* (ID 217 and 242 in Figure 7-4a and 242 and 251 in Figure 7-8a) were shown in the plots to be close to the LoFE, and the residuals are typically positive for large outcome counts and negative for small outcome counts, irrespective of whether the falls rate is consistent between the baseline and outcome periods. For the *NB-logged* models, the largest falls residuals (ID 232 and 204 in Figure 7-4b and 23 and 48 in Figure 7-8b) were cases far from the LoFE, that is, the subjects with inconsistent outcome and baseline rates, not necessarily with large counts.

Figure 7-5 and Figure 7-9 show the DFBETA from *NB-null* and *NB-logged*. These two plots indicate that in the *NB-null* models the estimation of the intervention effect was dominated by the large falls counts. One participant (ID 242) reported a relatively consistent falls rate across the baseline, intervention, and follow-up periods, but has large positive DFBETA in *NB-null*. A similar case is ID 251 in Figure 7-9: the plotting symbol for this subject lies exactly on the LoFE, that is, the follow-up falls rate is perfectly consistent with the baseline rate, but ID 251 also has a large positive DFBETA in *NB-null*. Although the outcome rates for the two subjects are consistent with the baseline rates, they are nevertheless influential in the estimation of the intervention effect and deleting either of them would result in a smaller intervention effect. These two subjects have small DFBETA in *NB-logged*. Furthermore, similar to the Anscombe residual plots, apart from ID 28 the two points that are far away from the LoFE have the largest DFBETA (ID 239 and 204 in Figure 7-5; ID 63 and 239 in Figure 7-9). This pattern indicates that the intervention effect estimated from the *NB-null* model was influenced for both the intervention and follow-up periods by a few large counts, whilst the estimator of intervention effect from *NB-logged* is more influenced by the participants reporting inconsistent falls rate across periods, which is anticipated since the baseline count is incorporated in *NB-logged*.

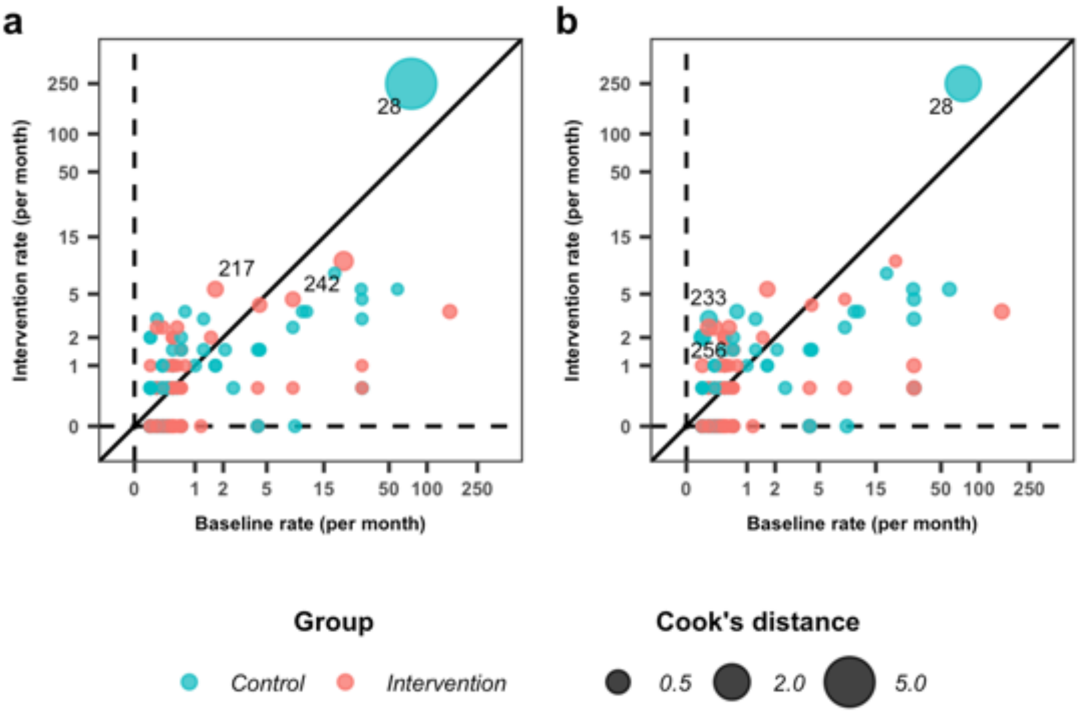


Figure 7-2 Cook’s distance from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the intervention count from the EXSart dataset (n=129).

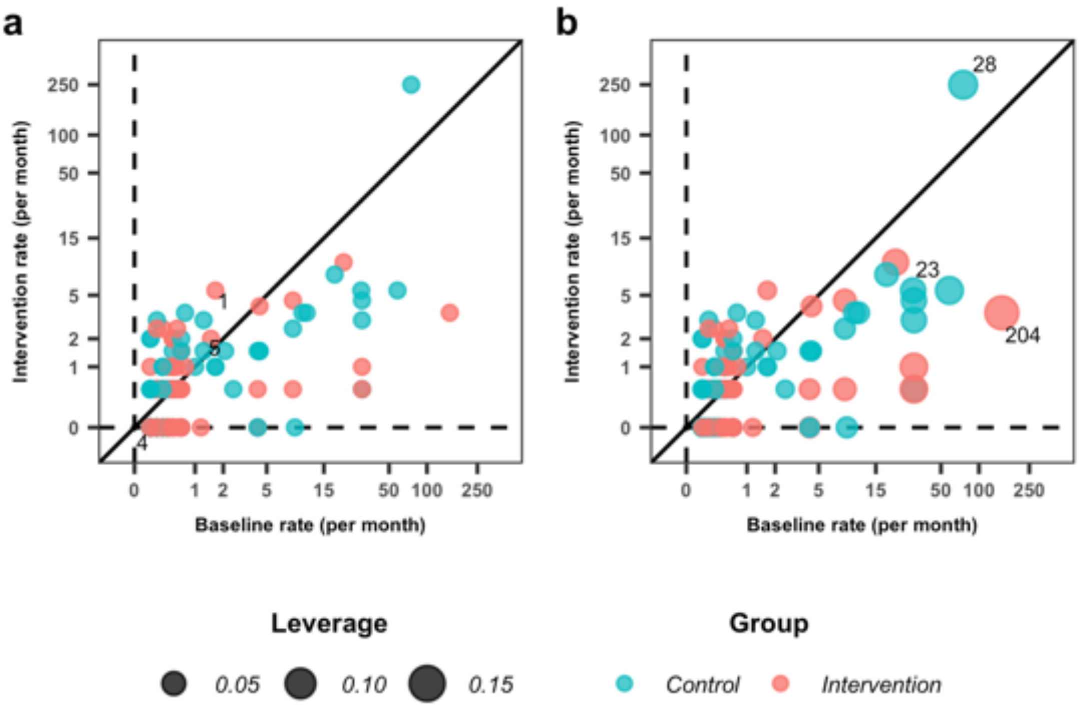


Figure 7-3 Leverage from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the intervention count from the EXSart dataset (n=129).

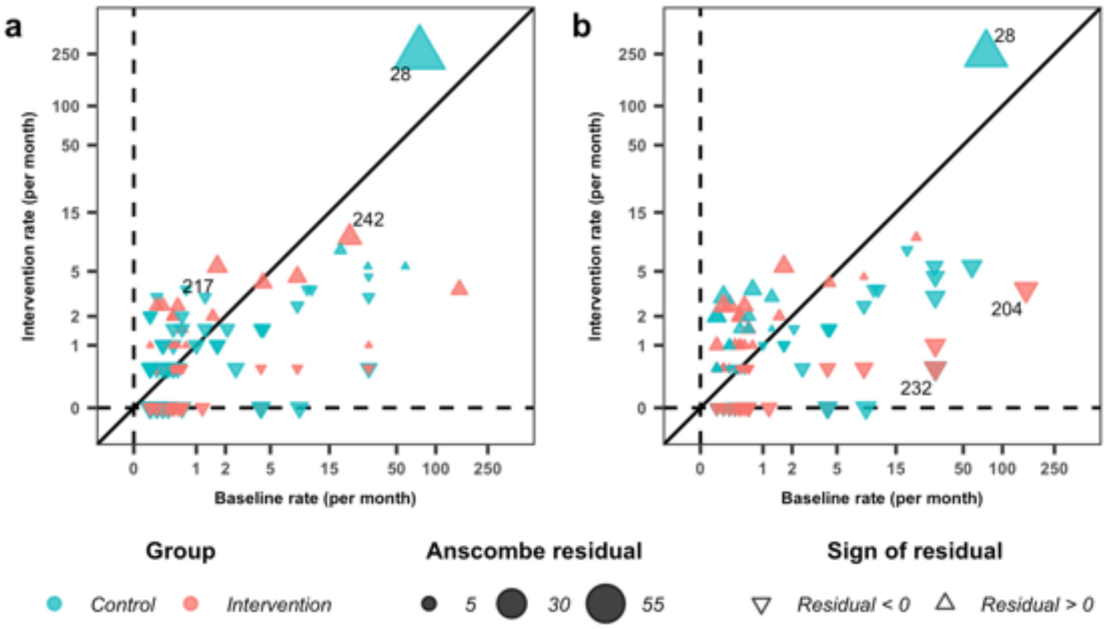


Figure 7-4 Anscombe residuals from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the intervention count from the EXSart dataset (n=129).

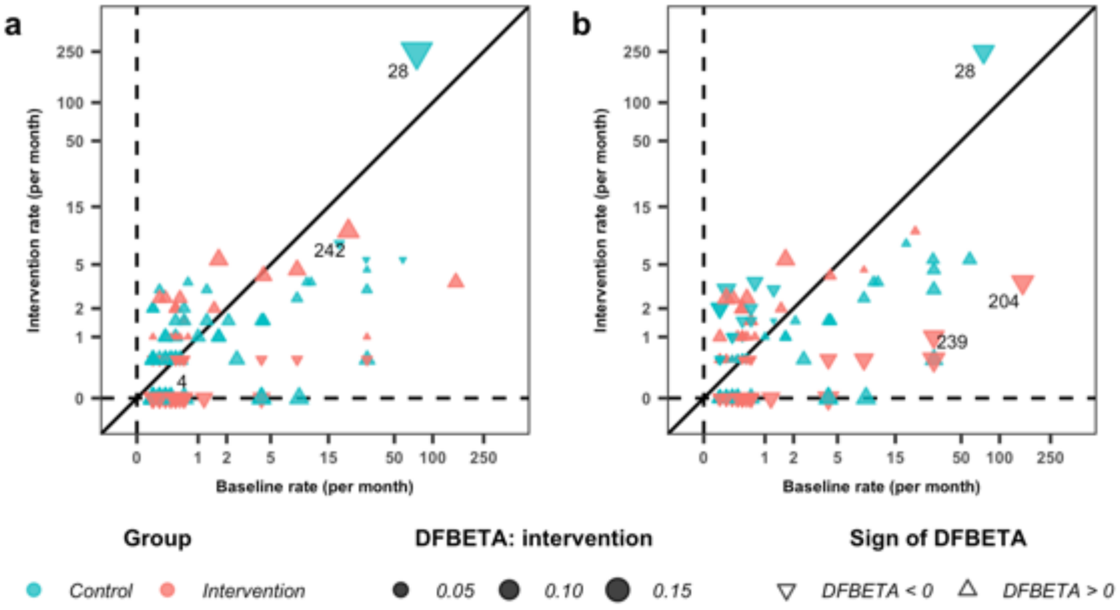


Figure 7-5 DFBETA for the intervention effect from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the intervention count from the EXSart dataset (n=129).

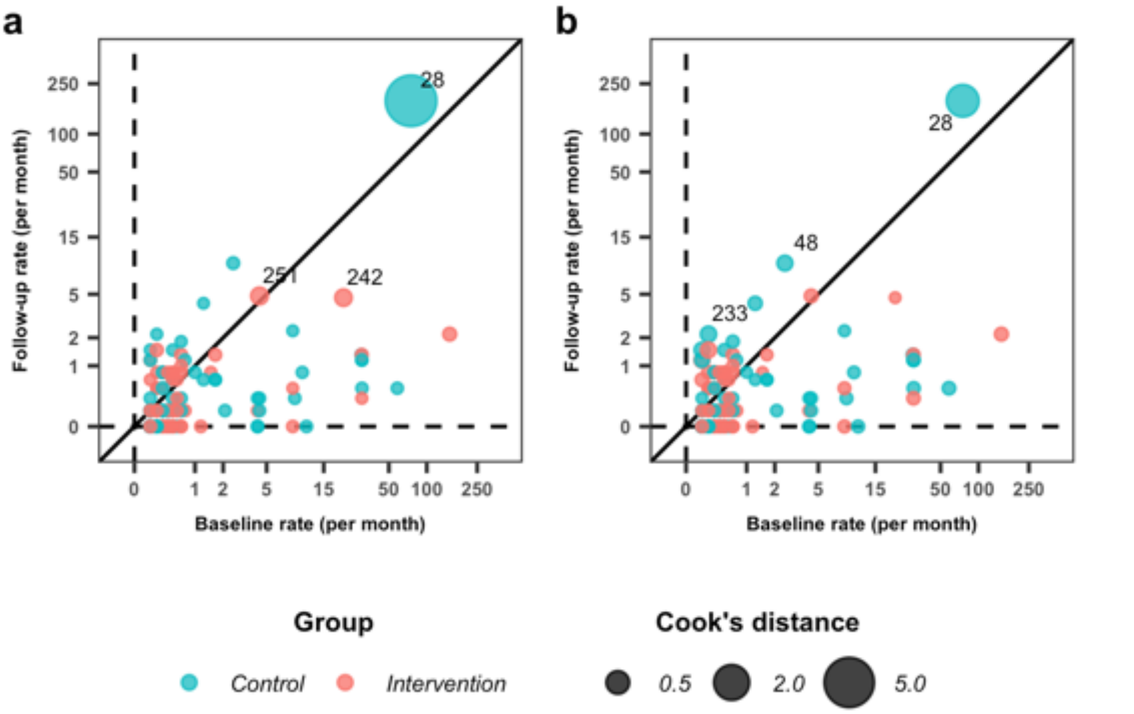


Figure 7-6 Cook's distance from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the follow-up count from the EXSart dataset (n=127).

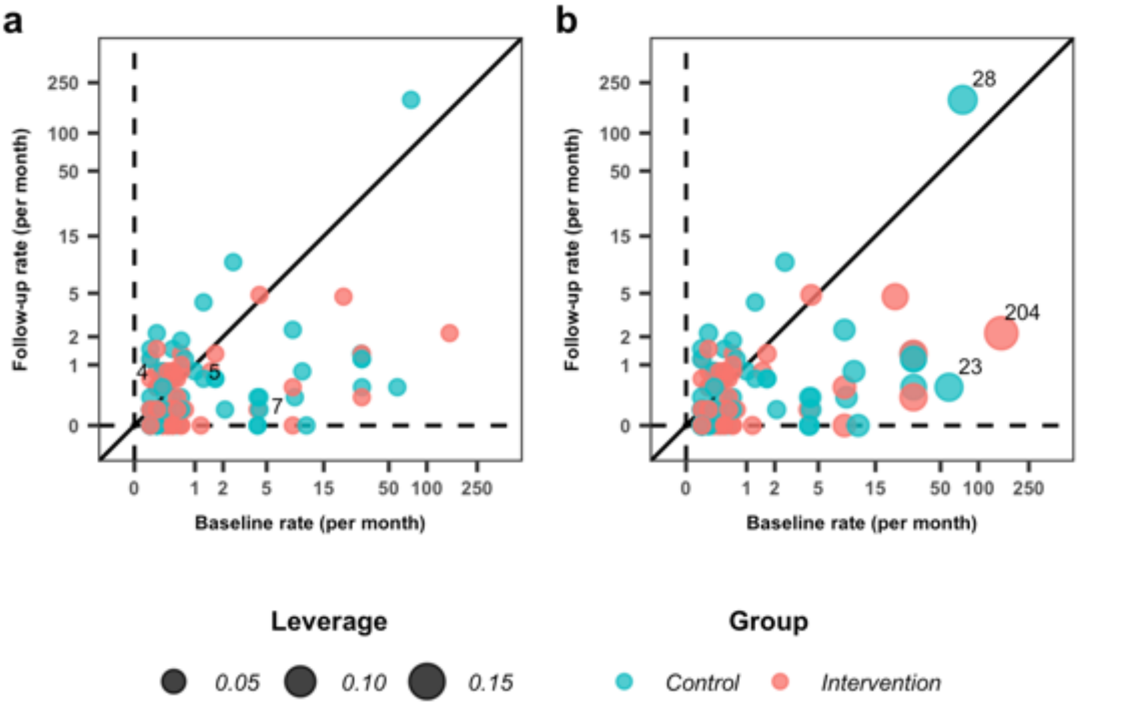


Figure 7-7 Leverage from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the follow-up count from the EXSart dataset (n=127).

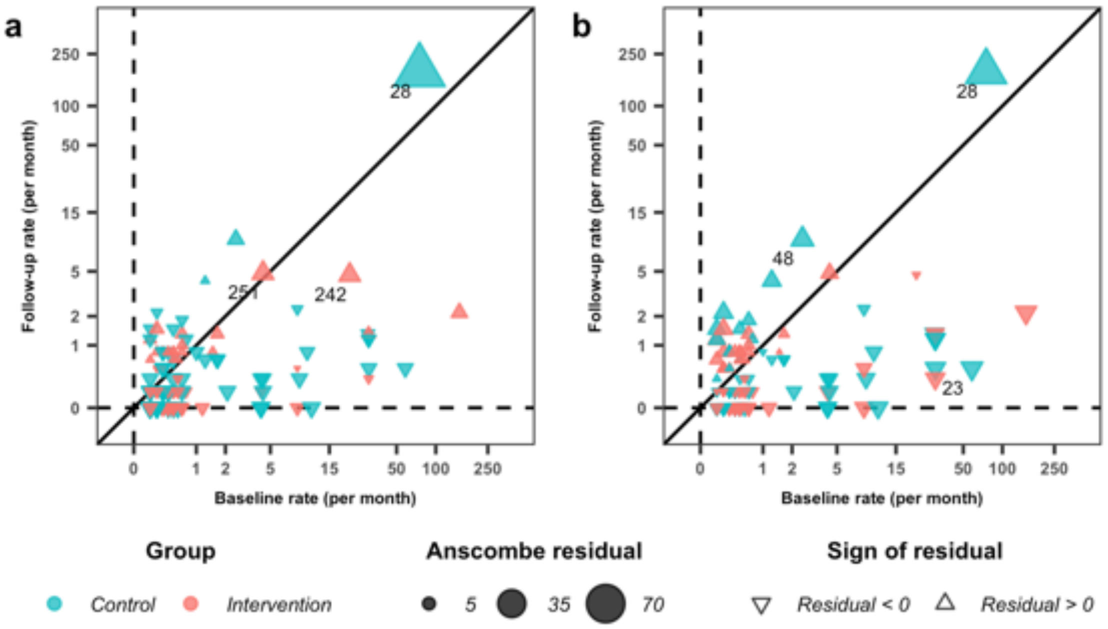


Figure 7-8 Anscombe residuals from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the follow-up count from the EXSart dataset (n=127).

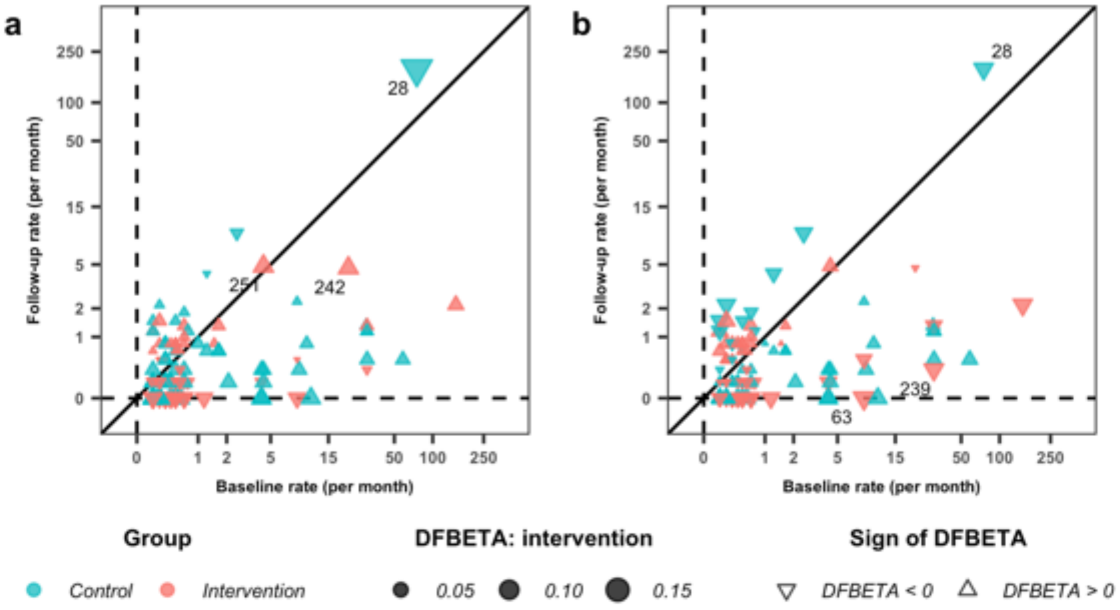


Figure 7-9 DFBETA for intervention effect from **(a)** *NB-null* versus **(b)** *NB-logged* fitted to the follow-up count from the EXSart dataset (n=127).

7.5 Simulation study

7.5.1 Simulation datasets

A simulation study was conducted to compare the performance of *NB-null*, *NB-logged*, *NB-offset*, and CNB models when the assumption of subject-specific heterogeneity underlying CNB does not hold (an example code is given in Appendix B).

Suppose m subjects are recruited in a falls prevention trial, and then randomized to an intervention (denoted by $x_i = 1$ for subject i) or control group (denoted by $x_i = 0$), with each group comprising $m/2$ subjects. The falls count during a baseline period (with constant length t_0) is denoted by y_{i0} and the count during an outcome period (with length t_1) is denoted by y_{i1}. The two counts were generated from:

$$y_{i0}|s_i \sim \text{Poisson}(s_i v_i \mu_0) \qquad (7\text{-}3)$$

$$y_{i1}|s_i \sim \text{Poisson}(s_i \mu_{i1}), \qquad (7\text{-}4)$$

where $\mu_0 = \lambda t_0$ and $\mu_{i1} = \lambda t_1 \exp{(\beta x_i)}$. The average falls rate during the baseline period is denoted by λ and assumed to be the same as the average falls rate in the control group during the follow-up period.

The subject-specific heterogeneity s_i was simulated from a gamma distribution with mean 1 and variance α, where α is a measure of the severity of the subject-specific heterogeneity. The first $m/2$ simulated subjects were assigned to the control group and the rest intervention group, without loss of generality. A perturbation v_i is introduced in (7-3) to create the inconsistency observed in the EXSart dataset, where v_i is simulated from a gamma distribution with mean 1 and variance ϵ, where ϵ is the level of perturbation. The v_i is included to increase the variability of Y_{i0} to mimic the lower precision expected from the retrospective method. Unlike the EXSart dataset, the baseline count was not truncated to be ≥ 2.

The simulation study was based on the falls counts during the intervention period of the EXSart dataset:

- As shown in Table 3-12, the average falls rate in the control group was relatively stable during the baseline (5.1 falls/month), intervention (5.1 falls/month), and follow-up (5.3 falls/month) periods, so the datasets were generated with falls rate λ=5.
- To resemble the lengths of the baseline and intervention periods in the EXSart trial, we set $t_0 = 12$ and $t_1 = 2$. When a baseline count is collected retrospectively, the length of the baseline period is usually chosen to be relatively long because 1) the accuracy of the recalled falls rate would be higher, and 2) choosing a long retrospective baseline period does not increase the cost of trial.
- Because the CNB model did not converge when fitted to the EXSart dataset (see Table 7-1), the HP of *NB-null* in the intervention period (3.593) was used as an approximate to the degree of heterogeneity (α) in the dataset.

A total of 72 scenarios spanning all combinations of:

- Sample sizes: $m = 50, 100, 200, \text{and } 500$;
- Intervention effects: $\beta = -0.4$ for a large effect, $\beta = -0.2$ for a small effect, and $\beta = 0$ for no effect;
- Degree of heterogeneity: $\alpha = 3.5$ to resemble the HP of *NB-null*, and $\alpha = 0.5$ for a smaller heterogeneity;
- Degree of perturbation in the baseline count: $\epsilon = 0$ for no perturbation, $\epsilon = 0.25$ for a small degree of perturbation, and $\epsilon = 0.5$ for a large degree of perturbation.

For each scenario, 2000 datasets were simulated. Three datasets were simulated with $\lambda = 5$, $\beta = -0.2$, $t_0 = 12$, $t_1 = 2$, $\alpha = 3.5$, $\epsilon = 0.5$, and $m = 130$, a scenario closest to EXSart, for visually checking that the simulated datasets resembled the EXSart dataset.

NB-null, *NB-logged*, *NB-offset*, and the CNB models were fitted to each dataset, and the same simulation-based statistics for the estimator of β and the model-based Wald tests included in section 6.5.1 were reported.

7.5.2 Simulation results

As shown in Figure 7-10, the three simulated datasets for visual check broadly resemble the pattern of the EXSart dataset during the intervention period as shown in Figure 7-1.

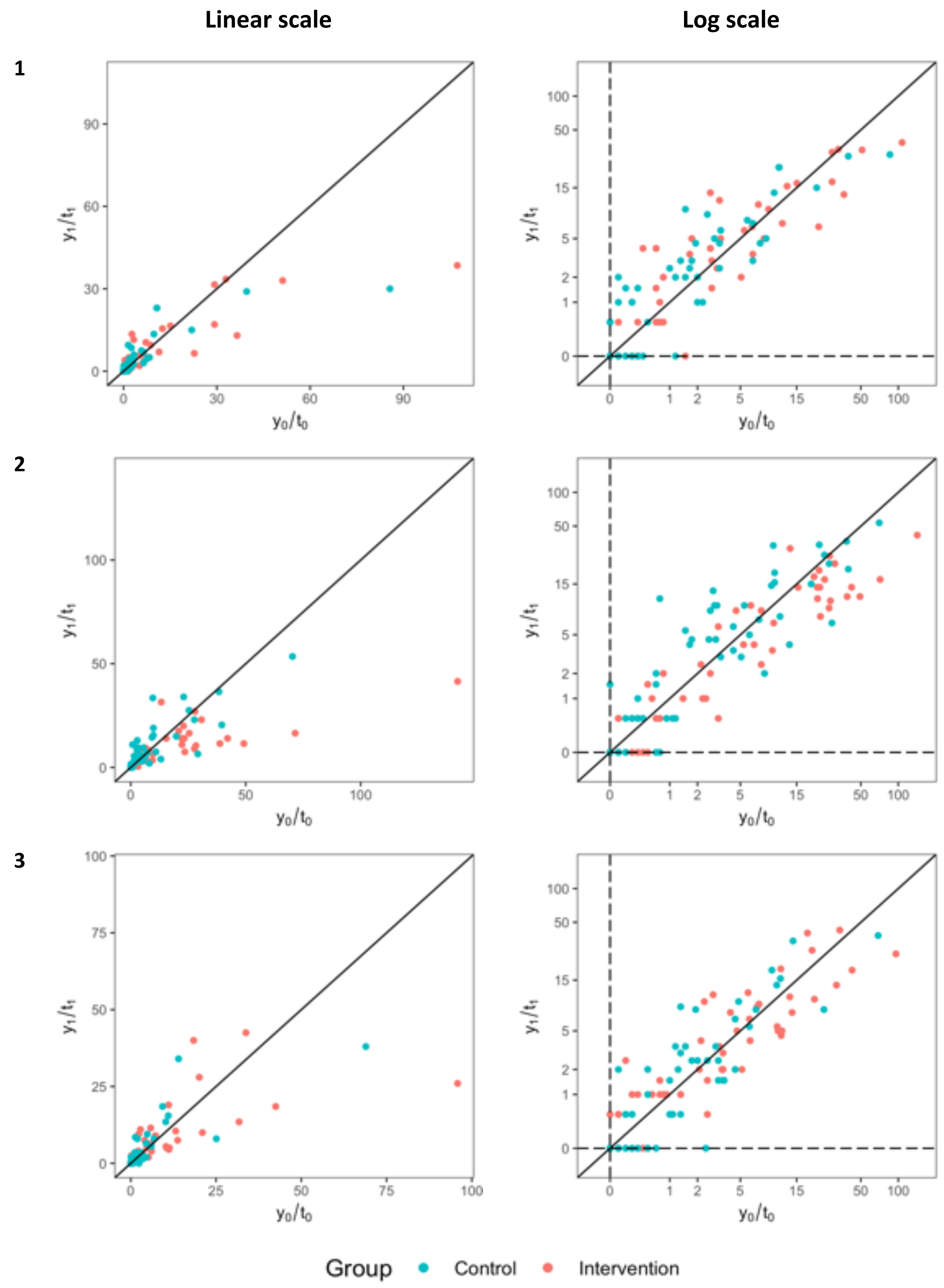


Figure 7-10 Three simulated datasets (n=130). Outcome rate versus baseline rate on the linear scale (left column) and the log scale (right column).

As shown in Table C-2 (see Appendix C), the four models converged successfully in most cases (including CNB). The estimated bias for $\hat{\beta}$, that is, $\widehat{\text{Bias}} = \text{av}(\hat{\beta}) - \beta$, is displayed in

Figure 7-11 to Figure 7-14 for $m = 50, 100, 200, \text{and } 500$. In general, $\hat{\beta}$ from *NB-null, NB-logged* and CNB are close to the underlying value β. In contrast, $\hat{\beta}$ from *NB-offset* are consistently lower than the underlying value, especially for a large intervention effect ($\beta = -0.4$) and small degree of subject-specific heterogeneity ($\alpha = 0.5$). The $\widehat{\text{Bias}}$ in *NB-offset* becomes larger as the level of perturbation ϵ increases, and it persists even for the largest sample size ($m = 500$). The error bars (calculated from MCError; see equation (6-24) in section 6.5.1) of *NB-null* are generally wider than *NB-logged* and CNB, especially when the subject-specific heterogeneity is high ($\alpha = 3.5$) and the sample size is less than 200, as shown in Figure 7-11 and Figure 7-12. The error bars of *NB-logged*, *NB-offset*, and CNB become wider as ϵ increases, but this does not affect *NB-null* because the baseline count is not included in the model.

Figure 7-15 to Figure 7-18 report the relative errors, defined in equation (6-25), in the model-based SE of $\hat{\beta}$: a positive relative error indicates the model-based SE is overestimated. As the baseline count is not included in *NB-null*, the SEs estimated from this model have low relative errors regardless of the level of ϵ as expected. The model-based SEs under *NB-offset* and CNB are considerably underestimated in all scenarios with perturbations ($\epsilon > 0$), and the relative errors for CNB are remarkably large even compared to *NB-offset*. In general, the *NB-logged* model has low relative errors when $\alpha = 0.5$, but the model-based SEs are smaller than the empirical SEs when $\alpha = 3.5$, especially for higher perturbation levels (ϵ). Compared to *NB-offset* and CNB, the *NB-logged* model has much smaller relative errors though, even for scenarios where $\alpha = 3.5$.

Figure 7-19 and Figure 7-20 show the empirical power and type I error rates of Wald tests of β. In accordance with the results of the previous simulations, the CNB model performed the best when there is no perturbation ($\epsilon = 0$), with higher power than the other models and type I error rates closer to the nominal level 0.05 (including the small sample size scenarios). However, CNB shows markedly inflated type I errors when $\epsilon > 0$. For scenarios with $\alpha = 0.5$ (Figure 7-19), the empirical type I error rates are around 0.30 for $\epsilon = 0.25$ and 0.45 for $\epsilon = 0.5$; for scenarios with $\alpha = 3.5$ (Figure 7-20), the empirical type I error rates are higher than 0.5: approximately 0.55 and 0.65 for $\epsilon = 0.25$ and $\epsilon = 0.5$, respectively. Though the type I error rates produced by *NB-offset* are also inflated, they are considerably closer to the nominal level. Similar to CNB, the empirical type I error rates for Wald tests based on *NB-offset* are inflated when $\epsilon > 0$, and the inflation rises as ϵ increases.

As shown in Figure 7-19b, *NB-logged* is more robust to perturbations than *NB-offset* and CNB — when $\alpha = 0.5$ the empirical type I error rates for *NB-logged* are generally close to 0.05. The empirical type I error rates for *NB-logged* are inflated when $\alpha = 3.5$ (Figure 7-20b), though still only one fifth of those for CNB. The empirical type I error rates are in line with the relative error plots in Figure 7-15 to Figure 7-18, which show *NB-logged* to have smaller relative errors than those from *NB-offset* and CNB.

Considering the enormous type I error rates of CNB when $\epsilon > 0$, its highest empirical power in all scenarios has little practical value. *NB-offset* has lower empirical power than *NB-logged* by a small to medium margin, and even yields lower empirical power than *NB-null* when $\epsilon = 0.5$, and $\alpha = 0.5$ (Figure 7-19). In comparison, *NB-logged* has higher empirical power than *NB-null* in all scenarios. The empirical power from *NB-logged* is considerably higher than that from *NB-null*, but the power gain from including y_{i0} is not as great for $\epsilon > 0$ as for $\epsilon = 0$.

As shown in Table 7-2 to Table 7-4, the $\hat{\alpha}$ from CNB are generally larger than the underlying values when there is perturbation ($\epsilon > 0$), especially when $\epsilon = 0.5$. As the level of perturbation becomes higher, the regression coefficient of the logged baseline count in *NB-logged* ($\hat{\phi}$) decreases. When the subject-specific heterogeneity is higher ($\alpha = 3.5$), $\hat{\phi}$ are closer to 1.

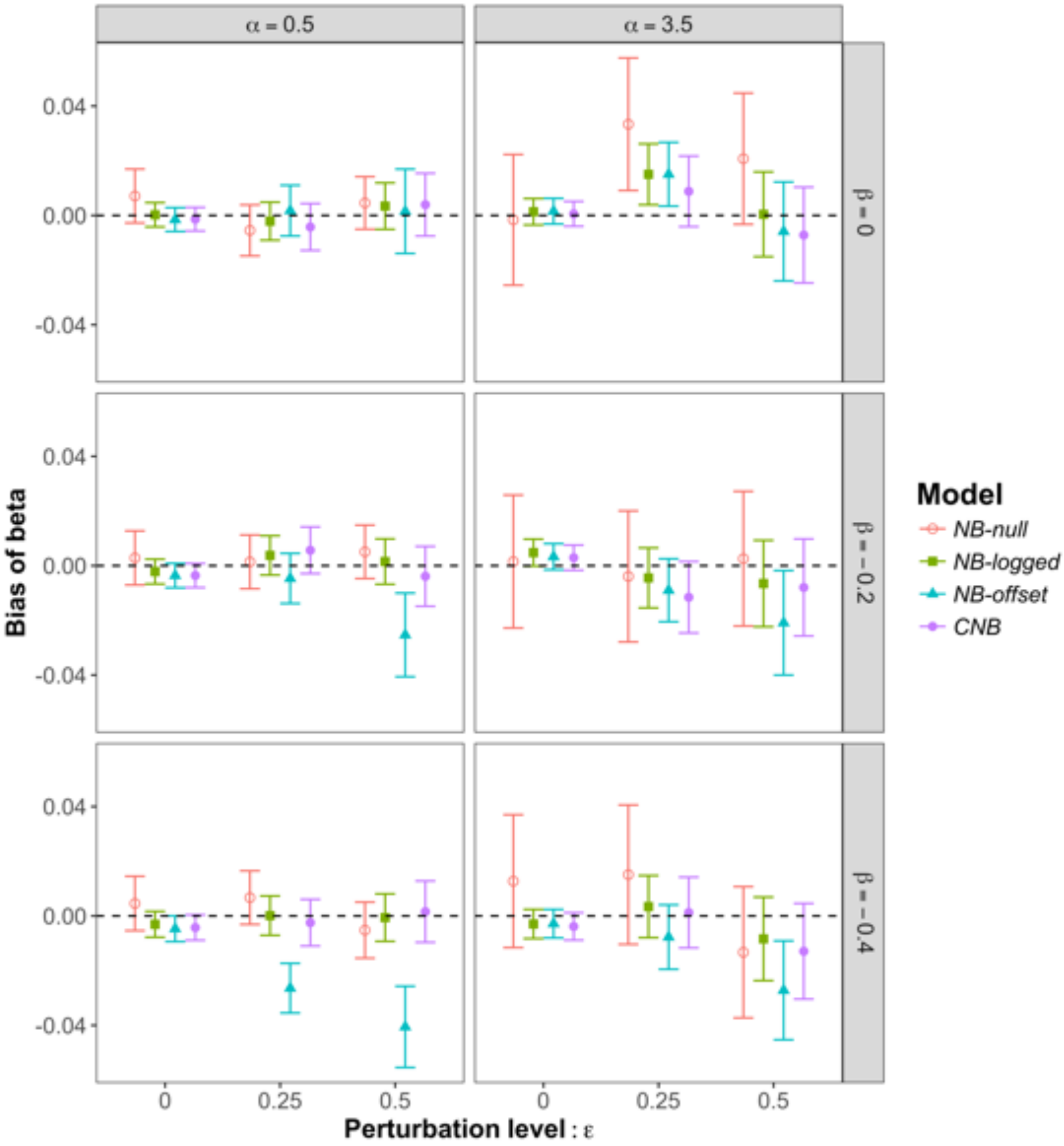


Figure 7-11 Bias plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 50$). The $\widehat{\text{Bias}}$ of $\hat{\beta}$ are shown as the points with error bars (the 95% CI calculated from the MCError of $\widehat{\text{Bias}}$).

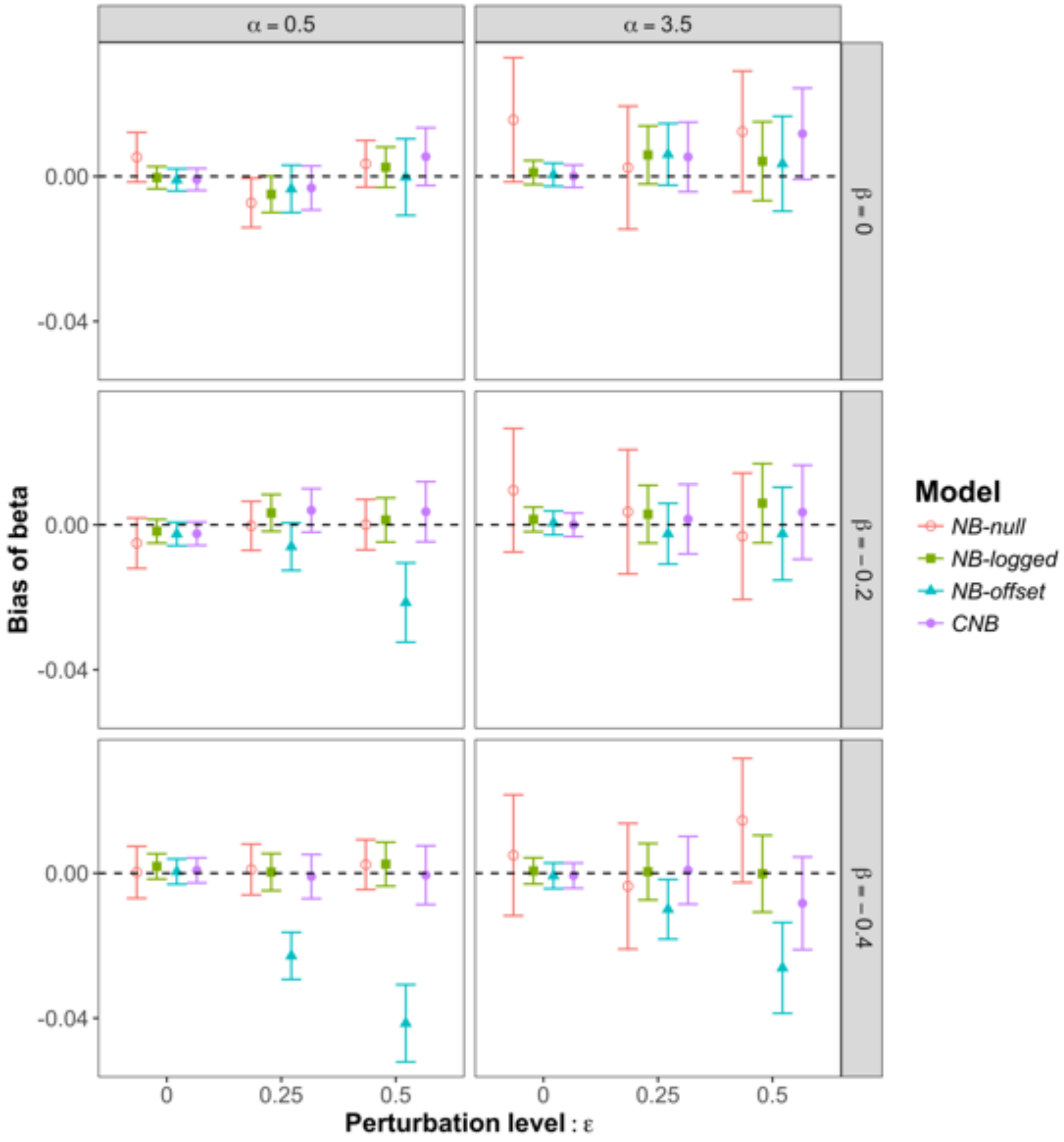


Figure 7-12 Bias plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 100$). The $\widehat{\mathrm{Bias}}$ of $\hat{\beta}$ are shown as the points with error bars (the 95% CI calculated from the MCError of $\widehat{\mathrm{Bias}}$).

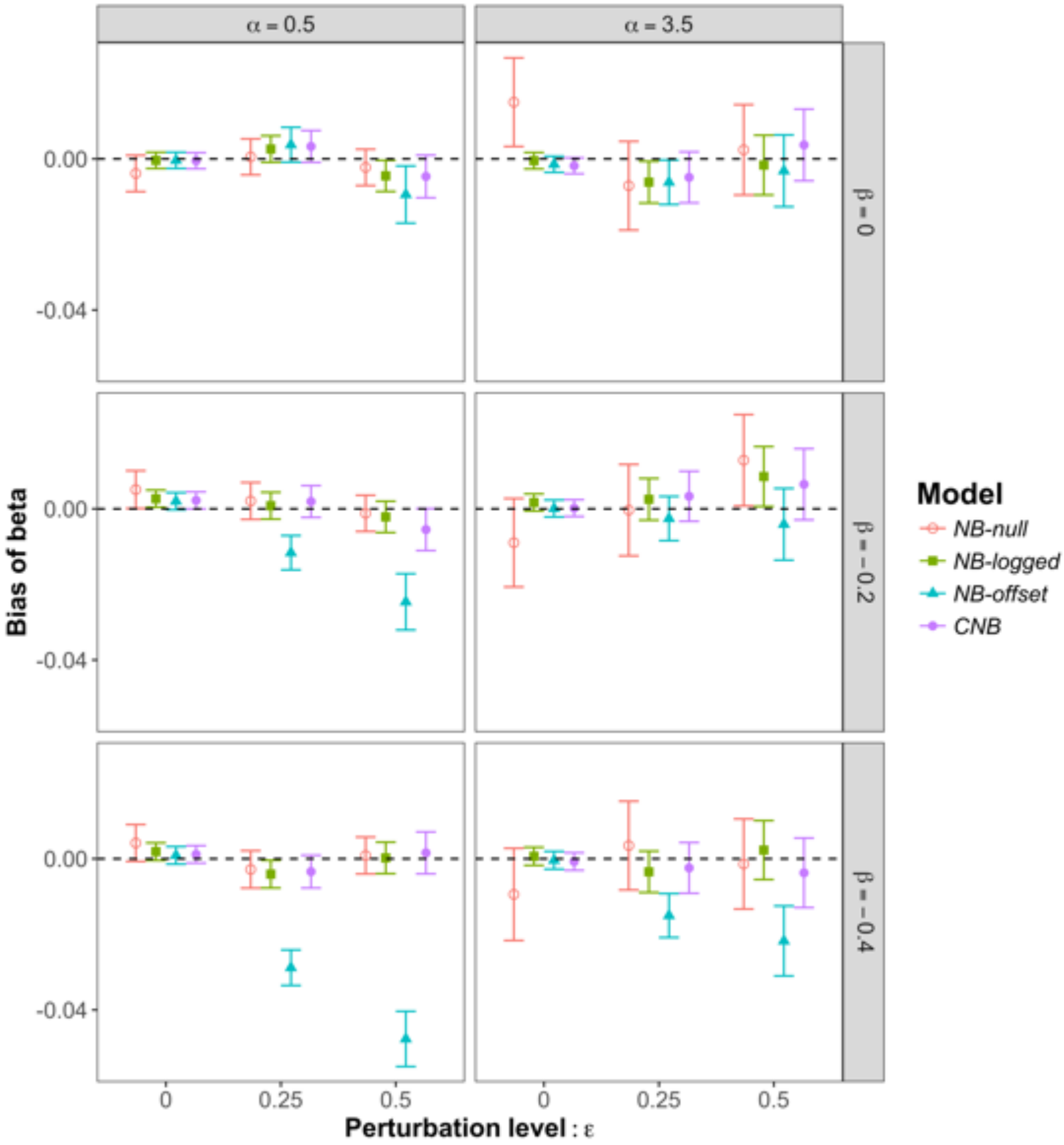


Figure 7-13 Bias plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 200$). The $\widehat{\mathrm{Bias}}$ of $\hat{\beta}$ are shown as the points with error bars (the 95% CI calculated from the MCError of $\widehat{\mathrm{Bias}}$).

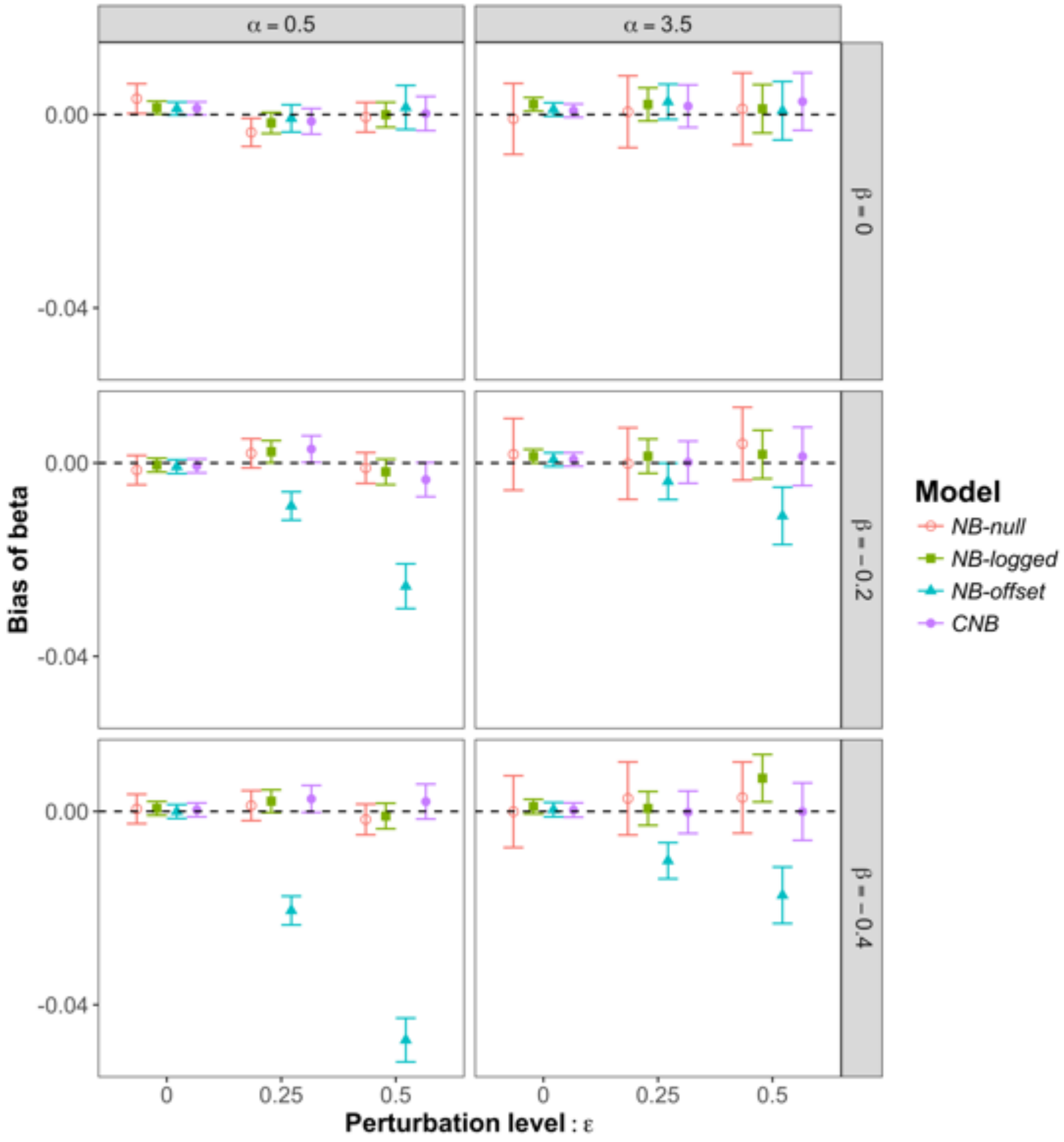


Figure 7-14 Bias plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 500$) The $\widehat{\text{Bias}}$ of $\hat{\beta}$ are shown as the points with error bars (the 95% CI calculated from the MCError of $\widehat{\text{Bias}}$).

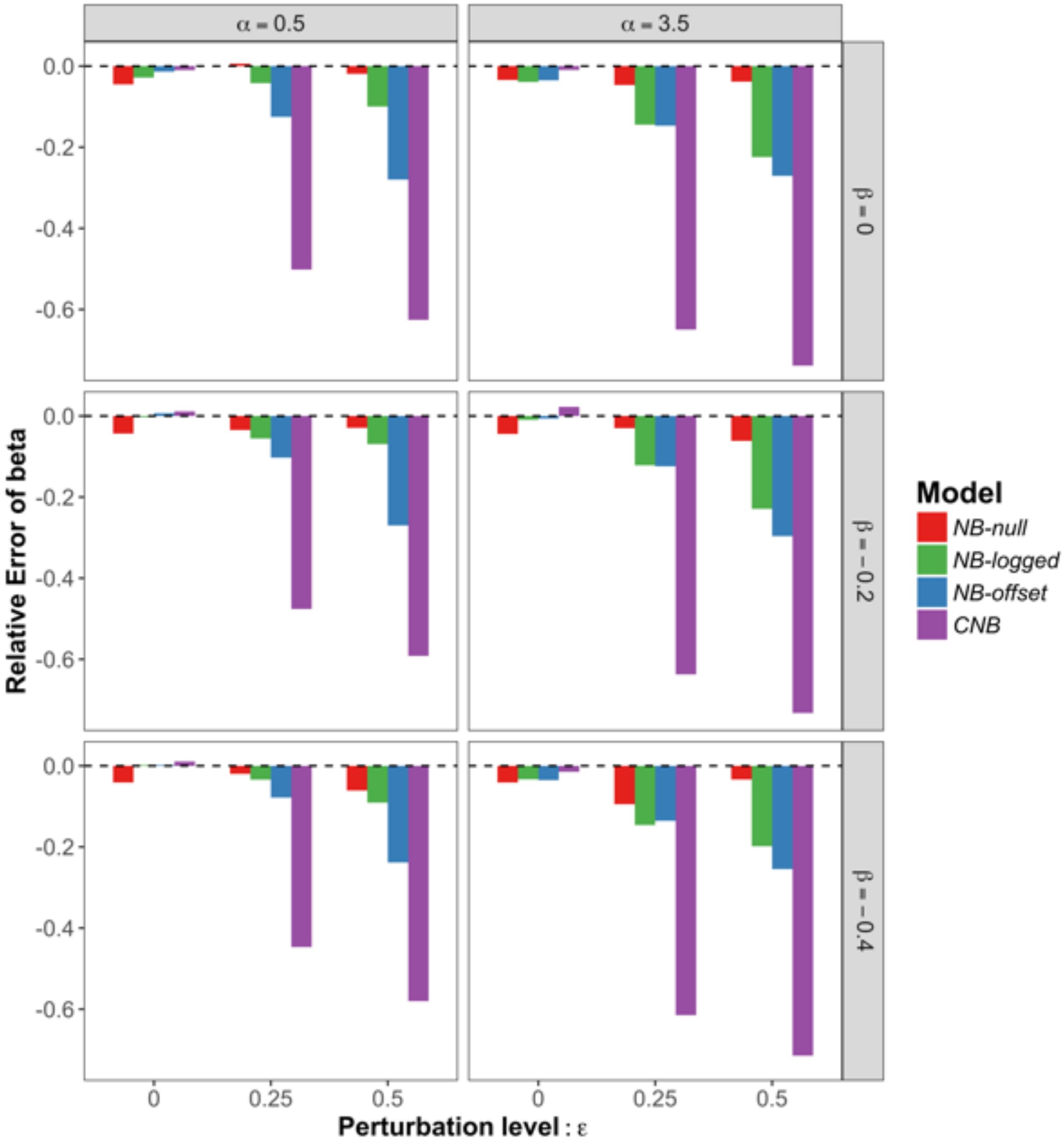


Figure 7-15 Relative error plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 50$).

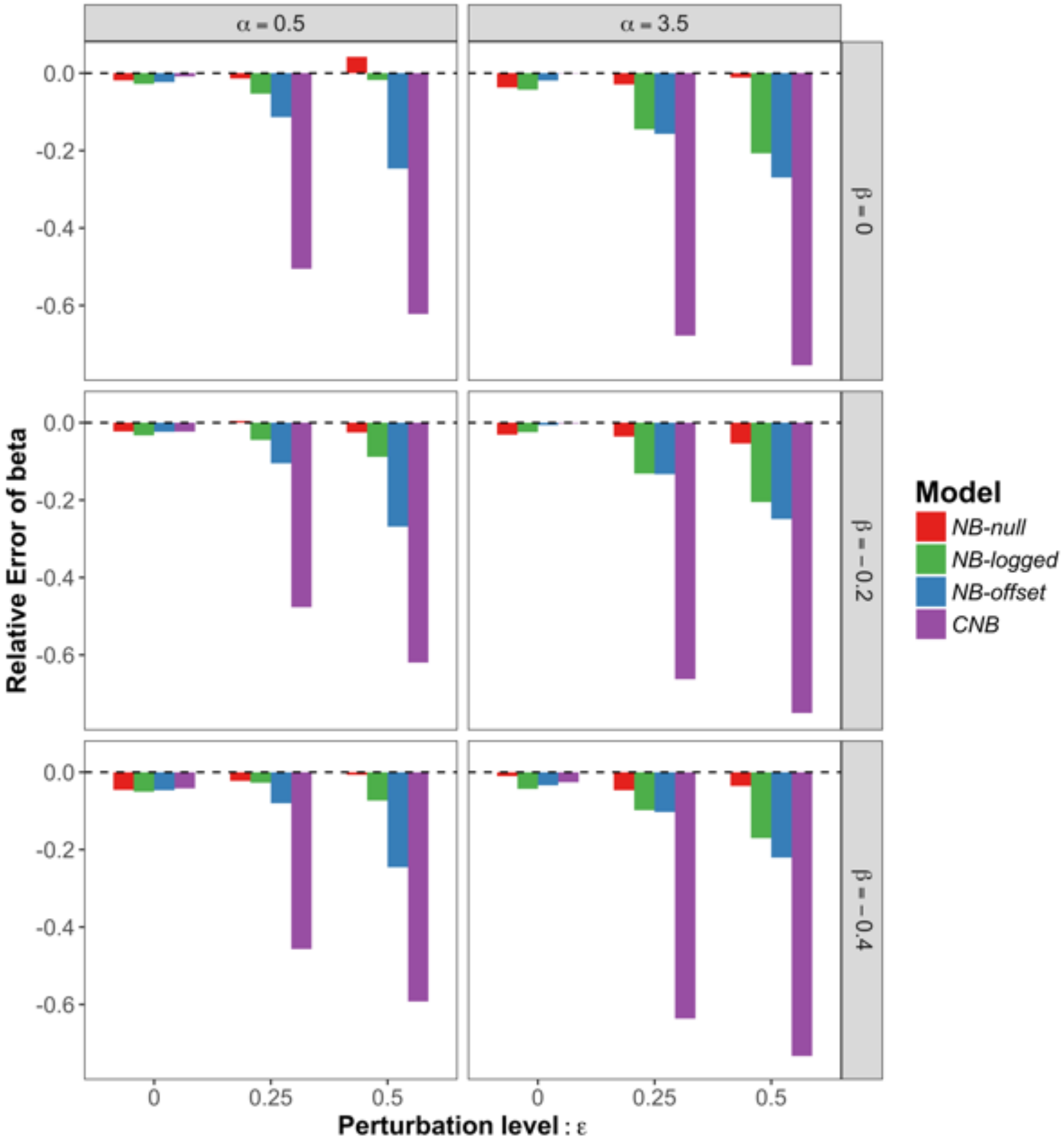


Figure 7-16 Relative error plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 100$).

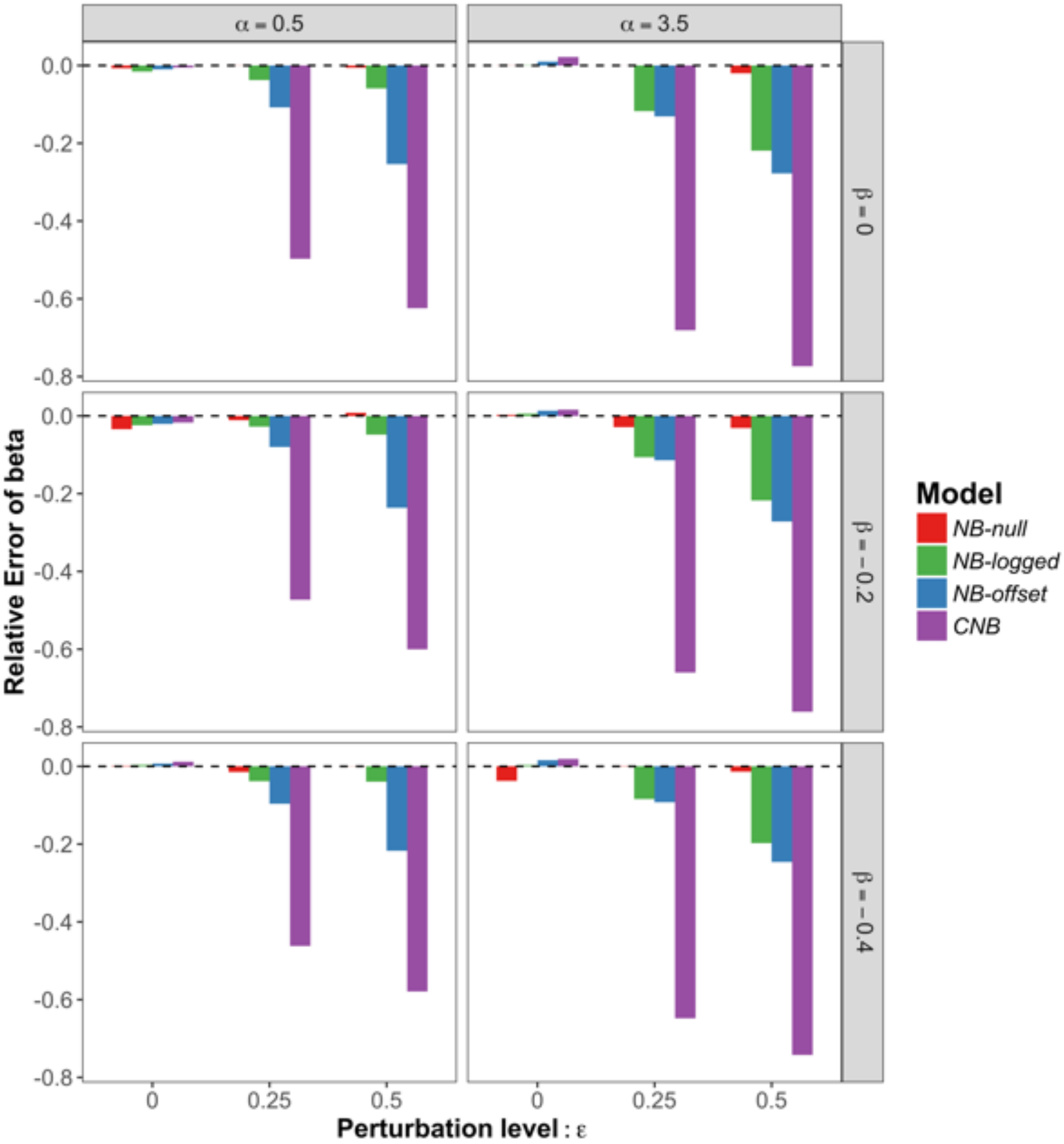


Figure 7-17 Relative error plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 200$).

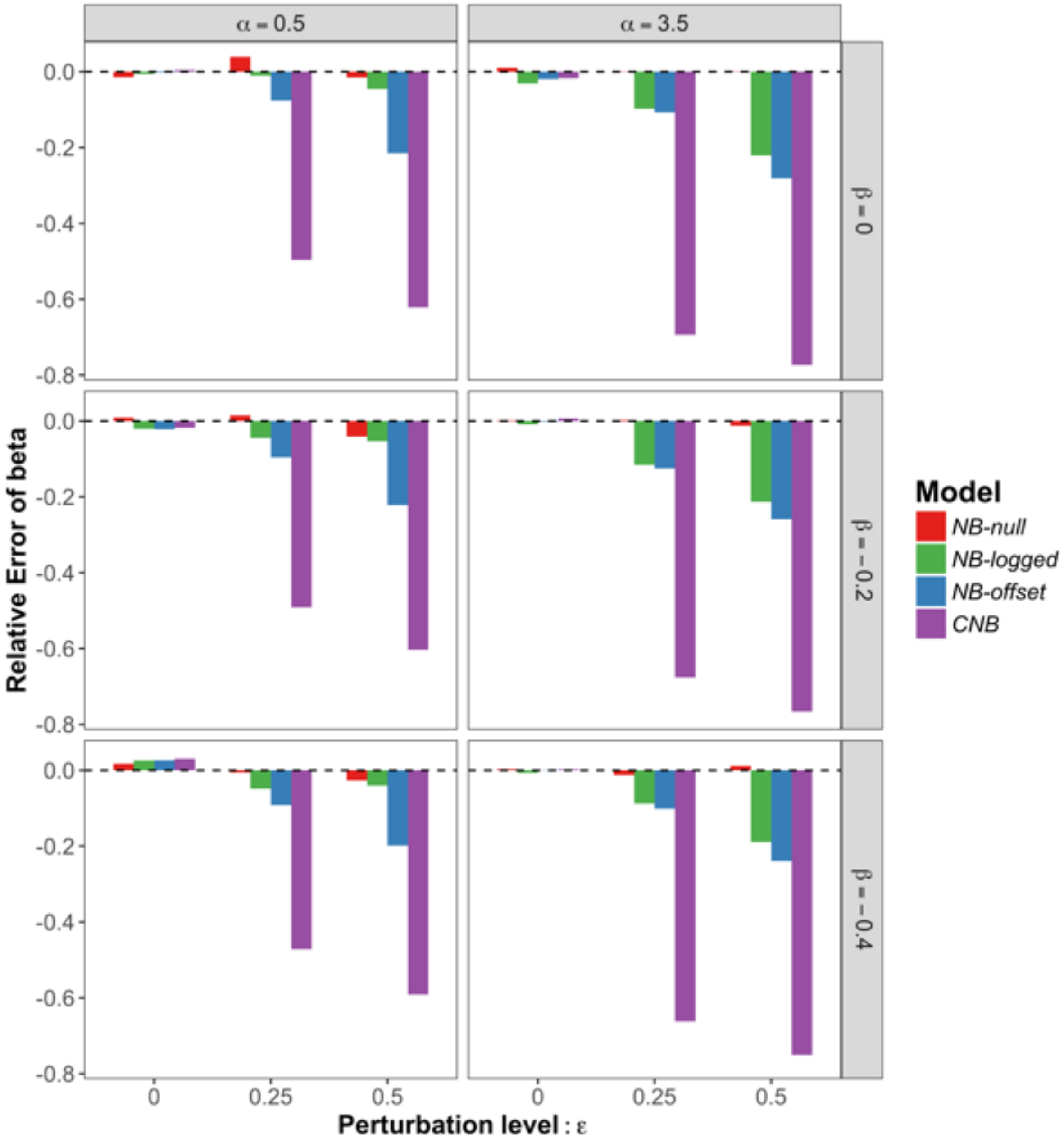


Figure 7-18 Relative error plots of *NB-null*, *NB-logged*, *NB-offset* and CNB with varying degrees of perturbation ($m = 500$).

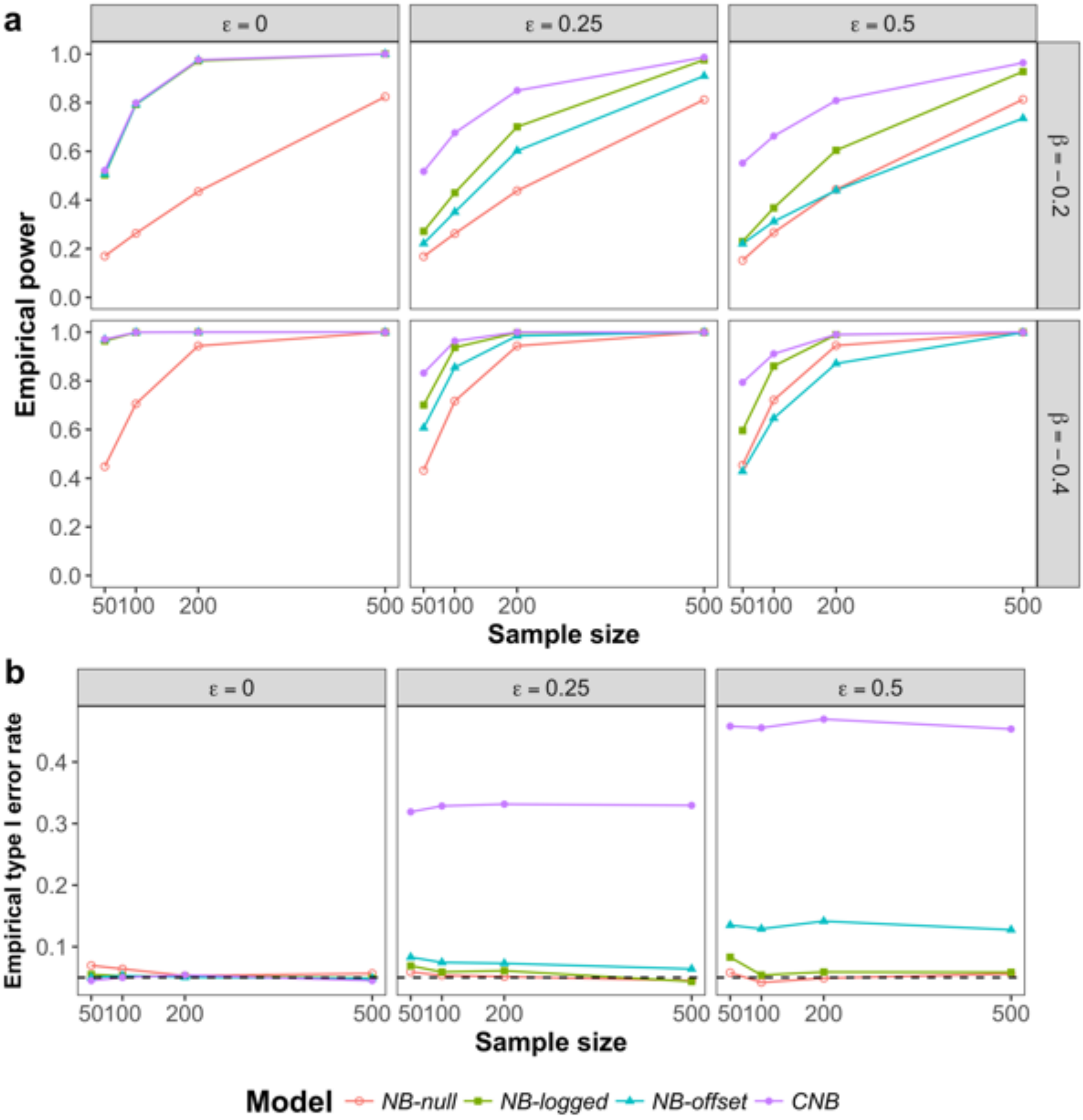


Figure 7-19 Performance of the Wald tests from *NB-null*, *NB-logged*, *NB-offset* and CNB in simulations with varying degrees of perturbation ($\alpha = 0.5$). **(a)** Empirical Power; **(b)** Empirical type I error rates.

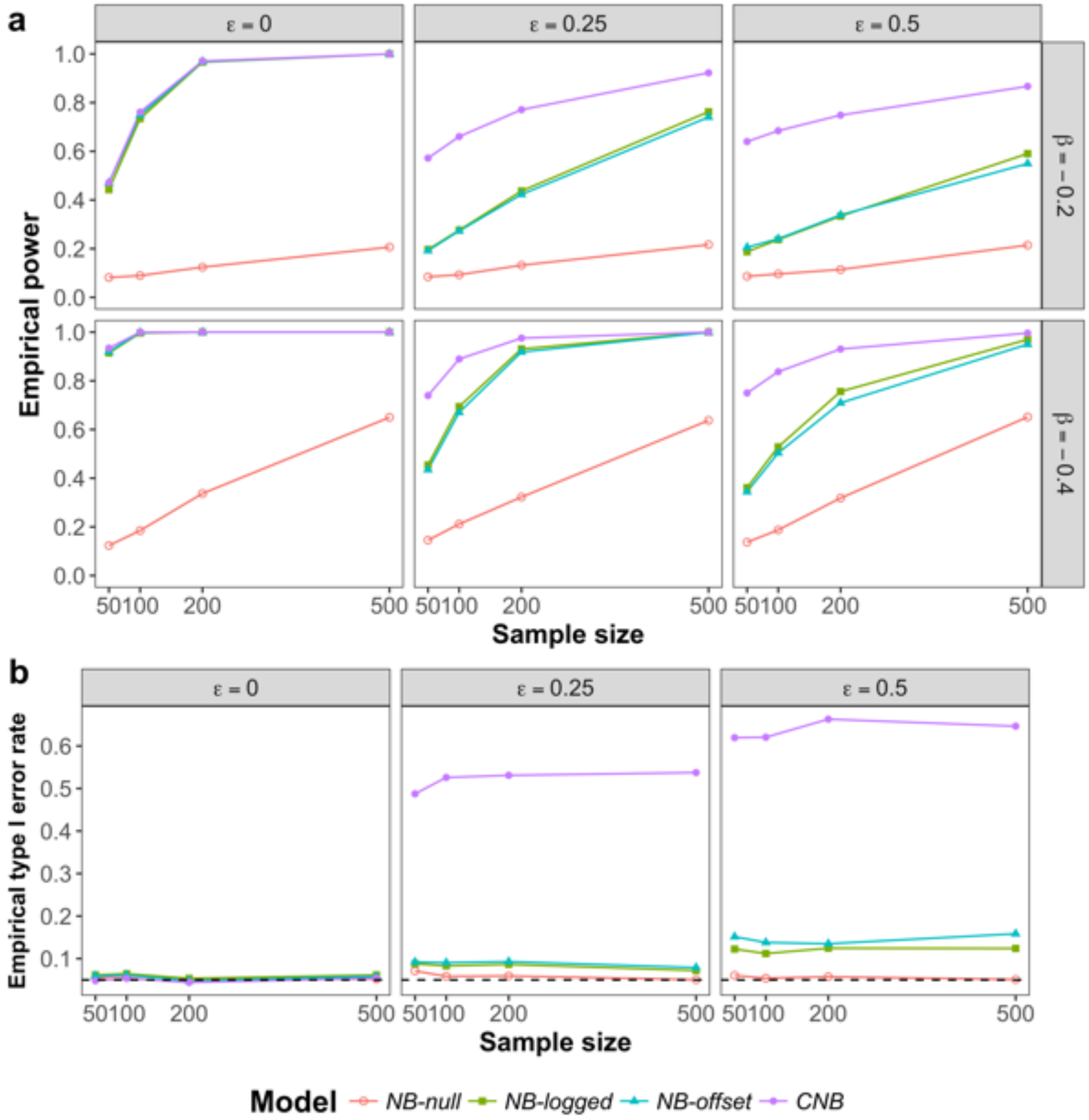


Figure 7-20 Performance of the Wald tests from *NB-null*, *NB-logged*, *NB-offset* and CNB in simulations with varying degrees of perturbation ($\alpha = 3.5$). **(a)** Empirical Power; **(b)** Empirical type I error rates. Note that the range of y-axis in subplot (b) is different to that in Figure 7-19

Table 7-2 Estimates of HP from NB models, $\hat{\alpha}$ from CNB, and $\hat{\phi}$ from *NB-logged* for $\epsilon = 0$.

			av(HP)			av($\hat{\alpha}$)	av($\hat{\phi}$)
α	β	m	*NB-null*	*NB-logged*	*NB-offset*	CNB	*NB-logged*
3.5	-0.4	50	3.433	0.006	0.008	3.507	1.014
		100	3.436	0.005	0.006	3.478	1.012
		200	3.470	0.005	0.005	3.495	1.011
		500	3.488	0.004	0.005	3.494	1.010
	-0.2	50	3.428	0.006	0.006	3.511	1.014
		100	3.454	0.005	0.006	3.494	1.011
		200	3.482	0.005	0.005	3.482	1.011
		500	3.491	0.005	0.005	3.498	1.010
	0	50	3.395	0.005	0.006	3.48	1.013
		100	3.461	0.005	0.006	3.494	1.012
		200	3.486	0.005	0.005	3.495	1.011
		500	3.481	0.004	0.005	3.489	1.010
0.5	-0.4	50	0.470	0.011	0.013	0.487	0.965
		100	0.487	0.012	0.013	0.495	0.968
		200	0.498	0.011	0.012	0.500	0.970
		500	0.498	0.011	0.012	0.499	0.968
	-0.2	50	0.475	0.012	0.014	0.486	0.972
		100	0.491	0.011	0.012	0.495	0.968
		200	0.493	0.011	0.012	0.497	0.967
		500	0.499	0.011	0.012	0.500	0.968
	0	50	0.479	0.011	0.013	0.490	0.970
		100	0.489	0.011	0.011	0.494	0.969
		200	0.495	0.011	0.012	0.498	0.968
		500	0.498	0.012	0.012	0.499	0.968

Table 7-3 Estimates of HP from NB models, $\hat{\alpha}$ from CNB, and $\hat{\phi}$ from *NB-logged* for $\epsilon = 0.25$.

			av(HP)			av($\hat{\alpha}$)	av($\hat{\phi}$)
α	β	m	*NB-null*	*NB-logged*	*NB-offset*	CNB	*NB-logged*
3.5	-0.4	50	3.359	0.225	0.256	3.622	0.927
		100	3.451	0.250	0.276	3.652	0.924
		200	3.476	0.260	0.284	3.661	0.920
		500	3.498	0.266	0.289	3.667	0.920
	-0.2	50	3.432	0.234	0.262	3.640	0.933
		100	3.460	0.251	0.277	3.660	0.924
		200	3.486	0.261	0.284	3.663	0.923
		500	3.488	0.267	0.289	3.658	0.921
	0	50	3.401	0.239	0.267	3.617	0.935
		100	3.449	0.259	0.283	3.647	0.927
		200	3.495	0.266	0.288	3.658	0.927
		500	3.490	0.269	0.291	3.657	0.924
0.5	-0.4	50	0.479	0.162	0.267	0.655	0.618
		100	0.489	0.172	0.273	0.661	0.615
		200	0.495	0.175	0.277	0.665	0.614
		500	0.497	0.179	0.281	0.666	0.611
	-0.2	50	0.479	0.163	0.268	0.652	0.619
		100	0.492	0.172	0.275	0.659	0.615
		200	0.495	0.176	0.278	0.661	0.614
		500	0.498	0.179	0.281	0.662	0.612
	0	50	0.477	0.165	0.27	0.644	0.615
		100	0.489	0.174	0.277	0.651	0.615
		200	0.492	0.176	0.279	0.654	0.612
		500	0.498	0.181	0.281	0.655	0.613

Table 7-4 Estimates of HP from NB models, $\hat{\alpha}$ from CNB, and $\hat{\phi}$ from *NB-logged* for $\epsilon = 0.5$.

			av(HP)			av($\hat{\alpha}$)	av($\hat{\phi}$)
α	β	m	*NB-null*	*NB-logged*	*NB-offset*	CNB	*NB-logged*
3.5	-0.4	50	3.396	0.479	0.573	3.785	0.857
		100	3.439	0.514	0.607	3.790	0.847
		200	3.480	0.531	0.623	3.819	0.843
		500	3.485	0.543	0.634	3.813	0.839
	-0.2	50	3.391	0.501	0.600	3.767	0.856
		100	3.445	0.518	0.610	3.793	0.850
		200	3.462	0.540	0.633	3.803	0.844
		500	3.496	0.549	0.639	3.816	0.842
	0	50	3.389	0.493	0.585	3.746	0.863
		100	3.452	0.533	0.624	3.779	0.854
		200	3.482	0.556	0.645	3.797	0.849
		500	3.498	0.555	0.644	3.809	0.845
0.5	-0.4	50	0.482	0.255	0.575	0.806	0.429
		100	0.486	0.264	0.581	0.807	0.422
		200	0.495	0.271	0.586	0.814	0.421
		500	0.497	0.274	0.591	0.818	0.418
	-0.2	50	0.479	0.252	0.574	0.796	0.429
		100	0.490	0.269	0.588	0.798	0.422
		200	0.496	0.274	0.595	0.809	0.420
		500	0.498	0.277	0.597	0.810	0.417
	0	50	0.480	0.257	0.582	0.784	0.429
		100	0.490	0.269	0.597	0.794	0.421
		200	0.494	0.275	0.600	0.796	0.418
		500	0.497	0.279	0.603	0.798	0.416

7.6 Discussion

In this chapter, the *NB-null*, *NB-logged, NB-offset,* and CNB models are examined in situations where the baseline and follow-up falls counts are collected using different methods. The four models were fitted to the EXSart dataset and compared in a simulation study, in which a perturbation was introduced into the baseline count to mimic datasets arising from the data collection methods in the EXSart dataset.

The CNB model assumes that the outcome and baseline counts are generated by a time-homogeneous Poisson process, that is, heterogeneity is subject-specific and does not change across periods. In Chapter 6, CNB models were shown to have great power and their type I error rates were close to target when correctly specified. However, if the baseline count is collected using a different method to the outcome count, this assumption does not hold and the CNB model performed poorly in Chapter 7. Firstly, it did not converge when fitted to either the intervention or follow-up falls count in the actual EXSart dataset. Secondly, it showed vastly inflated type I error rates in the simulation study, 50% higher than nominal levels when subject-specific heterogeneity was large.

The empirical type I error rates for *NB-offset* were closer to the nominal level 0.05 than for CNB but were still too high in all scenarios. Although the type I error rate from *NB-logged* model was also shown to be inflated for high subject-specific heterogeneity (that is, large α), it was close to 0.05 when $\alpha = 0.5$ irrespective of the level of perturbation (ϵ).

Because the baseline count is not included in *NB-null*, this model is not affected by the inconsistency between methods of collecting baseline and outcome counts. Although retrospective baseline counts are considered to have lower precision than prospectively collected baseline counts, it is still of value to incorporate the retrospective baseline counts in NB models. *NB-logged* showed higher power than *NB-null* in all scenarios and achieved better goodness of fit when fitted to the EXSart dataset. Another justification for including a retrospective baseline count is that the estimation of the intervention effect may be less influenced by very large outcome counts, as indicated by the DFBETA plots in Figure 7-5 and Figure 7-9, and the intervention effect estimated by *NB-logged* appeared more in line with the pattern shown in the BOE plots than the effect estimated in *NB-null*.

The simulations included scenarios resembling the EXSart dataset, and these results suggested that the Wald test of the intervention effect from the *NB-logged* in the actual data in Table 7-1 may be moderately liberal. However, in the actual data: 1) the Wald test did not indicate a significant intervention effect; 2) the simulations suggested that the effect estimated by *NB-logged* was unbiased; and 3) the model-based SE in *NB-logged* was only slightly smaller than the robust SE, which is a consistent SE estimator even when the model is incorrect (see section 2.3.7).

In conclusion, *NB-logged* is almost as powerful as CNB when CNB is correctly specified. When the baseline falls counts are collected retrospectively, *NB-logged* is likely to be preferable to CNB and *NB-offset* because it is more robust to increasing levels of perturbation in the baseline count. Though *NB-logged* accommodates the discrepancy to some extent, its type I error rate is inflated when the outcome counts are greatly overdispersed.

Chapter 8

Other count response models

In this chapter, alternative count response models to NB regression are fitted to the Goodwin et al. (2011) or EXSart (Ashburn et al., 2007) datasets, to examine their potential in modelling falls counts.

8.1 Poisson Inverse Gaussian model

Just as the NB model as a Poisson-gamma mixture, the Poisson Inverse Gaussian (PIG) model deals with overdispersion in a Poisson-inverse-Gaussian mixture. It is considered to have better performance for heavily skewed count data than NB models (Dean et al., 1989; Hilbe, 2014). The PIG model has previously been used by Canning et al. (2014) and Hauser et al. (2016) to analyse falls data from falls prevention trials in PwP.

Guo and Trivedi (2002) provided a parameterisation of the PIG model that is comparable to NB regression. The outcome Y is assumed to follow a distribution of:

$$\Pr(Y = y) = \int_0^\infty \Pr(Y = y|v) f(v) dv\,, \tag{8-1}$$

such that $Y|v \sim \mathrm{Poisson}(v)$, where the non-negative random variable v follows an inverse Gaussian distribution. The PDF of the inverse Gaussian is given by:

$$f(v; \tau, \mu) = \left(\frac{\tau}{2\pi v^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\tau(v-\mu)^2}{2\mu^2 v}\right), \tag{8-2}$$

where $\tau > 0$ is the shape parameter and $\mu > 0$ is the mean of v.

The mean and variance of the PIG model are μ and $\mu + \mu^3/\tau$ respectively. As with the NB model, the PIG variance may be parameterised with the reciprocal of τ, which yields the variance as $\mu + k\mu^3$ where $k = 1/\tau$. In this section, k is referred to as the Heterogeneity

Parameter (HP) of the PIG model. Because of the cubic form in the PIG variance function, the PIG model can accommodate more skewed data than the NB model (which has a variance $\mu + \alpha\mu^2$). If the distribution of the outcome falls count has a small mean but a long tail, the PIG method is potentially more suitable (Hilbe, 2014).

As in the NB model, the link function in the PIG model is the log function, therefore the approach of including a log-transformed baseline count as a covariate is appropriate for PIG models as well. Following the nomenclature earlier in Chapter 6, the PIG model ignoring the baseline count is referred to as *PIG-null*, and PIG including the logged baseline count is referred to as *PIG-logged*.

This section compares the performance of *NB-null* versus *PIG-null*, and *NB-logged* versus *PIG-logged* using the falls counts during the follow-up period of the EXSart trial (Ashburn et al., 2007). The EXSart dataset was chosen because the outcome count is heavily skewed, it was collected during a relatively long period (four months) so that the outliers have larger values.

The baseline count has no missing values, and the other baseline characteristics are not considered as covariates here because there are missing values in those variables for people reporting large follow-up counts. The statistical analysis was conducted in `Stata` using the `pigreg` command (Hardin and Hilbe, 2012).

Table 8-1 and Table 8-2 summaries the comparisons of *NB-null* versus *PIG-null* and *NB-logged* versus *PIG-logged*, respectively. The *NB-null* model is significantly overdispersed (P < 0.001), whilst no overdispersion test is available for PIG models. *PIG-null* has much smaller AIC than *NB-null* (627.7 versus 674.1). The HP from *PIG-null* (k) is greater than the HP (α) from *NB-null*. Because of the cubic form of its variance function, a PIG model tends to give a smaller estimate of μ (as indicated by the mean of predicted values), and a larger estimate of HP than the corresponding NB model.

As discussed in section 7.4, the intervention effect estimated from *NB-null* is too extreme (FRR = 0.147) and does not conform to the pattern shown in the BOE plots. In comparison, the FRR estimated from *PIG-null* (0.653) is closer to the effect size estimated from the *NB-logged* and *PIG-logged* models (see Table 8-2), both of which incorporate the baseline counts and are anticipated to yield more reliable estimates for the intervention effect. This indicates that the *PIG-null* model is less influenced by the outliers than *NB-null.*

As shown in Table 8-2, *PIG-logged* has smaller AIC than *NB-logged* (594.9 versus 604.2), but the difference is smaller than that between *PIG-null* and *NB-null*. Also, the intervention effect estimated from *NB-logged* and *PIG-logged* (FRR = 0.686 and 0.869) are also closer than the effect estimated from *NB-null* and *PIG-null*. This is because incorporating the baseline count reduces heterogeneity (*NB-logged* is not significantly overdispersed; P=0.079), and the advantage of the PIG model thus becomes smaller.

Table 8-1 EXSart dataset: *NB-null* and *PIG-null* models (follow-up period, n=127)

	NB-null				***PIG-null***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-1.916	0.357	0.147 (0.073, 0.296)	<0.001	-0.426	0.387	0.653 (0.306, 1.393)	0.270
HP	$\hat{\alpha}$ = 3.865				$\hat{k}$ = 15.803			
Mean(Predicted)	12.165				10.732			
NB overdispersion test	< 0.001				-			
AIC	674.1				627.7			

Table 8-2 EXSart dataset: *NB-logged* and *PIG-logged* models (follow-up period, n=127)

	NB-logged				***PIG-logged***			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-0.377	0.286	0.686 (0.391, 1.202)	0.188	-0.140	0.305	0.869 (0.478, 1.580)	0.646
Log(baseline count)	0.685	0.078	1.984 (1.701, 2.313)	<0.001	0.530	0.074	1.699 (1.470, 1.963)	<0.001
HP	1.958				4.110			
Mean(Predicted)	8.857				6.231			
NB overdispersion test	0.079				-			
AIC	604.2				594.9			

In conclusion, the PIG model may be an alternative of the NB model when outcome counts are heavily skewed, especially when a baseline count is not available, or NB model is significantly overdispersed.

8.2 Zero-inflated NB models

In this section Zero-Inflated NB (ZINB) models are considered in relation to the falls data from the Goodwin et al. (2011) trial. When there are excessive zero counts, some of them may be generated from a different process to the count process. In the context of falls prevention trials, zero-inflation may arise if participants fail to complete their falls diaries but still hand them in.

It is difficult to assess whether a large proportion of zeros is due to zero-inflation by inspecting the distribution of the outcome count. The ZINB model is compared with the standard NB model using the Goodwin et al. dataset. The baseline characteristics are included as covariates to improve the explanatory power of the model. One participant (ID 1) is not included in the models due to missing value in the baseline count.

The Vuong test was used to test for zero-inflation, and the AIC- and BIC-based corrections also carried out (section 2.3.3). In addition, the covariate-adjusted probability plot (Holling et al., 2016) was used to provide a graphical comparison of the fitted ZINB and NB models to assess whether zero-inflated models could be beneficial in modelling this dataset. The NB and ZINB models were fitted in `R` using the **MASS** and **pscl** packages (described in section 4.5). The results of the standard and AIC-/BIC- corrected Vuong tests were calculated using the `vuong()` function in the **pscl** package.

Table 8-3 and Table 8-4 display the results of ZINB and NB models. No inflation covariates (section 2.3.3) are included for the ZINB models (the ZINB models with inflation covariates were fitted but did not converge). The inflated zeros accounts for 3.1% and 7.8% of all zeros in the intervention and follow-up count, respectively.

In general, ZINB and NB models result in similar estimates. Although the NB models have marginally higher AIC than the ZINB models, the Vuong test of zero-inflation is not significant for either the intervention (P=0.239) or the follow-up counts (P=0.245). For the models fitted to the intervention count, the Vuong test with the AIC-based correction does not suggest significant zero-inflation (P=0.413), while the test with the BIC-based correction suggests that the standard NB model fits the data better (Z = -0.470), but the test result is also not significant (P = 0.319). For the models fitted to the follow-up count, the Vuong tests with both the AIC- and BIC- based correction suggests that the standard NB model fits the data significantly better than the ZINB model (AIC-based test: Z = -3.033, P = 0.001; BIC-based test: Z = -8.415, P < 0.001). As the Vuong test without correction is considered biased in favour of the zero-inflated models (see section 2.3.3), the results suggest that there is no evidence of zero-inflation for either the intervention or follow-up count.

Figure 8-1 and Figure 8-2 show the covariate-adjusted probabilities of the ZINB and NB models that are fitted to the intervention and follow-up counts, respectively. The plots suggest that the covariate-adjusted probability of NB models is close to the observed probability for zeros falls, and the ZINB models only have marginal improvements over the NB models.

Table 8-3 ZINB versus NB model fitted to the Goodwin et al. data: intervention period (n=124)

	ZINB				**NB**			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-0.416	0.151	0.660 (0.123, 3.105)	0.006	-0.441	0.155	0.643 (0.473, 0.874)	0.004
Log(baseline count+0.5)	0.933	0.054	2.542 (0.489, 0.890)	<0.001	0.946	0.056	2.574 (2.303, 2.878)	0.000
Female	-0.128	0.165	0.880 (2.284, 2.830)	0.438	-0.062	0.167	0.940 (0.675, 1.307)	0.708
Age	0.008	0.011	1.008 (0.634, 1.221)	0.476	0.012	0.010	1.012 (0.992, 1.033)	0.243
Years since diagnosis	0.033	0.014	1.034 (0.986, 1.029)	0.022	0.027	0.014	1.028 (1.000, 1.056)	0.046
Hoehn & Yahr								
Stage 1	0.040	0.262	1.041 (1.005, 1.064)	0.878	0.019	0.285	1.019 (0.580, 1.791)	0.947
Stage 2			1				1	
Stage 3	-0.279	0.190	0.756 (0.619, 1.751)	0.142	-0.233	0.191	0.792 (0.543, 1.155)	0.221
Stage 4	-0.096	0.250	0.909 (0.519, 1.103)	0.701	-0.070	0.249	0.932 (0.569, 1.526)	0.777
Living status								
With partner			1					
Alone	0.307	0.186	1.360 (0.554, 1.491)	0.099	0.284	0.194	1.328 (0.905, 1.948)	0.143
With family/friends	1.306	0.635	3.693 (0.940, 1.968)	0.040	1.421	0.629	4.141 (1.191, 14.392)	0.024
Residential home	-1.307	1.147	0.271 (1.050, 12.994)	0.254	-1.373	1.107	0.253 (0.028, 2.272)	0.215
			Percentage					
Zero-inflation (intercept)	-3.432	0.862	3.1%					
HP	0.384				0.468			
AIC	749.8				750.7			
Vuong test	Z = 0.710, P = 0.239							
Vuong test (AIC-corrected)	Z = 0.221, P = 0.413							
Vuong test (BIC-corrected)	Z = -0.470, P = 0.319							

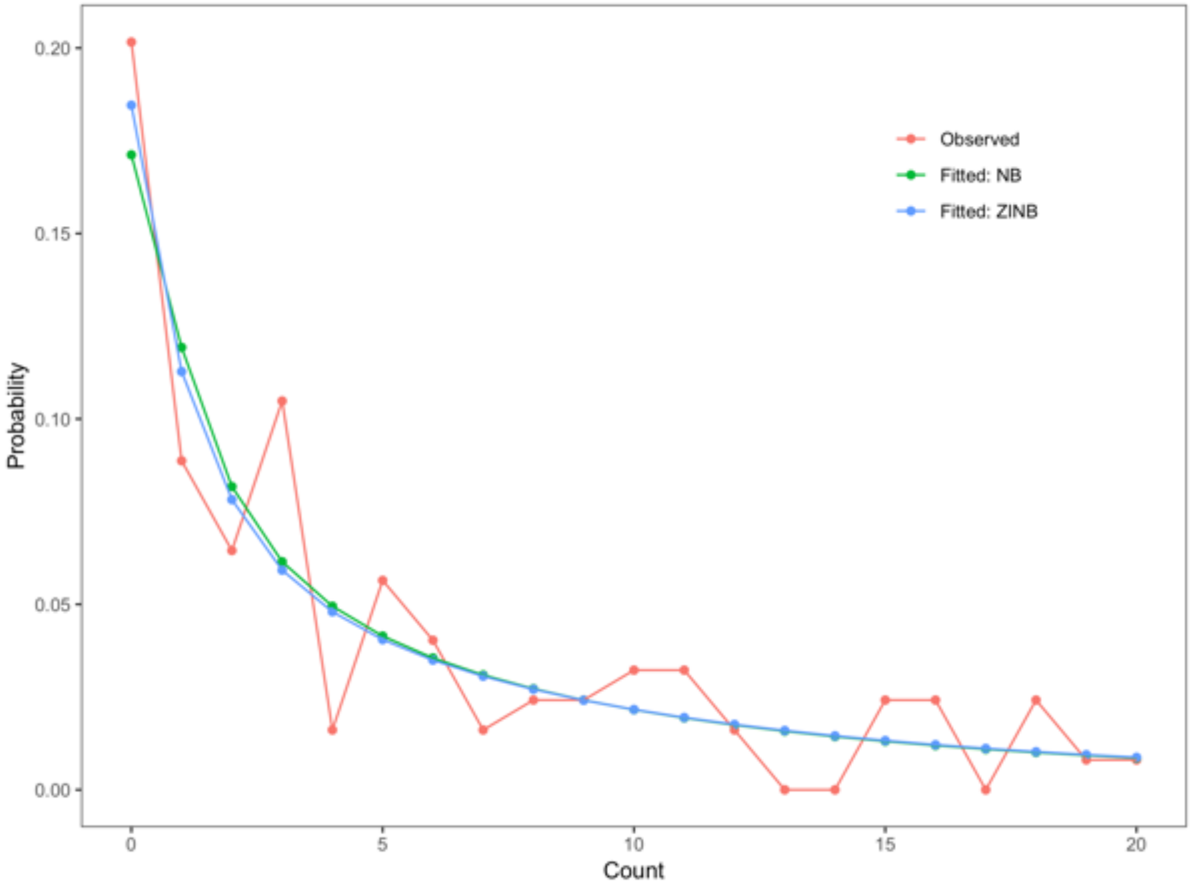


Figure 8-1 Covariate-adjusted probability plot: NB versus ZINB models fitted to the Goodwin et al. dataset (intervention period, n=124)

Table 8-4 ZINB versus NB model fitted to the Goodwin et al. data: follow-up period (n=115)

	ZINB				NB			
	Est.	SE	FRR (95% CI)	P	Est.	SE	FRR (95% CI)	P
Intervention	-0.249	0.228	0.780 (0.496, 1.226)	0.275	-0.297	0.227	0.743 (0.474, 1.165)	0.190
Log(baseline count+0.5)	0.955	0.077	2.599 (2.231, 3.029)	<0.001	0.968	0.082	2.632 (2.236, 3.099)	0.000
Female	0.110	0.231	1.116 (0.706, 1.765)	0.633	0.071	0.239	1.073 (0.668, 1.724)	0.768
Age	0.002	0.017	1.002 (0.970, 1.036)	0.897	0.007	0.015	1.007 (0.978, 1.038)	0.630
Years since diagnosis	0.023	0.020	1.023 (0.983, 1.065)	0.253	0.023	0.020	1.024 (0.985, 1.064)	0.233
Hoehn & Yahr								
Stage 1	0.518	0.351	1.679 (0.837, 3.370)	0.140	0.525	0.385	1.691 (0.789, 3.626)	0.172
Stage 2			1				1	
Stage 3	-0.180	0.277	0.835 (0.482, 1.446)	0.515	-0.163	0.286	0.850 (0.482, 1.498)	0.569
Stage 4	0.147	0.346	1.159 (0.583, 2.300)	0.670	0.148	0.360	1.160 (0.568, 2.369)	0.681
Living status								
With partner			1					
Alone	0.040	0.278	1.041 (0.599, 1.808)	0.885	0.014	0.279	1.014 (0.583, 1.764)	0.959
With family/friends	0.746	0.907	2.108 (0.349, 12.752)	0.411	1.008	0.905	2.740 (0.455, 16.499)	0.265
Residential home	-0.569	0.938	0.566 (0.088, 3.642)	0.544	-0.500	0.991	0.606 (0.085, 4.326)	0.614
			Percentage					
Zero-inflation (intercept)	-2.465	0.701	7.8%					
HP	0.789				1.065			
AIC	678.7				678.9			
Vuong test	Z = 0.889, P = 0.245							
Vuong test (AIC-corrected)	Z = -3.033, P = 0.001							
Vuong test (BIC-corrected)	Z = -8.415, P < 0.001							

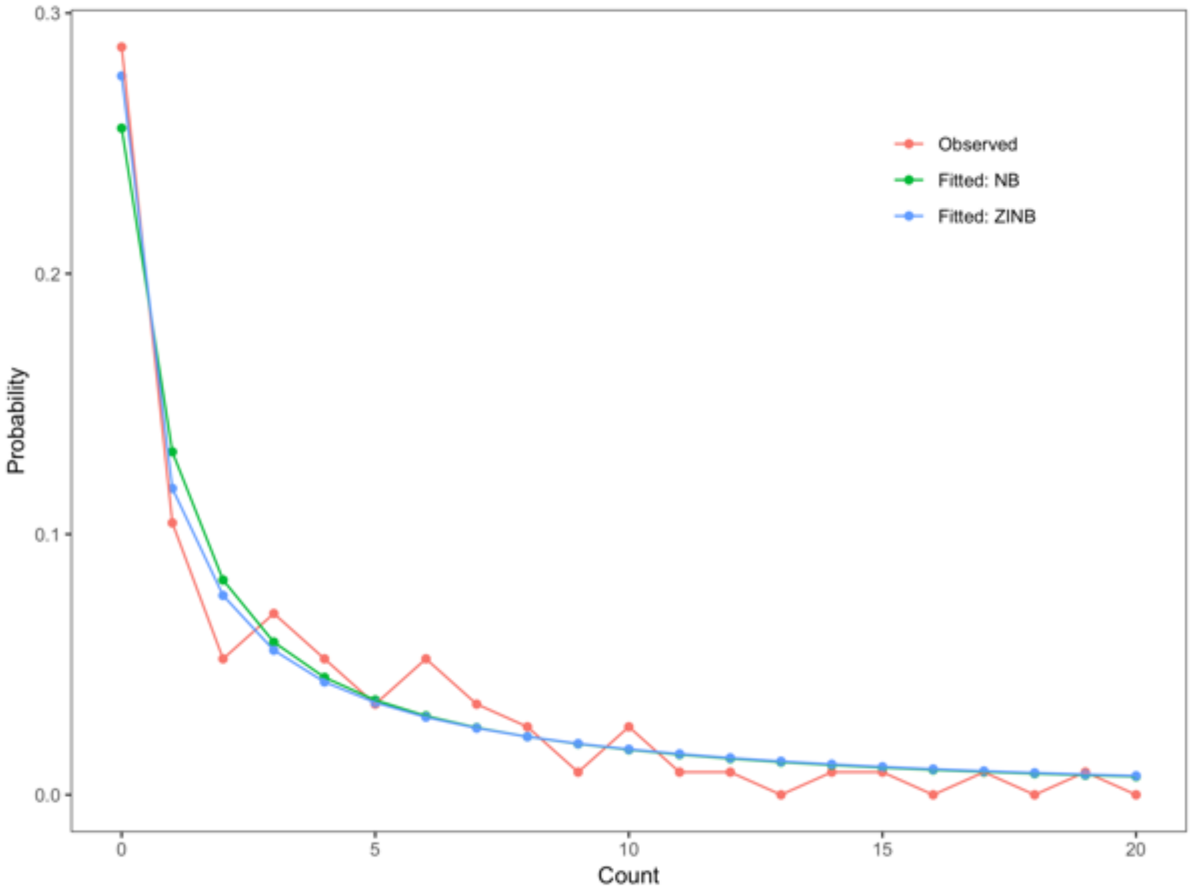


Figure 8-2 Covariate-adjusted probability plot: NB versus ZINB models fitted to the Goodwin et al. dataset (follow-up period, n=115)

In conclusion, the results showed that zero-inflation was not a major issue for the Goodwin et al. dataset. A possible reason is that, although the trial participants may skip recording falls when falling frequently, this is not considered a great issue for people who seldom fall, so the participants are not likely to skip recording falls altogether during an observation period. It is possible that ZINB might be useful for other datasets.

8.3 Right-censored and right-truncated NB models

As discussed in section 3.2.1 and Chapter 6, the Goodwin et al. (2011) dataset has consistent falls rates across periods, but some frequently falling participants in the intervention group dropped out after the intervention period, which resulted in group imbalance of frequent fallers during the follow-up period: among the ten participants who record the most falls during the follow-up period, only one was from the intervention group (see Table 3-3). Therefore, the NB model without including the baseline count yielded a large intervention effect for the follow-up period (see Table 3-7), but it was largely due to the group imbalance (the frequent fallers resulting in a higher average falls rate in the control group). Although the *NB-logged* model controls for the group imbalance of frequent fallers by adjusting for the baseline fall rate, a baseline count may not be available for some trials.

An approach for coping with the influential outliers is to choose a cut-point, drop the outliers that are greater than the cut-point, and then fit a standard NB model. This approach was adopted in a number of studies. A falls prevention trial in elderly people (Liu-Ambrose et al., 2008) reported that "a falls histogram revealed two outliers," and the NB model was fitted to the falls count with "these cases removed." Gill et al. (2009) compared regression models (including NB regression) for analysing the risk of falling using a dataset from a falls trial, in which one subject was excluded from analysis because the falls count recorded by him was too large. Another falls study (Stanaway et al., 2011) reported that "participants with a fall rate of 10 or more falls per year were excluded," and "next, a negative binomial multivariate analysis was carried out." A falls prevention trial (Cumming et al., 1999) fitted the NB model to the falls count, but three participants were excluded because they "reported more than 50 falls during follow-up." In a trial in PwP (Henderson et al., 2016), where falls were collected as a secondary outcome, the standard NB model was fitted to the falls count with an outlier excluded from the analysis, because this participant reported a large falls count (1122 falls) during the treatment period.

The concern about excluding large counts from modelling is that it results in a right-truncated distribution for the outcome count, and thus it would be more appropriate to fit a right-truncated NB model (see section 2.3.3), but in practice this is ignored.

Cameron and Trivedi (2013) mentioned another approach to the issue of outliers, that is to "downweight" the influence of the large counts by 1) reducing large counts by right-censoring at a chosen cut-point if the count is greater or equal to the cut-point, and 2) fitting a right-censored NB model (see section 2.3.3). This method keeps the individuals with large counts in the analysis, so that outliers are still included in the analysis but less influential than in a standard NB model.

An incorrect approach would be to revalue the large counts to the cut-point and then fit a standard NB model, but this has been done in practice. In a falls prevention trial for elderly people (Nikolaus and Bach, 2003), "the first five falls for each participant were used in this analysis rather than all falls (maximum 22) to avoid over-weighting by subjects who fell more than five times." In another falls prevention trial (Ryan et al., 2010), the NB model was fitted to the dataset in which "the number of falls by each person was truncated to an arbitrary threshold of 15, chosen as it corresponds to one fall per diary." In addition, the authors conducted "post-hoc sensitivity analyses to account for exclusion of the outlier from the calculation of fall rate," in which they "repeated the negative binomial regression model but including the outlier, who was assigned the next highest value of falls (number of falls plus one of next highest participant in that group)." It would be appropriate to use the right-censored NB model in these trials, but in practice the standard NB model was fitted.

In this section, right-censored and right-truncated models are fitted in the scenario where large outcome counts are not balanced between groups and a baseline falls count is not considered. Five models were fitted to the Goodwin et al. dataset for comparison:

- ***NB-null***: the standard NB model fitted to the complete dataset with only one covariate—group allocation. *NB-null* is included as a baseline model, as it was shown to be subject to group imbalance in section 3.2.1.
- ***NB-logged***: the standard NB model fitted to the complete dataset with two covariates—group allocation, and the logged baseline count (0.5 was added to include zero baseline counts). This model is included as a benchmark model, because it incorporates the baseline count, which controls for the group imbalance of large outcome counts. As the frequent fallers had consistent falls rates across periods, this model is anticipated to have a more reliable estimate of the intervention effect than *NB-null*.
- ***NB-reduced***: the standard *NB-null* model fitted to a reduced dataset: the outcome counts that are greater than the cut-point are dropped;
- ***NB-rt***: the right-truncated NB model with the same covariate as *NB-null* and fitted to the same reduced dataset as *NB-reduced*. The distribution of the outcome count in *NB*-rt is specified as right-truncated at the cut-point.

- ***NB-rc*****:** the right-censored NB model with the same covariate as *NB-null*. The outcome counts that are greater than the cut-point are regarded as right-censored at the cut-point.

To make *NB-logged* comparable to the other models, one participant (ID 1) with missing value in the baseline count was not included in the five models.

A cut-point was chosen as 200, based on the BOE plots of *NB-null* as shown in Figure 5-6 to Figure 5-9: the participants reporting more than 200 falls (20 falls/week) during the intervention/follow-up period were generally the most influential subjects to the estimation of *NB-null*, as indicated by their high Cook’s distances; also, their large DFBETA for the intervention effect suggests that these large counts may have a considerable influence in the estimation of the intervention effect. Another cut-point 60 was considered to show the difference between choosing a large and a small value for cut-points.

The statistical analysis was conducted in R (version 3.5.0), using the **MASS** package to fit the NB models, the **gamlss.cens** package to fit the right-censored NB models, and the **gamlss.tr** package to fit the right-truncated NB models (the three packages are described in section 4.5).

Table 8-5 and Table 8-6 summarise the comparisons of the models fitted to the intervention and follow-up falls counts in the Goodwin et al. dataset. As anticipated, the *NB-logged* models have smaller AIC than *NB-null*. For the same dataset and cut-point, the *NB-rt* models have smaller AIC than *NB-reduced*: the difference in AIC is only marginal for cut-point 200 but is larger for cut-point 60. This indicates that, by correctly specifying the distribution underlying the response variable, the *NB-rt* model achieves better goodness of fit than *NB-reduced*, which ignores the right-truncation.

For the intervention period, four participants reported a larger outcome count than the cut-point 200 (see Table 8-5). FRRs estimated by both *NB-rc* and *NB-rt* are closer to the FRR from *NB-logged* than that from *NB-null*, with *NB-rt* being closer than *NB-rc*. Both models also result in smaller HPs than *NB-null*, because the outliers are either censored or dropped.

Table 8-5 *NB-null*, *NB-logged*, *NB-reduced*, *NB-rt*, and *NB-rc* fitted to the intervention count in the Goodwin et al. dataset

Cut-point	Model	n	AIC*	$\widehat{\beta}$ (SE)	FRR (95% CI)	P	$\widehat{\phi}$ (SE)	HP
Original	*NB-null*	124	931.8	-0.571 (0.323)	0.565 (0.298, 1.071)	0.077		3.188
	NB-logged	124	744.3	-0.359 (0.155)	0.698 (0.946, 0.515)	<0.001	0.911 (0.051)	0.511
200	*NB-reduced*	120	826.7	-0.232 (0.284)	0.793 (0.452, 1.391)	0.414		2.339
	NB-rt	120	826.6	-0.238 (0.284)	0.788 (1.375, 0.452)	0.400		2.347
	NB-rc	124	875.5	-0.536 (0.309)	0.585 (1.072, 0.319)	0.085		2.898
60	*NB-reduced*	117	767.6	-0.343 (0.268)	0.710 (0.417, 1.206)	0.200		2.001
	NB-rt	117	761.0	-0.638 (0.282)	0.528 (0.304, 0.918)	0.026		2.257
	NB-rc	124	814.4	-0.525 (0.278)	0.592 (0.343, 1.020)	0.061		2.314

* Note: the AIC of *NB-rc* cannot be compared with the AIC of *NB-reduced* and *NB-rt* models

For the smaller cut-point 60, more subjects (seven participants) are regarded as right-censored or right-truncated in modelling the interventon count. Compared to the *NB-rc* model with cut-point 200*, NB-rc* with cut-point 60 yields an FRR that is closer to that estimated from *NB-logged*. However, *NB-rt* with cut-point 60 yields an FRR of 0.528, which is further away from the FRR estiamted from *NB-logged* (0.698) compared to the FRR from *NB-null* (0.565). This may because dropping the seven participant results in loss of information in *NB-rt*.

As shown in Table 8-6, the models fitted to the follow-up falls counts show similar results. Compared to the intervention period, there is a greater disparity between the FRRs estimated from *NB-null* and *NB-logged* (0.235 and 0.770, respectively). Only one participant has a greater follow-up count than the cut-point 200, and seven participants reported more follow-up falls than the cut-point 60. The FRR estimated from *NB-rt* are closer to the estimate from *NB-logged* than to that from *NB-rc*, while the FRRs from both *NB-rc* and *NB-rt* are less extreme than that from *NB-null*. For cut-point 60, the FRR from the *NB-rt* model (0.776, 95% CI: 0.405 to 1.487) is remarkably close to the FRR from *NB-logged* (0.770, 95% CI: 0.493, 1.201), with the 95% CI from *NB-rt* wider than *NB-logged*. This is different to the finding in Table 8-5: the cut-point 60 results in a more extreme FRR than the cut-point 200 when *NB-rt* is fitted the intervention count. A possible explanation is that the large falls counts during the follow-up periods are not balanced between groups, which results in an extreme FRR in *NB-null* (0.235), but the large counts are not included in *NB-rt*, which may remedy the issue of losing information.

Table 8-6 *NB-null*, *NB-logged*, *NB-reduced*, *NB-rt*, and *NB-rc* fitted to the follow-up count in the Goodwin et al. dataset

Cut-point	Model	n	AIC*	$\widehat{\beta}$ (SE)	FRR (95% CI)	P	$\widehat{\phi}$ (SE)	HP
Original	*NB-null*	115	771.6	-1.448 (0.352)	0.235 (0.117, 0.472)	<0.001		3.468
	NB-logged	115	666.5	-0.262 (0.227)	0.770 (0.493, 1.201)	0.249	0.943 (0.075)	1.142
200	*NB-reduced*	114	741.9	-1.028 (0.335)	0.358 (0.184, 0.696)	0.002		3.110
	NB-rt	114	740.9	-1.143 (0.339)	0.319 (0.164, 0.620)	0.001		2.875
	NB-rc	115	751.2	-1.208 (0.340)	0.299 (0.153, 0.582)	<0.001		3.034
60	*NB-reduced*	108	640.0	-0.149 (0.319)	0.861 (0.458, 1.621)	0.639		2.622
	NB-rt	108	636.0	-0.254 (0.332)	0.776 (0.405, 1.487)	0.446		2.622
	NB-rc	115	682.7	-0.995 (0.330)	0.370 (0.194, 0.706)	0.003		1.142

* Note: the AIC of *NB-rc* cannot be compared with the AIC of *NB-reduced* and *NB-rt* models

The results indicate that the right-censored and right-truncated NB models may be useful when the baseline count is not available, especially when the frequent fallers are not balanced between groups.

An advantage of the right-censored model is that choosing a smaller cut-point does not result in reduced sample size. However, if large outcome counts are not balanced between two groups (such as the follow-up periods in Goodwin et al. dataset), the right-censored model would only extenuate the effect. If there are a large number of influential outliers, censoring will not provide a complete remedy.

The right-truncated model obviates the effect of outliers by dropping them, thus would solve the issue of group imbalance. However, if the value of the cut-point is too small, many observations are dropped, which results in a loss of power.

8.4 Finite mixture Poisson models

A common assumption in statistical modelling is to describe the distribution of the outcome y in a sample using a distribution $f(y|\lambda)$ where λ is the parameter of the population. In practice, the assumption is often too strict because of unobserved population heterogeneity — the population of interest may consist of multiple subpopulations, each of which has a different parameter λ (see section 2.3.7). For example, subpopulations may respond to an intervention differently. The same intervention may have a different effect for a subgroup of PwP who fall frequently and a subgroup who seldomly fall. If the intervention effect is large for one subpopulation and mild for another, the marginal intervention effect may be moderate, but this estimated effect is misleading. Finite mixture models aim to estimate the proportion of each component in the population, and the effect of interest for each subpopulation (Schlattmann, 2009).

Finite mixture models are relevant to falls prevention trials for two reasons: first, the falls counts data may be heterogeneous in nature. The finite mixture model resembles the NB model in that they both aim to accommodate heterogeneity. The difference between them is that, unlike the NB model, the finite mixture model does not assume a continuous distribution for the heterogeneity, instead it uses a discrete unobserved structure for heterogeneity (Schlattmann, 2009), which may potentially make the finite mixture model useful for dealing with outliers. In this section a finite mixture Poisson model is fitted to the intervention falls count in the Goodwin et al. (2011) dataset to examine whether it could accommodate the frequent fallers as a subpopulation.

Three models are fitted and compared:

- *Poi-logged*: the standard *Poi-logged* model (see section 6.2.2) is fitted (using the **poisson** command in `Stata)` as a baseline model;

- *NB-logged*: the standard *NB-logged* model is fitted (using the **nbreg** command in Stata) as a Poisson mixture model with a continuous random subject effect to accommodate the subject heterogeneity;
- *Poi-logged* FMM: the finite mixture *Poi-logged* model is fitted to the dataset (using the **fmm** command in Stata). Two components were specified to examine whether the large counts could be accommodated in a component. The linear predictor of each component includes the same covariates: the group allocation and Log(baseline count + 0.5).

The estimates from the three models are summarised in Table 8-7. The *Poi-logged* FMM yields a much smaller AIC (855.3) than *Poi-logged*, but still larger than that of *NB-logged* (744.3). This suggests that the FMM partially controls for the heterogeneity in two components, but *NB-logged* better accommodates the heterogeneity in the gamma distributed subject effect. The *Poi-logged* FMM results in an estimate of the intervention effect for each subpopulation: FRR 0.411 for component 1 and FRR 0.633 for component 2.

Table 8-7 Estimates of *Poi-logged*, *NB-logged*, and *Poi-logged* FMM (n=124).

		Estimate	SE	FRR (95% CI)	P
Poi-logged	Intervention	-0.480	0.037	0.619 (0.575, 0.666)	< 0.001
	Log(baseline count + 0.5)	1.030	0.012	2.801 (2.735, 2.868)	< 0.001
	AIC	1131.5			
NB-logged	Intervention	-0.359	0.156	0.698 (0.514, 0.948)	0.022
	Log(baseline count + 0.5)	0.911	0.048	2.487 (2.263, 2.733)	< 0.001
	HP				
	AIC	744.3			
Poi-logged FMM					
Comp 1	Intervention	-0.890	0.121	0.411 (0.324, 0.520)	< 0.001
	Log(baseline count + 0.5)	1.025	0.070	2.787 (2.430, 3.200)	< 0.001
Comp 2	Intervention	-0.457	0.045	0.633 (0.579, 0.691)	< 0.001
	Log(baseline count + 0.5)	0.894	0.027	2.445 (2.316, 2.581)	< 0.001
	AIC	835.3			

As shown in Table 8-8, the proportions of components 1 and 2 are estimated to be 56.1% and 43.9%. This was not anticipated as 1) the component with a larger marginal mean (that is, component 2) accounts for more than forty percent of all participants, while there are only a few frequent fallers in the dataset; 2) the marginal mean of component 2 (33.251) is not as large as expected.

Table 8-8 Estimated proportions and marginal means for components 1 and 2 from the *Poi-logged* FMM (n=124).

	Proportion			**Marginal mean**		
	Estimate	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**
Comp 1	0.561	0.100	(0.366, 0.738)	14.895	0.884	(13.162, 16.627)
Comp 2	0.439	0.100	(0.262, 0.634)	33.251	0.464	(30.382, 36.120)

Figure 8-3 shows a density histogram of the predicted values from components 1 and 2 of *Poi-logged* FMM. Although component 2 has a larger marginal mean than component 1 (see Table 8-8), it is mostly made up of falls counts between 0 and 50.

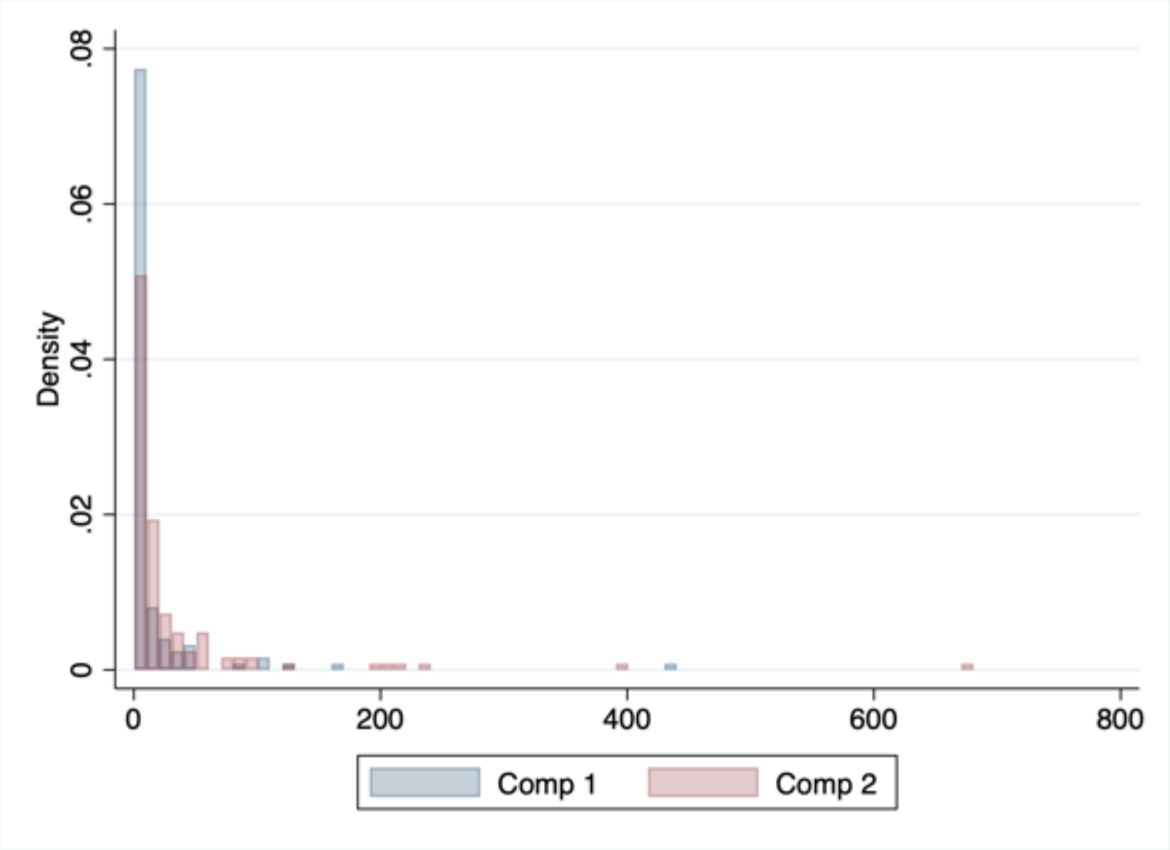


Figure 8-3 Density histogram of the predicted values from components 1 and 2 of *Poi-logged* FMM (n=124).

In summary, *Poi-logged* FMM did not appear to outperform *NB-logged* in the Goodwin et al. dataset. However, this may be due to the scarcity of large counts. If the frequent fallers indeed belong to a subpopulation, it is possible that an FMM could capture the large counts in a component when the sample size is large. In such case, it would be interesting to examine whether this subpopulation of frequent fallers respond to the same intervention differently compared to other subpopulation.

8.5 Analysis of longitudinal falls data with random-effects NB models

Datasets from falls prevention trials are often longitudinal in nature. When the falls counts are collected prospectively using falls diaries, they may be made available as a longitudinal dataset, with each row containing the falls counts per week recorded by a participant, for example. The participant and the week are usually indicated by the variables *id* and *time* respectively.

The random-effects NB model, which assumes observations are independent between participants within a week but allowed to be correlated over time within a participant (Cameron and Trivedi, 2013; Hausman et al., 1984) and introduced in section 2.3.8, is considered in this section to examine the longitudinal structure of the Goodwin et al. (2011) data (2011) in more detail.

This section focuses on the estimates of intervention and time effects, as well as their interactions. In addition to the logged baseline count, the same baseline characteristics such as demographic characteristics and severity rating of Parkinson's that were considered in section 3.2.1 are included in all the models in this chapter to improve statistical power and control for heterogeneity, but we shall forego discussing the estimates and test results of the baseline characteristics in detail.

As introduced in section 1.2.1, the falls count in the Goodwin et al. trial was recorded prospectively by each participant in a falls diary, and made available to this project as the number of falls experienced by each participant during each of the 30 weeks of observation window, which comprises 10 weeks for each of the baseline, intervention, and follow-up periods. Previously weekly count data were aggregated to the period level. The random-effects NB model is considered in this section for analysing the longitudinal dataset of weekly falls counts.

A random-effects NB model (see section 2.3.8) was fitted to the 10 weekly falls counts during the intervention and follow-up periods separately. The models were fitted in `Stata` (version 15) using the `xtnbreg, re` command (described in section 4.2). The group allocation was included in the model as a factor giving an estimate of the intervention effect. The logged baseline count (the aggregated falls counts during the baseline period + 0.5) was also included. Time (counted in weeks) from the beginning of week 11, when the intervention started, was included in the model as a regressor to examine whether the falls rate changed linearly over time in the control group. The same baseline characteristics as in section 3.2.1 were also included. The P values from the Wald test for each covariate were reported. This model is referred to as the *Linear Time* (LT) model in the following text.

Another goal of this section is to see whether there is an interaction effect between the intervention and time, by 1) fitting a model including the interaction term between intervention and linear time in addition to the two main effects and baseline characteristics; and 2) testing the interaction effect using the LR test. This model is referred to as the *Linear Time Interaction* (LTI) model.

To check the linearity assumptions in the LT and LTI models, the following random-effects NB models were examined:

- *Factorial Time* (FT) model: The FT model includes the same covariates as the LT model, except that time is included as a factor instead of as a continuous variable. The FT model does not assume a linear time effect, and it gives an estimate of the time effect specific to each week compared to the first week. This model was examined to check the assumption of a linear time effect in the LT models.
- *Factorial Time Interaction* (FTI) model: The FTI model includes the same covariates as the LTI model, except that time is included as a factor in both the main effect and the interaction between interventions and time. The FTI model does not assume the intervention effect changes linearly with time, and it gives an estimate of the intervention effect specific to each week. This model was examined to check the linear assumption in the LTI models.

For both the FT and FTI models, the first week of the intervention/follow-up period was specified as the reference level for the factor variable time.

All the fitted random-effects NB models were tested against the NB1 models (which were fitted to the panel data but the panel structure is ignored; see section 2.3.8) using the default LR test produced in Stata to examine if accounting for the panel structure significantly improves the goodness of fit of the random-effects NB model.

The matrix of correlation coefficients between outcome weekly counts were examined. Because the distribution of falls count is typically skewed, the Spearman correlation coefficient ρ was used.

Figure 8-4 and Figure 8-5 present Spearman correlation coefficients between falls counts recorded by the same participant during each week within the intervention and follow-up periods respectively. The weekly falls counts show strong correlation at the subject-level, with the Spearman ρ lying between 0.549 and 0.797 for the intervention period, and between 0.538 and 0.817 for the follow-up period (the majority of ρ being over 0.6). No clear pattern of autocorrelation over time is found in either period.

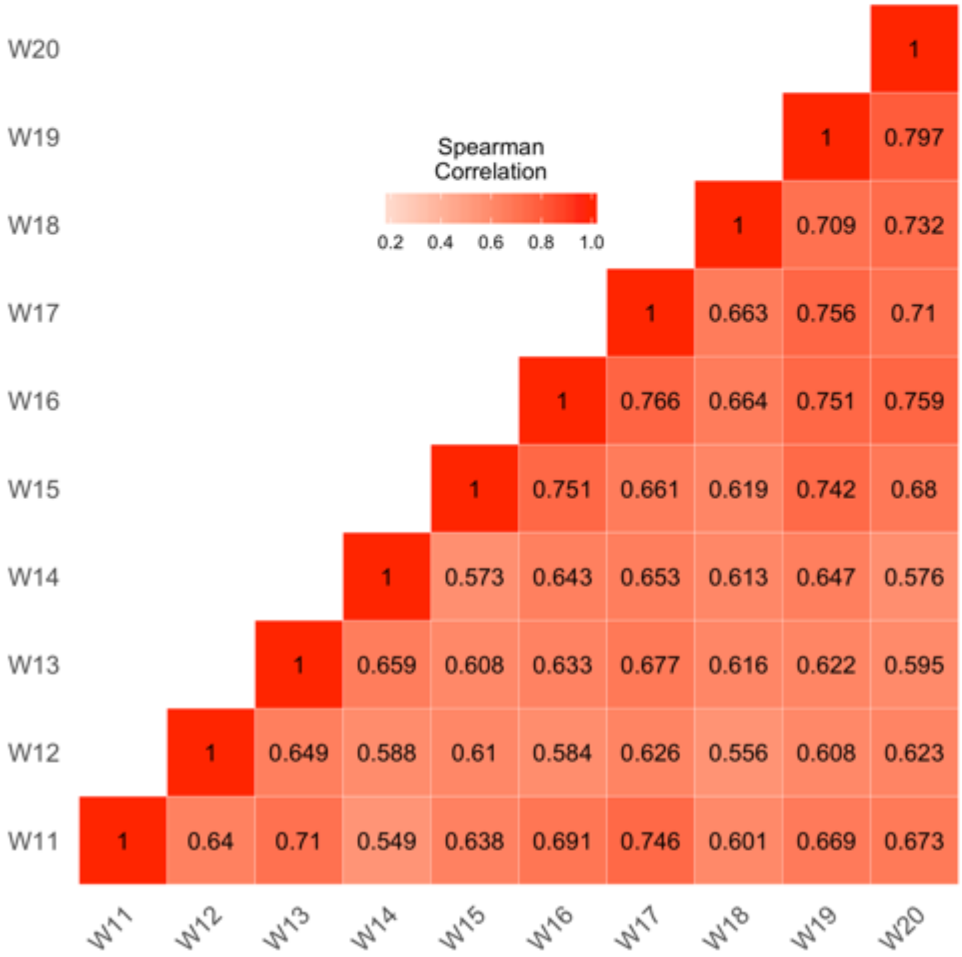
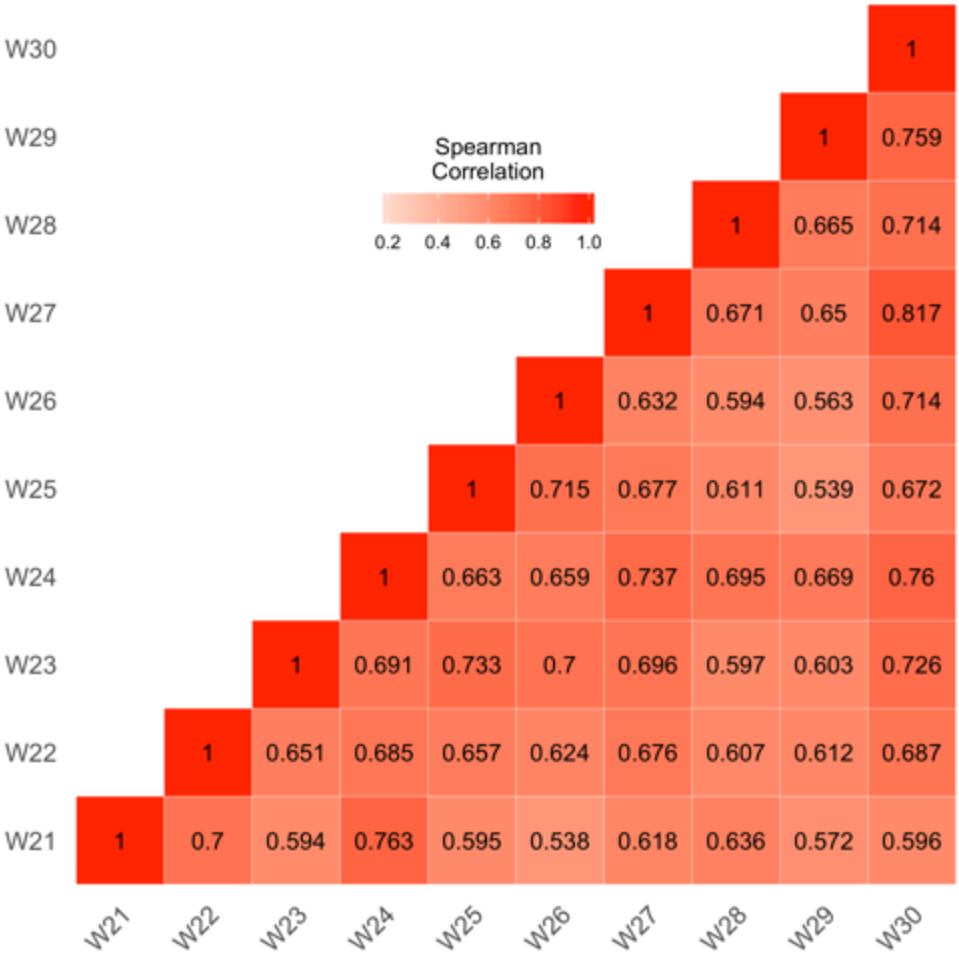


Figure 8-4 Correlation of the number of falls occurred each week (weeks 11-20)



	W21	W22	W23	W24	W25	W26	W27	W28	W29	W30
W30										1
W29									1	0.759
W28								1	0.665	0.714
W27							1	0.671	0.65	0.817
W26						1	0.632	0.594	0.563	0.714
W25					1	0.715	0.677	0.611	0.539	0.672
W24				1	0.663	0.659	0.737	0.695	0.669	0.76
W23			1	0.691	0.733	0.7	0.696	0.597	0.603	0.726
W22		1	0.651	0.685	0.657	0.624	0.676	0.607	0.612	0.687
W21	1	0.7	0.594	0.763	0.595	0.538	0.618	0.636	0.572	0.596

Figure 8-5 Correlation of the number of falls occurred each week (weeks 21-30)

The random-effects NB models fitted to the falls count during each week of the intervention (weeks 11-20) and follow-up (weeks 21-30) periods are summarized in Table 8-9 to Table 8-14. The LR tests indicate that all the random-effects models have significantly better goodness of fit than the NB1 model fitted to the same weekly dataset (P < 0.001).

As shown in Table 8-9, during the intervention period the intervention significantly (P < 0.001) reduced the falls rate by half (FRR: 0.505, 95% CI: 0.360 to 0.710), and the effect of time is not significant (P = 0.805). The FRRs for the time effect were examined in the FT model (Table 8-11), and they do not show patterns contraindicating the assumption of linear time effect.

Table 8-9 Goodwin et al. dataset: LT model (intervention period: weeks 11-20; n=124; obs=1240)

	Estimate	SE	FRR (95% CI)	P
Intervention	-0.683	0.173	0.505 (0.360, 0.710)	< 0.001
Time (weeks)	-0.002	0.009	0.998 (0.981, 1.015)	0.805
Log(baseline count + 0.5)	0.632	0.065	1.882 (1.658, 2.136)	< 0.001
Female	-0.542	0.203	0.582 (0.391, 0.866)	0.008
Age	-0.017	0.012	0.983 (0.959, 1.007)	0.166
Years since diagnosis	0.068	0.018	1.070 (1.034, 1.108)	0.000
Hoehn & Yahr				
Stage 1	0.296	0.330	1.344 (0.704, 2.565)	0.370
Stage 2			1	
Stage 3	-0.148	0.230	0.863 (0.550, 1.353)	0.520
Stage 4	-1.059	0.277	0.347 (0.202, 0.597)	< 0.001
Living status				
With partner			1	
Alone	0.751	0.233	2.120 (1.344, 3.343)	0.001
With family/friends	1.147	0.690	3.150 (0.815, 12.175)	0.096
Residential home	-0.616	1.188	0.540 (0.053, 5.539)	0.604
r	3.966			
s	1.613			
AIC	3093.3			
LR test: random-effects NB versus NB1: P < 0.001				

Table 8-10 shows that the interaction of intervention and linear time is not significant (P = 0.971) in the LTI model. The FTI model was examined and compared with the LTI model (both interaction terms are summarised in Table 8-11). The interaction of intervention and factorial time is also not significant (P = 0.069). The pattern of weekly specific FRRs in Table 8-11 does not suggest an obvious alternative to the LTI model, and the FTI model has a larger AIC than the LTI model (3107.6 versus 3095.3).

Because the interaction of intervention and time is not significant for the intervention counts, we drop the interaction term and conclude that during the intervention period, the intervention reduced the risk of falling by half and the effect was reasonably constant over time.

Table 8-10 Goodwin et al. dataset: LTI model (intervention period: weeks 11-20; n=124; obs=1240)

	Estimate	SE	FRR (95% CI)	P
Intervention	-0.679	0.204	0.507 (0.340, 0.757)	0.001
Time (weeks)	-0.002	0.010	0.998 (0.978, 1.019)	0.847
Intervention x Time (weeks)	-0.001	0.020	0.999 (0.961, 1.039)	0.971
Log(baseline count + 0.5)	0.632	0.065	1.882 (1.657, 2.137)	< 0.001
Female	-0.541	0.203	0.582 (0.391, 0.866)	0.008
Age	-0.017	0.013	0.983 (0.959, 1.007)	0.166
Years since diagnosis	0.068	0.018	1.070 (1.034, 1.108)	0.000
Hoehn & Yahr				
Stage 1	0.296	0.330	1.344 (0.704, 2.566)	0.370
Stage 2			1	
Stage 3	-0.148	0.230	0.863 (0.550, 1.353)	0.520
Stage 4	-1.060	0.279	0.347 (0.201, 0.598)	< 0.001
Living status				
With partner			1	
Alone	0.752	0.233	2.121 (1.343, 3.348)	0.001
With family/friends	1.147	0.690	3.148 (0.814, 12.179)	0.097
Residential home	-0.616	1.188	0.540 (0.053, 5.543)	0.604
r	3.964			
s	1.612			
AIC	3095.3			
LR test: random-effects NB versus NB1				P < 0.001
LR test: the interaction between the intervention and time				P = 0.971

Table 8-11 Time and intervention effect and during each week of intervention periods of Goodwin et al. dataset (weeks 11-20; n=124; obs=1240)

	Time effect		Intervention effect	
Week	**LT: FRR**	**FT: FRR (95% CI)**	**LTI: FRR**	**FTI: FRR (95% CI)**
11	1	1	0.551	0.510 (0.324, 0.802)
12	0.998	1.026 (0.829, 1.27)	0.550	0.945 (0.596, 1.500)
13	0.996	0.889 (0.712, 1.11)	0.550	0.974 (0.604, 1.568)
14	0.994	0.948 (0.762, 1.18)	0.549	0.724 (0.442, 1.186)
15	0.992	1.012 (0.814, 1.258)	0.549	1.409 (0.892, 2.225)
16	0.990	0.997 (0.800, 1.243)	0.548	1.247 (0.779, 1.995)
17	0.988	0.980 (0.784, 1.223)	0.548	1.166 (0.719, 1.892)
18	0.986	0.886 (0.709, 1.107)	0.547	0.614 (0.371, 1.018)
19	0.984	0.972 (0.780, 1.211)	0.547	1.145 (0.713, 1.840)
20	0.982	1.005 (0.808, 1.249)	0.546	1.078 (0.674, 1.726)
AIC (df)	3093.3 (15)	3105.6 (23)	3095.3 (16)	3107.7 (32)
LR test of interaction: Intervention x Time	-	-	P=0.971 (df=1)	P=0.069 (df=9)

Table 8-12 shows the estimates from the random-effects NB models fitted to the falls count during the follow-up period. The intervention effect is still significant (P = 0.010), and its estimate (FRR: 0.551, 95% CI: 0.351 to 0.866) is similar to that during the intervention periods (FRR: 0.505, 95% CI: 0.360 to 0.710). Again, the effect of time is not statistically significant, and the FT model does not indicate obvious alternative to the assumption of linear time effect (see Table 8-14).

Table 8-12 Goodwin et al. dataset: LT model (follow-up period: week 21-30; n=119; obs=1160)

	Estimate	SE	FRR (95% CI)	P
Intervention	-0.595	0.230	0.551 (0.351, 0.866)	0.010
Time (weeks)	-0.001	0.009	0.999 (0.982, 1.017)	0.946
Log(baseline count + 0.5)	0.901	0.076	2.462 (2.122, 2.857)	0.000
Female	0.265	0.236	1.304 (0.821, 2.070)	0.260
Age	0.016	0.015	1.016 (0.985, 1.047)	0.315
Years since diagnosis	0.031	0.020	1.032 (0.993, 1.072)	0.110
Hoehn & Yahr				
Stage 1	-0.025	0.359	0.975 (0.482, 1.970)	0.944
Stage 2			1	
Stage 3	-0.289	0.280	0.749 (0.432, 1.298)	0.303
Stage 4	0.303	0.351	1.354 (0.681, 2.694)	0.388
Living status				
With partner			1	
Alone	0.621	0.284	1.862 (1.068, 3.246)	0.028
With family/friends	1.683	0.833	5.383 (1.053, 27.524)	0.043
Residential home	-0.060	1.001	0.942 (0.132, 6.707)	0.953
r	4.562			
s	1.160			
AIC	2544.9			
LR test: random-effects NB versus NB1				P < 0.001

Table 8-13 shows results from the LTI model for the follow-up period. Now the interaction between intervention and time is statistically significant (LR test: P = 0.002), with the FRR of the interaction estimated to be 0.930 (95% CI: 0.890 to 0.972), that is, the intervention effect in preventing falls increases over weeks. The FTI model was fitted to check the linearity assumption and the week specific FRRs are shown in Table 8-14 from both models. Basically speaking, they show a pattern of declining FRR (reduced from approximately 1.00 to approximately 0.50) supporting modelling the interaction with a linear time. The weekly FRRs calculated from LTI decreases by 7% per week and falls from 0.798 to 0.417 (Table 8-14). In accordance with the results in the LTI model, the interaction between the intervention and time is statistically significant (P = 0.001) in FTI model. The FTI model again results in larger AIC than LTI.

Note that although the FRR for the main effect of intervention is estimated to be 1.760, this estimates the intervention effect for week 0, which falls outside the timeline of weeks 11-20.

In summary, during the follow-up period the time did not have a significant effect on falls rates; the intervention reduced the falls rate increasingly by 7% per week, with the overall intervention estimated to reduce falls rate by 45% during the follow-up period.

Table 8-13 Goodwin et al. dataset: LTI model (follow-up period: week 21-30; n=119; obs=1160)

	Estimate	SE	FRR (95% CI)	P
Intervention	0.566	0.418	1.760 (0.775, 3.998)	0.177
Time (weeks)	0.012	0.010	1.013 (0.994, 1.032)	0.194
Intervention x Time (weeks)	-0.072	0.023	0.930 (0.890, 0.972)	0.001
Log(baseline count + 0.5)	0.920	0.077	2.509 (2.160, 2.915)	< 0.001
Female	0.258	0.238	1.294 (0.811, 2.063)	0.279
Age	0.015	0.016	1.015 (0.985, 1.047)	0.333
Years since diagnosis	0.033	0.020	1.033 (0.993, 1.075)	0.104
Hoehn & Yahr				
Stage 1	0.024	0.364	1.024 (0.502, 2.089)	0.948
Stage 2			1	
Stage 3	-0.248	0.282	0.781 (0.449, 1.357)	0.380
Stage 4	0.277	0.353	1.320 (0.661, 2.636)	0.432
Living status				
With partner			1	
Alone	0.577	0.284	1.780 (1.020, 3.106)	0.042
With family/friends	1.637	0.847	5.141 (0.978, 27.021)	0.053
Residential home	-0.069	0.999	0.933 (0.132, 6.607)	0.945
r	4.979			
s	1.144			
AIC	2536.9			
LR test: random-effects NB versus NB1				P < 0.001
LR test: the interaction between the intervention and time				P = 0.002

Table 8-14 Time and intervention effect and during each week of follow-up periods of Goodwin et al. dataset (weeks 21-30; n=119; obs=1160)

	Time effect		Intervention effect	
Week	**LT: FRR**	**FT: FRR (95% CI)**	**LTI: FRR**	**FTI: FRR (95% CI)**
21	1	1	0.798	1.035 (0.601, 1.780)
22	0.999	0.911 (0.74, 1.122)	0.742	1.151 (0.755, 1.755)
23	0.998	0.933 (0.755, 1.155)	0.691	0.443 (0.268, 0.732)
24	0.997	0.921 (0.744, 1.141)	0.643	0.515 (0.316, 0.839)
25	0.996	1.003 (0.812, 1.239)	0.598	0.561 (0.342, 0.919)
26	0.995	0.968 (0.781, 1.2)	0.557	0.487 (0.291, 0.814)
27	0.994	0.964 (0.779, 1.194)	0.518	0.511 (0.307, 0.850)
28	0.993	0.947 (0.764, 1.174)	0.482	0.507 (0.304, 0.845)
29	0.992	0.825 (0.659, 1.034)	0.448	0.485 (0.282, 0.834)
30	0.991	1.042 (0.844, 1.287)	0.417	0.581 (0.357, 0.947)
AIC (df)	2544.9 (15)	2555.1 (23)	2536.9 (16)	2544.9 (32)
LR test of interaction: Intervention x Time	-	-	P=0.002 (df=1)	P=0.001 (df=9)

NB models have become widely used for analysing data from falls prevention trials, but few studies have considered longitudinal NB models. In this section, random-effects models were fitted to the Goodwin et al. dataset, to explore how to model falls data in longitudinal format.

An advantage of random-effects NB models is that a time variable can be included as a covariate, so that the model allows a check on the assumptions of a constant falls rate and a constant intervention effect. Compared to studies with outcomes measured repeatedly during a short period, falls prevention trials are different in two aspects: first, when falls counts are collected prospectively, data collection takes weeks or even months to complete, and each record has a relatively long gap with the previous record (in the Goodwin et al. trial the gap was one week); second, participants may experience worsening ability to maintain body balance during the observation period due to disease progression, which results in higher falls rates as time goes on; another possibility is that participants may start restricting their activities to avoid falling, which results in lower falls rates as time goes on. For these two reasons, the time effect should be checked for falls prevention trials and modelling this effect might be desirable. In the Goodwin et al. dataset, the time effect was not significant, which is possibly due to the short length of each period (10 weeks).

A step forward from modelling a constant intervention effect is to include a two-way interaction between the intervention and time. This enables checking whether the intervention effect is constant over time. The intervention did not have a significant interaction with time during the intervention period. However, during the follow-up period, the interaction between interventions and time was significant, and the estimated interaction effect suggest that during each week of the follow-up period the FRR of the intervention effect deceased by 7% (that is, the falls rate in the intervention group deviated further from the rate in the control group as time passes).

Chapter 9

Discussion

In this chapter, the main findings of the project are summarised and discussed. We also address limitations and suggest future directions for research concerning the analysis of data from falls prevention trials and counts of falls more generally.

9.1 General discussion

Falling is a common and often recurrent event for PwP. The quality of life for PwP is compromised due to their high risk of falling, as falls may lead to injuries and loss of independence in daily activities. Falls prevention trials aim to find an effective treatment to prevent or reduce falling, but the interventions rarely show statistically significant effect in published trials. This is possibly due to the low power of the statistical methods for analysing falls counts, which hinders the adoption of potentially effective treatments.

Fundamentally, falls prevention trials aim to answer two questions. First, is the intervention effective in reducing falls rate? Second, what is the size of the intervention effect? A natural approach to answering the two questions is to fit a count response model, which yields an effect size as an FRR and a P value from a model-based hypothesis test of the intervention effect (such as the Wald, LR, and score test). However, there are two major challenges faced by researchers.

The first challenge is outliers in the outcome count. Some participants with Parkinson's in falls prevention trials report very large falls count—this was found in all three motivating datasets used in this project, including the Martin et al. (2015) dataset, which consists of only 21 participants, among whom one participant reported 1599 falls during the 20-week outcome period. Large outcome counts often have great influence in model estimation and may considerably influence the estimated intervention effect, especially when a baseline count is not included in the model. The reason that outliers influence estimation is that if the large counts are not balanced between groups, the group with more large counts, or even just one extremely large count, may substantially increase the group mean, and consequently, the group difference in falls rates could be dominated by a few large counts. For example, suppose a trial is conducted to study the effect of an intervention, which is assumed to moderately reduce the falls rate. By chance if there are more frequent fallers in the intervention group, the large falls counts may increase the average falls rate in the intervention group, so that the estimated intervention effect is smaller than the true effect and is less likely to achieve statistical significance in the analysis of outcome counts alone, that is, ignoring the baseline count. In a more extreme case, the average falls rate in the intervention group may surpass the rate in the control group due to the imbalance in large counts, so that the model yields a positive FRR for the intervention effect, suggesting that the intervention increases the risk of falling. Conversely, if there are more large counts in the control group, the model may yield an overestimate of the intervention effect. Because of the skewed distribution of falls counts, large counts only account for a small proportion of the sample. Therefore, a very large sample size is required to achieve group balance in frequent fallers via randomisation.

Another challenge in analysing falls counts using count response models is overdispersion, which arises when heterogeneity is not fully accommodated in the model. The risk of falling is usually considered to be related to multiple risk factors, and as a result the mechanism of Parkinson's induced falling is not thoroughly understood by researchers, so it is likely that important prognostic variables are not observed, not observable, or not incorporated in the model. Thus, unobserved heterogeneity is common in falls data and is expected to result in model overdispersion.

Overdispersion leads to underestimation of the model-based SEs, especially for Poisson regression, which assumes equidispersion and does not accommodate any degree of overdispersion. The SE of the estimated intervention effect provides a measure of the precision of the estimator and is as important as the estimator itself, because an estimate has little value if its precision is not given (Fisher, 1956). When the SE of an estimator is underestimated, the corresponding hypothesis test generally has inflated type I error rates.

The main goal of a falls prevention trial is to study whether an intervention is effective in preventing falls, but unobserved prognostic variables, though their effects may not be among the research questions, perturb the model-based test of the intervention effect via overdispersion. Therefore, unless the effects of the unobserved variables or heterogeneity are sufficiently accounted for in the model, researchers cannot give credence to the P values from the model-based hypothesis tests. Poisson regression is almost always overdispersed in modelling falls counts, and for this reason widely understood to be unsuitable for analysing data from falls prevention trials.

The two challenges in analysing falls counts are intertwined rather than distinct. If large counts cannot be sufficiently accommodated in a count response model, they may result in overdispersion. A model that better controls for overdispersion may also better accommodate outliers reducing their influence on the model fit. For the analysis of falls prevention trials, the NB model has become a standard approach in modelling falls counts, because it has better performance in coping with outliers and overdispersion than Poisson models.

NB regression is an extension of Poisson regression that aims to accommodate heterogeneity in a gamma-distributed subject effect. As Box (1979) put it, "all models are wrong but some are useful." NB regression fits this description perfectly in that the distribution of heterogeneity is unknown, but the gamma distributed subject effect may be flexible enough to accommodate heterogeneity, if adequate prognostic variables are included in the model linear predictor as covariates. Another advantage of NB regression is that the underlying NB distribution is more skewed than the Poisson distribution given the same mean, so that it fits count data better than Poisson regression when there are outliers.

Although NB regression mitigates the effects of outliers and overdispersion, it is not a panacea in all cases. NB models are still subject to the influence of outliers, and NB models themselves can be overdispersed if the gamma-distributed subject effect does not sufficiently accommodate the heterogeneity. The dissertation focuses on models and diagnostic plots to address the issues of outliers and overdispersion, especially for NB models.

Incorporating a baseline count in the NB-based models

It is not uncommon in a falls prevention trial to collect a falls count during a pre-randomisation baseline period, but researchers do not seem to be fully aware of the benefit of incorporating it in analysis. In previous studies, a baseline count was sometimes used only as an eligibility criterion and excluded from further analysis; or, in some trials dichotomised into a *history of falling* (< 1 versus $\geqslant 1$) or a *history of frequent falling* ($< n$ versus $\geqslant n$ where $n > 1$) indicator before inclusion in the model. These two approaches both result in a great loss of power.

Despite being overlooked in practice, the baseline count is actually essential to the analysis of falls counts. A primary goal of the thesis is to investigate how to incorporate the baseline count into the analysis of a falls prevention trial using count response models. As discussed above, there may be latent variables related to the risk of falling that are not observed in a trial. Although collecting more covariates and including them in the analysis may control for heterogeneity, it is impossible to capture everything.

A possible solution to the problem heterogeneity is to control for the latent variables via a proxy variable — the baseline count. The logic of this approach is as follows. If heterogeneity is brought about by subject-level latent variables, the variables should be correlated with both the outcome and baseline counts. Thus, incorporating the baseline count in modelling should at least partially control for the heterogeneity. Cook and Wei (2003) accommodated heterogeneity in the joint distribution by assuming both baseline and outcome falls counts follow a Poisson distribution with a gamma-distributed random subject effect shared in the two count variables. By conditioning on the baseline count, the authors derived the Conditional NB (CNB) model from the joint distribution accounting for heterogeneity in the outcome. In Chapter 6, the CNB model was fitted to datasets simulated from the underlying joint distribution. The results showed that CNB indeed had good performance when it was correctly specified: the empirical power was high regardless of varying degrees of heterogeneity, and the type I error rate was close to the nominal level 0.05 even for small sample sizes. Compared with the NB model without including the baseline count (referred to as *NB-null*), CNB achieved much higher empirical power, and the power gain increased as the heterogeneity became greater. This demonstrates that conditioning on the baseline count improves statistical power and makes the model less subject to heterogeneity.

An alternative to the CNB model is to include the logged baseline count in an NB regression model as a covariate (referred to as the *NB-logged* model in the thesis), or as an offset (referred to as *NB-offset*). These two models are based on the same joint distribution underlying CNB. The log-transformation is applied to the baseline count because of the log link-function in NB models, so that the baseline count is on the same scale as the outcome, but this may be neglected in practice. Only one study (Aeberhard et al., 2017) was found to include the logged pre-randomisation count in NB modelling. The simulations in Chapter 6 showed that *NB-logged* and *NB-offset* had similar empirical power to CNB when data met the assumption of subject-specific heterogeneity underlying CNB. Though the type I error rates from *NB-logged* and *NB-offset* were modestly inflated for small sample sizes (total number in trial = 50 or 100), the rates converged to the nominal level 0.05 as the sample size increased.

Incorporating the baseline count is also a solution to group imbalance in relation to large outcome counts, as participants tend to report a consistent falls rate across baseline and outcome periods. For the *NB-null* model, the large outcome counts may have a great influence on the estimation of intervention effect. This was shown in the diagnostic plots in Chapter 5: take the Goodwin et al. dataset as an example (see Figure 5-9), most of the frequent fallers during the follow-up period were in the control group, they all showed relatively large negative DFBETA for the intervention effect. This resulted in an extreme intervention effect (FRR=0.287; Table 3-7) not in line with the general pattern shown in the plot. In comparison, the large counts were not influential in *NB-logged*, which yielded an FRR of 0.716 (see Table C-2 in Appendix C). The simulations in section 6.5.2 showed that, when the distribution of the outcome count is skewed, incorporating the baseline count reduced the influence of the large counts on the estimation of the intervention effect.

In practice the untransformed baseline count may be included as a covariate, ignoring the underlying scaling of these variables. This is referred to as the *NB-unlogged* model in the thesis. The simulations in Chapter 6 showed that *NB-unlogged* considerably overestimated the SE of the regression coefficient across simulated datasets. The tendency for *NB-unlogged* to overestimate SEs is in line with its deflated empirical type I error rate, which was lower than the nominal level even for large sample sizes. The description of the statistical analysis from published falls studies suggests that the baseline count is often treated as an untransformed regressor, though it is often unclear exactly what has been done. As the test of intervention effect based on *NB-unlogged* is conservative, it is possible that the analysis used in previous falls studies missed effects that might have proved significant had the baseline count been appropriately incorporated in analysis.

When the baseline and outcome counts are both collected prospectively, they are likely to be strongly correlated. However, in some trials the baseline count was obtained via a retrospective question, while the outcome count was collected prospectively because this method is thought to have a better precision. A retrospective baseline count and prospective outcome count would be expected to be correlated, but to a lesser extent because of the different data collection methods. A typical example of this is the EXSart trial (Ashburn et al., 2007), with weaker correlation between the baseline and outcome rates than the Goodwin et al. (2011) or Martin et al. (2015) trials (both with prospective baseline counts). This design violates the assumption underlying CNB because each count is obtained subject to a different measurement error, which is not accommodated in the shared gamma component.

In addition to the baseline counts being collected retrospectively, there may be other reasons leading to violation of the CNB assumption. For example, the risk of falling may be correlated with the progression of Parkinson’s. Parkinson’s is an irreversible progressive neurological disease, and if there is a wide gap between the baseline and outcome periods, the assumption underlying CNB that heterogeneity is at the subject-level may not hold. The rate of disease progression may differ across participants, also resulting in a discrepancy between the baseline and outcome counts within subjects.

To introduce a discrepancy between baseline and outcome counts, the simulations in Chapter 7 added a perturbation term when generating the baseline count. The *NB-logged* model allows the coefficient of the logged baseline count to vary and so it was more robust to perturbations than the CNB or *NB-offset* models, while the CNB model performed very poorly, with type I error rates greater than 0.5. The estimates of the intervention effect from *NB-logged* were generally close to the underlying value. Unlike the simulations in

Chapter 6, the regression coefficient of the logged baseline count was generally lower than 1, possibly tuning down the influence of the baseline count to remedy the discrepancy.

When the subject-specific heterogeneity is large, the hypothesis test of the intervention effect based on *NB-logged* is liberal, but to a lesser extent than CNB and *NB-offset*, thus the *NB-logged* is still preferable to these two models. However, researchers should be cautious with the result from the test of intervention effect from *NB-logged* when 1) different methods are used to collect the baseline and outcome counts, and 2) the HP of *NB-null* indicates great heterogeneity. A potential solution to the inflated type I error rate is to perform a test of the intervention based on the robust SE instead of on the model-based SE. A number of authors (Hardin, 2003; Hardin and Hilbe, 2007; White, 1980) have shown that the robust SE produces a consistent SE estimator even when the model is misspecified. Freedman (2006) argued that although the robust SE may "help on the variance side", the estimator of interested may be biased if the model is incorrect. King and Roberts (2015) expressed a similar opinion: if the robust SE considerably differs from the model-based SE, this should be considered as an indicator that the model is inappropriate and that the estimate of the effect of interest could be biased. Although these points are valid, this is not an issue in our case because the simulations in section 7.5 showed that the estimator of the intervention effect was unbiased, and so the robust standard SE might be a useful approach to investigate.

One might argue that, since a difference in methodologies for collecting the baseline and outcome counts may lead to inflated type I error rate, a retrospectively collected baseline count should be dropped altogether and *NB-null* should be preferred over the other models. This may be fair if the discrepancy is so great that the baseline and outcome counts are virtually uncorrelated, so that including the baseline count has little benefit. However, including the retrospective falls count in modelling can be justified on two counts: first, the simulations showed a great power gain; and second, as discussed before, *NB-null* may result in an extreme estimate of the intervention effect when large outcome counts are not balanced between groups, which may be a bigger issue than the inflated type I error rate.

The CNB and *NB-offset* models both showed poor performance in simulations when a perturbation violates the underlying assumption. This issue of CNB and *NB-offset* has a striking parallel with the issue of overdispersion for Poisson regression: the variance is restricted to be equal to the mean in Poisson regression, so the model is not flexible enough to accommodate overdispersion. When the assumption of equidispersion is violated, the type I error rate of the model-based hypothesis test becomes inflated. Similarly, the CNB model assumes that all heterogeneity is at the subject-level, so it is not flexible enough to accommodate discrepancy between the baseline and outcome counts, which also results in inflated type I error rates.

To summarise the results from Chapter 6 and 7, *NB-logged* appears to be a better way of incorporating baseline counts in analysis than CNB and *NB-offset*: 1) when CNB is correctly specified, *NB-logged* has comparable performance to CNB; and 2) *NB-logged* is more robust to perturbations in the simulation than CNB or *NB-offset*. Another advantage of *NB-logged* over CNB is that it is widely supported in statistical packages (as listed in Chapter 4), while the CNB model, at the time of writing, is not currently supported in any package.

Benefit of collecting a baseline count for designing a falls prevention trial

Zhu and Lakkis (2014) proposed three formulae for power calculations related to *NB-null*, and their simulations showed them to have good performance. The simulations in section 6.6 showed that Tango's (2009) formula for calculating the sample size for the conditional score test may be used as an approximation of the sample size required for *NB-logged*. When data have mild heterogeneity, the sample size calculated using Tango's formula achieved the specified power level of 80% for *NB-logged* in the scenarios examined, but when the degree of heterogeneity is large, the empirical power from *NB-logged* at sample sizes suggested by the formula to achieve 80% power is moderately lower than the nominal level.

The sample sizes calculated from Zhu and Lakkis's formulae and Tango's formula were compared in Table 6-5, and show the *NB-logged* model to require a much smaller sample size than *NB-null*, especially for small intervention effects, where average outcome count is large, and there is great heterogeneity. In order to discuss the benefit of collecting the baseline count in the context of designing a falls prevention trial in PwP, we now consider a representative trial setting.

In the Goodwin et al. trial, the participants in the control group reported 32.25 falls on average during the 10-week intervention period and 31.88 during the 10-week follow-up periods (Table 3-1). In the EXSart trial, the average falls count in the control group was 10.12 during the 8-week intervention period and 21.33 during the 18-week follow-up period (Table 3-12). These numbers are reasonable for a prospectively collected falls count in a falls prevention trial, because if the observation period is longer, the falls count will be larger, but the drop-out rate is anticipated to increase. As shown in sections 3.2.1 and 3.2.3, HP in *NB-null* fitted to outcome falls count in Goodwin et al. and EXSart dataset were generally larger than 3. Based on these results, if we suppose planning a trial that the average outcome count in the control group is 30, the intervention has a relatively large effect in reducing falls rate (say, by 26%), and we assume $\alpha = 3$, to achieve 80% power, the required sample size is over 1000 for *NB-null* but only 26 from *NB-logged* and *NB-offset* (see Scenario 6 in Table 6-5). Considering that the number 26 was an approximation from the Tango (2009) formula, the actual sample size required is larger than this, but even doubling the number to 52 results in a reasonable trial size. A Cochrane review (Gillespie et al., 2012) included 159 falls prevention trials in the general elderly living in the community, and among them only five trials (Day et al., 2002; Hornbrook et al., 1994; Reid et al., 2006; Sanders et al., 2010; Stevens et al., 2001) have both prospectively collected outcome falls counts and a sample size greater than 1000. For trials in PwP the recruitment of participants is more difficult than recruiting to trials in the general elderly, so it would be very challenging indeed to recruit 1000 PwP into a falls prevention trial in PwP.

If the intervention effect is smaller, say, reducing falls rate by 9%, the sample size would be somewhat over 216 for *NB-logged* or *NB-offset*, and 9526 to 9530 for *NB-null* (see Scenario 4 in Table 6-5). The largest RCT (Smith et al., 2007) in Gillespie et al.'s (2012) Cochrane review recruited 9440 people, which is still smaller than the required sample size for *NB-null* with a small intervention effect and 80% power. In the Smith et al. (2007) trial, outcome falls were collected retrospectively as a binary outcome (falling or not), considerably less expensive than collecting falls prospectively. Such a large trial with prospectively collected fall counts is unlikely to be carried out.

NB Diagnostic plots in the context of falls prevention trials

Because of the potential for outliers to be greatly influential in NB models, assessing the diagnostic statistics for each observation is essential for NB modelling.

Large counts may be deemed to be outliers because of their anticipated impact on model fit. Nevertheless, a large response count *per se* does not warrant labelling a count as an outlier without examination of model diagnostics. The reason is that large counts are ubiquitous in falls data in Parkinson’s and should be accommodated by heavily skewed distribution in modelling so that they are well fitted. The great influence of the large counts underscores the importance of the diagnostic plots for NB models, especially in the context of falls prevention trials.

A set of diagnostic plots were described in Chapter 5 and referred to as the Baseline/Outcome Event (BOE) plots in the thesis. BOE plots present the diagnostic statistics from NB models where a baseline count is available (though not necessarily included in the model) in a scatter plot, with y-axis of the logged outcome falls rate and x-axis of the logged baseline rate. Where the length of the periods of data collection for the counts do not differ over participants, counts can be used instead of rates. The size of a plotting symbol is proportional to the chosen diagnostic statistic. A Line of Falls Equity (LoFE) is plotted in the BOE plot as a reference line, which shows whether the falls rate is constant across periods. It also shows whether a participant reported a lower, higher, or similar outcome falls rate compared to their baseline rate.

The baseline rate can be seen in the BOE plots and provides valuable information on each subject as well as the whole trial. At the subject level, the plot shows whether a large outcome rate is consistent with a similarly large baseline rate, in which case, the large outcome is anticipated to be successfully accommodated by a model appropriately including the baseline count.

At the trial level, the plots show the correlation between the outcome and baseline rates, and they facilitate a visual evaluation of the discrepancy between the two. In addition, the LoFE provides a visual check for the period effect. If the plotting symbols from the control group are symmetric around the LoFE, the falls rate is relatively stable across periods. If the symbols are above the line, the falls rate increases over time, possibly because of disease progression.

The estimator of the intervention effect is often dominated by large counts, but because they account for only a small proportion of a dataset, a trial requires a large sample size to achieve group balance with respect to large outcome counts via randomisation. In order to check the influence of the subjects with large outcome counts or inconsistent falls rates, DFBETA corresponding to the outcome and baseline rates may be examined in a BOE plot. In addition, because the trial groups are indicated by different colours, we obtain a visual impression of the size of the intervention effect, so that if the estimation is influenced by a few outliers, we can see that clearly.

Another useful aspect of the BOE plots is that they are based on scatter plots on a logarithmic scale so that the positioning of each plotting symbol is evenly scattered. Hence, the diagnostic statistics of each subject can be easily compared, especially for the large counts.

Poisson Inverse Gaussian model

The Poisson Inverse Gaussian (PIG) model is an alternative way of dealing with Poisson overdispersion to the NB model. It fits counts with highly skewed distribution better than the NB model, because the PIG variance function ($\mu + k\mu^3$) is parameterised is based on a cubic form of μ, compared with the quadratic form in the NB variance function ($\mu + \alpha\mu^2$).

The PIG model was fitted to the follow-up falls count in EXSart dataset and compared to the NB models. The follow-up period of the EXSart trial was relatively long (four months), so there were a few very large outcome counts (maximum 1099), which results in the falls count being particularly skewed. The fitted *NB-null* model was significantly overdispersed ($\mathrm{P} < 0.001$ from the NB overdispersion score test). For such trials the PIG model may be more suitable. As the PIG model has the same log link function as the NB model, the approach of including the logged baseline count as a covariate is transferable to PIG models (referred to as *PIG-logged*).

Compared with *NB-null*, which yielded an intervention effect that was not in line with the BOE plot, the estimation of the intervention effect from *PIG-null* model appeared to be less influenced by the large counts and was closer to that from the *NB-logged* model. This indicates that the *NB-logged* model is probably not overdispersed if a baseline count is available and correctly included, because it may adequately account for heterogeneity. If a baseline count is not available, the PIG model may be more suitable because it copes with large counts better and may accommodate NB overdispersion.

A limitation of PIG model is that it is not as widely supported in statistical packages. This is possibly why it is rarely used in falls prevention trials. Canning et al. (2014) used the PIG model because

> *"a blind review of the falls data revealed that the negative binomial model was not flexible enough to capture both the nonfallers and the large number of multiple fallers. In contrast, the Poisson inverse gaussian (PIG) distribution gave a good fit."*

However, they included the baseline count after dichotomising it to multiple baseline fallers ($\geqslant 10$ or < 10 in the previous 12 months, retrospective question), not logged as a regressor, and so had not adequately accommodated heterogeneity.

Right-censored versus right-truncated models

Although including the baseline count may control for large outcome counts, the baseline count is not necessarily consistent with the outcome count, or in some trials, a baseline count may not be collected. In these situations, a cut-point may be used and participants who report a greater outcome count than the cut-point are either excluded from the NB model, or revalued to the cut-point. Both methods are problematic, because the distribution underlying the standard NB regression has a range from zero to infinity. If the outcome counts are excluded at a chosen cut-point, the underlying distribution of the outcome is right-truncated at the cut-point. Revaluing the outcome count to a small value the cut-point is also inappropriate.

The right-censored NB model has an advantage over the right-truncated NB model — the individuals reporting large outcome counts are not excluded, preserving sample size, but their influence on model estimation is reduced. In some trials, the large counts remain influential on the model estimation even when censored. In this case the right-truncated NB model may perform better, but power is lost because the large counts are excluded.

An issue for the right-censored and right-truncated models is that the model estimation is based on the chosen cut-point. Therefore, it may be difficult to justify the chosen value. The results from the follow-up period of the Goodwin et al. dataset indicates a trade-off in choosing the cut-point: with a smaller cut-point, the large counts were less influential in the estimation of the right-censored NB model, and the right-truncated model gave an estimate of the intervention effect that was closer to the estimate from *NB-logged*; however, during the process more outcome counts were labelled as right-censored, or dropped from the truncated model.

For falls prevention trials without a baseline count, a sensible approach for examining the sensitivity of the estimated intervention effect from an NB models might be to:

1. Fit an NB model;
2. Check the diagnostic statistics for the NB model, especially the DFBETA and Cook's distance;
3. Choose a cut-point based on the diagnostic statistics, and fit a right-censored and right-truncated NB model;
4. Consider the PIG model if the large counts are overly influential.

If the FRRs from the right-censored and right-truncated model are very different to that from NB model, the estimator of the intervention effect is probably sensitive to the outliers.

ZINB

In some falls prevention trials there are a large proportion of zero outcome counts, which raises the question as to whether there could be zero-inflation in the dataset. Excessive zeros may result from an additional process to the count process, that is, some trial participants may report zero counts while actually experiencing one or more falls.

Because the distribution of falls count is usually skewed, a considerable number of zeros are anticipated even when there is no zero-inflation, especially when the average falls count is small. Hence, a histogram of the outcome count may show a large proportion of zeros, but this does not necessarily mean that there is zero-inflation.

A possible approach dealing with potential zero-inflation is to fit a Zero-Inflated Negative Binomial (ZINB) model, which accounts for the process of excessive zeros in a binary component. To examine the issue of zero-inflation in real falls data, the ZINB model was fitted to the intervention and follow-up counts from the Goodwin el al. trial, and compared with the NB models to examine zero-inflation.

A strength of this study is that the zero-inflation was examined using two diagnostic tools —the Vuong test and the covariate-adjusted probability plot. The original Vuong test is commonly used for testing zero-inflation via comparing the goodness of fit of a ZINB model and an NB model, but it is biased in favour of ZINB models. Desmarais and Harden (2013) proposed AIC- and BIC-based corrections for the Vuong test. However, neither the original Vuong test nor the tests with AIC/BIC corrections provide an unbiased test result of zero-inflation in NB models. Therefore, the P values from all three tests were produced for the Goodwin et al. dataset. The results showed reasonable agreement: none of the tests suggested that the ZINB model had significantly better fit than the NB model for either the intervention or follow-up count.

The covariate-adjusted probability plot proposed by Holling et al. (2016) was considered in section 8.2 as a diagnostic plot for zero-inflation. It proved to be a valuable tool to visualise possible zero-inflation, by inspecting whether the covariate-adjusted probability of zero from the NB model was considerably different from the observed probability. In addition, the plot provides a graphical comparison of the NB and ZINB models regarding the fit to zeros, which can be used to double check for zero-inflation. The plots for the intervention and follow-up counts were in line with the test results from the Vuong test: the NB models provided a reasonable fit to the zero counts, and the covariate-adjusted probabilities from the ZINB models were very close to those from the NB models.

No evidence of zero-inflation was found for the Goodwin et al. dataset, and this might be related to the prospective falls collection method in the trial. If the outcome falls count was obtained retrospectively in a trial, it is possible that the participants do not recall the exact time of each fall, and therefore, not certain whether a fall they experienced occurred was within the study period, so that they may report no falls whilst actually haven fallen during the period. Although zero-inflation is not significant in the Goodwin et al. dataset, this issue should be checked in future trials.

Random-effects NB model for the longitudinal dataset

The random-effects NB model provides a new dimension for studying the intervention effect: the interaction between an intervention and time can be included in the model so that the intervention effect can be checked to see whether it is constant over time or not. This is particularly relevant to falls studies in Parkinson's. The intervention in such trials is usually a physiotherapy program that aims to enhance strength or body balance. It is reasonable to assume that the intervention requires some time to work, and it may take some time for the trial participants to master the physiotherapy program. In both cases the intervention may have an increasing intervention effect over time in reducing falls rate. Another possibility is that the participants may become less willing to carry on with the intervention, especially after the intervention program has ended.

The Goodwin et al. dataset is available in longitudinal form with weekly counts. An interesting finding is that, during the follow-up period after the intervention had ended, there was a significant interaction between intervention and time, with the intervention effect increasing over weeks in preventing falls, while the intervention effect did not materially change during the intervention period. This pattern is the opposite to that anticipated.

Strategies for analysing falls counts

To address the common issues in analysing falls counts, the findings presented in the thesis can be combined as a set of strategies for analysing falls counts from a falls prevention trial in PwP:

1. When a baseline count is available, fit an *NB-logged* model, because it has satisfying statistical power and is relatively robust to discrepancy between the outcome and baseline count (especially when mixed collection methods are used for the two counts).

2. Produce the BOE plots to check
    a. whether the estimated intervention effect might be overly influenced by a few outliers;
    b. whether there is a peculiar pattern suggesting model inadequacy.

3. Check for zero-inflation by fitting an ZINB model and compare it to the standard NB model. This can be done using the Vuong test and corrected versions. Because the Vuong-based tests may still be biased, the covariate-adjusted probability plot can be used to check the result from the Vuong test.

4. If a baseline count is not available, the large counts are likely to be highly influential. If this is indicated by model diagnostics, two approaches may be used:
    a. PIG models: they are less subject to the influence of the large counts;
    b. Right-truncated/-censored NB model: a cut-point may be chosen to drop outcome counts that are greater than the cut-point, or reduce the influence of the large counts.

5. If the falls counts are available in a longitudinal dataset, the random-effects NB model may be fitted to study the trend of intervention effect over time.

9.2 Strengths

The dissertation has the following strengths:

First, current and widely used statistical methods for analysing counts from falls prevention trials are examined. A broad range of work has been done examining the characteristics of the data, statistical modelling, diagnostic plots and software available in facilitating the analysis of falls counts.

Second, statistical models were compared using actual datasets from falls prevention trials, as well as simulated data based on features found in real-life falls prevention trials.

9.3 Limitations

There are a number of limitations to the research.

Firstly, although *NB-logged* utilises the baseline count in modelling and is relatively robust to discrepancy in methods for collecting the baseline and outcome counts, there is currently no formula to calculate the exact sample size required to achieve a given power in the model. In section 6.6, the formula for calculating the sample size of the conditional score test (Tango, 2009) was assessed as an approximation for the *NB-logged* and *NB-offset* models, but the simulation suggested that when there is great heterogeneity in the dataset, the formula underestimates the sample size required. Tango's formula may be used in planning a trial where the outcome is a count and little heterogeneity is anticipated (based on previous studies of the same event). For data with greater heterogeneity, simulation-based methods may be used to obtain the sample size required to achieve a specified power.

Secondly, for all the datasets considered in this thesis, the participants who dropped out of trials were assumed to be missing at random, but this may not be the case in practice. It is possible that frequent fallers may drop out because recording falls in dairies is gruelling and time-consuming for them. For the Goodwin et al. trial, among the ten participants who reported the largest falls counts during the intervention period, only two were in the intervention group. This is possibly because the intervention effectively reduced falls rate in this group, as the large counts were balanced between groups during the baseline period and no one dropped out of the trial before the intervention period. However, the two frequent fallers in the intervention group dropped out at the end of the intervention period, which resulted in a group imbalance of large falls count: the nine most frequent fallers during the follow-up period were all in the control group. The largest follow-up count was 49 in the intervention group, but 678 in the control group. This may happen by pure chance, but the possibility of informative missingness cannot be ruled out, and this should be considered in relation to falls prevention trials.

Thirdly, the right-censored and right-truncated NB models were evaluated using the Goodwin el al. dataset in section 9.3. They both appeared to be good solutions for reducing the influence of outliers, but there are two limitations for this study: 1) the potential benefit over standard NB modelling has not been examined in simulation studies; and, 2) the two models are conditional on a cut-point, and different cut-points result in different estimates. In this study, the cut-point was chosen based on the BOE plots in section 5.4.1, by inspecting the diagnostic statistics and the corresponding outcome counts. However, the chosen cut-point nevertheless involves subjectivity and may be hard to justify. This is the reason the two models may be more appropriate considered as sensitivity analysis. Two cut-points were compared and the results demonstrated the trade-off in choosing a smaller value. Further investigation is required regarding an objective approach in choosing a cut-point for the right-censored and right-truncated NB models.

Fourthly, the PIG model showed good performance in the EXSart dataset when the baseline count is ignored, but it was only compared with the NB model regarding the goodness of fit for this single dataset. Work remains to be done to compare the two models more systematically.

Finally, findings from modelling falls counts in the three motivating datasets may not be extrapolated to other studies, because they may be influenced by the design or other aspects of each trial, for example, the ZINB model did not show evidence of zero-inflation in the Goodwin et al. dataset, but it would be interesting to re-examine this issue for a study with retrospectively collected outcome counts. Another limitation is that data from only three trials were examined in detail.

9.4 Future research

In addition to the models considered in the thesis, the Generalised Additive Model (GAM) merits further investigations in modelling falls counts. GAM is an extension to GLM, with a sum of smoothing functions for some covariates included in the linear predictor (Wood, 2017). The smoothing functions are used in GAMs to span the space of transformation for the covariate. The idea of GAMs is that in a GLM, a relationship is assumed between the response variable and the covariates, via the link function and/or transformation of the covariates, while for GAMs the relationship is dictated by the data. Despite the greater flexibility, there are two issues for GAMs regarding statistical inference: 1) A more flexible model is generally less interpretable; and 2) As Wood (2017) pointed out, for a model with higher flexibility, "the methods for inference become less well founded," and generalising from GLMs to GAMs, "penalization lowers the convergence rates of estimators, hypothesis testing is only approximate, and satisfactory interval estimation seems to require the adoption of a Bayesian approach." Falls prevention trials generally aim to study whether an intervention reduces the incidence of falling, thus the issue of inference limits the application of the GAM in this context. However, for future falls studies GAMs could be valuable in predicting the patient-specific risk of falling for PwP based on a model fitted to a training dataset. The discussion of GAMs regarding the issue of inference and the potential in predicting future falls extends to the Generalized Additive Model for Location, Scale and Shape (GAMLSS) model, which fits a predictor (with or without smoothing terms) for each of the expectation, location, scale, and shape parameters (Stasinopoulos et al., 2017). The flexibility may improve the prediction accuracy of a model, but too much flexibility leads to overfit in the training dataset and leads to lower accuracy in prediction (James et al., 2013).

The falls rate is commonly used as the outcome in falls prevention trials (Gillespie et al., 2012), and it is collected as the falls count during a study period, either via the retrospective or prospective methods. As discussed in section 2.2.1, both methods have intrinsic characteristics that potentially result in underreporting or overreporting. New technologies have been developed to record falls using wearable sensors or other devices, and they have the potential to solve the issues around self-reporting. An example of such a device is the smart watch, which has become increasingly popular and affordable in recent years. With multiple activity sensors integrated into a compact and portable device, the smart watch has drawn interest from the research community for use in recording falls (Ghayvat et al., 2015). In 2018, Apple released the Apple Watch Series 4, and provided a fall detection function (Apple Inc., 2018): If the watch detects a fall, it shows a notification asking if the person has just fallen. The person will also be asked if emergency services should be contacted. If the person is irresponsive to the query and has been immobile for one minute, the watch automatically makes an emergency call. Falls are recorded unless the person chooses the option “I’m OK” provided in the prompt notification after detection of a possible fall. Such procedures are expected to reduce misclassification of falls. Also, using this device may make recruiting trial participants easier because 1) wearing an Apple Watch has safety benefits, such as calling for medical help when participants are unconscious due to a fall-related injury; and 2) recording falls using sensors considerably alleviates the burden on participants compared to recording falls in diaries.

Another benefit of smart devices is that they may record different measurements of health data multiple times a day. The high dimensional data improve modelling and enable predicting disease progression and falls rate in the future. For example, the mPower study (Bot et al., 2016) collected data from a total of 9520 participants using an iPhone app, which makes it the largest Parkinson’s study so far, and the accessibility of the app and smart phones contributed to the large sample size. The app notified the study participants to complete a walking activity three times daily during which gait and balance were evaluated. Although the mPower app did not record falls, it set a precedent for using mobile/wearable devices to collect data in PwP.

In addition, wearable devices may record the sedentary and in-bed time (this is already possible for the Apple Watch by measuring movements and pulse), and this time could be excluded from exposure time (that is, the length of time during which a fall may occur) because it is impossible to fall when a person is sitting in a chair or lying in bed. A caveat is the misclassification of the lying and sedentary stages. An alternative is to track the activity time using wearable sensors. Srulijes et al. (2019) tracked Physical Activity (PA) of 88 people who had been diagnosed with a neurodegenerative disease (amongst them 14 were PwP) and proposed the measurement of "falls per individual PA exposure time." The benefit of this measurement is that for people who restrict their physical activity to avoid falling (for example people with more severe Parkinson's), their falls rate is usually low, but after adjusting for the PA exposure time, the falls rate better reflects their risk of falling. This is confirmed by the authors' finding that PwP with low walking PA had higher falls per individual PA exposure time, suggesting that PwP tend to walk less to avoid falling.

Wearable devices typically record each fall at the exact time of occurrence. This enables modelling falls based on time between each fall event. Cox regression is sometimes used to analyse the time between the start of an intervention to the occurrence of the first fall. A similar approach is to analyse the time of each fall using the Andersen-Gill model (Andersen and Gill, 1982). This model has an interesting link to the Poisson and NB models in that they are all derived from the Poisson process, which assumes falls occur randomly such that the falls counts during nonoverlapping intervals are independent within subjects (Cook and Lawless, 2007). The Andersen-Gill model is anticipated to perform better than the Poisson and NB models when the falls are generated from a time-dependent Poisson process, though Jahn-Eimermacher (2008) compared the three models in simulations and concluded that the NB model performed better when data were generated from this process.

In addition to modelling falls, there are alternative directions to pursue in the future to widen the scope of falls prevention trials. One possibility is to study fall-related injuries, which are usually reported as adverse events in falls prevention trials and a serious safety issue for PwP, but the current studies in this field are limited: in RCTs the fall-related injury is usually not the main outcome so that the sample size may not provide sufficient power for analysing the rate of injuries (Gardner et al., 2000; Province et al., 1995). A challenge for conducting an RCT of fall-related injuries is that each fall only has a small chance of leading to an injury, so a fall-rated injury is an event with lower incidence than a fall. If the eligibility criterion is set to PwP with a history of fall-related hospitalisation, the participants could be recruited from hospitals, but the participants may restrict their daily activity because of their history of serious injuries, resulting in few injuries during a follow-up period. Another issue with the low incidence is that participants need to be followed up for a longer periods and the sample size has to be much bigger as well.

To solve these problems, a possible approach is to differentiate falls into *causing injury* and *not causing injury*. This was referred to as the multitype recurrent event by Cook and Lawless (2007), and a parametric model (Ng and Cook, 1999), or a semiparametric model in which the baseline mean function is unspecified (Cai and Schaubel, 2004), can be used to analyse the risk of fall-related injury. These models may also be used to study different types of fall-related injuries that were found to be related to Parkinson's (Cook and Lawless, 2002), such as hip and Colles fractures (O. et al., 1992; Vestergaard et al., 2007). Because a fall-related injury is conditional on a fall, modelling falls as a multitype recurrent event should improve the statistical power, compared with modelling the fall-related injury only, so that the required sample size would be smaller.

9.5 Summary and main findings

The main findings and contributions of the research can be summarised as follows:

- The baseline count is essential in falls modelling, but applied researchers may not be aware of its central role. Incorporating the baseline count was found to have two benefits: 1) correctly including the baseline count in an NB or CNB model largely accounts for heterogeneity and considerably increases the statistical power of the model to detect an intervention effect, as indicted by the simulations in Chapter 6 and 7; and 2) adjusting for the baseline rate controls for group imbalance in large outcome counts, which is likely to happen for small to medium sized trials and even trials considered in the area to be of large size, and may result in a misleading estimate of the intervention effect. Another finding is that, if the logged baseline count is included as a covariate in NB regression, the model is relatively robust to the differences in methodologies used to collect the outcome and baseline counts, for example, the baseline count being obtained retrospectively while the outcome count is obtained prospectively. Though examined in the context of falls prevention trials, the lessons carry over to counts of other events and to study designs other than RCTs where initial and subsequent counts are available, especially where counts can extend to large values.

- Five commonly used statistical packages were reviewed regarding their functionality for fitting NB and NB-related models. This work facilitates researchers in choosing a statistical package that best meets the analysis planed for a particular dataset, as each package supports different post-estimation statistics.
- A set of diagnostic plots for the NB model in the context of falls count data from a falls prevention trial were developed so that patterns in diagnostic statistics related to the distribution of the outcome and baseline counts can be easily identified. An R package named **NBDiagnostics** was written to automate the production of the diagnostic plots for an NB model, specifically for situations where the count of a recurring event is available during both an outcome and a baseline period. The covariate-adjusted probability plot, an existing diagnostic plot, was studied in the context of falls data focusing on the visual presentation of overdispersion in the fitted model.
- Zero-inflation was examined in the Goodwin et al. dataset and found not to be an issue. The Zero Inflated NB (ZINB) model was compared with the standard NB model using multiple statistical tools, which appeared effective in testing and visualising zero-inflation in count data.
- The Poisson Inverse Gaussian (PIG) model fitted the extremely skewed data from the EXSart trial better than the NB model. The PIG model was less subject to the influence of outliers than the NB model when the baseline count is not included in both models. This highlights the potential of the model, especially for extremely skewed counts or when important covariates are unobserved.

- The right-censored and right-truncated NB models were considered for reducing the influence of large counts. These two models were shown to be potentially useful ways of coping with the large counts in the Goodwin et al. dataset.
- The finite mixture models have the potential to model the frequent fallers as a subpopulation, but this requires further examinations for trials with larger sample sizes
- The random-effects NB model was fitted to the Goodwin et al. dataset, which showed the potential of longitudinal modelling.

Appendix A

The **NBDiagnostics** R package

The **NBDiagnostics** package introduced in Chapter 5 can be installed from its GitHub repository (https://github.com/AlexZHENGH/NBDiagnostics). This can be done in `R` using the following commands:

```
install.packages("devtools")
library(devtools)
install_github("AlexZHENGH/NBDiagnostics")
```

The package contains a function `nbdiagnostics()` that 1) fits an NB model using the `glm.nb()` function in the **MASS** package (Venables and Ripley, 2002) with the same syntax, and 2) specify the names of the following four variables: the outcome event rate (`outcome_varname`), the baseline event rate (`baseline_varname`), the group allocation (`group_varname`), and ID (`id_varname`).

The `nbdiagnostics()` function returns an "NBDiagnostics" object, which is an NB model fitted to the dataset. An "NBDiagnostics" object can be passed into the `model` argument in the `boeplot()` function to produce the BOE plots introduced in section 5.2. In `boeplot()`, the `diagnostic_stat` argument can be specified as "cookd", "leverage", "anscombe_resid", or "dfbeta" to present the Cook's distance, leverage, Anscombe residuals, and DFBETA, respectively.

The `caprob_nb()` and `caprob_nb_poi()` functions produce the covariate-adjusted probability plots (Holling et al., 2016) for an NB model, and NB versus Poisson models, respectively.

Appendix B

Example **R** code for simulating mixed Poisson distribution

Core code for simulations in section 6.5

The following `R` code demonstrates the data generating mechanism (time-homogeneous Poisson process) for simulating the baseline and outcome falls counts in section 6.5.

```
N <- 500
alpha <- 3

## Simulate the gamma-distributed subject-specific
heterogeneity
s <- rgamma(N, shape = 1/alpha, scale = alpha)

## Simulate the baseline counts
mu <- 20
mu_0 <- mu * s
y_0 <- rpois(n=N, lambda=mu_0)

## Allocate subjects to two group
group <- c(rep(1, N/2), rep(0, N/2))
beta <- -0.2

## Simulate the outcome counts
mu_1 <- exp(beta * group) * mu * s
y_1 <- rpois(n=N, lambda=mu_1)
```

Core code for simulations in section 7.5

The following `R` code demonstrates the simulations of the baseline and outcome falls counts in section 7.5. Because a gamma-distributed perturbation is introduced in the baseline count, the assumption of the subject-specific heterogeneity does not hold.

```
N <- 500
```

```
alpha <- 3.5
epsilon <- 0.5
lambda <- 5
t0 <- 12
t1 <- 2

## Simulate the gamma-distributed subject-specific
heterogeneity
s <- rgamma(N, shape = 1/alpha, scale = alpha)

## Simulate the gamma-distributed perturbation
v <- rgamma(N, shape = 1/ epsilon, scale = epsilon)

## Simulate the baseline counts

mu_0 <- lambda * s * v * t0
y_0 <- rpois(n=N, lambda=mu_0)

## Allocate subjects to two group
group <- c(rep(1, N/2), rep(0, N/2))
beta <- -0.2

## Simulate the outcome counts
mu_1 <- exp(beta * group) * lambda * s * t1
y_1 <- rpois(n=N, lambda=mu_1)
```

Appendix C

Additional results

Table C-1 Number of successful replicates for the simulation study in section 6.5

			Number of included replicates				
α	β	m	*NB-null*	*NB-unlogged*	*NB-logged*	*NB-offset*	CNB
3	-0.4	50	2000	2000	1999	2000	1999
		100	2000	2000	2000	2000	2000
		200	2000	2000	2000	2000	2000
		500	2000	2000	2000	2000	2000
	-0.2	50	2000	2000	2000	2000	1990
		100	2000	2000	2000	2000	2000
		200	2000	2000	2000	2000	2000
		500	2000	2000	2000	2000	2000
	0	50	2000	2000	1998	1999	1996
		100	2000	2000	2000	2000	1999
		200	2000	2000	2000	2000	2000
		500	2000	2000	2000	2000	2000
0.5	-0.4	50	2000	2000	2000	2000	2000
		100	2000	2000	2000	2000	2000
		200	2000	2000	2000	2000	2000
		500	2000	2000	2000	2000	2000
	-0.2	50	2000	2000	2000	2000	1990
		100	2000	2000	2000	2000	1995
		200	2000	2000	2000	2000	1999
		500	2000	2000	2000	2000	2000
	0	50	2000	2000	2000	1999	1991
		100	2000	2000	2000	2000	1998
		200	2000	2000	2000	2000	2000
		500	2000	2000	2000	2000	2000

Table C-2 Number of successful replicates for the simulation study in section 7.5.

				Number of included replicates (2000 in total)			
ϵ	α	β	m	*NB-null*	*NB-logged*	*NB-offset*	CNB
0.5	3.5	-0.4	50	2000	2000	2000	2000
0.5	3.5	-0.4	100	2000	2000	2000	2000
0.5	3.5	-0.4	200	2000	2000	2000	2000
0.5	3.5	-0.4	500	2000	2000	2000	2000
0.5	3.5	-0.2	50	2000	2000	2000	2000
0.5	3.5	-0.2	100	2000	2000	2000	2000
0.5	3.5	-0.2	200	2000	2000	2000	2000
0.5	3.5	-0.2	500	2000	2000	2000	2000
0.5	3.5	0	50	2000	2000	2000	2000

0.5	3.5	0	100	2000	2000	2000	2000
0.5	3.5	0	200	2000	2000	2000	2000
0.5	3.5	0	500	2000	2000	2000	2000
0.5	0.5	-0.4	50	2000	2000	2000	2000
0.5	0.5	-0.4	100	2000	2000	2000	2000
0.5	0.5	-0.4	200	2000	2000	2000	2000
0.5	0.5	-0.4	500	2000	2000	2000	2000
0.5	0.5	-0.2	50	2000	2000	2000	2000
0.5	0.5	-0.2	100	2000	2000	2000	2000
0.5	0.5	-0.2	200	2000	2000	2000	2000
0.5	0.5	-0.2	500	2000	2000	2000	2000
0.5	0.5	0	50	2000	2000	2000	2000
0.5	0.5	0	100	2000	2000	2000	2000
0.5	0.5	0	200	2000	2000	2000	2000
0.5	0.5	0	500	2000	2000	2000	2000
0.25	3.5	-0.4	50	2000	2000	2000	2000
0.25	3.5	-0.4	100	2000	2000	2000	2000
0.25	3.5	-0.4	200	2000	2000	2000	2000
0.25	3.5	-0.4	500	2000	2000	2000	2000
0.25	3.5	-0.2	50	2000	2000	2000	2000
0.25	3.5	-0.2	100	2000	2000	2000	2000
0.25	3.5	-0.2	200	2000	2000	2000	2000
0.25	3.5	-0.2	500	2000	2000	2000	2000
0.25	3.5	0	50	2000	2000	2000	2000
0.25	3.5	0	100	2000	2000	2000	2000
0.25	3.5	0	200	2000	2000	2000	2000
0.25	3.5	0	500	2000	2000	2000	2000
0.25	0.5	-0.4	50	2000	2000	2000	2000
0.25	0.5	-0.4	100	2000	2000	2000	2000
0.25	0.5	-0.4	200	2000	2000	2000	2000
0.25	0.5	-0.4	500	2000	2000	2000	2000
0.25	0.5	-0.2	50	2000	2000	2000	2000
0.25	0.5	-0.2	100	2000	2000	2000	2000
0.25	0.5	-0.2	200	2000	2000	2000	2000
0.25	0.5	-0.2	500	2000	2000	2000	2000
0.25	0.5	0	50	2000	2000	2000	2000
0.25	0.5	0	100	2000	2000	2000	2000
0.25	0.5	0	200	2000	2000	2000	2000
0.25	0.5	0	500	2000	2000	2000	2000
0	3.5	-0.4	50	2000	1996	1999	2000
0	3.5	-0.4	100	2000	1999	1998	2000
0	3.5	-0.4	200	2000	2000	2000	2000
0	3.5	-0.4	500	2000	2000	2000	2000
0	3.5	-0.2	50	2000	1999	1996	2000
0	3.5	-0.2	100	2000	1999	1999	2000
0	3.5	-0.2	200	2000	1997	1998	2000
0	3.5	-0.2	500	2000	2000	2000	2000
0	3.5	0	50	2000	1993	1999	2000
0	3.5	0	100	2000	1998	1996	2000
0	3.5	0	200	2000	1999	1999	2000
0	3.5	0	500	2000	1999	1999	2000
0	0.5	-0.4	50	2000	2000	1998	2000
0	0.5	-0.4	100	2000	1999	1998	2000
0	0.5	-0.4	200	2000	2000	1998	2000
0	0.5	-0.4	500	2000	2000	2000	2000
0	0.5	-0.2	50	2000	2000	1999	2000
0	0.5	-0.2	100	2000	2000	1999	2000
0	0.5	-0.2	200	2000	2000	2000	2000
0	0.5	-0.2	500	2000	2000	2000	2000
0	0.5	0	50	2000	1998	1998	2000
0	0.5	0	100	2000	1999	2000	2000
0	0.5	0	200	2000	2000	2000	2000
0	0.5	0	500	2000	2000	2000	2000

Table C-3 Goodwin et al. dataset: *NB-logged* including baseline characteristics fitted to the intervention count (n=125)

	Estimate	**SE**	**FRR (95% CI)**	**P**
Intervention	-0.441	0.155	0.643 (0.473, 0.874)	0.004
Log(baseline count + 0.5)	0.946	0.056	2.574 (2.303, 2.878)	< 0.001
Female	-0.062	0.167	0.940 (0.675, 1.307)	0.708
Age	0.012	0.010	1.012 (0.992, 1.033)	0.243
Years since diagnosis	0.027	0.014	1.028 (1.000, 1.056)	0.046
Hoehn & Yahr				
Stage 1	0.019	0.285	1.019 (0.580, 1.791)	0.947
Stage 2			1	
Stage 3	-0.233	0.191	0.792 (0.543, 1.155)	0.221
Stage 4	-0.070	0.249	0.932 (0.569, 1.526)	0.777
Living status				
With partner				
Alone	0.284	0.194	1.328 (0.905, 1.948)	0.143
With family/friends	1.421	0.629	4.141 (1.191, 14.392)	0.024
Residential home	-1.373	1.107	0.253 (0.028, 2.272)	0.215
HP	0.468			
Dispersion	1.1			
AIC	750.7			
Overdispersion test	0.641			

Table C-4 Goodwin et al. dataset: *NB-logged* fitted to the follow-up count (n=120)

	Estimate	**SE**	**FRR (95% CI)**	**P**
Intervention	-0.335	0.219	0.716 (0.464, 1.103)	0.126
Log(baseline count + 0.5)	0.927	0.068	2.528 (2.208, 2.894)	< 0.001
HP	1.105			
Dispersion	1.1			
AIC	700.4			
Overdispersion test	0.305			

Table C-5 Goodwin et al. dataset: *NB-logged* including baseline characteristics fitted to the follow-up count (n=120)

	Est.	SE	FRR (95% CI)	P
Intervention	-0.369	0.220	0.691 (0.447, 1.070)	0.094
Log(baseline count + 0.5)	0.942	0.076	2.566 (2.206, 2.984)	< 0.001
Female	0.028	0.234	1.029 (0.646, 1.638)	0.903
Age	0.003	0.014	1.003 (0.975, 1.032)	0.840
Years since diagnosis	0.026	0.019	1.026 (0.988, 1.066)	0.183
Hoehn & Yahr				
Stage 1	0.525	0.379	1.690 (0.798, 3.581)	0.166
Stage 2				
Stage 3	-0.109	0.277	0.896 (0.517, 1.553)	0.693
Stage 4	0.103	0.349	1.109 (0.554, 2.216)	0.768
Living status				
With partner				
Alone	0.048	0.274	1.049 (0.609, 1.806)	0.862
With family/friends	0.917	0.893	2.501 (0.426, 14.677)	0.304
Residential home	-0.418	0.981	0.658 (0.094, 4.602)	0.670
HP	1.040			
Dispersion	1.1			
AIC	713.4			
Overdispersion test	0.510			

Table C-6 EXSart dataset: *NB-logged* including baseline characteristics fitted to the intervention count (n=126)

	Est.	SE	FRR (95% CI)	P
Intervention	-0.091	0.209	0.913 (0.603, 1.381)	0.663
Log(baseline count)	0.415	0.065	1.514 (1.332, 1.721)	0.000
Female	-0.279	0.224	0.756 (0.485, 1.179)	0.213
Age	0.010	0.013	1.010 (0.985, 1.036)	0.431
Years since diagnosis	0.022	0.017	1.022 (0.989, 1.057)	0.189
Hoehn & Yahr				
Stage 2	-0.837	0.427	0.433 (0.186, 1.009)	0.050
Stage 3			1	
Stage 4	0.165	0.286	1.179 (0.669, 2.076)	0.565
UPDRS	-0.007	0.012	0.993 (0.970, 1.016)	0.540
Living status				
With partner				
Alone	-0.350	0.278	0.704 (0.406, 1.222)	0.208
With family/friends /others	-0.232	0.487	0.793 (0.302, 2.083)	0.635
HP	0.587			
Dispersion	1.0			
AIC	450.1			
Overdispersion test	P = 0.552			

Table C-7 EXSart dataset: *NB-logged* including baseline characteristics fitted to the follow-up count (n=126)

	Est.	SE	FRR (95% CI)	P
Intervention	-0.254	0.236	0.776 (0.486, 1.239)	0.283
Log(baseline count)	0.240	0.078	1.271 (1.089, 1.484)	0.002
Female	-0.341	0.247	0.711 (0.436, 1.159)	0.167
Age	-0.006	0.015	0.994 (0.965, 1.023)	0.675
Years since diagnosis	0.030	0.020	1.031 (0.990, 1.073)	0.136
Hoehn & Yahr				
Stage 2	-1.910	0.532	0.148 (0.052, 0.425)	0.000
Stage 3				
Stage 4	0.754	0.325	2.125 (1.116, 4.045)	0.020
UPDRS	-0.022	0.014	0.978 (0.952, 1.005)	0.101
Living status				
With partner			1	
Alone	-0.416	0.307	0.659 (0.359, 1.211)	0.175
With family/friends /others	-0.192	0.537	0.826 (0.285, 2.393)	0.721
HP	1.132			
Dispersion	1.1			
AIC	550.1			
Overdispersion test	P = 0.283			

References

Aban, I.B., Cutter, G.R., Mavinga, N., 2009. Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. Comput. Stat. Data Anal. 53, 820–833.

Aeberhard, W.H., 2016. robNB: Robust estimation and tests for negative binomial regression.

Aeberhard, W.H., Cantoni, E., Heritier, S., 2017. Saddlepoint tests for accurate and robust inference on overdispersed count data. Comput. Stat. Data Anal. 107, 162–175.

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. 19, 716–723.

Allcock, L.M., Rowan, E.N., Steen, I.N., Wesnes, K., Kenny, R.A., Burn, D.J., 2009. Impaired attention predicts falling in Parkinson's disease. Park. Relat. Disord. 15, 110–115.

Andersen, P.K., Gill, R.D., 1982. Cox 's regression model for counting processes : A large sample study. Ann. Stat. 10, 1100–1120.

Anscombe, F.J., 1972. Contribution to the discussion of H. Hotelling's paper. J. R. Stat. Soc. – Ser. B 15, 229–230.

Apple Inc., 2018. Use fall detection with Apple Watch Series 4 [WWW Document]. URL https://support.apple.com/en-gb/HT208944

Ashburn, A., Fazakarley, L., Ballinger, C., Pickering, R., McLellan, L.D., 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.

Ashburn, A., Stack, E., Pickering, R.M., Ward, C.D., 2001. A community-dwelling sample of people with Parkinson's disesase: Characteristics of fallers and non-fallers. Age Ageing 30, 47–52.

Assmann, S.F., Pocock, S.J., Enos, L.E., Kasten, L.E., 2000. Subgroup analysis and other (mis) uses of baseline data in clinical trials. Lancet 355, 1064–1069.

Atkinson, A., Riani, M., 2012. Robust diagnostic regression analysis. Springer Science & Business Media.

Balakrishnan, N., 2014. Methods and Applications of Statistics in Clinical Trials, Volume 2: Planning, Analysis, and Inferential Methods. John Wiley & Sons.

Belsley, D.A., Kuh, E., Welch, R.E., 1980. Regression diagnostics: identifying data and sources of colinearity. New York: J. NY John Wiley Sons Inc.

Böhning, D., 1999. Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others. CRC press.

Böhning, D., Seidel, W., 2003. Editorial: Recent developments in mixture models. Comput. Stat. Data Anal. 41, 349–357.

Boston College Department of Economics, n.d. Statistical Software Components [WWW Document]. URL https://ideas.repec.org/s/boc/bocode.html (accessed 1.19.19).

Bot, B.M., Suver, C., Neto, E.C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E.R., Friend, S.H., Trister, A.D., 2016. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci. Data 3, 160011.

Brännäs, K., 1992. Limited dependent Poisson regression. TheJournal R. Stat. Soc. Ser. D (The Stat. 41, 413–423.

Breslow, N., 1990. Tests of Hypotheses in Overdispersed Poisson Regression and other Quasi-Likelihood Models. J. Am. Stat. Assoc. 85, 565–571.

Breslow, N.E., 1996. Generalized Linear Models: Checking for assumtions and strengthening conclusions. Prep. Congr. Naz.

Cai, J., Schaubel, D.E., 2004. Marginal means/rates models for multiple type recurrent event data. Lifetime Data Anal. 10, 121–138.

Cameron, A., Trivedi, P., 1986. Econometric Models Based on Count Data-Comparisons and Applications of Some Estimators and Tests. J. Appl. Econom. 1, 29–53.

Cameron, A.C., Trivedi, P.K., 2013. Regression analysis of count data. Cambridge university press.

Canning, C.G., Sherrington, C., Lord, S.R., Close, J.C.T., Heller, G.Z., Howard, K., Allen, N.E., Latt, M.D., Murray, S.M., Rourke, S.D.O., Paul, S.S., 2014. Exercise for falls prevention in Parkinson disease A randomized controlled trial. Am. Acad. Neurol. 84, 304–312.

Castañeda, J., Gerritse, B., 2010. Appraisal of Several Methods to Model Time to Multiple Events per Subject: Modelling Time to Hospitalizations and Death. Rev. Colomb. Estad. 33, 43–61.

Clark, R.D., Lord, S.R., Webster, I.W., 1993. Clinical Parameters Associated with Falls in an Elderly Population. Gerontology 39, 117–123.

Cook, R.D., 1977. Detection of Influential Observation in Linear Regression. Technometrics 19, 15–18.

Cook, R.D., Weisberg, S., 1982. Residuals and influence in regression, Monographs on statistics and applied probability. Chapman and Hall, New York.

Cook, R.J., Lawless, J., 2007. The statistical analysis of recurrent events. Springer Science &

Business Media.

Cook, R.J., Lawless, J.F., 2002. Analysis of repeated events. Stat. Methods Med. Res. 11, 141–166.

Cook, R.J., Wei, W., 2003. Conditional analysis of mixed Poisson processes with baseline counts: implications for trial design and analysis. Biostatistics 4, 479–494.

Cook, R.J., Wei, W., Yi, G.Y., 2005. Robust tests for treatment effects based on censored recurrent event data observed over multiple periods. Biometrics 61, 692–701.

Cui, J., 2007. QIC program and model selection in GEE analyses. Stata J. 7, 209.

Cumming, R.G., Kelsey, J.L., Nevitt, M.C., 1990. Methodologic issues in the study of frequent and recurrent health problems falls in the elderly. Ann. Epidemiol. 1, 49–56.

Cumming, R.G., Thomas, M., Szonyi, G., Salkeld, G., O'Neill, E., Westbury, C., Frampton, G., 1999. Home visits by an occupational therapist for assessment and modification of environmental hazards: a randomized trial of falls prevention. J. Am. Geriatr. Soc. 47, 1397–1402.

Cummings, S.R., Nevitt, M.C., Kidd, S., 1988. Forgetting Falls. J. Am. Geriatr. Soc. 36, 613–616.

Davison, A.C., 2003. Statistical models. Cambridge University Press.

Day, L., Fildes, B., Gordon, I., Fitzharris, M., Flamer, H., Lord, S., 2002. Randomised factorial trial of falls prevention among older people living in their own homes. BMJ 325, 128–128.

Dean, C., Lawless, J.F., Willmot, G.E., 1989. A mixed Poisson–inverse-Gaussian regression model. Can. J. Stat. 17, 171–182.

Deandrea, S., Lucenteforte, E., Bravi, F., Foschi, R., La Vecchia, C., Negri, E., 2010. Risk factors for falls in community-dwelling older people: a systematic review and meta-analysis. Epidemiology 21, 658–668.

Deane, K.H.O., Ellis-Hill, C., Jones, D., Whurr, R., Ben-Shlomo, Y., Playford, E.D., Clarke, C.E., 2002. Systematic review of paramedical therapies for Parkinson's disease. Mov. Disord. 17, 984–991.

Dembe, A.E., Partridge, J.S., Geist, L.C., 2011. Statistical software applications used in health services research: Analysis of published studies in the U.S. BMC Health Serv. Res. 11, 252.

Desmarais, B.A., Harden, J.J., 2013. Testing for zero inflation in count models: Bias correction for the Vuong test. Stata J. 13, 810–835.

Donaldson, M.G., Sobolev, B., Cook, W.L., Janssen, P. a., Khan, K.M., 2009. Analysis of recurrent events: A systematic review of randomised controlled trials of interventions to prevent falls. Age Ageing 38, 151–155.

Everitt, B.S., 1995. The Cambridge dictionary of statistics in the medical sciences. Cambridge University Press Cambridge.

Fahn, S., Oakes, D., Shoulson, I., Kieburtz, K., Rudolph, A., Lang, A., Olanow, C.W., Tanner, C., Marek, K., 2004. Levodopa and the progression of Parkinson’s disease. N. Engl. J. Med. 351, 2498–2508.

Fisher, R.A., 1956. Statistical methods and scientific inference.

Foongsathaporn, C., Panyakaew, P., Jitkritsadakul, O., Bhidayasiri, R., 2016. What daily activities increase the risk of falling in Parkinson patients? An analysis of the utility of the ABC-16 scale. J. Neurol. Sci. 364, 183–187.

Freedman, D.A., 2006. On the so-called “Huber Sandwich Estimator” and “robust standard errors.” Am. Stat. 60, 299–302.

Gardner, M.M., Roberstson, M.C., Campbell, A.J., 2000. Exercise in preventing falls and fall related injuries in older people. Br. J. Sports Med. 34, 7–17.

Genever, R.W., Downes, T.W., Medcalf, P., 2005. Fracture rates in Parkinson’s disease compared with age- and gender-matched controls: a retrospective cohort study. Age Ageing 34, 21–24.

Ghayvat, H., Liu, J., Mukhopadhyay, S.C., Gui, X., 2015. Wellness Sensor Networks: A Proposal and Implementation for Smart Home for Assisted Living. IEEE Sens. J. 15, 7341–7348.

Gill, D.P., Zou, G.Y., Jones, G.R., Speechley, M., 2009. Comparison of Regression Models for the Analysis of Fall Risk Factors in Older Veterans. Ann. Epidemiol. 19, 523–530.

Gillespie, L.D., Robertson, M.C., Gillespie, W.J., Sherrington, C., Gates, S., Clemson, L.M., Lamb, S.E., 2012. Interventions for preventing falls in older people living in the community, in: Gillespie, L.D. (Ed.), Cochrane Database of Systematic Reviews. John Wiley & Sons, Ltd, Chichester, UK, p. CD007146.

Glynn, R.J., Buring, J.E., 1996. Ways of measuring rates of recurrent events. BMJ 312, 364–367.

Goetz, C.G., Pal, G., 2014. Initial management of Parkinson’s disease. BMJ 349, g6258–g6258.

Goetz, C.G., Poewe, W., Rascol, O., Sampaio, C., Stebbins, G.T., Counsell, C., Giladi, N., Holloway, R.G., Moore, C.G., Wenning, G.K., Yahr, M.D., Seidl, L., 2004. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations. Mov. Disord. 19, 1020–1028.

Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A.E., Lees, A., Leurgans, S., LeWitt, P. a., Nyenhuis, D., Olanow, C.W., Rascol, O., Schrag, A., Teresi, J. a., van Hilten, J.J., LaPelle, N., Agarwal, P., Athar, S., Bordelan, Y., Bronte-Stewart, H.M., Camicioli, R., Chou, K., Cole, W., Dalvi, A., Delgado, H., Diamond, A., Dick, J.P., Duda, J., Elble, R.J., Evans, C., Evidente, V.G., Fernandez, H.H., Fox, S., Friedman, J.H., Fross, R.D., Gallagher, D., Goetz, C.G., Hall, D., Hermanowicz, N., Hinson, V., Horn, S., Hurtig, H., Kang, U.J., Kleiner-Fisman, G., Klepitskaya, O., Kompoliti, K., Lai, E.C., Leehey, M.L., Leroi, I., Lyons, K.E., McClain, T., Metzer, S.W., Miyasaki, J., Morgan, J.C., Nance, M., Nemeth, J., Pahwa, R., Parashos, S. a., Schneider, J.S.J.S., Schrag, A., Sethi, K., Shulman, L.M., Siderowf, A., Silverdale, M., Simuni, T., Stacy, M., Stern, M.B., Stewart, R.M., Sullivan, K., Swope, D.M., Wadia, P.M., Walker, R.W., Walker, R., Weiner, W.J., Wiener, J., Wilkinson, J., Wojcieszek, J.M., Wolfrath, S., Wooten, F., Wu, A., Zesiewicz, T. a., Zweig, R.M., 2008. Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov. Disord. 23, 2129–2170.

Goodwin, V.A., Richards, S.H., Henley, W., Ewings, P., Taylor, A.H., 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.

Gray, P., Hildebrand, K., 2000. Fall risk factors in Parkinson's disease. J. Neurosci. Nurs. 32, 222–228.

Grogger, J.T., Carson, R.T., 1988a. Models for counts from choice based samples. Dep. Econ. Work. pap., Univ. Calif.

Grogger, J.T., Carson, R.T., 1988b. Truncated counts. Univ. California, San Diego.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions. Oxf. Bull. Econ. Stat. 64, 63-82+2.

Gurmu, S., Trivedi, P.K., 1992. Overdispersion tests for truncated Poisson regression models. J. Econom. 54, 347–370.

Hardin, J.W., 2003. The sandwich estimate of variance, in: Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later. Emerald Group Publishing Limited, pp. 45–73.

Hardin, J.W., Hilbe, J.M., 2012. pigreg.

Hardin, J.W., Hilbe, J.M., 2007. Generalized linear models and extensions. Stata press.

Hardin, J.W., Hilbe, J.M., 2002. Generalized estimating equations. Chapman and Hall/CRC.

Hauser, R.A., Heritier, S., Rowse, G.J., Hewitt, L.A., Isaacson, S.H., 2016. Droxidopa and reduced falls in a trial of Parkinson disease patients with neurogenic orthostatic hypotension. Clin. Neuropharmacol. 39, 220–226.

Hausman, J.A., Hall, B.H., Griliches, Z., 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica 52.

Henderson, E.J., Lord, S.R., Brodie, M.A., Gaunt, D.M., Lawrence, A.D., Close, J.C.T., Whone, A.L., Ben-Shlomo, Y., 2016. Rivastigmine for gait stability in patients with Parkinson's disease (ReSPonD): a randomised, double-blind, placebo-controlled, phase 2 trial. Lancet Neurol. 15, 249–258.

Hilbe, J.M., 2014. Modeling count data. Cambridge University Press.

Hilbe, J.M., 2011. Negative Binomial Regression, 2nd ed, Cambridge University Press.

Hilbe, J.M., Robinson, A., 2016. Methods of statistical model estimation. Chapman and Hall/CRC.

Hilbe, J.M., Robinson, A., 2014. msme: Functions and Datasets for “Methods of Statistical Model Estimation”.

Hin, L.-Y., Wang, Y.-G., 2009. Working-correlation-structure identification in generalized estimating equations. Stat. Med. 28, 642–658.

Hoehn, M.M., Yahr, M.D., 1967. Parkinsonism: onset, progression, and mortality. Neurology 17, 427 LP – 427.

Holling, H., Böhning, W., Böhning, D., Formann, A.K., 2016. The covariate-adjusted frequency plot. Stat. Methods Med. Res. 25, 902–916.

Hornbrook, M.C., Stevens, V.J., Wingfield, D.J., Hollis, J.F., Greenlick, M.R., Ory, M.G., 1994. Preventing falls among community-dwelling older persons: results from a randomized trial. Gerontologist 34, 16–23.

Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. Proc. fifth Berkeley Symp. Math. Stat. Probab. Vol. 1. Berkeley 221–233.

IBM Corp., 2017. IBM SPSS Statistics for Windows.

Jackman, S., Kleiber, C., Zeileis, A., 2007. Regression Models for Count Data in R.

Jahn-Eimermacher, A., 2008. Comparison of the Andersen-Gill model with poisson and negative binomial regression on recurrent event data. Comput. Stat. Data Anal. 52, 4989–4997.

Jakobsen, J., Tamborrino, M., Winkel, P., Haase, N., Perner, A., Wetterslev, J., Gluud, C., 2015. Count Data Analysis in Randomised Clinical Trials. Biometrics Biostat. 6, 1–5.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Springer.

Jørstad, E.C., Hauer, K., Becker, C., Lamb, S.E., 2005. Measuring the psychological outcomes of falling: A systematic review. J. Am. Geriatr. Soc. 53, 501–510.

Keus, S.H.J., Bloem, B.R., Verbaan, D., de Jonge, P.A., Hofman, M., van Hilten, B.J., Munneke, M., 2004. Physiotherapy in Parkinson’s disease: utilisation and patient satisfaction. J. Neurol. 251, 680–687.

Kim, J., Lee, W., 2018. On testing the hidden heterogeneity in negative binomial regression models. Metrika.

King, G., Roberts, M.E., 2015. How robust standard errors expose methodological problems they do not fix, and what to do about it. Polit. Anal. 23, 159–179.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46, 33–50.

Lamb, S.E., Jørstad-Stein, E.C., Hauer, K., Becker, C., 2005. Development of a Common Outcome Data Set for Fall Injury Prevention Trials: The Prevention of Falls Network Europe Consensus. J. Am. Geriatr. Soc. 53, 1618–1622.

Lambert, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Latt, M.D., Lord, S.R., Morris, J.G., Fung, V.S., 2009. Clinical and physiological assessments for elucidating falls risk in Parkinson's disease. Mov. Disord. 24, 1280–1289.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Can. J. Stat. 15, 209–225.

Lesnoff, M., Lancelot, R., 2012. aod: Analysis of Overdispersed Data.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 73, 13–22.

Liu-Ambrose, T., Donaldson, M.G., Ahamed, Y., Graf, P., Cook, W.L., Close, J., Lord, S.R., Khan, K.M., 2008. Otago home-based strength and balance retraining improves executive functioning in older fallers: A randomized controlled trial. J. Am. Geriatr. Soc. 56, 1821–1830.

Long, J.S., Freese, J., 2006. Regression models for categorical dependent variables using Stata. Stata press.

Machado, J.A.F., Santos Silva, J.M.C., 2005. Quantiles for counts. J. Am. Stat. Assoc. 100, 1226–1237.

Mackenzie, L., Byles, J., D'Este, C., 2006. Validation of self-reported fall events in intervention studies. Clin. Rehabil. 20, 331–339.

Madsen, H., Thyregod, P., 2010. Introduction to general and generalized linear models. CRC Press.

Martin, T., Weatherall, M., Anderson, T.J., 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.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models.

McKinney, W., 2010. Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference. Austin, TX, pp. 51–56.

McLennan, J.E., Nakano, K., Tyler, H.R., Schwab, R.S., 1972. Micrographia in Parkinson's disease. J. Neurol. Sci. 15, 141–152.

Melton, L.J., Leibson, C.L., Achenbach, S.J., Bower, J.H., Maraganore, D.M., Oberg, A.L., Rocca, W.A., 2006. Fracture risk after the diagnosis of Parkinson's disease: Influence of concomitant dementia. Mov. Disord. 21, 1361–1367.

National Collaborating Centre for Chronic Conditions, 2006. Parkinson's disease: national clinical guideline for diagnosis and management in primary and secondary care, Royal College of Physicians. Royal College of Physicians, London.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. J. R. Stat. Soc. Ser. A J. R. Stat. Soc. Ser. A (General J. R. Stat. Soc. A 13517213, 370–384.

Ng, E., Cook, R.J., 1999. Robust inference for bivariate point processes. Can. J. Stat. 27, 509–524.

Nieuwboer, A., 2008. Cueing for freezing of gait in patients with Parkinson's disease: A rehabilitation perspective. Mov. Disord. 23.

Nijkrake, M.J., Keus, S.H.J., Kalf, J.G., Sturkenboom, I.H.W.M., Munneke, M., Kappelle, A.C., Bloem, B.R., 2007. Allied health care interventions and complementary therapies in Parkinson's disease. Parkinsonism Relat. Disord. 13, S488–S494.

Nikolaus, T., Bach, M., 2003. Preventing falls in community-dwelling frail older people using a home intervention team (HIT): Results from the randomized falls-HIT trial. J. Am. Geriatr. Soc. 51, 300–305.

Nyström, H., Nordström, A., 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.

O., J., L.J., M.I.I.I., E.J., A., W.M., O., L.T., K., 1992. Fracture risk in patients with Parkinsonism: A population-based study in Olmsted County, Minnesota. Age Ageing 21, 32–38.

Oliphant, T.E., 2006. A guide to NumPy. Trelgol Publishing USA.

Pan, W., 2001. Akaike's information criterion in generalized estimating equations. Biometrics 57, 120–125.

Parkinson, J., 1817. An Essay on the Shaking Palsy, Sherwood, Neeley and Jones.

Paul, S.S., Harvey, L., Canning, C.G., Boufous, S., Lord, S.R., Close, J.C.T., Sherrington, C., 2017. Fall-related hospitalization in people with Parkinson's disease. Eur. J. Neurol. 24, 523–529.

Peel, N., 2000. Validating recall of falls by older people. Accid. Anal. Prev. 32, 371–372.

Pickering, R.M., Grimbergen, Y.A.M., Rigney, U., Ashburn, A., Mazibrada, G., Wood, B., Gray, P., Kerr, G., Bloem, B.R., 2007. A meta-analysis of six prospective studies of falling in Parkinson's disease. Mov. Disord. 22, 1892–1900.

Province, M.A., Hadley, E.C., Hornbrook, M.C., Lipsitz, L.A., Miller, J.P., Mulrow, C.D., Ory, M.G., Sattin, R.W., Tinetti, M.E., Wolf, S.L., 1995. The effects of exercise on falls in elderly patients: a preplanned meta-analysis of the FICSIT trials. Jama 273, 1341–1347.

Python Core Team, 2015. Python: A dynamic, open source programming language.

R-Foundation, 2015. The R Foundation [WWW Document]. URL https://www.r-project.org/foundation/

R Core Team, 2016. R: A Language and Environment for Statistical Computing.

Rascol, O., Payoux, P., Ferreira, J., Brefel-Courbon, C., 2002. The management of patients with early Parkinson's disease. Parkinsonism Relat. Disord. 9, 61–67.

Reid, I.R., Mason, B., Horne, A., Ames, R., Reid, H.E., Bava, U., Bolland, M.J., Gamble, G.D., 2006. Randomized Controlled Trial of Calcium in Healthy Older Women. Am. J. Med. 119, 777–785.

Ridout, M., Hinde, J., DeméAtrio, C.G.B., 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57, 219–223.

Rigby, R.A., Stasinopoulos, D.M., Heller, G.Z., De Bastiani, F., 2017. Distributions for modelling location, scale, and shape: using GAMLSS in R. URL www. gamlss. org.(last accessed 5 March 2018).

Robertson, M.C., Campbell, A.J., Herbison, P., 2005. Statistical analysis of efficacy in falls prevention trials. J. Gerontol. A. Biol. Sci. Med. Sci. 60, 530–534.

Rocha, P.A., Porfírio, G.M., Ferraz, H.B., Trevisani, V.F.M., 2014. Effects of external cues on gait parameters of Parkinson's disease patients: A systematic review. Clin. Neurol. Neurosurg. 124, 127–134.

Rubenis, J., 2007. A rehabilitational approach to the management of Parkinson's disease. Parkinsonism Relat. Disord. 13, S495–S497.

Ryan, D.J., Nick, S., Colette, S.M., Roseanne, K., 2010. Carotid sinus syndrome, should we pace? A multicentre, randomised control trial (Safepace 2). Heart 96, 347–351.

Sanders, K.M., Stuart, A.L., Williamson, E.J., Simpson, J.A., Kotowicz, M.A., Young, D., Nicholson, G.C., 2010. Annual High-Dose Oral Vitamin D and Falls and Fractures in Older Women. Jama 303, 1815.

SAS Institute Inc., 2013. SAS for Windows.

Schlattmann, P., 2009. Medical applications of finite mixture models. Springer.

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464.

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python, in: Proceedings of the 9th Python in Science Conference. SciPy society Austin, p. 61.

Smith, H., Anderson, F., Raphael, H., Maslin, P., Crozier, S., Cooper, C., 2007. Effect of annual intramuscular vitamin D on fracture risk in elderly men and women - A population-based, randomized, double-blind, placebo-controlled trial. Rheumatology 46, 1852–1857.

Spaulding, S.J., Barber, B., Colby, M., Cormack, B., Mick, T., Jenkins, M.E., 2013. Cueing and gait improvement among people with Parkinson's disease: A meta-analysis. Arch. Phys. Med. Rehabil. 94, 562–570.

Spineli, L.M., Jenz, E., Großhennig, A., Koch, A., 2017. Critical appraisal of arguments for the delayed-start design proposed as alternative to the parallel-group randomized clinical trial design in the field of rare disease. Orphanet J. Rare Dis. 12, 1–7.

Sroka, C.J., Nagaraja, H.N., 2018. Odds ratios from logistic, geometric, Poisson, and negative binomial regression models. BMC Med. Res. Methodol. 18, 112.

Srulijes, K., Klenk, J., Schwenk, M., Schatton, C., Schwickert, L., Teubner-Liepert, K., Meyer, M., K C, S., Maetzler, W., Becker, C., Synofzik, M., 2019. Fall Risk in Relation to Individual Physical Activity Exposure in Patients with Different Neurodegenerative Diseases: a Pilot Study. Cerebellum.

Stanaway, F.F., Cumming, R.G., Naganathan, V., Blyth, F.M., Handelsman, D.J., Le Couteur, D.G., Waite, L.M., Creasey, H.M., Seibel, M.J., Sambrook, P.N., 2011. Ethnicity and falls in older men: Low rate of falls in Italian-born men in Australia. Age Ageing 40, 595–601.

Stasinopoulos, M. (London M.U., Rigby, R.A. (London metropolitan U., 2007. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. J. Stat. Softw. 223, 1–46.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible regression and smoothing: using gamlss in r. CRC Press.

StataCorp, 2015. Stata 14 Base Reference Manual. Stata Press., College Station, TX.

Stevens, M., Holman, C.D.J., Bennett, N., De Klerk, N., 2001. Preventing falls in older people: outcome evaluation of a randomized controlled trial. J. Am. Geriatr. Soc. 49, 1448–1455.

Tango, T., 2009. Sample size formula for randomized controlled trials with counts of recurrent events. Stat. Probab. Lett. 79, 466–472.

Terza, J. V., 1985. A Tobit-type estimator for the censored Poisson regression model. Econ. Lett. 18, 361–365.

Thas, O., Neve, J. De, Clement, L., Ottoy, J.P., 2012. Probabilistic index models. J. R. Stat. Soc. Ser. B Stat. Methodol. 74, 623–671.

Tinetti, M.E., Speechley, M., Ginter, S.F., 1988. Risk factors for falls among elderly persons living in the community. N. Engl. J. Med. 319, 1701–1707.

Tomlinson, C.L., Herd, C.P., Clarke, C.E., Meek, C., Patel, S., Stowe, R., Deane, K.H., Shah, L., Sackley, C.M., Wheatley, K., Ives, N., 2014. Physiotherapy for Parkinson's disease: a comparison of techniques, in: Tomlinson, C.L. (Ed.), Cochrane Database of Systematic

Reviews. John Wiley & Sons, Ltd, Chichester, UK.

Tysnes, O.B., Storstein, A., 2017. Epidemiology of Parkinson's disease. J. Neural Transm. 124, 901–905.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, Fourth. ed. Springer, New York.

Venkataraman, N., Shankar, V., Blum, J., Hariharan, B., Hong, J., 2016. Transferability Analysis of Heterogeneous Overdispersion Parameter Negative Binomial Crash Models. Transp. Res. Rec. J. Transp. Res. Board 2583, 99–109.

Vermeulen, K., Thas, O., Vansteelandt, S., 2015. Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment. Stat. Med. 34, 1012–1030.

Vestergaard, P., Rejnmark, L., Mosekilde, L., 2007. Fracture Risk Associated with Parkinsonism and Anti-Parkinson Drugs. Calcif. Tissue Int. 81, 153–161.

Vickers, A.J., Altman, D.G., 2001. Analysing controlled trials with baseline and follow up measurements. BMJ 323, 1123–1124.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econom. J. Econom. Soc. 57, 307–333.

Wade, D.T., 1992. Measurement in neurological rehabilitation. Oxford University Press, New York.

Wang, H.C., Lin, C.C., Lau, C.I., Chang, A., Sung, F.C., Kao, C.H., 2014. Risk of accidental injuries amongst Parkinson disease patients. Eur. J. Neurol. 21, 907–913.

White, H., 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica 48, 817–828.

White, I.R., 2010. simsum: Analyses of simulation studies including Monte Carlo error. Stata J. 10, 369.

Williams, D.A., 1987. Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. Appl. Stat. 36, 181.

Winkelmann, R., 2008. Econometric Analysis of Count Data. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wood, B.H., Bilclough, J. a, Bowron, a, Walker, R.W., 2002. Incidence and prediction of falls in Parkinson's disease: a prospective multidisciplinary study. J. Neurol. Neurosurg. Psychiatry 72, 721–725.

Wood, S.N., 2017. Generalized additive models: an introduction with R. CRC press.

Xue, D., Deddens, J.A., 1992. Overdispersed negative binomial regression models. Commun. Stat. - Theory Methods 21, 2215–2226.

Yardley, L., 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.

Yau, K.K.W., Wang, K., Lee, A.H., 2003. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical J. 45, 437–452.

Zheng, H., Kimber, A., Goodwin, V.A., Pickering, R.M., 2018. A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials. Biometrical J. 60, 66–78.

Zhu, H., Lakkis, H., 2014. Sample size calculation for comparing two negative binomial rates. Stat. Med. 33, 376–387.