Maximizing statistical power in neurological trials by covariate adjustment, exploiting ordinality and repeated assessments: A simulation study and application to GBS

Ruiter, S.C. de

Maximizing statistical power in neurological trials by covariate adjustment, exploiting ordinality and repeated assessments

A simulation study and application to GBS

Sophie de Ruiter (s1853457)

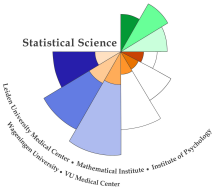
First supervisor: Prof. Dr. Ewout Steyerberg

Second supervisor: Prof. Dr. Bart Jacobs

MASTER THESIS

Defended on December 10, 2021

Specialization: Statistical Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Foreword

First of all, I would like to thank my supervisors Prof. Dr. Ewout Steyerberg and Prof. Dr. Bart Jacobs for guiding me through the research activities. Ewout, thank you for the weekly meetings where I could discuss my doubts and decisions. Thank you also for the opportunities you provided me. Examples are the interactions with a medical student to advice on the statistics of an academic study and the involvement in weekly meetings with biostatisticians from a pharmaceutical company overseas. These experiences taught me a lot, not only about statistics and its real-world application, but it also strengthened my personal skills. Bart, thank you for providing me insight in the current challenges in neurological trials and for the useful clinical input in reviewing my thesis. Understanding the relevance of the current study, made it an even more special experience.

Thank you Linda, for thinking along and for your helpful comments on my writings about the medical aspects. Thanks to the other PhD students and "colleagues" at the 22^{nd} floor at Erasmus Medical Center for creating a pleasant working environment.

Finally, I would like to express my gratitude to my friends and family for their emotional support during the last couple of months. Special thanks to my sister Anne, who helped me to put things to perspective, and to my roommates who served as essential office-mates during COVID-times.

A general research question for this project was set up by Bart and Ewout in the past. The data used in this study is obtained from a previous study carried out at Erasmus Medical Center and is confidential. The main body of this thesis is written in a style resembling that of a scientific paper, because this was found to be the most useful in preparation for a potential career in academia.

Contents

Abstract

Objective
Randomized controlled trials (RCTs) for rare neurological diseases, such as the Guillain-Barré syndrome (GBS), have a disappointing lack of success, possibly due to inefficient statistical analysis. We aimed to evaluate the impact of covariate adjustment for baseline characteristics, ordinal analysis and repeated assessments on statistical power in randomized controlled trials with ordinal scales as outcome measure.

Methods
We re-analysed a previous trial in GBS (the IVIg + placebo vs IVIg + Methylprednisolone trial, $n = 221$) and conducted power simulations to assess performance of different approaches for analysis of ordinal scales such as the GBS Disability Scale under different conditions. The approaches consist of binary logistic regression and proportional odds logistic regression, with and without covariate adjustment for important prognostic factors (MRC sum score and days since onset of weakness to randomisation). The conditions consist of satisfaction of the proportional odds assumption, the use of weaker prognostic baseline characteristics, and quantitative versus qualitative violation of the proportional odds assumption. We extended these approaches to a longitudinal proportional odds model. Simulations varied in sample size and treatment effect.

Results
Covariate adjustment led to an increased estimated treatment effect and increased standard error in the GBS trial. Proportional odds analysis decreased the standard error in comparison to a binary logistic regression analysis, indicating a more sensitive analysis. The longitudinal proportional odds resulted in a larger standard error as compared to single time point proportional odds analyses. Simulations for analysis of continuous data with a linear mixed model confirmed that a longitudinal approach does not increase power as compared to a single time point analysis in case of a low within-subject variance, as was observed for the GBS trial. In simulations we focused on the effect of covariate adjustment and ordinal analysis. Simulations indicated that Type I errors were generally around 5%. A small gain in power was achieved by covariate adjustment for two known prognostic factors in GBS, and a larger gain by exploiting ordinality instead of dichotomizing the ordinal scale. The gains translated to a gain in power of up to 7 and 13% points by covariate adjustment and exploiting ordinality respectively. The gains in power were only slightly smaller under violation of the proportional odds assumption and with smaller prognostic effects of the covariates.

Conclusion
Optimal analysis of ordinal scales should adjust for baseline characteristics (covariate adjustment) and should respect the ordinality of the outcome measure. A longitudinal proportional odds model for analysis of repeated assessments may not have added benefit as compared to a single time point proportional odds model. Further research should confirm that the use of a longitudinal proportional odds model is only beneficial when the observed disease course within patients is more variable over time.

Chapter 1

Introduction

Randomized controlled trials (RCTs *) are regarded as the gold standard for evaluation of effectiveness of treatments and interventions. Each year, many trials are performed for neurological diseases. For Guillain-Barré Syndrome (GBS), a rare acute immune-mediated disease of the peripheral nervous system, however, no trial since the 1990s has indicated a benefit of new treatments for patients' recovery. A potential contributing factor to this lack of convincing positive findings might be suboptimal statistical analysis.

A first challenge in the analysis of RCTs for neurological diseases, among others, is dealing with ordinal outcomes. Outcome measures in trials in neurology are mostly ordinal scales quantifying the degree of recovery. The analysis of ordinal outcomes is not straight-forward. One might dichotomize the ordinal scale in favorable versus unfavorable outcomes, analyse the ordinal scale as if it was continuous or use the proportional odds model for ordinal analysis. Even though many studies advised to use the proportional odds model [1, 22, 31, 35], dichotomization is still common and seems appealing to researchers for a variety of reasons such, as its straightforward interpretation and its ease of use.

Another obstacle in RCTs for rare diseases with a heterogeneous patient population is imbalance. Especially in small trials, imbalance in baseline characteristics between treatment arms might exist by chance, despite the random allocation of treatment in RCTs [9, 32]. Covariate adjustment corrects for such imbalance and multiple studies support the use of covariate adjustment [17, 22]. Moreover, in other acute monophasic neurological diseases such as stroke and traumatic brain injury, covariate adjustment is regularly applied. With covariate adjustment, a conditional treatment effect is estimated. This yields a clinically relevant profile-specific estimate of the treatment effect [9]. Another benefit of covariate adjustment is its ability to increase statistical power [29, 33].

Also, the timing of evaluation for treatment effect forms a challenge. Especially in diseases with a heterogeneous time course, it is arbitrary which time point during the follow-up is considered as the most representative and informative about the effect of treatment on clinical recovery. Most trials in neurology analyse data of one time point only, which discards potentially valuable information. Modelling multiple repeated assessments in a longitudinal approach is expected to increase power since it extracts information over the complete patient's trajectory.

In this study, we aimed to find the most sensitive method for analysis of RCTs in neurology that deal with repeated assessments of ordinal outcomes. Hereto, we performed power simulations to compare different approaches: binary logistic regression and proportional odds logistic regression, with and without covariate adjustment. Simulations were conducted for different

*Abbreviations in Appendix A, page 26

treatment effect sizes and absence of treatment effect (to assess the type I error), for different sample sizes, for qualitative and quantitative violations of the proportional odds assumption, and for weak to strong covariate effects. For illustration, we demonstrated application of these approaches in the re-analysis of a previous trial in GBS.

Chapter 2

Background

2.1 Guillain-Barré syndrome

Guillain-Barré Syndrome is a severe immune-mediated disease of the peripheral nervous system, affecting 200 to 300 patients per year in the Netherlands [11]. Although patients differ considerably in their clinical presentation, common symptoms of GBS are acute limb weakness and sensory disturbances. Some patients experience facial paralysis and around 20% of the patients require mechanical ventilation [16]. GBS is an acute monophasic disease. The acute progressive phase typically is two to four weeks, followed by disease stabilization in a plateau phase and a recovery phase that can last from months to years. The established options for treatment of GBS are either intravenous immunoglobulin (IVIg) or plasmapheresis [16]. The use of methylprednisolone (MP) as adjuvant to IVIg was found to be not beneficial in a previously conducted RCT [34]. Despite treatment, trajectories of recovery are extremely heterogeneous [11]. Some patients may recover quickly, while others may require mechanical ventilation and observation at the intensive care unit for several months and remain severely disabled or might even die.

The primary outcome in most GBS trials is the following 7-category ordinal outcome, named the GBS Disability Scale (GBS DS):

0. Healthy state
1. Minor symptoms and capable of running
2. Able to walk 10m or more without assistance but unable to run
3. Able to walk 10m across an open space with help
4. Bedridden or chairbound
5. Requiring assisted ventilation for at least part of the day
6. Dead

2.2 Statistical Analysis

Other acute monophasic neurological diseases are comparable to GBS in its disease course, heterogeneity of the patient population, and the ordinality of the commonly used outcome measures. Examples are traumatic brain injury (TBI), stroke, subarachnoid haemorrhage (SAH), and meningitis. Different approaches exist to analyse the treatment effect for RCTs in such

diseases. Binary logistic regression at a single time point is a reference analysis strategy, while a variety of more advanced techniques may better exploit properties of the data (table 2.3). More techniques exist to improve efficiency, such as generalized estimation equations or repeated measures ANOVA for exploiting repeated assessments. The following sections describe the techniques in more detail.

As an example, we present two previous studies in neurology. The first one being the DECRA trial in TBI, studying the effects of decompressive craniectomy ([12]) and the second one being the MR CLEAN trial, studying the effects of intraarterial treatment after stroke [2]. Both studies were published in a top medical journal (the New England Journal of Medicine). The outcome measures for these studies are the Glasgow Outcome Scale Extended (GOSE) and the modified Rankin Scale (mRS), respectively. Distribution of the outcome measure is presented in raw numbers, and in a so-called Grotta chart (table 2.1, table 2.2, Appendix B, page 27 (figure 5.1)).

Table 2.1: Distribution on the Glasgow Outcome Scale Extended (GOSE) at 6 months in the DECRA trial for TBI [12]

	GOSE						
	1	**2**	**3**	**4**	**5**	**6**	**7/8**
Regular Medical Care	92	4	27	15	19	18	13
Craniectomy	54	17	44	31	20	27	8

Table 2.2: Distribution on the modified Modified Rankin Scale (mRS) at 90 days in the MR CLEAN trial for stroke [2].

	mRS					
	1	**2**	**3**	**4**	**5**	**6**
Usual care	16	35	44	81	32	59
Intraarterial treatment	27	49	43	52	13	49

2.2.1 Binary logistic regression (conventional analysis)

In exploring novel techniques, we use the unadjusted binary logistic regression as a reference technique. Often, ordinal scales can be dichotomized into favorable versus unfavorable outcomes. Subsequently, the collapsed ordinal scale is analysed with binary logistic regression. This approach is most useful when the main interest is in the treatment effect for one specific state (e.g. mortality) and when shifts in other outcomes (e.g. from slight clinical symptoms to a healthy state) or individualized shift is of less interest. However, dichotomization is difficult in case of disagreement about the most relevant cut-off. Also, mortality may be relatively rare, inducing a shift in focus of the analysis towards differences in functional outcome among those who survive. Also, ignoring ordinality results in a loss of statistical power [22, 29, 35].

In both studies discussed above, we can estimate a treatment effect for a specific dichotomy. For example, for the DECRA trial, we can dichotomize the GOSE scale as an unfavorable outcome (GOSE < 2, death) versus a favorable outcome (GOSE ≥ 2). We find a significant unadjusted treatment effect for mortality (OR, 2.61, 95%, 1.71 to 4.00). For the MR CLEAN trial, we can dichotomize the mRS scale as an unfavorable outcome (mRS > 5, death) versus a favorable outcome (mRS ≤ 5). We find a non-significant unadjusted treatment effect for mortality (OR, 1.07, 95% CI, 0.69 to 1.63). A chi-squared test could be performed for such a dichotomization. Different dichotomizations lead to different estimates of the treatment effect (figure 2.1). In the MR CLEAN trial, ORs per cut-off are more or less similar. In the DECRA trial, surgery has a positive effect on survival (27% in the surgical group vs 49% in the medical group, OR 2.61 for GOSE > 1). For higher cut-offs at the GOSE we note a consistent trend towards a lower OR, with a detrimental effect for being alive with good recovery (4% in the surgical group vs 10% in the medical group, OR 0.65 for GOSE > 6, combining the two small good recovery categories).

Table 2.3: Overview of techniques to optimize statistical analysis as compared to unadjusted binary logistic regression analysis at a single time point. The overview presents techniques that are explored in this study and is non-exhaustive.

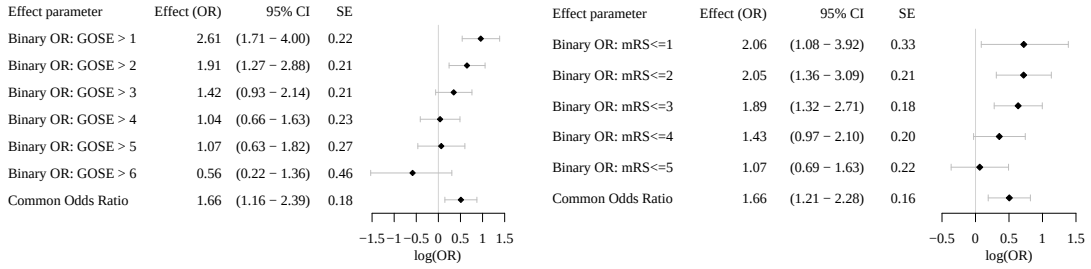
Options to optimize analysis	**How**	**Assumption**	**Test for assumption**	**Interpretation under satisfaction**	**Interpretation under violation**
Exploit heterogeneity (correct for baseline imbalance)	Covariate adjustment (covariates define the patient profiles)	Common effect irrespective of covariate values	LRT for significance of interaction terms, subgroup analyses	Interpret as a conditional, covariate specific treatment effect.	Interpret as the typical effect across covariate profiles. Provide effects per covariate profile for transparency.
Exploit ordinality	Proportional odds logistic regression	Proportional odds (common effect at each possible cut-offs)	Brant test, visual inspection of logits for each intercept	Interpret common OR as a summary measure, equal to the binary OR for each of the potential cut-offs.	Interpret common OR as a summary measure of shift in outcome distribution, do not interpret as binary OR for the individual cut-offs. Provide binary ORs for transparency.
Exploit repeated assessments	Longitudinal dichotomous or proportional odds logistic regression (generalized linear mixed models)	Common effect at all time points if no interaction effect between time point and treatment is specified	Test for significance of interaction terms of treatment indicator by time and other covariates by time	Interpret aggregate common OR as a summary measure, equal to the ORs for each of the time points.	Interpret aggregate common OR as a summary measure of shift in outcome distribution over time, do not interpret as common OR for the individual time points. Provide time-specific ORs for transparency.

2.2.2 Covariate adjustment

One way to increase the statistical power of a RCT is to adjust for baseline characteristics. Covariate adjusted analyses take the heterogeneity of patients into account and provide the clinically most relevant profile-specific (i.e. individualized) treatment effect estimates [9, 24, 27]. Also, covariate adjustment corrects for imbalance in the included baseline characteristics between treatment arms [10].

2.2.3 Proportional odds logistic regression

Another way to increase statistical power is by exploiting ordinality instead of dichotomizing the ordinal scale [1, 29, 33]. The proportional odds model [18] respects the ordinal nature of outcome measures. The model is stated as follows, for an ordinal response variable y consisting of levels 0, 1, 2, ..., k:



(a) Glasgow Outcome Scale Extended at 6 months **(b)** modified Rankin Scale at 90 days

Figure 2.1: Results of re-analysis of (a) the DECRA trial [12] and (b) the MR CLEAN trial [2]. Forest plots show the ORs and confidence intervals obtained from binary logistic regression for each possible cut-off for the ordinal scale and the common OR obtained from proportional odds logistic regression.

$$P(Y \geq c|X) = \frac{1}{1 + \exp(-1(\beta_{0c} + X\beta))} \tag{2.1}$$

where $c = 1, 2, \ldots, k$. There are k intercepts (β_{0c}s). The proportional odds model is parsimonious in its use of a single vector of regression coefficients β connecting the probabilities for varying cut-off values c. A common OR is conceptually obtained from a maximum likelihood estimation of the best fit for each possible cut-off for dichotomizing the ordinal scale [20]. The common OR can be viewed as an overall shift in the outcome across the complete ordinal scale. For this reason, the proportional odds model is also known as a "shift analysis" [30]. The shift idea is similar to the use of U statistics such as in the Mann Whitney test. Interpretation of the proportional odds model (i.e. with one treatment effect β) comes at the cost of the proportional odds assumption. It is assumed that the OR for a better versus worse outcome is equal for each dichotomization. Or, put differently, the regression coefficients β are independent of c. For this reason, exploiting ordinality is most useful when the treatment effect is expected to be in the same direction over the full scale and when all transitions are equally important.

The proportional odds assumption calls for caution in the interpretation of the treatment effect. In case of violation of the proportional odds assumption, the common OR cannot be interpretated as if it applies equally for each cut-off. However, the common OR can still be interpreted as the average shift over the ordinal scale caused by the treatment. Many consider it a useful summary measure [20, 29, 31]. Taking the studies in neurology as an example again, the common OR can be seen as a pooled estimate of the estimated binary ORs for each cut-off (figure 2.1). This is the case for both the DECRA and MR CLEAN trial. Even if the binary ORs are too much separated for the proportional odds assumption to hold, such as in the DECRA trial, where the goodness-of-fit test rejected the proportional odds assumption ($\chi^2 = 22.86$, 5 df, $p < 0.001$), the common OR provides a summary of the shift in distribution.

For testing and estimation of the treatment effect, violation of the proportional odds assumption does not form a problem. This is because the OR obtained from an unadjusted proportional odds model for comparing two groups is equivalent to the concordance probability c of a Wilcoxon-Mann Whitney U two-sample test, regardless of the satisfaction of the proportional odds assumption ([6]; [7]).

2.2.4 Longitudinal proportional odds logistic regression

A longitudinal proportional odds model exploits both the ordinality of the outcome measure as the repeated assessments during the follow-up. As such, it allows for evaluation of the treatment effect in a broad range of disease severity and over the entire clinically relevant time course. It also allows to study disease progression and the effect of the treatment on this progression. These properties are appealing from a conceptual point of view. From a statistical point of view, this method is appealing as well, as integrating a larger amount of data should increase power. Since repeated measurements within patients are expected to be correlated, the fundamental assumption of uncorrelated errors in regular linear regression is violated. A longitudinal analysis requires the use of a mixed model. Within a mixed model, the treatment effect per time point is estimated conditionally on the random effect per patient. The treatment effect that corrects for the correlation between repeated measurements within a patient is found by averaging out the conditional estimates. The longitudinal proportional odds logistic regression model is stated as follows, for an ordinal response variable Y consisting of levels 0, 1, 2, ..., k for individual i at time point j:

$$P(Y_{ij} \geq c|X) = \frac{1}{1 + \exp\big(- 1(\beta_{0c} + \beta_1 treat_i + \beta_2 time_{ij} + b_{i0c} + b_{i1} time_{ij})\big)}, \qquad (2.2)$$

where $c = 1, 2, \ldots, k$. There are k intercepts (β_{0c}s). β_1 and β_2 represent the fixed regression parameters, $treat_i$ and $time_{ij}$ are the treatment and time point indicators, respectively, b_{i0c} and b_{i1} are the random effects term for random intercepts and random slopes, respectively. Estimation of parameters can be done by maximum likelihood estimation of a generalized linear mixed model, using a generalized estimation equation, or with use of a Bayesian framework. In this study, a Bayesian framework will be used to enable more accurate and interpretable estimations.

Chapter 3

Methods

3.1 The "IVIG vs IVIG-MP" trial

We re-analysed data from a previous RCT in GBS [34]. This trial investigated the effectiveness of methylprednisolone (MP) when added to the standard treatment with intravenous immunoglobulin (IVIg). The data consists of 221 complete cases (i.e. observations without missing values), of which 111 cases were assigned to the control group (IVIg) and 110 cases were assigned to the treatment group (IVIg+MP). The primary outcome in the original trial was improvement by one or more grades on the GBS disability score (GBS DS) after 4 weeks. Data was collected at 12 time points between six days and 26 weeks after randomisation. In this study, we combined both the healthy state (0) and minor symptoms (1), as well as requiring ventilation for at least a part of the day (5) and death (6) because of small numbers in the extreme categories. We reversed the GBS DS, such that a higher score stands for a healthier state. In this way, an OR larger than one corresponds to a positive treatment effect. Detailed information on the original trial, its data collection, and patient characteristics has been described previously [34].

We estimated the treatment effect using four different approaches. These approaches were either exploiting the ordinality, the repeated assessments, both the ordinality and the repeated assessments or none of these properties. Each method was performed with and without covariate adjustment. For this adjustment, we used two pre-specified, clinically important prognostic factors: the Medical Research Council (MRC) sum score at baseline and days since onset of GBS symptoms to randomisation [13, 36].

3.1.1 Approach I: Binary logistic regression

As a reference, we performed dichotomous analysis for each cut-off in GBS DS at each time point separately. A treatment effect binary OR for each of the four possible cut-offs at each of the 12 time points is obtained for a total of 48 treatment effect estimates using the lrm() function from the *rms* package in R.

3.1.2 Approach II: Mixed effects logistic regression

To exploit the repeated assessments, we performed longitudinal dichotomous analysis for each possible cut-off in GBS DS across time points of the follow-up. An OR across time points for each

cut of the four possible cut-offs is obtained using the glmer() function from the *lme4* package in R. The steps we took for model building are:

1. We specified an elaborate fixed effects part, including interaction terms with the treatment indicator.
2. We decided on the number of quadrature points.
3. We decided on the random effects that are included (random intercepts and slopes or random intercepts only).
4. Eventually, we simplified the fixed effects part.
5. The final model that was suggested based on statistical criteria was verified with experts in the neurological field to ensure practical usefulness of the model.

Each of the final models contains a term for treatment group and week of assessment (treated as a continuous variable) to model the distributions of the ordinal outcome measure in both treatment arms over the 6 days to 26 weeks follow-up (Appendix C, page 28). The covariate adjusted models contained a term for MRC sum score at baseline and days since onset of weakness as well. Random terms were included, so that the intercepts and the relationship between time and GBS DS could be different for each subject.

3.1.3 Approach III: Proportional odds logistic regression

To exploit ordinality, we used a proportional odds model to model the overall shift in GBS DS at each time point separately. A common OR for treatment effect at each of the 12 time points was obtained with the lrm() function from the *rms* package in R.

3.1.4 Approach IV: Longitudinal proportional odds logistic regression

To exploit both the ordinality and the repeated assessments, we performed longitudinal ordinal analysis of the overall shift in GBS DS across time points during the follow-up. An aggregate common OR is obtained: a single summary measure on overall shift across the complete ordinal scale across time points during the follow-up. The models are fitted within a Bayesian framework, using the function blrm() from the package *brms* in R. This functon uses a STAN implementation for ordinal logistic regression.

As the function blrm() does not allow for specification of correlation structures and variance covariance structures, we only decide on the fixed effects (main effects and interactions) and the random effects (random intercepts and random slopes). We compared three approaches for model building, because each approach makes different assumptions and none of these approaches is unquestionably right. The final model is the model that is optimal according to the majority of these approaches (Appendix D, page 31). In the first approach, we assume that the best fit for a mixed binary logistic regression is the best fit for a mixed ordinal logistic regression outcome as well. In the second approach, we simplify the analysis by assuming the ordinal outcome measure to be continuous and follow steps for model building of a linear mixed effects model rather than a generalized linear mixed effects model. In the third approach, various models are compared in their goodness of fit to the data, using the compareBmods() function in R which compares Bayesian model fits based on leave-one-out-cross-validation [19].

The final model to model the distributions of the ordinal outcome measure in both treatment arms over the six days to 26 weeks follow-up is stated as follows, for an ordinal response variable y consisting of levels 0, 1, 2, ..., k for individual i at time point j:

$$\mathrm{P}(\mathrm{Y}_{ij} \geq c|X) = \frac{1}{1+\exp\left(-1(\beta_{0c}+\beta_1 treat_i+\beta_2 time_{ij}+\beta_3 mrcss_i+\beta_4 weakonset_i+b_{i0}+b_{i1}time_{ij})\right)} \quad (3.1)$$

where $c = 1, 2, \ldots, k$. There are k intercepts (β_{0c}s). We assume an equal spacing between consecutive intercepts across subjects. β_1, β_2, β_3 and β_4 represent the fixed regression parameters, $treat_i$ and $time_{ij}$ the treatment and time point indicators, respectively. $time_{ij}$ is treated as a continuous variable, with square root transformation. The terms β_3 and β_4 were equal to zero for the unadjusted case. b_{i0} and b_{i1} are the random effect terms.

To execute our model in STAN, different specifications need to be done. We follow Lambert's recommendations on executing a STAN program in R [14]. For the prior distribution of the intercepts, we use a Dirichlet distribution for the cell probabilities. We run iterations on four parallel chains. Each chain samples 5000 posterior samples. We use random initial values and use half of the iterations as burn in. Basic STAN diagnostics and trace plots were inspected to assure convergence and symmetry. Mean βs were interpreted as point estimates of treatment effect.

3.2 Simulation study

We conducted a simulation study based on the "IVIG vs IVIG-MP" trial [34]. We follow systematic steps for planning and reporting of the simulation study (algorithm 1, [21]).

Aims
The aim of the simulation study is to evaluate the impact of ordinal analysis and covariate adjustment on treatment effect estimates, standard error, statistical power, and potential reduction in sample size. We calculated the potential reduction in sample size, because it is a practical expression of the gain in efficiency for each of the models that facilitates comparison of the gains in power under different conditions. The formula that is used is described as a performance measure below (equation 3.4).

Data-generating mechanisms
We draw random samples from the X matrix with replacement, to assure independence between simulated draws. We used simulation based on resampling instead of simulating from a specific parametric model or using a closed form expression, because resampling preserves the empirical distributions of covariates and their correlations. This will yield datasets that are close to reality. Furthermore, closed form expressions to approximate power assume that the model is exactly consistent with the data generating mechanism, which cannot be true.

The samples are assigned with an equal allocation ratio to either the control or treatment arm. GBS DS scores for each patient at week 4 are generated using the coefficients obtained from a fit to the original dataset. One treatment effect β_{treat} or two treatment effects $\beta_{treat,01vs23456}$ and $\beta_{treat,remainingcut-offs}$ were simulated in order to obtain a different distribution between treatment arms. Simulations were done for four different conditions, as described below.

- *Condition A: Reference*
  Data is generated using βs for two strong prognostic factors (MRC sum score and days since onset of weakness) estimated from a fit to the GBS trial. We assume proportional

odds.

- *Condition B: Weaker prognostic effects*
  Data is generated using βs for two weaker prognostic factors (age and preceding diarrhoea) estimated from a fit to the GBS trial. We assume proportional odds.

- *Condition C: Qualitative non-proportionality*
  Data is generated using βs for two strong prognostic factors (MRC sum score and days since onset of weakness) estimated from a fit to the GBS trial. Qualitative non-proportionality was introduced for the treatment OR. Instead of using a single treatment effect that applies to the whole set of dichotomies, we specify two treatment effects, that are reversed in direction. Using a grid search, we identified a combination of a treatment effect for the two worst outcome categories versus better outcomes $\beta_{treat,01vs23456}$ and a treatment effect for each of the remaining cut-offs $\beta_{treat,remainingcut-offs}$ in which the treatment had a negative effect on the two worst outcomes and a positive effect on the remaining categories, providing the overall β_{treat} is still equal to the overall β_{treat}.

- *Condition D: Quantitative non-proportionality*
  Data is generated using βs for two strong prognostic factors (MRC sum score and days since onset of weakness) estimated from a fit to the GBS trial. Quantitative non-proportionality was introduced for the treatment OR. We specify two treatment effects, that are different but in the same direction. Again, a grid search was done to identify a combination of a treatment effect for the two worst outcome categories versus better outcomes $\beta_{treat,01vs23456}$ and a treatment effect for each of the remaining cut-offs $\beta_{treat,remainingcut-offs}$ in which the treatment had a neutral effect on the two worst outcomes and a positive effect on the remaining categories, providing the overall β_{treat} is still equal to the overall β_{treat}. In this case $\beta_{treat,01vs23456}$ was restricted to be one.

Estimands
Our estimand is the treatment effect OR e^{β}, the standard error of the treatment effect estimation SE_{β}, the rejection of the null, and the z value defined as $z = \frac{\beta}{SE}$.

Algorithm 1 An algorithm with caption

1: Estimate parameters (alphas, betas) from a training dataset (previous GBS trial)
2: Set a seed for reproducability
3: **for** $M \in \{1, ..., 10000\}$ **do** ▷ Generate data
4: From the sample $\{X_1, \ldots, X_n\}$, draw a sample size n at random with replacement, say, $X_1^*, \ldots, X_n^*$
5: Let $treat_i \in (0, 1)$ be an indicator denoting assignment to treatment. Assign samples with an equal allocation ratio to simulate a balanced treatment design.
6: Calculate for each subject its predicted marginal probabilities in terms of their baseline covariates, using: $\text{logit}[Pr(Y_i \leq c|x_i)] = \beta_{0c} + \beta_{treat} treat_i + \beta_{MRCss} MRCSS_i + \beta_{weakonset} weakonset_i$ *
7: Simulate an outcome for each subject, utilizing these estimated probabilities in a multinomial normal distribution
8: **for** each method **do** ▷ Fit models
9: Regress $GBSDS_i$ on the treatment indicator $treat_i$ (and any covariates) using the data generated in step 5-7
10: Store the point estimate of treatment effect ($\hat{\beta}$), its standard error ($\widehat{SE}(\hat{\beta})$), the z value ($\hat{\beta}/\widehat{SE}$) and indication of hypothesis rejection (0/1) for the test $\beta_{treat} = 0$.
11: **end for**
12: **end for**
13: For each method, compute power (the proportion of p-values smaller than 0.05), $\bar{Z}$, $\bar{\beta}$, $\overline{SE}$, the reduction in sample size as compared to unadjusted dichotomous analysis using equation (3.4) and compute the Monte Carlo SE using equation (3.2).
14: Repeat steps 3 through 13 for each simulation condition A - D.

Methods
Each simulated dataset is analysed in four ways:

1. Unadjusted dichotomous analysis for a commonly used primary outcome in GBS: favorable outcome (GBS DS 0-2) versus unfavorable outcome (GBS DS 3-6) at week 4.

2. Adjusted dichotomous analysis for a commonly used primary outcome in GBS: favorable outcome (GBS DS 0-2) versus unfavorable outcome (GBS DS 3-6) at week 4.

3. Unadjusted ordinal analysis of overall shift in GBS DS at week 4.

4. Adjusted ordinal analysis of overall shift in GBS DS at week 4.

Again, adjustment is done for the two prognostic factors: the MRC sum score at baseline and days since onset of GBS symptoms to randomisation. Monte Carlo simulations were performed for conditions A through D described above. For each of these four conditions, we varied the total sample size n and the (overall) OR $e^{\beta_{treat}}$ relating the treatment to the outcome. We allowed n to take on the values of 120 and 200 and overall $e^{\beta_{treat}}$ to take on the values of 1, 1.5, 2.1 and 2.7. We thus simulated under 4*2*4 = 32 different scenarios. The overall OR equal to one served as a type I error check. Simulations were performed under a full factorial design with each 10000 replications.

*For simulation of proportional odds violated datasets (condition C and D), we used seperate formulas. For the cut-off 01 vs 23456, we used $\text{logit}[Pr(Y_i \leq c|x_i)] = \beta_{0c} + \beta_{treat,01vs23456} treat_i + \beta_{MRCss} MRCSS_i + \beta_{weakonset} weakonset_i$. For the remaining cut-offs we used the formula as stated in the algorithm.

Performance Measures
We assessed the type I error when there was no treatment effect (OR = 1) and the power (and potential reduction in sample size) when there truly was a treatment effect (OR >1) in each of the simulation scenarios for each approach. The type I error and power were calculated as the rate of rejection for a statistically significant estimate according to the Wald statistic (p-value lower than 0.05). We computed the Monte Carlo standard error for the power estimates to express the uncertainty of our simulations as follows:

$$\sqrt{\frac{\widehat{\text{power}} \text{ x } (1 - \widehat{\text{power}})}{n_{sim}}} \tag{3.2}$$

For verification, we performed the same simulations relying on the Likelihood Ratio Test (LRT) statistic p-value. This was obtained by estimating two models for each of the approaches under study: one model with treatment effect included (the complex model), one model with treatment excluded (the nested model). Consequently, the LR test compares the log likelihoods of these two models. The LR test statistic is calculated as follows:

$$LR = -2\ln\left(\frac{L(m_{nested})}{L(m_{complex})}\right) = 2\big(\text{loglik}(m_{complex}) - \text{loglik}(m_{nested})\big) \tag{3.3}$$

Where $L(m_*)$ denotes the likelihood of the respective model and $\text{loglik}(m_*)$ the natural logarithm of the model's likelihood. The LR test statistic is chi-squared distributed with one degree of freedom.

The reduction in sample size (RSS) can be attractive to express the gain in power. The RSS as compared to the reference technique (unadjusted dichotomous analysis) is defined as:

$$RSS = 100 - 100 * \left(\frac{z_r}{z_a}\right)^2 \tag{3.4}$$

where the mean standardized unadjusted z value $z_r = \sum_{i=1}^{nsim}\left(\frac{\beta_{r_i}}{SE_{r_i}}\right)$ and the mean standardized advanced (adjusted and/or ordinal) z value is $z_a = \sum_{i=1}^{nsim}\left(\frac{\beta_{a_i}}{SE_{a_i}}\right)$ [23].

Software
All analysis and simulations were performed using R 4.0.5, using the *rms* package for the proportional odds analysis, the *brms* package for longitudinal proportional odds analysis, the *lme4* package for generalized linear mixed model (longitudinal dichotomous analysis) [8, 25]. The code is provided as a Supplement. The input seed for random generation is "1". We use a seed to assure reproducibility.

Chapter 4

Results

The results section is divided in two parts. First, we present results from the re-analysis of the RCT in GBS. Second, we compare the characteristics of unadjusted dichotomous analysis to adjusted dichotomous analysis, unadjusted proportional odds, and adjusted proportional odds analysis by simulation.

4.1 Re-analysis "IVIG vs IVIG-MP" trial

We analysed data from 221 patients in the IVIg vs IVIg+MP trial. The distribution of GBS DS at 4 weeks after randomisation is presented in table 4.1. A plot of GBS DS scores at each time point per subject is provided in Appendix E (page 32).

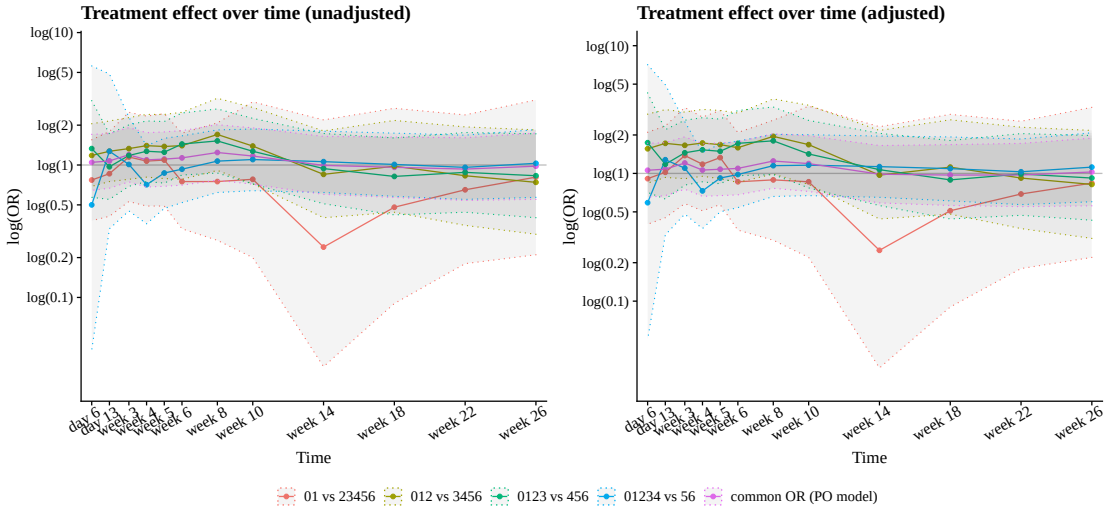
4.1.1 Covariate adjustment

With covariate adjustment, the estimated treatment effect was larger than without covariate adjustment (Appendix F, page 33 and Appendix G, page 34). This was the case for all four approaches we compared. For longitudinal approaches, the SE did slightly decrease or remained equal after adjustment. For single time point approaches, the SE did slightly increase or remained

Table 4.1: Distribution of baseline covariates and GBS DS outcome at 4 weeks after randomization to IVIg or IVIg+MP in the GBS trial.

	Total (n = 221)	**Control (IVIg) (n = 111)**	**Treatment (IVIg + MP) (n = 110)**
MRC sum score (Median, Interquartile Range $25^{th} - 75^{th}$ Percentile)	44 (36 - 48)	46 (38 - 49)	44 (34 - 48)
Days from onset of weakness to randomisation (Median, Interquartile Range $25^{th} - 75^{th}$ Percentile)	5 (3 - 8)	6 (1 - 7)	4 (1 - 8)
Preceding diarrhea	60 (27%)	30 (27%)	30 (27%)
Age (Median, Interquartile Range $25^{th} - 75^{th}$ Percentile)	55 (35 - 67)	52 (35 - 67)	57 (34 - 68)
GBS Disability score after 4 weeks			
0 – healthy state	5 (2%)	0 (0%)	5 (5%)
1 – minor symptoms and capable of running	37 (17%)	24 (22%)	13 (12%)
2 – able to walk 10m or more without assistance but unable to run	74 (34%)	31 (28%)	43 (39%)
3 – able to walk 10m across an open space with help	22 (10%)	10 (9%)	12 (11%)
4 – bedridden or chair bound	54 (24%)	31 (28%)	23 (21%)
5 – requiring assisted ventilation for at least a part of the day	26 (12%)	14 (13%)	12 (11%)
6 – dead	3 (1%)	1 (1%)	2 (2%)

Figure 4.1: Binary OR and common OR per time point (in days or weeks after randomisation) obtained from a binary logistic regression or proportional odds logistic regression, respectively. Binary logistic regressions were performed for four different possible cut-offs of the GBS DS. Right panel shows analyses adjusted for baseline MRC sum score and number of days since onset of weakness. Left panel shows unadjusted analyses.



equal after adjustment. For example, in the adjusted proportional odds analysis at week 4 the common OR was 1.19 (SE = 0.24, 95% 0.73 – 1.92), while in the unadjusted proportional odds at week 4 analysis the common OR was 1.09 (SE = 0.24, 95% 0.68 – 1.75).

4.1.2 Exploiting ordinality

The common OR obtained from the PO model (OR=1.19 at week 4, adjusted) can be interpreted as a pooled estimate of the binary ORs obtained from each dichotomy at that time point. Indeed, the estimate was in the range of the binary ORs for the possible cut-offs at each time point (figure 4.1). Exploiting ordinality decreased the SEs of the estimates. With unadjusted ordinal analysis of the outcome measure, the SE was 0.24 at 4 weeks, which was smaller than the SE with unadjusted binary logistic regression for each possible dichotomy (SE typically around 0.32). Also in the longitudinal ordinal models, the SEs were lower (0.31 and 0.29 for unadjusted and adjusted analyses, respectively) than the longitudinal binary models (typically around 1 for both unadjusted and adjusted analyses).

4.1.3 Exploiting repeated assessments

The binary ORs and common ORs varied over time (figure 4.1). Integrating all available data from the six days to 26 weeks follow-up resulted in an aggregated common OR of 1.05 and 1.25 for the unadjusted and adjusted analyses, respectively. This aggregate common OR can be interpreted as an overall shift in GBS DS over time, or as a pooled estimate of the common ORs at each time point in the follow-up. The SE of the estimated treatment effect from the longitudinal proportional odds model was smaller than the SE for the longitudinal binary logistic regressions

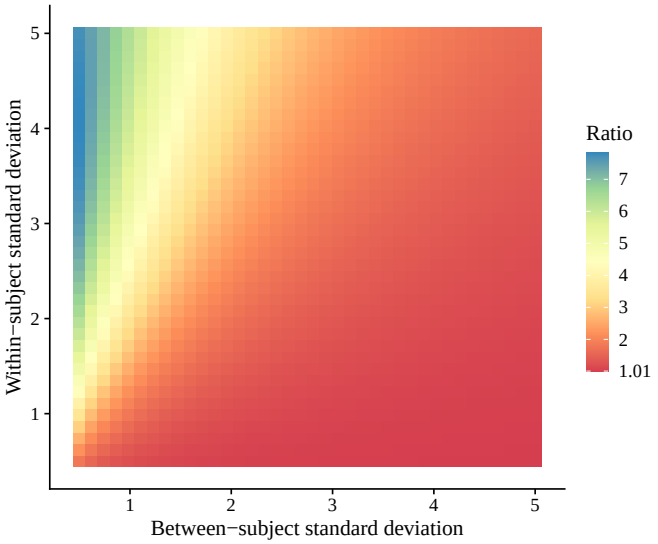
for each cut off. Thus, for the longitudinal approaches, exploiting ordinality increased accuracy. Surprisingly, the SE of the estimated treatment effect from the longitudinal proportional odds model was either increased (in the unadjusted analyses) or similar (in the adjusted analyses) in comparison to the SEs of single time point ORs. For the ordinal analysis, taking into account the repeated assessments, does not increase accuracy.

4.2 Simulation of longitudinal continuous data

As the results of the longitudinal proportional odds model were against our expectations, we did additional analyses in order to understand under which circumstances a longitudinal approach does have benefits in comparison to a single time point approach. For this exploration, we studied the performance of a linear mixed effects model compared to a single time point linear model for continuous data. Simulating longitudinal continuous data is attractive as it readily allows to tune the within- and between-subject variance, while tuning the within-subject variance in longitudinal ordinal data is complicated by the non-linear link function.

In this simulation, we generated longitudinal continuous data for 200 patients on ten time points. Each patient had its own random intercept, drawn from a normal distribution with various between-subject variance terms. Each time point had its random time point effect, drawn from a normal distribution with various within-subject variance terms. A fixed time effect and treatment effect were specified as well. The estimand was the squared z value, defined as $\left(\frac{\beta}{SE}\right)^2$. Each simulated dataset was analysed with a linear mixed effects model integrating all data from ten time points and a linear model for data of one of the time points. The performance measure was the average squared z value, defined as $z = \sum_{i=1}^{nsim}\left(\frac{\beta_i}{SE_i}\right)$. z_{lme} is the average squared z value for the linear mixed model, z_{lm} is the average squared z value for the linear model. We expressed the gain in efficiency as the ratio of z values (z_{lme}/z_{lm}).

Figure 4.2: Results from simulations of longitudinal continuous data (10 time points), analysed with both a linear mixed model and a linear model. The larger the ratio of z values, the larger the gains in efficiency by using an longitudinal approach.



The gain in efficiency obtained by longitudinal analysis appears to be related to the ratio of between- and within-subject variance (figure 4.2). We found that linear mixed effects models were at least as efficient as linear models for single time points (a ratio ≥ 1). In case of a large between-subject variance and a low within-subject variance, the gains are only small. In contrast, a small between-subject variance in combination with a large within-subject variance is related to a large gain in efficiency.

4.3 Simulation study

Tables 4.2 to 4.5 compare the power of binary logistic regression and proportional odds logistic regression, considering covariate adjustment for MRC sum score and number of days since onset of weakness. The power (or type I error rate), average OR, average SE, and reduction in sample size were calculated for the different conditions, overall ORs relating the treatment to the outcome measure, and sample sizes. The type I error was rather close to the nominal value of 5%.

With covariate adjustment, the estimated treatment effects are further from zero and the standard error is slightly increased. This effect was noted for each condition but was less pronounced in condition B with weaker prognostic factors. For both the binary logistic model and the proportional odds model, adjusting for baseline MRC sum score and number of days since onset of weakness universally raised power by about 1.1 - 7.2%. In the condition with adjustment for weaker prognostic factors, covariate adjustment raised power by about 1.1 - 3.0% points. Covariate adjustment in a dichotomous analysis led to potential reductions in sample size between 4.9 and 11%. These reductions were a little smaller (3.1 - 4.3%) in the case of weak prognostic factors. Also in ordinal analysis, covariate adjustment leads to potential reductions in sample size.

With ordinal analysis instead of collapsing the ordinal scale into a dichotomy, the treatment effects were generally closer to zero and the SEs were smaller. This effect was even stronger under non-proportionality. For both adjusted and unadjusted analyses, exploiting ordinality raised power by about 1.9 - 13% points across all conditions with strong prognostic factors. Given that the overall treatment effect was maintained, exploiting ordinality under qualitative or quantitative violation of the proportional odds assumption still led to some gain in power. As expected, the gain was less than without violation of the PO assumption. In case of violation of the proportional odds assumption, either qualitatively or quantitatively, the gains in power obtained by exploiting ordinality were about 0.6 - 13% points. Sample size reductions by performing adjusted ordinal analysis ranged from about 22 to 26% points (as compared to unadjusted dichotomous analysis), and these reductions were larger when adjustment for baseline characteristics was done as well. In case of non-proportionality, the sample size reductions were less systematic and varied depending on overall OR.

The Monte Carlo standard errors for the simulations were all lower than 0.005, indicating a low simulation uncertainty. All gains in efficiency were obtained without considerably increasing the type I error. Results did not change when the likelihood ratio test was used for significance testing instead of the computationally faster Wald test.

Table 4.2: Condition A. Results of simulated balanced RCTs estimating treatment effect. Adjusted analyses corrected for baseline MRC sum score and number of days between onset of GBS symptoms and randomisation. All simulations are based on 10000 iterations. Performance of different models were evaluated using the average treatment effect (OR, coefficient), the average standard error (SE) (averaging over iterations), the rejection rate (percentage of the 10000 simulated datasets in which the Wald test statistic for the estimated treatment effect was significant at the 0.05 level, interpreted as the type I error or power depending on the true OR) and the reduction in sample size as compared to unadjusted dichotomous analysis.

	Treatment effect (true OR)										
	OR = 1		OR = 1.5			OR = 2.1			OR = 2.7		
Model and sample size	Coefficient (SE)	Type I error (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)
n = 120 (60/arm)											
Unadjusted dichotomous	1.0 (0.37)	5.2	1.5 (0.37)	17	reference	2.1 (0.38)	44	reference	2.6 (0.39)	64	reference
Adjusted dichotomous	1.1 (0.40)	5.0	1.7 (0.40)	18	11	2.4 (0.41)	48	9.6	3.1 (0.42)	68	9.0
Unadjusted ordinal	1.1 (0.33)	5.0	1.5 (0.33)	20	24	2.1 (0.34)	54	24	2.6 (0.34)	77	27
Adjusted ordinal	1.1 (0.34)	5.4	1.6 (0.34)	23	37	2.3 (0.34)	61	36	3.0 (0.35)	84	38
n = 200 (100/arm)											
Unadjusted dichotomous	1.0 (0.28)	5.7	1.5 (0.29)	26	reference	2.0 (0.29)	65	reference	2.6 (0.30)	86	reference
Adjusted dichotomous	1.1 (0.30)	5.2	1.6 (0.31)	27	10	2.3 (0.32)	68	10	2.9 (0.32)	89	8.7
Unadjusted ordinal	1.0 (0.25)	5.2	1.5 (0.26)	31	22	2.0 (0.26)	75	24	2.6 (0.26)	94	26
Adjusted ordinal	1.0 (0.26)	5.3	1.6 (0.26)	35	34	2.2 (0.26)	82	36	2.9 (0.27)	97	37

Table 4.3: Condition B. Results of simulated balanced RCTs estimating treatment effect. Adjusted analyses corrected for a set of covariates with a *weaker prognostic value* (age (years) and preceding diarrhoea (yes/no)). All simulations are based on 10000 iterations. Performance of different models were evaluated using the average treatment effect (OR, coefficient), the average standard error (SE) (averaging over iterations), the rejection rate (percentage of the 10000 simulated datasets in which the Wald test statistic for the estimated treatment effect was significant at the 0.05 level, interpreted as the type I error or power depending on the true OR) and the reduction in sample size as compared to unadjusted dichotomous analysis.

	Treatment effect (true OR)											
	OR = 1		OR = 1.5			OR = 2.1			OR = 2.7			
Model and sample size	Coefficient (SE)	Type I error (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)
n = 120 (60/arm)											
Unadjusted dichotomous	1.1 (0.37)	5.2	1.6 (0.37)	19	reference	2.2 (0.38)	47	reference	2.9 (0.40)	69	reference
Adjusted dichotomous	1.1 (0.38)	4.8	1.6 (0.39)	19	4.3	2.3 (0.40)	48	3.0	3.1 (0.41)	71	3.1
Unadjusted ordinal	1.1 (0.33)	5.0	1.6 (0.33)	22	23	2.2 (0.34)	58	25	2.8 (0.34)	82	26
Adjusted ordinal	1.2 (0.33)	5.2	1.6 (0.34)	23	28	2.3 (0.34)	60	29	2.9 (0.35)	84	30
n = 200 (100/arm)											
Unadjusted dichotomous	1.0 (0.28)	5.0	1.6 (0.29)	28	reference	2.2 (0.30)	69	reference	2.8 (0.31)	89	reference
Adjusted dichotomous	1. (0.29)	4.6	1.6 (0.30)	28	3.9	2.3 (0.31)	72	3.8	2.9 (0.32)	91	3.1
Unadjusted ordinal	1.0 (0.25)	4.8	1.5 (0.26)	34	23	2.1 (0.26)	80	23	2.7 (0.26)	96	26
Adjusted ordinal	1.0 (0.26)	4.7	1.6 (0.26)	36	28	2.2 (0.26)	83	28	2.9 (0.27)	97	30

Table 4.4: Condition C. Results of simulated balanced RCTs estimating treatment effect under *qualitative violation of the proportional odds assumption.* Adjusted analyses corrected for baseline MRC sum score and number of days between onset of GBS symptoms and randomisation. All simulations are based on 10000 iterations. Performance of different models were evaluated using the average treatment effect (OR, coefficient), the average standard error (SE) (averaging over iterations), the rejection rate (percentage of the 10000 simulated datasets in which the Wald test statistic for the estimated treatment effect was significant at the 0.05 level, interpreted as the type I error or power depending on the true OR) and the reduction in sample size as compared to unadjusted dichotomous analysis.

	Treatment effect (true OR)											
	Overall OR = 1 ORmortality = 0.80; ORnon-mortality = 1.05		Overall OR = 1.5 ORmortality = 0.70; ORnon-mortality = 1.75			Overall OR = 2.1 ORmortality = 0.70; ORnon-mortality = 2.60			Overall OR = 2.7 ORmortality = 0.95; ORnon-mortality = 3.15		
Model and sample size	Coefficient (SE)	Type I error (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)
n = 120 (60/arm)											
Unadjusted dichotomous	1.2 (0.50)	5.2	2.0 (0.46)	20	reference	2.9 (0.45)	56	reference	3.4 (0.44)	71	reference
Adjusted dichotomous	1.3 (0.52)	5.2	2.1 (0.49)	21	5.0	3.1 (0.47)	58	5.9	3.8 (0.47)	74	6.3
Unadjusted ordinal	1.1 (0.33)	5.0	1.6 (0.33)	22	1.4	2.2 (0.34)	59	10	2.7 (0.34)	79	19
Adjusted ordinal	1.1 (0.34)	5.0	1.6 (0.34)	24	9.16	2.4 (0.34)	63	20	3.0 (0.35)	83	28
n = 200 (100/arm)											
Unadjusted dichotomous	1.1 (0.38)	5.3	1.8 (0.35)	33	reference	2.6 (0.34)	76	reference	3.1 (0.34)	91	reference
Adjusted dichotomous	1.2 (0.39)	5.7	1.9 (0.37)	35	5.5	2.9 (0.36)	79	6.1	3.5 (0.36)	93	5.7
Unadjusted ordinal	1.0 (0.25)	5.4	1.5 (0.26)	34	1.9	2.1 (0.26)	78	9.3	2.6 (0.26)	94	18
Adjusted ordinal	1.0 (0.26)	5.7	1.6 (0.26)	37	10	2.2 (0.26)	83	19	2.9 (0.27)	97	27

Table 4.5: Condition D. Results of simulated balanced RCTs estimating treatment effect under *quantitative violation of the proportional odds assumption.* Adjusted analyses corrected for baseline MRC sum score and number of days between onset of GBS symptoms and randomisation. All simulations are based on 10000 iterations. Performance of different models were evaluated using the average treatment effect (OR, coefficient), the average standard error (SE) (averaging over iterations), the rejection rate (percentage of the 10000 simulated datasets in which the Wald test statistic for the estimated treatment effect was significant at the 0.05 level, interpreted as the type I error or power depending on the true OR) and the reduction in sample size as compared to unadjusted dichotomous analysis.

	Treatment effect (true OR)											
	Overall OR = 1 ORmortality = 1.0; ORnon-mortality = 1.1			Overall OR = 1.5 ORmortality = 1.0; ORnon-mortality = 1.6			Overall OR = 2.1 ORmortality = 1.0; ORnon-mortality = 2.4			Overall OR = 2.7 ORmortality = 1.0; ORnon-mortality = 3.1		
Model and sample size	Coefficient (SE)	Type I error (%)		Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)	Coefficient (SE)	Power (%)	Reduction in sample size (%)
n = 120 (60/arm)												
Unadjusted dichotomous	1.2 (0.49)	4.4		1.8 (0.47)	16	reference	2.6 (0.45)	45	reference	3.4 (0.44)	73	reference
Adjusted dichotomous	1.2 (0.52)	4.6		1.9 (0.49)	16	6.2	2.8 (0.48)	47	4.9	3.8 (0.47)	75	5.9
Unadjusted ordinal	1.1 (0.33)	4.9		1.5 (0.33)	20	28	2.1 (0.34)	55	24	2.7 (0.34)	79	19
Adjusted ordinal	1.1 (0.34)	5.1		1.6 (0.34)	23	37	2.3 (0.35)	60	32	3.0 (0.35)	83	28
n = 200 (100/arm)												
Unadjusted dichotomous	1.1 (0.38)	5.0		1.7 (0.36)	24	reference	2.4 (0.34)	67	reference	3.2 (0.34)	91	reference
Adjusted dichotomous	1.1 (0.39)	5.4		1.8 (0.37)	26	6.1	2.6 (0.36)	69	6.1	3.5 (0.36)	93	6.0
Unadjusted ordinal	1.1 (0.25)	5.6		1.5 (0.26)	31	25	2.1 (0.26)	75	21	2.7 (0.26)	95	19
Adjusted ordinal	1.1 (0.26)	5.5		1.6 (0.26)	35	35	2.2 (0.27)	80	31	2.9 (0.27)	97	28

Chapter 5

Discussion

Covariate adjustment and ordinal analysis (as compared to dichotomizing) increase the statistical power in randomized controlled trials, without inflation of type I error. Adjusting for baseline characteristics and exploiting ordinality allow smaller treatment effects to be detected. Or, put differently, the same treatment effects can be detected with equal power in smaller trials (reducing the sample size up to 38%).

The gains in power obtained by exploiting ordinality are larger (between 2 and 13% points) than the gains obtained by covariate adjustment (between 1 and 7% points). The combination of both covariate adjustment and respecting ordinality resulted in the largest efficiency as compared to the reference technique of unadjusted dichotomous analysis (around 18% points gain in power). The gains in efficiency are only slightly smaller under suboptimal conditions such as qualitative and quantitative violation of the proportional odds assumption and weak prognostic effects. Under both qualitative and quantitative violation of the proportional odds assumption, ordinal analysis still increases the power and reduces the required sample size as compared to dichotomous analysis. The gains are smaller, but non-proportionality does not hamper the gains in efficiency. The same applies to covariate adjustment for weak prognostic effects. The gains are smaller in comparison to adjustment for strong prognostic factors, but adjustment for any covariates is still more efficient in terms of statistical power and sample size requirements than no adjustment. However, it is practice to adjust for no more than a few covariates according to EMA guidance [3].

Covariate adjustment increases power due to the larger increase in effect size than the increase in SE [28]. It corrects for imbalance and results in a clinically relevant profile-specific treatment effect estimates instead of population-averaged treatment effect estimates. One should note that treatment effect estimates in adjusted non-linear models are non-collapsible [9]. In linear models, covariate adjustment affects the precision of the estimate only, while in nonlinear models, omitting covariates from the analyses lead to a change in the estimated treatment effect. Thus, the average conditional estimate is not equal to the marginal effect if heterogeneity is present.

Proportional odds analysis is recommended to increase efficiency. Dichotomization is suboptimal from a statistical point of view, because it results in a loss of power [1] and from a conceptual view, because the decision on the cut-off value is arbitrary. However, the proportional odds assumption asks for careful interpretation. The common OR that is obtained from the proportional odds analysis can always be interpreted as an overall shift in outcome distribution. It can only be interpreted as if it is equal to the OR for each dichotomy if the PO assumption holds.

A longitudinal proportional odds analysis is from a conceptual point of view particularly useful when the main interest is in the treatment effect over the complete follow-up rather than the treatment effect at one specific time point. For example, in GBS, patients' trajectories of recovery are extremely heterogeneous [11]. In such diseases, there is no clear best time point

to estimate treatment effect. A fixed time point analysis might suffer from loss of power if the time point is not chosen well [4]. Therefore, integrating information from multiple time points provides a more robust evaluation of the treatment effect and circumvents the decision at which specific time point outcomes need to be evaluated.

Surprisingly, no clear benefit in efficiency was found in the re-analysis of a previous trial in GBS. Results were against our expectations, as the standard error for treatment effect estimate was increased in the longitudinal approach. Simulations in continuous cases suggested that the gains of longitudinal approaches are related to the ratio of between- and within-subject variance. In cases of low within-subject variance, analysing patients at multiple time points seems to provide limited additional accuracy. When implementing the longitudinal proportional odds model, one should consider its computational burden and its exact interpretation. The OR obtained can be seen as an aggregate common OR, summarized over each cut-off and each time point.

A previous methodological study in GBS also advised to use a proportional odds model and covariate adjustment [35]. However, it was unknown how much power gains and potential sample size reductions are obtained in GBS through these approaches. Our findings are in agreement with and provide extra support for these recommendations. Both real patient data and simulated datasets demonstrated efficiency gains through covariate adjustment and ordinal analysis. Similar study designs for trial design in TBI found larger power gains [10]. This can for example be attributed to different covariate strengths. Also, although findings based on GBS are presumably relevant to other neurological diseases that also deal with heterogeneous populations, ordinal outcomes and monophasic trajectories, it is uncertain to what extent different diseases are similar to GBS. The same analyses might have different effects on different trials.

Two important notes regarding statistical power have to be made. First, the aim of using more advanced statistical analysis techniques should be to increase the efficiency of trials. Although the potential reduction in sample size was used as a measure to quantify gains in efficiency, the aim should not be to make trials smaller. The aim should be to optimally utilize resources and increase the chance of detecting treatment effects that are truly present.

Moreover, one should differentiate between clinical significance and statistical significance [26]. Exploiting ordinality increases power, or, in other words, it increases the chance to detect a positive effect when there indeed is an effect. Statistically significant results are more likely, but these should not be misinterpreted as clinically important results. Statistical significance quantifies the degree of compatibility of the data to the pattern predicted by the test hypothesis and the assumptions used in the test, while clinical significance is about its implications on clinical practice, such as the extent of change on patients' health, consumer acceptability and the ease of implementation and [5, 15].

This work has some limitations. First, our simulation study was based on resampling from a previous trial in GBS, which make the results relevant for at least the study at hand. Results are applicable to an infinite population with the exact characteristics of that dataset, but it is unknown how far results can be extrapolated to future trials or other diseases [21]. Also, one should take into account that mixed models and fixed time point analyses provide answers to different research questions. Performance of both approaches were presented side by side because it was our interest to compare logical extensions that could improve efficiency to conventional methods.

In conclusion, statistical analysis of RCTs should adjust for prognostic factors and exploit ordinality of the ordinal scale. This will increase power of trials and allow potential reductions in sample size, without inflation of type I error. Further research should explore under which conditions a longitudinal proportional odds model is more efficient than a regular proportional odds model.

Bibliography

[1] Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, *332*(7549), 1080.

[2] Berkhemer, O. A., Fransen, P. S., Beumer, D., Van Den Berg, L. A., Lingsma, H. F., Yoo, A. J., Schonewille, W. J., Vos, J. A., Nederkoorn, P. J., Wermer, M. J. et al. (2015). A randomized trial of intraarterial treatment for acute ischemic stroke. *n Engl J Med*, *372*, 11–20.

[3] Committee for Medicinal Products for Human Use. (2015). https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf

[4] Dodd, L. E., Follmann, D., Wang, J., Koenig, F., Korn, L. L., Schoergenhofer, C., Proschan, M., Hunsberger, S., Bonnett, T., Makowski, M. et al. (2020). Endpoints for randomized controlled clinical trials for covid-19 treatments. *Clinical Trials*, *17*(5), 472–482.

[5] Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337–350.

[6] Harrell, F. (2001). *Regression modeling strategies*. Springer.

[7] Harrell, F. (2021a). *If you like the wilcoxon test you must like the proportional odds model*. Retrieved November 10, 2021, from https://www.fharrell.com/post/wpo/

[8] Harrell, F. (2021b). *Rms: Regression modeling strategies*. Retrieved November 10, 2021, from https://github.com/harrelfe/rms

[9] Hauck, W. W., Anderson, S., & Marcus, S. M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled clinical trials*, *19*(3), 249–256.

[10] Hernández, A. V., Steyerberg, E. W., & Habbema, J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of clinical epidemiology*, *57*(5), 454–460.

[11] Hughes, R. A., Swan, A. V., Raphaël, J.-C., Annane, D., van Koningsveld, R., & van Doorn, P. A. (2007). Immunotherapy for guillain-barré syndrome: A systematic review. *Brain*, *130*(9), 2245–2257.

[12] Hutchinson, P. J., Kolias, A. G., Timofeev, I. S., Corteen, E. A., Czosnyka, M., Timothy, J., Anderson, I., Bulters, D. O., Belli, A., Eynon, C. A. et al. (2016). Trial of decompressive craniectomy for traumatic intracranial hypertension. *N Engl J Med*, *375*, 1119–1130.

[13] Kleyweg, R. P., Van Der Meché, F. G., & Schmitz, P. I. (1991). Interobserver agreement in the assessment of muscle strength and functional abilities in guillain-barré syndrome. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, *14*(11), 1103–1109.

[14] Lambert, B. (2018). *A student's guide to bayesian statistics*. Sage.

[15] LeFort, S. M. (1993). The statistical versus clinical significance debate. *Image: the journal of nursing scholarship*, *25*(1), 57–62.

[16] Leonhard, S. E., Mandarakas, M. R., Gondim, F. A., Bateman, K., Ferreira, M. L., Cornblath, D. R., van Doorn, P. A., Dourado, M. E., Hughes, R. A., Islam, B. et al. (2019). Diagnosis and management of guillain–barré syndrome in ten steps. *Nature Reviews Neurology*, *15*(11), 671–683.

[17] Maas, A. I., Steyerberg, E. W., Marmarou, A., McHugh, G. S., Lingsma, H. F., Butcher, I., Lu, J., Weir, J., Roozenbeek, B., & Murray, G. D. (2010). Impact recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics*, *7*(1), 127–134.

[18] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127.

[19] McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

[20] McHugh, G. S., Butcher, I., Steyerberg, E. W., Marmarou, A., Lu, J., Lingsma, H. F., Weir, J., Maas, A. I., & Murray, G. D. (2010). A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: Results from the impact project. *Clinical trials*, *7*(1), 44–57.

[21] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102.

[22] of Stroke Trials (OAST) Collaboration, O. A. (2007). Can we improve the statistical analysis of stroke trials? statistical reanalysis of functional outcomes in stroke trials. *Stroke*, *38*(6), 1911–1915.

[23] Piantadosi, S. (2017). *Clinical trials: A methodologic perspective*. John Wiley & Sons.

[24] Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practiceand problems. *Statistics in medicine*, *21*(19), 2917–2930.

[25] R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

[26] Ranganathan, P., Pramesh, C., & Buyse, M. (2015). Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspectives in clinical research*, *6*(3), 169.

[27] Roberts, C., & Torgerson, D. J. (1999). Baseline imbalance in randomised controlled trials. *Bmj*, *319*(7203), 185.

[28] Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, 227–240.

[29] Roozenbeek, B., Lingsma, H. F., Perel, P., Edwards, P., Roberts, I., Murray, G. D., Maas, A. I., & Steyerberg, E. W. (2011). The added value of ordinal analysis in clinical trials: An example in traumatic brain injury. *Critical Care*, *15*(3), 1–7.

[30] Saver, J. L. (2007). Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. *Stroke*, *38*(11), 3055–3062.

[31] Senn, S., & Julious, S. (2009). Measurement in clinical trials: A neglected issue for statisticians? *Statistics in medicine*, *28*(26), 3189–3209.

[32] Thompson, D. D., Lingsma, H. F., Whiteley, W. N., Murray, G. D., & Steyerberg, E. W. (2015). Covariate adjustment had similar benefits in small and large randomized controlled trials. *Journal of clinical epidemiology*, *68*(9), 1068–1075.

[33] Valenta, Z., Pitha, J., & Poledne, R. (2006). Proportional odds logistic regression—effective means of dealing with limited uncertainty in dichotomizing clinical outcomes. *Statistics in medicine*, *25*(24), 4227–4234.

[34] van Koningsveld, R., Schmitz, P., Van der Meche, F., Visser, L., Meulstee, J., Van Doorn, P., Group, D. G. S. et al. (2004). Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for guillain-barré syndrome: Randomised trial. *The Lancet*, *363*(9404), 192–196.

[35] van Leeuwen, N., Walgaard, C., van Doorn, P. A., Jacobs, B. C., Steyerberg, E. W., & Lingsma, H. F. (2019). Efficient design and analysis of randomized controlled trials in rare neurological diseases: An example in guillain-barre syndrome. *PloS one*, *14*(2), e0211404.

[36] Walgaard, C., Lingsma, H. F., Ruts, L., Drenthen, J., van Koningsveld, R., Garssen, M. J., van Doorn, P. A., Steyerberg, E. W., & Jacobs, B. C. (2010). Prediction of respiratory insufficiency in guillain-barre syndrome. *Annals of neurology*, *67*(6), 781–787.

Appendix

Appendix A

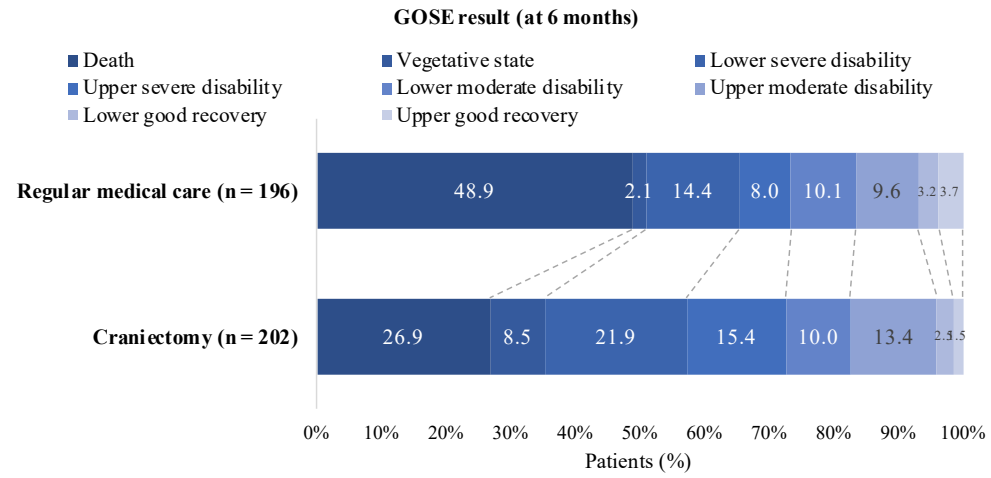
Abbreviations

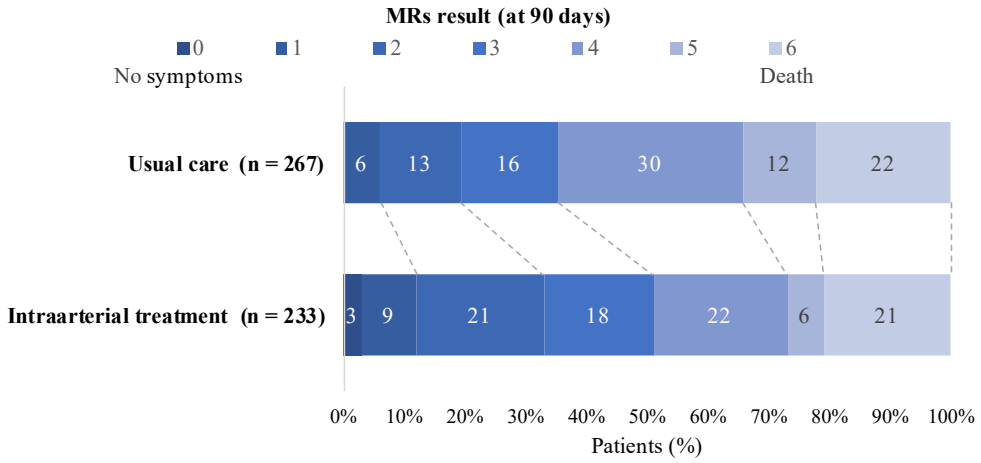
GBS	Guillain-Barré Syndrome
GBS DS	Guillain-Barré Syndrome Disability Scale
MRC	Medical Research Council
OR	Odds Ratio
PO	Proportional Odds
RCT	Randomized Controlled Trial
SE	Standard Error
TBI	Traumatic Brain Injury

Appendix B

Grotta charts NEJM trials

Figure 5.1: Grotta charts of the distribution on the Glasgow Outcome Scale Extended (GOSE) at 6 months in a TBI trial (upper panel) and the modified Rankin Scale (mRS) at 90 days in the MrCLEAN trial (lower panel). Both outcome measures assess the functional status [2, 12]





Appendix C

Model building longitudinal binary regression

Unadjusted models

Table 5.1: Building the unadjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 01 vs 23456

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, Treat * Time)	(1) **Random intercepts and random slopes**	-349.58			
		(2) Random intercepts	-425.58	152	2	<.0001
With or without interaction effect?	Random intercepts and random slopes	(1) Treat, Time, Treat * Time	-349.58			
		(2) **Treat, Time**	-349.57	0.006	1	0.93

Table 5.2: Building the unadjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 012 vs 3456

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, Treat * Time)	(1) **Random intercepts and random slopes**	-565.52			
		(2) Random intercepts	-677.20	223.36	2	<.0001
With or without interaction effect?	Random intercepts and random slopes	(1) Treat, Time, Treat * Time	-565.52			
		(2) **Treat, Time**	-565.53	0.0203	1	0.89

Table 5.3: Building the unadjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS DS 0123 vs 456

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, Treat * Time)	(1) **Random intercepts and random slopes**	-630.55			
		(2) Random intercepts	-803.42	345.74	2	<.0001
With or without interaction effect?	Random intercepts and random slopes	(1) Treat, Time, Treat * Time	-630.55			
		(2) **Treat, Time**	-630.53	0	1	1

Table 5.4: Building the unadjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 01234 vs 56

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, Treat * Time)	(1) **Random intercepts and random slopes**	-570.73			
		(2) Random intercepts	-631.54	121.63	2	<.0001
With or without interaction effect?	Random intercepts and random slopes	(1) Treat, Time, Treat * Time	-570.73			
		(2) **Treat, Time**	-570.91	0.356	1	0.55

Adjusted models

Table 5.5: Building the adjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 01 vs 23456

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset)	(1) **Random intercepts and random slopes**	-320.65			
		(2) Random intercepts	-397.23	153.16	2	<.0001
With or without interaction effect for treat by time?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset	-320.65			
		(2) **Treat, Time, MRCss, Weakonset, Treat * MRCss, Treat * Weakonset**	-320.65	0	1	1
With or without interaction effect for treat by mrcss?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * MRCss, Treat * Weakonset	-320.65			
		(2) **Treat, Time, MRCss, Weakonset, Treat * Weakonset**	-320.64	0	1	1
With or without interaction effect for treat by weakonset?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * Weakonset	-320.64			
		(2) **Treat, Time, MRCss, Weakonset**	-321.32	1.36	1	0.24

Table 5.6: Building the adjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 012 vs 3456

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset)	(1) **Random intercepts and random slopes**	-534.8			
		(2) Random intercepts	-643.9	218.2	2	<.0001
With or without interaction effect for treat by time?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset	-534.8			
		(2) **Treat, Time, MRCss, Weakonset, Treat * MRCss, Treat * Weakonset**	-534.8	0	1	1
With or without interaction effect for treat by weakonset?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * MRCss, Treat * Weakonset	-534.8			
		(2) **Treat, Time, MRCss, Weakonset, Treat * MRCss**	-534.8	0	1	1
With or without interaction effect for treat by MRCss?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * MRCss	-534.8			
		(2) **Treat, Time, MRCss, Weakonset**	-535.3	0.9	1	0.34

Table 5.7: Building the adjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 0123 vs 456

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset)	(1) **Random intercepts and random slopes**	-606.45			
		(2) Random intercepts	-777.50	342.1	2	<.0001
With or without interaction effect for treat by time?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset	-606.45			
		(2) **Treat, Time, MRCss, Weakonset, Treat * MRCss, Treat * Weakonset**	-606.42	0	1	1
With or without interaction effect for treat by weaksonset?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * MRCss, Treat * Weakonset	-606.42			
		(2) **Treat, Time, MRCss, Weakonset, Treat * MRCss**	-606.52	0	1	0.65
With or without interaction effect for treat by mrcss?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * MRCss	-606.52			
		(2) **Treat, Time, MRCss, Weakonset**	-606.75	0.46	1	0.50

Table 5.8: Building the adjusted longitudinal binary model, selected options in bold. Dichotomy: GBS DS 01234 vs 56

Decision to be made	Model as starting point	Models under comparison; selected option underlined	loglikelihood	χ^2	df	p-value
How many random effects?	Elaborate model (Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset)	(1) **Random intercepts and random slopes**	-567.06			
		(2) Random intercepts	-618.67	103.21	2	<.0001
With or without interaction effect for treat by weakonset?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss, Treat * Weakonset	-567.06			
		(2) **Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss**	-567.07	0	1	1
With or without interaction effect for treat by time?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * Time, Treat * MRCss	-567.07			
		(2) **Treat, Time, MRCss, Weakonset, Treat * MRCss**	-567.13	0.13	1	0.72
With or without interaction effect for treat by MRCss?	Random intercepts and random slopes	(1) Treat, Time, MRCss, Weakonset, Treat * MRCss	-567.13			
		(2) **Treat, Time, MRCss, Weakonset**	-567.84.32	1.41	1	0.23

Appendix D

Model building longitudinal proportional odds model

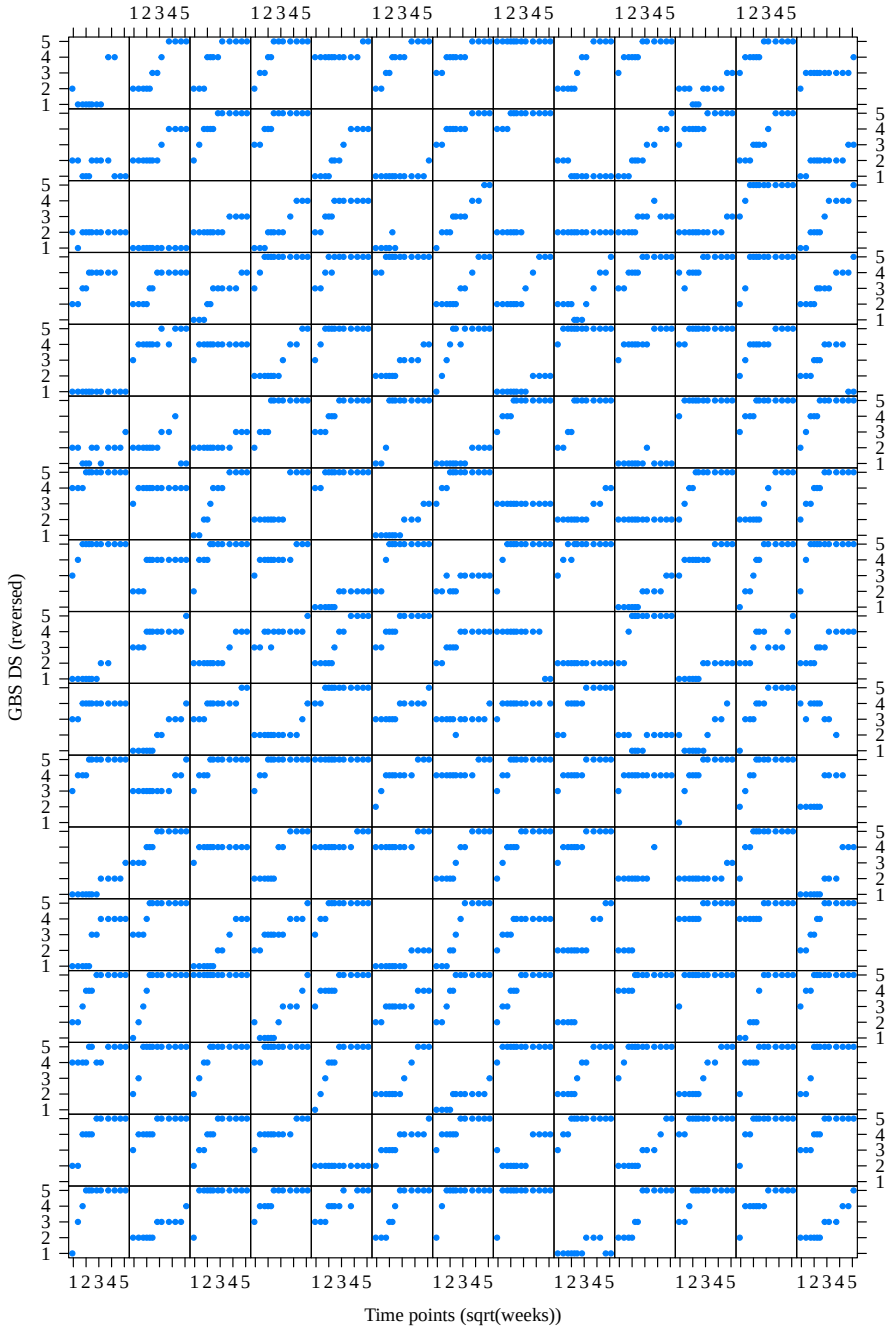
Table 5.9: Various approaches for longitudinal proportional odds model building and its result for unadjusted and adjusted analyses. In bold the model specifications that were eventually used for the longitudinal proportional odds model.

	Adjusted longitudinal proportional odds model	Unadjusted longitudinal proportional odds model
(1) Same model as mixed binary logistic regression	**- No interaction terms, main effects only** **- Random intercepts and random slopes**	**- No interaction terms, main effects only** **- Random intercepts and random slopes**
(2) Structure same as what would be best for linear mixed effect model	- No interaction terms, main effects only - Random intercepts and random slopes - Continuous autocorrelation correlation structure - Unstructured variance covariance matrix	- No interaction terms, main effects only - Random intercepts and random slopes - Continuous autocorrelation correlation structure - Unstructured variance covariance matrix
(3) Model specification that is the best fit according to the CompareBmods() function (based on LOO)	- All interaction terms included: treatment by weeks, and treatment by each covariate - Random intercepts only	- Interaction terms weeks by treat - Random intercepts only - Specification of variance covariance matrix and correlation structures not possible

Appendix E

Visualisation raw data GBS trial

Figure 5.2: GBS DS score at each time point, plotted per subject. GBS DS was reversed, such that a higher score stands for a healthier state [34]



Appendix F

Unadjusted binary logistic, proportional odds and longitudinal regression analysis for the outcome GBS DS

Table 5.10: Results of re-analysis of an RCT in GBS [34]. GSB DS was reversed such that an OR larger than 1 indicates a positive treatment effect. Point estimates of treatment effect across time points (last column) were obtained from mixed models with random intercepts per subject fitted to data from all time points in the follow-up. Point estimates of treatment effect across the complete ordinale scale (common OR, last row) were obtained from a PO model. Point estimates of the treatment effect across the complete ordinal scale and across all time points (aggregate common OR, bottom right cell) were obtained from a longitudinal PO model. Binary logistic regression was performed for each time point in the follow-up and each possible dichotomization in the GBS DS (inner cells). All analysis were done without covariate adjustment.

	Day 6	**Day 13**	**Week 3**	**Week 4**	**Week 5**	**Week 6**
Binary OR (GBS DS) 23456 vs 01	0.77	0.86	1.15	1.07	1.09	0.75
(95% CI)	(0.38 - 1.55)	(0.41 – 1.79)	(0.53 – 2.50)	(0.48 – 2.38)	(0.49 - 2.46)	(0.32 - 1.73)
SE	0.36	0.37	0.39	0.40	0.42	0.42
Binary OR (GBS DS) 3456 vs 012	1.18	1.27	1.33	1.40	1.39	1.40
(95% CI)	(0.69– 2.05)	(0.75 – 2.15)	(0.78 - 2.28)	(0.81 – 2.44)	(0.79 – 2.46)	(0.78 – 2.55)
SE	0.28	0.27	0.27	0.28	0.29	0.30
Binary OR (GBS DS) 456 vs 0123	1.33	0.97	1.18	1.27	1.25	1.44
(95% CI)	(0.58– 3.13)	(0.55 – 1.72)	(0.69 – 2.01)	(0.75 – 2.17)	(0.74 – 2.16)	(0.84 - 2.48)
SE	0.43	0.29	0.27	0.27	0.27	0.27
Binary OR (GBS DS) 56 vs 01234	0.50	1.27	1.01	0.71	0.87	0.92
(95% CI)	(0.02-5.29)	(0.33 – 4.87)	(0.45 – 2.29)	(0.37 – 1.38)	(0.47 – 1.60)	(0.52 - 1.62)
SE	1.23	0.68	0.42	0.34	0.31	0.29
Common OR	1.05	1.07	1.20	1.09	1.11	1.13
(95% CI)	(0.64-1.70)	(0.67-1.72)	(0.75-1.92)	(0.68-1.75)	(0.69-1.78)	(0.70-1.82)
SE	0.25	0.24	0.24	0.24	0.24	0.24

	Week 8	**Week 10**	**Week 14**	**Week 18**	**Week 22**	**Week 26**	Across time points
	0.74	0.78	0.24	0.47	0.64	0.81	0.83
	(0.27 – 2.06)	(0.20 – 3.03)	(0.02 – 1.62)	(0.07 – 2.45)	(0.16 - 2.30)	(0.21 – 3.10)	(0.09 – 7.37)
	0.52	0.69	1.07	0.88	0.66	0.69	1.12
	1.71	1.39	0.85	0.99	0.83	0.74	3.38
	(0.90 – 3.24)	(0.73 – 2.74)	(0.39 – 1.81)	(0.43 – 2.25)	(0.35 - 1.94)	(0.30 – 1.84)	(0.22 - 52.90)
	0.32	0.34	0.39	0.42	0.44	0.46	1.40
	1.52	1.28	0.94	0.82	0.89	0.83	1.99
	(0.88 – 2.66)	(0.72 – 2.25)	(0.52 – 1.74)	(0.42 – 1.59)	(0.45 - 1.73)	(0.40 – 1.72)	(0.29 – 13.88)
	0.28	0.29	0.31	0.34	0.34	0.37	0.99
	1.07	1.10	1.05	1.01	0.96	1.03	0.58
	(0.62 – 1.83)	(0.65 – 1.87)	(0.62 – 1.78)	(0.58 – 1.74)	(0.55 - 1.70)	(0.57 – 1.85)	(0.07 - 4.29)
	0.28	0.27	0.27	0.28	0.29	0.30	1.02
	1.24	1.17	0.99	0.96	0.93	0.98	1.05
	(0.77-2.01)	(0.72-1.91)	(0.60-1.64)	(0.57-1.63)	(0.54-1.61)	(0.55-1.74)	(0.58 – 1.89)
	0.25	0.25	0.26	0.27	0.28	0.29	0.31

Appendix G

Adjusted binary logistic, proportional odds and longitudinal regression analysis for the outcome GBS DS

Table 5.11: Results of re-analysis of an RCT in GBS [34]. GSB DS was reversed such that an OR larger than 1 indicates a positive treatment effect. Point estimates of treatment effect across time points (last column) were obtained from mixed models with random intercepts per subject fitted to data from all time points in the follow-up. Point estimates of treatment effect across the complete ordinale scale (common OR, last row) were obtained from a PO model. Point estimates of the treatment effect across the complete ordinal scale and across all time points (aggregate common OR, bottom right cell) were obtained from a longitudinal PO model. Binary logistic regression was performed for each time point in the follow-up and each possible dichotomization in the GBS DS (inner cells). All analysis were done with adjustment for baseline MRC sum score and days from onset of weakness to randomisation.

	Day 6	**Day 13**	**Week 3**	**Week 4**	**Week 5**	**Week 6**	
Binary OR (GBS DS) 23456 vs 01	0.91	1.02	1.38	1.18	1.33	0.86	
(95% CI)	(0.40 - 2.10)	(0.45 - 2.35)	(0.58 - 3.28)	(0.51 - 2.76)	(0.57 - 3.11)	(0.36 - 2.09)	
SE	0.42	0.42	0.44	0.43	0.43	0.45	
Binary OR (GBS DS) 3456 vs 012	1.56	1.72	1.66	1.73	1.68	1.59	
(95% CI)	(0.84 - 2.91)	(0.93 - 3.16)	(0.91 - 3.02)	(0.95 - 3.17)	(0.91 - 3.11)	(0.85 - 2.97)	
SE	0.32	0.31	0.30	0.31	0.31	0.32	
Binary OR (GBS DS) 456 vs 0123	1.74	1.18	1.45	1.53	1.49	1.72	
(95% CI)	(0.71 - 4.27)	(0.63 - 2.20)	(0.81 - 2.61)	(0.86 - 2.73)	(0.83 - 2.66)	(0.96 - 3.11)	
SE	0.46	0.32	0.30	0.29	0.30	0.30	
Binary OR (GBS DS) 56 vs 01234	0.59	1.28	1.10	0.73	0.92	0.98	
(95% CI)	(0.05 - 7.10)	(0.33 - 4.95)	(0.48 - 2.54)	(0.37 - 1.46)	(0.50 - 1.70)	(0.54 - 1.79)	
SE	1.27	0.69	0.42	0.35	0.31	0.30	
Common OR	1.06	1.08	1.21	1.06	1.08	1.09	
(95% CI)	(0.65 – 1.72)	(0.67 – 1.74)	(0.75 – 1.93)	(0.66 – 1.71)	(0.67 – 1.73)	(0.68 – 1.76)	
SE	0.25	0.24	0.24	0.24	0.24	0.24	

	Week 8	**Week 10**	**Week 14**	**Week 18**	**Week 22**	**Week 26**	Across time points
	0.89	0.86	0.25	0.51	0.69	0.84	1.18
	(0.30 - 2.61)	(0.22 - 3.36)	(0.03 - 2.32)	(0.09 - 2.90)	(0.18 - 2.56)	(0.22 - 3.29)	(0.14 – 10.17)
	0.55	0.70	1.14	0.88	0.67	0.69	1.10
	1.95	1.68	0.97	1.12	0.92	0.82	4.96
	(0.99 - 3.83)	(0.82 - 3.42)	(0.44 - 2.17)	(0.48 - 2.63)	(0.37 - 2.30)	(0.31 - 2.17)	(0.41 – 59.78)
	0.35	0.36	0.41	0.44	0.47	0.49	1.27
	1.80	1.42	1.07	0.89	0.99	0.92	3.01
	(0.98 - 3.31)	(0.77- 2.60)	(0.56 - 2.06)	(0.44 - 1.81)	(0.47 - 2.05)	(0.43 - 1.98)	(0.51 – 17.88)
	0.31	0.31	0.33	0.36	0.37	0.39	0.91
	1.15	1.16	1.13	1.09	1.03	1.12	0.62
	(0.66 - 2.02)	(0.67 - 2.01)	(0.65 - 1.96)	(0.61 - 1.93)	(0.57 - 1.86)	(0.60 - 2.08)	(0.08 - 4.62)
	0.29	0.28	0.28	0.29	0.30	0.32	1.02
	1.25	1.19	0.99	0.97	0.97	1.03	1.24
	(0.76 – 2.04)	(0.72 – 1.97)	(0.59 – 1.65)	(0.56 – 1.68)	(0.55 – 1.72)	(0.56 – 1.89)	(0.69 – 2.15)
	0.25	0.26	0.26	0.28	0.29	0.31	0.29

Appendix H

Overview Github files All files (table 5.12) to reproduce the current study can be found at: https://github.com/SCdeRuiter/Thesis.git

Table 5.12: Overview of files in folder 'Thesis' at GitHub

File name	Description
Analysis_NEJM.Rmd	Code to analyse MR CLEAN and DECRA trial
Sim_LongCont.Rmd	Code to run simulation of longitudinal continuous data
Sim_CondABCD_CovAdj_PO.R	Code to run simulations for condition A through D, computing gains in efficiency by covariate adjustment and/or exploiting ordinality
Analysis_MPtrial.Rmd	Code to analyse the previous GBS trial (including code for a binary logistic regression, PO model, mixed effect binary logistic regression and longitudinal PO model)