Modelling Multivariate Nonlinear Vaccine Induced Immune Responses

Brendon M. Lapham
University of Cape Town

March 15, 2020

Dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Statistics

To Roxie and Ben

DECLARATION

I, Brendon Michael Lapham, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgments indicate otherwise) and that neither the whole work nor any part of it has been, is being or is to be submitted for another degree in this or any other University. I empower the University of Cape Town to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

BM Lapham

ABSTRACT

Interpretable statistical models for multivariate vaccine induced immune response data are important as they provide a rigorous means of deciding which vaccine candidates should be advanced in the clinical trials process. We consider applications of several different statistical models to a vaccine data set which contains multivariate immune responses for several novel Tuberculosis vaccines and the current BCG vaccine. The immune responses in the data set have several features which the models need to account for. In particular, the models need to account for the multivariate repeated measures for the subjects, the nonlinear profiles of the immune responses, and the zero-inflated skew distributions of the immune responses. We find that Tweedie multivariate generalised linear mixed effect and latent variable models with cubic B-splines perform well for this data set relative to linear, nonlinear, and univariate Tweedie generalised linear mixed effect models. In addition, the Tweedie multivariate generalised linear mixed effect and latent variable models have several advantages over the other models we consider and are also capable of interpretation; importantly, we are able to draw clinical conclusions about which novel TB vaccine candidates appear to be the most promising.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Associate Professor Francesca Little, whose guidance and support were invaluable.

I would also like to thank my colleagues at RGA Reinsurance Company of South Africa who have supported me in my studies over the past two years. I am very grateful to RGA for the financial support that they provided.

On a personal note, I am forever indebted to my wife Roxie for her unwavering support, encouragement, and love. Without this, the work here would not have been possible. To both my son, Ben, and my wife, I owe considerable amounts of family time.

CONTENTS

1

INTRODUCTION

1.1 BACKGROUND

Tuberculosis (TB) is the leading cause of death from a single infectious agent (World Health Organisation, 2019). In 2018, it is estimated that 10.0 million people fell ill with TB, and that there were 1.2 million TB deaths from HIV-negative people and 251 000 deaths from HIV-positive people (World Health Organisation, 2019). These numbers, particularly the number of deaths, have significantly reduced in the last two decades, but many regions are not on track to reach the 2020 milestones set out in the World Health Organisation's (WHO's) End TB Strategy (World Health Organisation, 2015, 2019). Figure 1 illustrates the targets of the WHO's End TB Strategy with the ultimate goal of ending the TB epidemic by 2035.

The WHO identifies effective vaccines as part of their End TB Strategy and outlines that new vaccines will be an essential tool to break the trajectory of the TB epidemic as can be seen in Figure 1 (World Health Organisation, 2015, p. 58). Currently the bacille Calmette-Guérin (BCG) vaccine is the only licensed vaccine for TB. However, in response to the need for new vaccines, there are several TB vaccine candidates which are in development, of which only a few will be able to be advanced through the clinical trials process due to limited resources (Rodo *et al.*, 2019a).

As only a few vaccines can be advanced, it is important to be able to objectively identify those which are the most promising. By looking at the antigen-specific T cell responses induced by several candidate vaccines and using statistically rigorous methods, the work of Rodo *et al.* (2019a) aims to identify the candidate vaccines that should be advanced. They argue that examining the antigen-specific immune response induced by the candidate vaccines is an objective and data driven means of comparing and prioritising the vaccines which should be advanced. The work that is carried out by Rodo *et al.* (2019a) aims to compare the 'memory responses' of the candidate vaccines and as such is a cross-sectional study of the vaccines. We take an alternative approach in the work here and attempt to model and compare the longitudinal profiles of the antigen-specific immune responses induced by candidate vaccines.

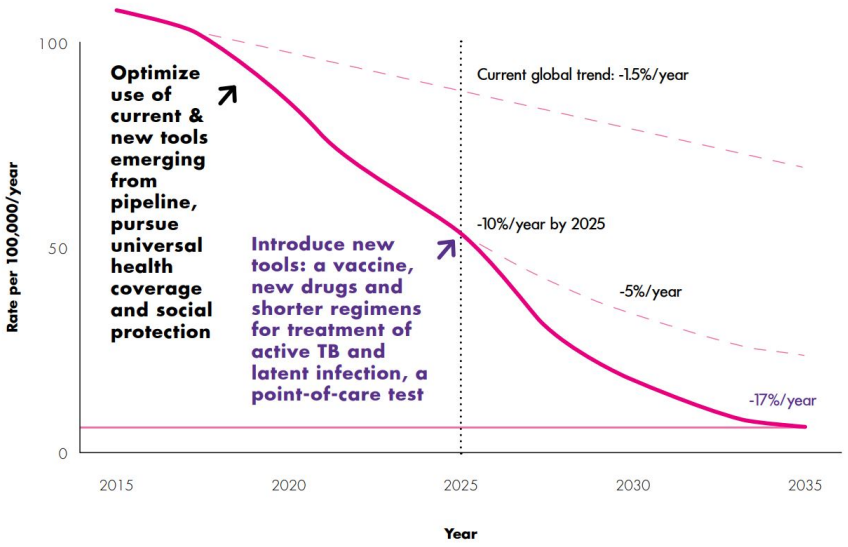


Figure 1: Illustration of the WHO's desired reduction in the global TB incidence rates to reach the 2035 goal of ending the TB epidemic (World Health Organisation, 2015, p. 18).

The vaccine data that we have available has repeated measures of the subjects' antigen-specific immune responses and as such is amenable to longitudinal data analysis methods. In particular, the longitudinal methods that we use allow us to control for the variation across the subjects (inter-subject variation), the variation within subjects, and account for the correlation arising from repeated measures on the same subject. The methods that we use allow us to perform inference and as such we can objectively identify the vaccine candidates that are the most promising.

1.2 LONGITUDINAL VACCINE-INDUCED T CELL RESPONSES

In this section we provide further details regarding TB infections and vaccines to add further context to the data that we use in this study.

An infection by the *M*ycobacterium tuberculosis (*M.tb*) is a precursor to a person developing the TB disease. People who are infected with *M.tb* will not all develop the disease; about 5-10% of lives with an *M.tb* infection will develop the TB disease and the other lives will live with a latent *M.tb* infection (World Health Organisation, 2019). It is estimated that the global prevalence of latent *M.tb* infection in 2014 was 23.0%, with poorer regions such as South East Asia (30.8%), the Western Pacific Region (27.9%), and Africa (22.4%) having the highest prevalences (Houben and Dodd, 2016).

As mentioned new vaccines are key to meeting the milestones and goals of the WHO's End TB Strategy. In 2017, there were thirteen novel vaccine candidates being investigated and in various stages of

trials (Voss *et al.*, 2018). Typically these trials would measure the CD4 and or the CD8 T cells expressing the cytokines IFN-γ, TNF-α, and IL-2 which are induced by the administration of the vaccine. The reason that these cytokines are measured is because they are deemed to be necessary, but not sufficient, for immunity to *M.tb* infection, though there may be limited evidence for this (Jasenosky *et al.*, 2015; Lewinsohn *et al.*, 2017; Rodo *et al.*, 2019a). The expression of cytokines by T cells is the immune response to the antigens contained by the vaccine candidates.

For a vaccine candidate with a single initial dose, it is expected that the immune response, the frequency of CD4 or CD8 T cells expressing certain combinations of cytokines, will initially increase from a baseline starting point and then peak as the immune system responds to the presence of the antigens, after which it will decrease until it reaches some memory response or level. This memory response is expected to be higher than the initial baseline response. This is similar to the concentration curves seen in pharmacokinetic work; e.g. see the textbook examples given by Pinheiro and Bates (2000, p. 350).

An example of an immune response is given in Figure 2 where we see the pattern described above; the figure shows the frequency of CD4 T cells expressing the cytokines IFN-γ, TNF-α, and not IL-2 over time following an initial dose of the BCG vaccine for a single subject. The baseline and memory states of the immune response to the vaccine have been marked in the figure, and we can see that the memory response is greater than the baseline response. It is believed that a vaccine which has the largest prolonged immune response will convey the greatest immunity, or immunological memory, and that the memory response is the best measure of this (Rodo, 2017; Rodo *et al.*, 2019a). The immunological memory in the context of TB is the T cells' ability to recognise the antigens of *M.tb* and effectively mount an immune response.

The immune responses from a candidate vaccine are induced by the presence of antigens in the vaccine candidate. For example, the antigen in the BCG vaccine is BCG, but other vaccine candidates may contain several antigens which are used to induce the immune response. The CD4 and CD8 T cell responses and the recorded frequencies of the cytokine combinations are antigen specific. Vaccine candidates typically contain varied and distinct antigens so as to explore the possible immune responses and immunological memory induced by the vaccine. It is important that vaccine candidates give rise to distinct immunological responses to ensure that the space of possible induced immune responses which may convey immunological memory for *M.tb* is explored (Rodo *et al.*, 2019a).

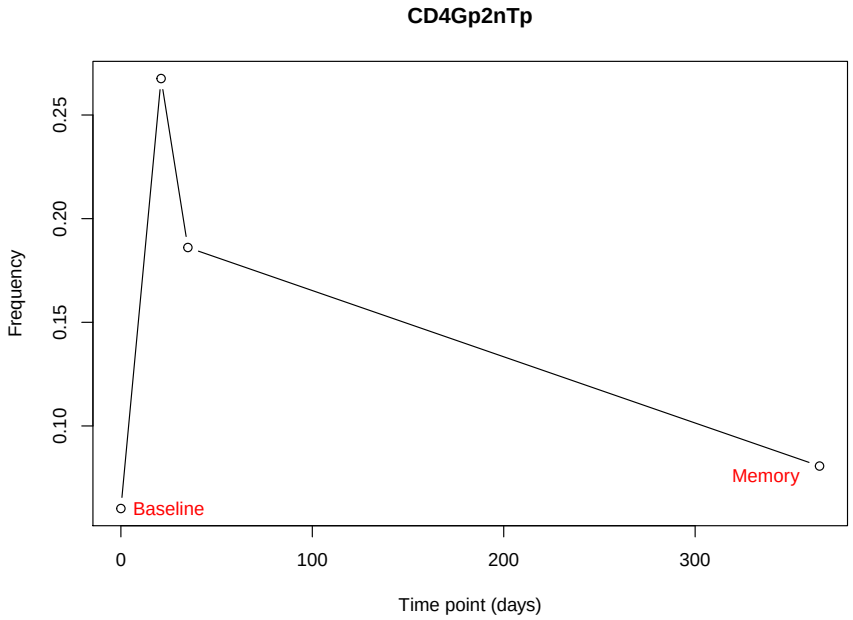


Figure 2: Vaccine induced immune response showing the baseline and memory responses. The curve shows the frequency of CD_4 T cells expressing the cytokines IFN-γ, TNF-α, and not IL-2.

1.3 VACCINE CANDIDATE DATA

1.3.1 *Data Sources*

The data that we use for the study here are the publicly available data provided by Rodo *et al.* (2019b). This data are analyzed extensively by Rodo (2017) and Rodo *et al.* (2019a). This data originally come from several vaccine trials carried out by the South African TB Vaccine Initiative and the details of the vaccine candidates are given in Table 1. In the work that follows we will refer to the vaccines by their Vaccine Number.

The details of the available data are shown in the Figure 3. We can see that the trials were not standardised as they had different measurement times and different dose strategies. We also see that the data do not contain all of the antigen-specific T cell frequencies when compared to those listed in Table 1. The time point fields refer to the days following the initial administration of the vaccine.

1.3.2 *Ethics Approval*

As the data used here is anonymised third party data, explicit ethics approval was not required. The work of Rodo (2017) and Rodo *et al.* (2019a) received ethics approval from the University of Cape Town's

Table 1: Details of the TB vaccine candidates and their associated antigens (Rodo *et al.*, 2019a).

Vaccine No.	Vaccine	Citation	Vaccine Antigen	Dose
1	AERAS-402	Abel *et al.* (2010)	Ag85A Ag85B TB10.4	3×10^{10} viral particles
2	H56:IC31	Suliman *et al.* (2019)	Ag85B ESAT6 Rv2660	$5\mu g$
3	M72/AS01$_E$	Day *et al.* (2013)	Mtb39a Mtb32a	$10\mu g$
4	MVA85A	Scriba *et al.* (2010) and Scriba *et al.* (2012)	Ag85A	5×10^7 plaque forming units
5	H1:IC31	Mearns *et al.* (2017)	Ag85B ESAT6	$15\mu g$
6	ID93+GLA-SE	Penn-Nicholson *et al.* (2018)	Rv1813 Rv2608 Rv3619 Rv3620	$10\mu g$ + $2\mu g$ GLA-SE
7	BCG	Suliman *et al.* (2016)	BCG	BCG ($2 - 8 \times 10^5$ CFU)

					Time points																						
Vaccine	Antigen	Infected	No. Vaccine Admin.	Patients	0	7	14	21	28	30	35	37	42	56	60	70	84	112	126	168	182	196	210	224	292	294	364	365
1	ag85	0	1	9	V	M			M								M				M							
1	tb104	0	1	9	V	M			M								M				M							
2	ag85b	0	2	15	V		M							V		M									M			
2	ag85b	1	2	12	V		M							V		M									M			
2	esat6	0	2	15	V		M							V		M									M			
2	esat6	1	2	12	V		M							V		M									M			
2	rv2660	0	2	15	V		M							V		M									M			
2	rv2660	1	2	12	V		M							V		M									M			
3	m72ppl	0	2	37	V	M				V		M			M								M					
3	m72ppl	1	2	46	V	M				V		M			M								M					
4	ag85a	0	1	12	V	M			M											M								
4	ag85a	1	1	12	V	M			M								M										M	
5	ag85b	0	2	34	V		M							V		M								M				
5	ag85b	1	2	24	V		M							V		M								M				
5	esat6	0	2	34	V		M							V		M								M				
5	esat6	1	2	24	V		M							V		M								M				
6	rv1813	0	3	10	V		M		V (no M)				M					V	M			M				M		
6	rv1813	1	3	14	V		M		V (no M)				M					V	M			M				M		
6	rv2608	0	3	10	V		M		V (no M)				M					V	M			M				M		
6	rv2608	1	3	14	V		M		V (no M)				M					V	M			M				M		
6	rv3619	0	3	10	V		M		V (no M)				M					V	M			M				M		
6	rv3619	1	3	14	V		M		V (no M)				M					V	M			M				M		
6	rv3620	0	3	10	V		M		V (no M)				M					V	M			M				M		
6	rv3620	1	3	14	V		M		V (no M)				M					V	M			M				M		
7	bcg	1	1	27	V			M			M																	M

Figure 3: A table describing the available data for each vaccine, antigen, and initial *M.tb* infection status. A "V" indicates that a vaccine was administered and a measurement was taken at the associated time point, a "V (no M)" indicates that a vaccine was administered but no measurement was taken at the associated time point, and an "M" indicates that only a measurement was taken at the associated time point. The time points are measured in days from the initial administration of the relevant vaccine.

Human Research Ethics Committee (HREC). The relevant HREC ethics approval number was 039/2017.

1.3.3 *Dose Strategy Selection*

As described by Rodo (2017), the trials for vaccines 1, 2, 5, and 6 were dose-finding trials where several different dose strategies were considered. Dose strategies can vary by the size of the dose, the number of administrations of the vaccine, and the timing of the administration of the vaccine. The individuals included in the data by Rodo (2017) and Rodo *et al.* (2019a) were those with the 'optimal' dose strategy, where optimal is defined as those with the largest immune response. The trials for vaccines 3, 4, and 7 all had the same dose strategy for all of the participants and so all individuals were included. The details of the dose sizes are given in Table 1, and the number and timing of the doses are given in Figure 3.

As the doses varied across the vaccines the work here is effectively a vaccine and dose comparison.

1.3.4 *Baseline* M.tb *Infection Status*

In several of the vaccine trials, as can be seen in Figure 3, there were both subjects who had an existing *M.tb* infection and those that did not have an *M.tb* infection. In general, it is of interest to consider the effect of baseline *M.tb* infection on the induced immune responses for the different vaccine candidates. However, in the study here we focus on the subjects with an existing *M.tb* infection.

1.3.5 *Cytokine Frequencies*

The data provides the background subtracted frequencies of antigen-specific CD4 and CD8 T cells expressing combinations of the cytokines:

- interferon-γ (IFN-γ),
- interleukin-2 (IL-2),
- interleukin 17A (IL-17), and
- tumour necrosis factor-α (TNF-α).

By background subtracted we mean that the frequency of the antigen-specific T cells before stimulation by the vaccine candidates is deducted from the frequency after the stimulation by the vaccine candidate. As noted by Rodo (2017), some of the background subtracted antigen-specific frequencies had to be set to zero when the frequency before stimulation exceeded the frequency after stimulation.

Similar to the work carried out by Rodo (2017) and Rodo *et al.* (2019a), we focus on IFN-γ, IL-2, and TNF-α. We sum over the IL-17 cytokine to get frequencies for the different combinations of the cytokines IFN-γ, IL-2, and TNF-α.

In the work that follows, we make use of the shorthand notation G[p/n]2[p/n]T[p/n] to identify the cytokine combination expressed in a particular T cell. The G, 2, and T stand for IFN-γ, IL-2, and TNF-α respectively, and the presence of a "p" indicates that the preceding cytokine was expressed while the presence of an "n" indicates that the preceding cytokine was not expressed. For example, Figure 2 shows the frequency of CD4 T cells expressing Gp2nTp; i.e. the frequency of CD4 T cells expressing IFN-γ, TNF-α, and not IL-2. In total we have seven possible combinations as shown below:

- Gp2pTp,
- Gn2pTp,
- Gp2nTp,
- Gp2pTn,
- Gn2nTp,
- Gn2pTn, and
- Gp2nTn.

Figure 4 shows plots of the frequencies for the CD4 T cells expressing different combinations of the IFN-γ, IL-2, and TNF-α cytokines for Vaccines 1 and 2. We will refer to these frequencies as immune responses and the curves as immune response profiles. Plots for the other vaccines are shown in the Appendix A. There are several features that we can note from the immune response profiles. These are:

- there is significant variation in the immune response profiles for the different subjects,
- the different antigens can induce quite different magnitude immune responses,
- the immune response profiles do not all conform to the description we gave above and differ to the example shown in Figure 2. This comment applies to both the vaccines with a single dose and those with more than one dose. Most notably the vaccines with multiple dose times tend to have immune response profiles with multiple peaks; this can be seen for Vaccine 2 in Figure 4 (b).
- there are clear outliers in the data which have profiles quite different to the others.

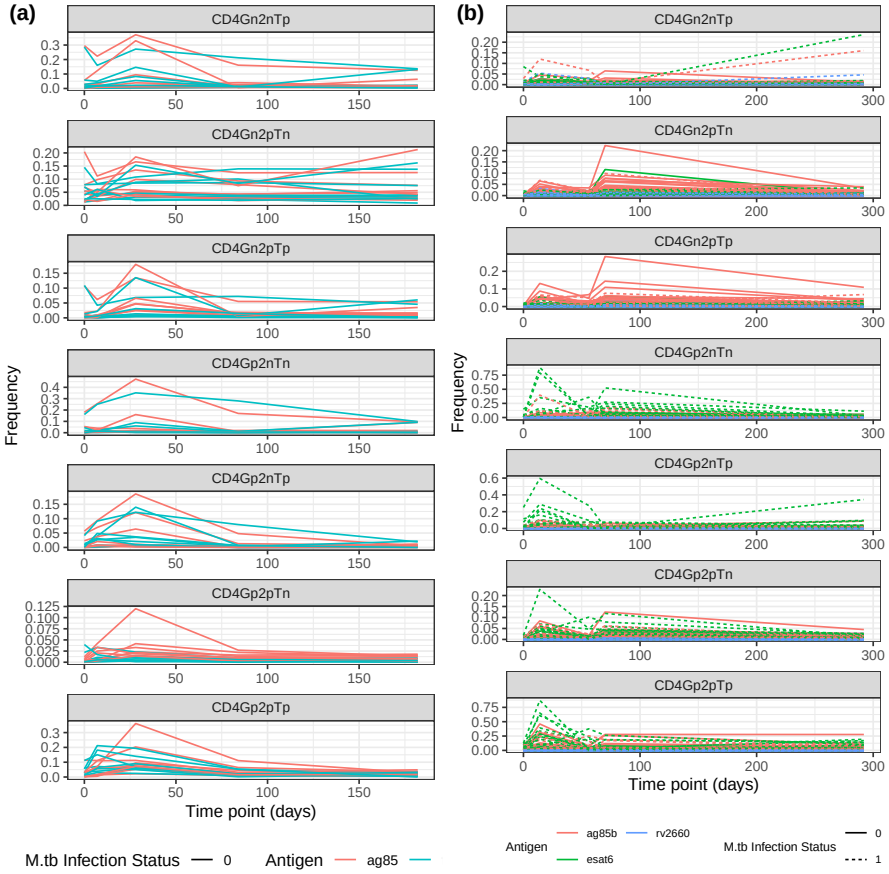


Figure 4: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for (a) Vaccine 1 and (b) Vaccine 2.

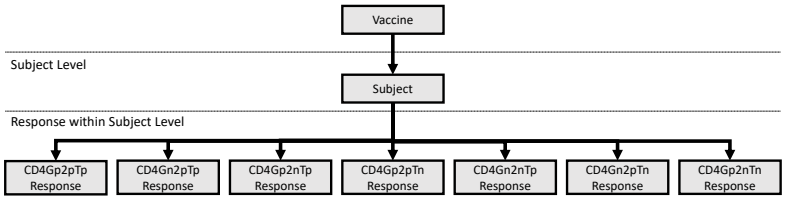


Figure 5: Illustration of the multivariate multilevel nature of the vaccine data for a specific time point for the CD4 cell frequencies for combinations of the cytokines IFN-γ, IL-2, and TNF-α.

1.3.6 *Multivariate and Multilevel Nature of the Data*

The vaccine data that are available are multivariate in nature. At each time point when an immune response is recorded we have antigen-specific frequencies for multiple combinations of cytokines for both CD4 and CD8 cells. For example, if we consider a vaccine with one antigen, the three cytokines IFN-γ, IL-2, and TNF-α, and only CD4 cells, we have 7 responses for each subject at each time point. In this case, the list of combinations of cytokines are as listed above.

The data are also multilevel in nature. This is illustrated in Figure 5 for a specific subject at a specific time point. Specifically, for each vaccine candidate, we have several subjects to whom the vaccine was admitted to. These subjects represent a level within the data. Then for each subject at each time point we have a set of observations, the immune responses. In the illustration, we have shown the observations as the CD4 cell frequencies for combinations of cytokines IFN-γ, IL-2, and TNF-α. Each of these levels within the data are potentially a source of variability. As we have repeated measurements of the same variables over time for the same subjects, the data are repeated measures data.

The modelling methods that we use will need to account for the multilevel multivariate repeated measure nature of the data.

1.3.7 *Data Used in the Applications*

We use a subset of the data considered by Rodo (2017), Rodo *et al.* (2019a), and provided by Rodo *et al.* (2019b). The main reason for considering a subset of the data is that we wanted vaccine induced immune response profiles which were amenable to being modelled by using the methods outlined in Chapter 2. This primarily required that the immune response profiles broadly followed the pattern described above and were not too dissimilar to that shown in Figure 2. This was important as some of the models, particularly the non-linear mixed effect models, are difficult to fit if the profiles are not consistent.

To achieve consistent profiles we performed the following preprocessing steps:

1. We summed over the cytokine IL-17 to provide frequencies for the combinations of the cytokines IFN-γ, IL-2, and TNF-α.

2. We considered only the CD4 T cell frequencies as the CD8 T cell frequencies were significantly smaller and as such more volatile. The reduced frequencies seen for CD8 T cells is noted by Scriba *et al.* (2010), Scriba *et al.* (2012), Suliman *et al.* (2016), Mearns *et al.* (2017), Penn-Nicholson *et al.* (2018), and Abel *et al.* (2010) at the time made a comment that this is a feature of most new TB vaccines.

3. We only considered the immune responses from Vaccines 3, 4, 5, and 7. These vaccine candidates had immune responses across several subjects which were relatively consistent.

4. We also only considered subjects who had an initial *M.tb* infection; this makes the comparison across the selected vaccines comparable as the trial for Vaccine 7, BCG, only considered subjects with an initial *M.tb* infection.

5. As described in Figure 3, Vaccines 3 and 5 both had multiple doses and as a result their profiles had multiple peaks. This can be seen in Figures 6 and 7. To smooth out the profiles, we deleted the observations at the time of the second dose where the immune responses typically dipped before peaking for the second time. The justification for this is that we are interested in the overall immune response as opposed to specific features of the response.

6. We also removed subjects with outlying observations. The subjects with outlying observations were identified systematically as those with observations at any time point which fell outside of the range defined by the lower and upper whiskers of a box-and-whisker plot for that time point. The lower whisker being calculated as the first quartile less 1.5 times the inter-quartile range, and the upper whisker being defined as the third quartile plus 1.5 times the inter-quartile range.

This preprocessing makes significant modifications to the data set, but as the ultimate goal is to investigate whether the models described in Chapter 2 allow us to model the long-term differences in the immune responses between the vaccines, the "smoothing" at earlier time points is required for us to be able to fit comparable models for the different vaccines. It should be noted that the exclusion of outliers in step 6 reduces the sample sizes significantly. After the preprocessing, we are left with the following numbers of subjects per vaccine:

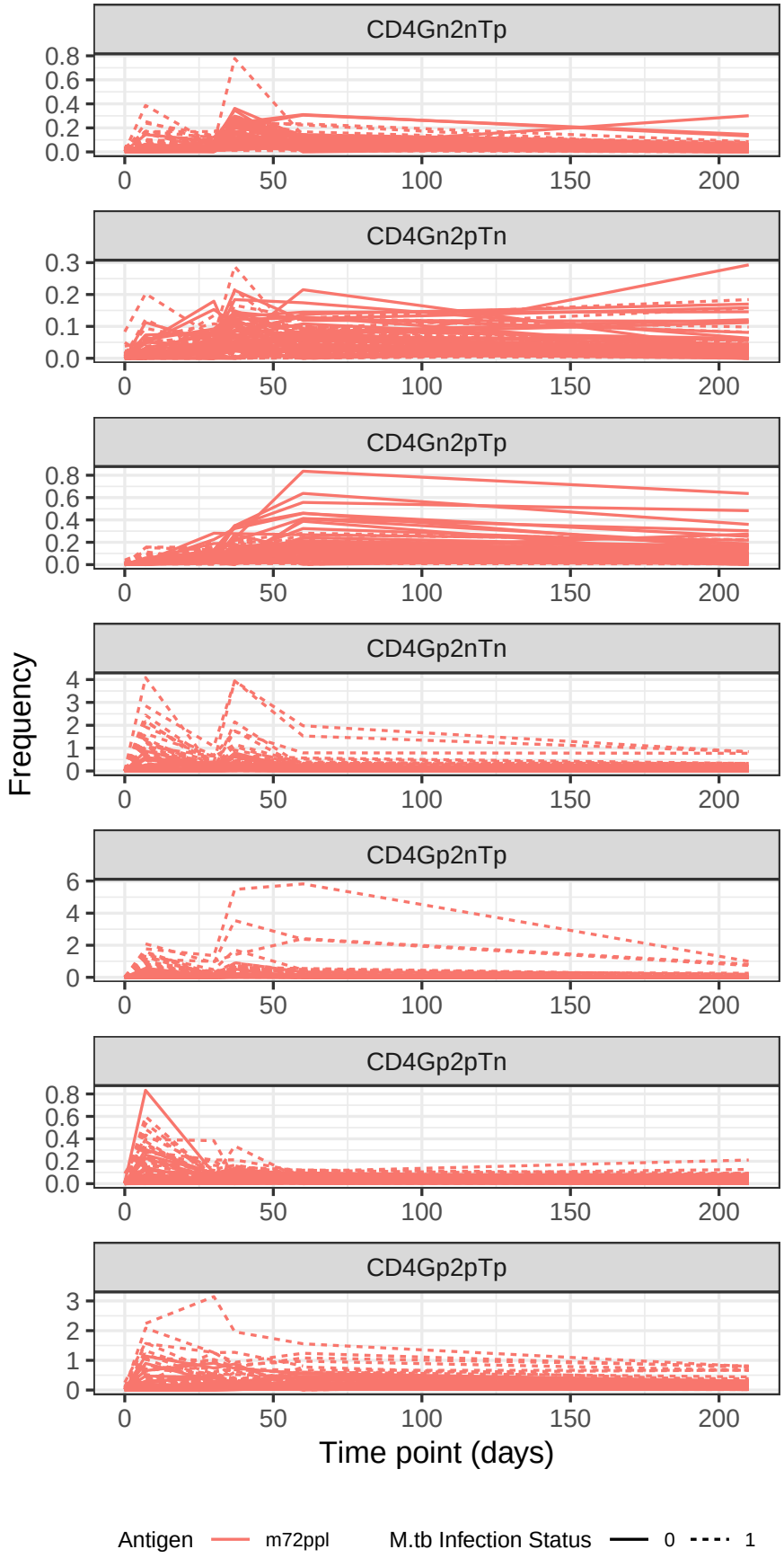


Figure 6: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 3.

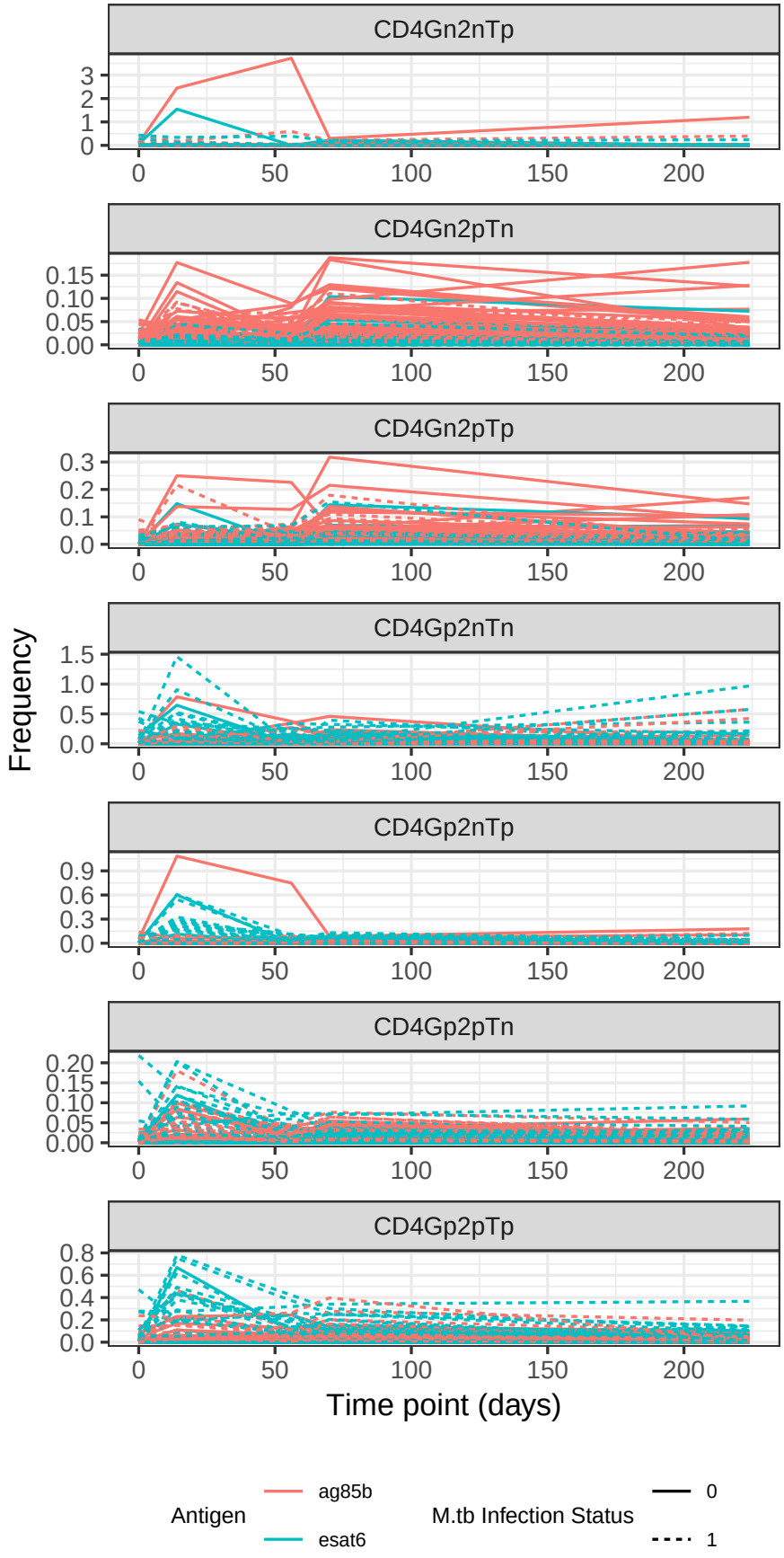


Figure 7: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 5.

- Vaccine 3: 15 subjects,
- Vaccine 4: 3 subjects,
- Vaccine 5: 5 subjects, and
- Vaccine 7: 14 subjects.

The significantly reduced sample sizes are likely to make some of the model fitting and interpretation difficult, particularly for Vaccine 4 and Vaccine 5.

Figure 8 presents box-and-whisker plots of the final data set that we use in the modelling. We see that the lines which join the medians at each time point for each vaccine follow the general pattern that we described earlier; the immune responses initially increase from their baseline response, peak as the immune system responds to the presence of the antigens, and then decrease until they reach some memory response level. This suggests that the profiles are now fairly consistent and amenable to being modelled using the techniques we outline in the next chapter.

Figure 9 presents histograms of the processed data of the observed frequency of cytokine combinations. An important feature to notice from the histograms is that there are a large number of zeros for the frequencies. In addition to this, we also see that the distributions are positively skewed. These features are something that we will need to take into account when we are building the models.

For each of the vaccines that we include in the analysis, we do not model responses from different antigens separately. Instead we treat the measurements for different antigens as repeated measurements for the subject at the particular time point. This affects only Vaccine 5 where we have multiple antigens. This may not be the best approach to modelling the immune responses for the vaccines and, in future work, alternative approaches, such as summing the responses across antigens for vaccines with multiple antigens, could be investigated.

1.4 RESEARCH AIMS

As outlined by Voss *et al.* (2018), there are currently no unanimously agreed criteria for deciding which vaccine candidates to advance through the clinical trials process. This can make it difficult for vaccines to move through the different stages of development.

The work of Rodo *et al.* (2019a) broadly aimed to provide a data-driven and statistically rigorous means of comparing vaccine candidates and identifying those that should be advanced. Our aim is similar in that we want to investigate whether we can use statistical models to identify the candidate vaccines to advance. The vaccine candidates which should be advanced are those which induce the largest

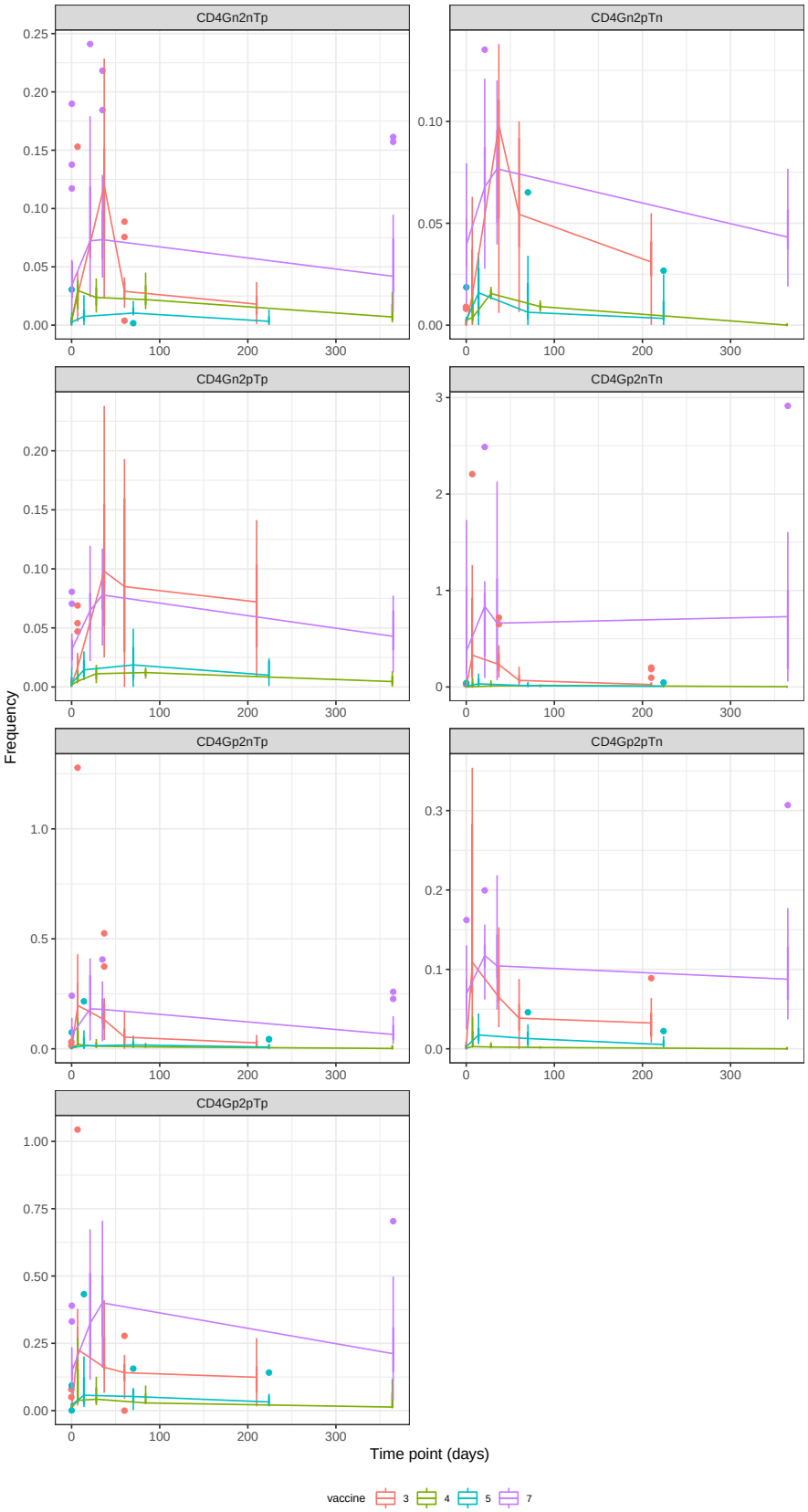


Figure 8: Box-and-whisker plots of the frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccines 3, 4, 5, and 7 after the preprocessing. Lines join the medians at each time point for each vaccine.

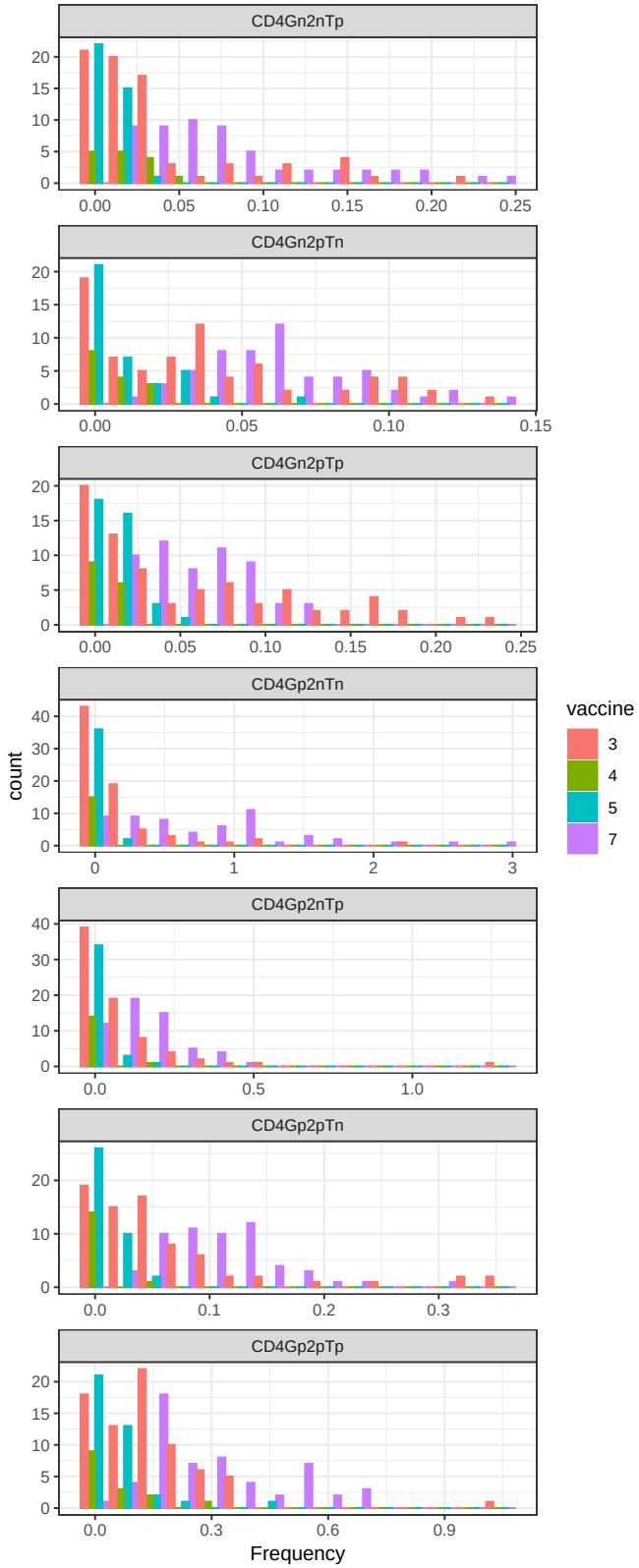


Figure 9: Histograms for the frequencies of CD4 T cells expressing the seven different cytokine combinations for Vaccines 3, 4, 5, and 7 after the preprocessing.

long term immune response as this conveys the greatest immunological memory. The models that we investigate will need to account for the features of the data that we have outlined above; specifically they will need to account for

1. the repeated measures for each of the subjects,
2. the zero-inflated skew distributions of the immune responses,
3. the multiple immune responses in the data, and
4. the nonlinear immune response profiles.

We want to use the fitted models for inference to identify the vaccine candidates that should be advanced. However, our aims are primarily methodological. We were less concerned with the exact vaccines and vaccine sub-groups included in the study, and rather we are concerned with demonstrating that the models we investigate can be used to identify the candidate vaccines to advance.

The primary difference between our work here and that of Rodo (2017) and Rodo *et al.* (2019a) is that the methods that we use here model the full immune response profiles over time to identify those vaccines to advance whereas the work of Rodo (2017) and Rodo *et al.* (2019a) only investigated the differences in the final memory responses for the vaccine candidates.

1.5 LIMITATIONS AND ASSUMPTIONS

1.5.1 *Subject Profile*

The subjects that were included in the trials that make up the data set from Rodo *et al.* (2019b) are from Worcester in the Western Cape, South Africa. The subjects are HIV-negative adults and adolescents. The results here are likely to be generalisable to lives who have similar social, biological, and environmental profiles to these subjects.

1.5.2 *Immune Responses*

The immune responses that we consider in the applications here are a subset of the full immune response to a TB vaccine. Thus we are assuming that the frequencies of the antigen-specific CD4 T cells expressing the cytokines IFN-γ, IL-2, and TNF-α are sufficient to characterise the immune response of a TB vaccine candidate and to identify the vaccine candidates to advance.

The vaccine candidates we consider here all had different designs as can be seen in Figure 3. We have tried to standardise the vaccines that we consider by the preprocessing we have outlined above, and as such we are assuming that the vaccines have comparable immune response profiles.

1.5.3 *Vaccine Selection and Sub-group Selection*

As mentioned earlier, our aims are primarily methodological and so we are less concerned with the exact vaccines and vaccine sub-groups included in the study. However, we are still making the implicit assumption that the vaccines and the vaccine sub-groups that we have included have immune response profiles over time that are representative of immune responses of vaccines in general. Provided this assumption holds, our work can then be generalised to other TB vaccine candidates and vaccines in general.

2

METHODOLOGY

In this chapter we outline the models, and their associated methods, which we use in our applications in Chapter 3. The models that we consider are intended to capture most of the features of the immune response data that we outlined in the previous chapter. Specifically the models will need to capture

1. the repeated measures for each of the subjects which induces correlation between observations for an individual,
2. the zero-inflated skew distributions of the immune responses,
3. the multiple immune responses in the data, and
4. the nonlinear immune response profiles.

The models that we consider here attempt to capture most of these features, but in some cases the models do not capture all of the features. In the cases where the models are not capturing all the features of the data, we are attempting to investigate whether the simpler models are still suitable.

We consider three different classes of models. The first class of models that we consider is the class of linear mixed effect models (LMEMs) and the theory that we present follows that given by Pinheiro and Bates (2000). LMEMs can be used to model the nonlinear vaccine response profiles that we saw in Section 1.3 by using polynomials and splines. The second class of models that we consider is nonlinear mixed effect models (NLMEMs) and the theory we present follows that of Pinheiro and Bates (2000) and Davidian and Giltinan (2003). NLMEMs explicitly capture the non-linear profile that we observed in Section 1.3 by using a nonlinear function. The third class of models that we consider is the generalised linear mixed effect models (GLMEMs). We consider both univariate and multivariate Tweedie GLMEMs. For the multivariate Tweedie GLMEMs, in addition to random effects, we also include latent variables. The theory we present follows that presented by Fitzmaurice *et al.* (2008), Zhang (2013), and Hui (2016). The nonlinearity of the vaccine response profiles is captured again by using splines.

The LMEMs and NLMEMs that we use assume that the within-group errors are normally distributed. It is anticipated that this assumption will be violated when we fit the models to the data described

in Section 1.3. This is because the immune response data are zero-inflated and skew. The univariate and multivariate Tweedie GLMEMs are expected to perform better than the LMEMs and NLMEMs as the Tweedie distribution has a probability mass at zero and can accommodate skew distributions.

2.1 LINEAR MIXED EFFECT MODELS

The theory presented here is a summary of that presented in Chapters 1–5 of Pinheiro and Bates (2000).

LMEMs are models that include both fixed effects and random effects which are associated linearly with the response variable. The fixed effects are included to capture the effects which would be repeated if the experiment was to be repeated; e.g. in our case the vaccine types are fixed effects as these would be included if the experiments were to be redone. The random effects are included to capture the effects which arise from us drawing a random sample from a population; e.g. in our case the subjects included in the experiment are drawn from a population of potential subjects and if we were to repeat the experiment we would use a different random sample of subjects. The random effects are intended to capture the inter-subject variation that arises from the subjects having different reactions to the vaccines, and in addition the random effects also capture the correlation in the subjects' measurements over time. The theory presented below starts by presenting linear models and then progressing to discuss LMEMs.

The typical linear model relates the n_i-dimensional dependent vector $\boldsymbol{y}_i$ for subject i to a $n_i \times p$ matrix of covariates, independent variables or predictors, $\boldsymbol{X}_i$ as follows

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\beta}$ is the p-dimensional parameter vector of fixed effects which is to be estimated and where typically $\boldsymbol{\epsilon}_i : n_i \times 1 \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$. This model assumes that the observations are independent and normally distributed, i.e.

$$\boldsymbol{y}_i \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{n_i}).$$

The model does not capture the covariance structure induced by the grouping which is typical of longitudinal data; i.e. each of the observations for subject i is assumed to be independent in the above model whereas in a longitudinal setting we would expect them to be correlated.

2.1.1 *Single Level of Grouping*

LMEMs extend the above linear model by including random effects. The models are specified in two stages with the random effects becoming apparent in the second stage.

Suppose the grouping is at the subject level, then the stage one (observation level) formulation of the model for the observations from subject i, y_{ij} for $j = 1, 2, \ldots, n_i$, is given by

$$y_{ij} = \boldsymbol{z}_{ij}^{\intercal}\boldsymbol{\beta}_i + \epsilon_{ij}$$

where y_{ij} is the observation for subject i at time j, $\boldsymbol{z}_{ij}$ is a $q \times 1$ vector of observation level covariates, $\boldsymbol{\beta}_i$ is a $q \times 1$ vector of subject specific coefficients, and ϵ_{ij} is the within-subject error. The vector $\boldsymbol{z}_{ij}$ typically has a one in the first position and then subsequent entries in the vector are for variables which can vary within the subject which in our case will be the time variable. The within-subject error is typically assumed to be such that $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$.

The second stage of the model formulation is given by

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\phi} + \boldsymbol{b}_i$$

where $\boldsymbol{A}_i$ is a matrix of subject level covariates which affect the means of the coefficients, $\boldsymbol{\phi}$ is a vector of coefficients to be estimated, and $\boldsymbol{b}_i$ is the vector of subject specific random effects which give the random deviations from the mean. Typically it is assumed that $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi})$. These random effects model the between subject variability.

The two stages can be combined into a single formulation as follows. Let $\boldsymbol{y}_i$ be the $n_i \times 1$ vector of observations y_{ij}, then we have

$$\boldsymbol{y}_i = \boldsymbol{Z}_i\boldsymbol{A}_i\boldsymbol{\phi} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{Z}_i$ is the matrix with jth row equal to $\boldsymbol{z}_{ij}^{\intercal}$ and $\boldsymbol{\epsilon}_i$ is the vector of within-subject errors ϵ_{ij}.

The above specification can be generalised by using an arbitrary design matrix for the fixed effects. Specifically we can write

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\phi} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{X}_i$ is the $n_i \times p$ design matrix for the fixed effects $\boldsymbol{\phi}$.

We can show that the inclusion of the random effects induces correlation between the observations for subject i by finding the covariance between the jth and kth observations for subject i, $\mathrm{cov}(y_{ij}, y_{ik})$, as follows

$$\begin{aligned}
\mathrm{cov}(y_{ij}, y_{ik}) &= \mathrm{cov}(\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\phi} + \boldsymbol{z}_{ij}^{\intercal}\boldsymbol{b}_i + \epsilon_{ij}, \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\phi} + \boldsymbol{z}_{ik}^{\intercal}\boldsymbol{b}_i + \epsilon_{ik}) \\
&= \mathrm{cov}(\boldsymbol{z}_{ij}^{\intercal}\boldsymbol{b}_i + \epsilon_{ij}, \boldsymbol{z}_{ik}^{\intercal}\boldsymbol{b}_i + \epsilon_{ik}) \\
&= \mathrm{cov}(\boldsymbol{z}_{ij}^{\intercal}\boldsymbol{b}_i, \boldsymbol{z}_{ik}^{\intercal}\boldsymbol{b}_i) + \mathrm{cov}(\epsilon_{ij}, \epsilon_{ik}) \\
&= \begin{cases} \psi + \sigma^2 & \text{if } j = k \\ \psi & \text{if } j \neq k \end{cases} \qquad (1)
\end{aligned}$$

where, in the last line, we assumed that $\boldsymbol{b}_i$ has one element and that $\boldsymbol{Z}_i$ is a column of ones. Under the original linear model that we

introduced, the covariance between different observations for subject i would have been zero, and as such we can see that the introduction of the random effects has induced correlation between the different observations of subject i.

The within-subject error $\boldsymbol{\epsilon}_i$ can have its distribution made more general to allow for heteroscedasticity and correlation. Specifically the within-subject or intra-subject variability is captured by $\boldsymbol{\epsilon}_i$ which we now assume to have a more general multivariate normal distribution, i.e.

$$\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$$

where the $\boldsymbol{\Lambda}_i$ are positive-definite matrices which are parameterised by a small number of parameters. This extension is intended to allow for the intra-subject errors to be heteroscedastic and or correlated. For $\boldsymbol{\Lambda}_i$, we consider the following form $\boldsymbol{\Lambda}_i = \boldsymbol{V}_i \boldsymbol{C}_i \boldsymbol{V}_i$ which allows us to split out the variance and correlation structures into $\boldsymbol{V}_i$ and $\boldsymbol{C}_i$ respectively. The matrix $\boldsymbol{V}_i$, the variance matrix, is a positive diagonal matrix and $\boldsymbol{C}_i$, the correlation matrix, is a positive definite matrix with all diagonal elements equal to one.

There are various possible forms for $\boldsymbol{V}_i$ and $\boldsymbol{C}_i$. The general form of $\boldsymbol{V}_i$ is given by

$$\mathrm{var}(\epsilon_{ij}) = \sigma^2 g(\mu_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta})$$

where $\mu_{ij} = \mathrm{E}(y_{ij}|\boldsymbol{b}_i)$, $\boldsymbol{v}_{ij}$ is a vector of covariates, $\boldsymbol{\delta}$ is a vector of variance parameters, and $g(\cdot)$ is a variance function. The form that we consider in our applications is:

$$\mathrm{var}(\epsilon_{ij}) = \sigma^2 \mu_{ij},$$

which allows the variance to vary with the expected mean value of the response. Various other structures for variance matrix and structures for the correlation matrix are available; see Chapter 5 of Pinheiro and Bates (2000).

2.1.2 *Extensions to Multiple Responses*

Our approach to modelling the multiple immune responses in the data that we have is to stack the multiple responses into a single column vector. The particular response is indicated by including an additional factor covariate that identifies the response. In this way we have the same model structure as outlined above for the single-level grouping except with an additional covariate which gives the immune response we are referring to; i.e. we now have the response y_{ijk} which is for individual i, response j, and occasion k where $i = 1, \ldots, 37$, $j =$ `CD4Gn2nTp`, `CD4Gn2pTn`, `CD4Gn2pTp`, `CD4Gp2nTn`, `CD4Gp2nTp`, `CD4Gp2pTn`, and `CD4Gp2pTp`, and $k = 1, \ldots, n_{ij}$, and a covariate indicating the value of j in the covariate matrix $\boldsymbol{X}_{ij}$.

The inclusion of the j is not necessary, but we do so as this makes the additional level within the data clear; i.e. the data described above

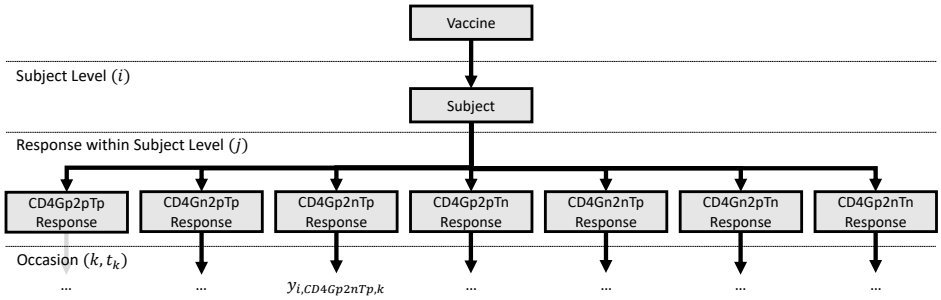


Figure 10: A detailed illustration of the multilevel nature of the vaccine data for a specific time point for the CD4 cell frequencies for combinations of the cytokines IFN-γ, IL-2, and TNF-α.

now has multiple levels. Specifically, the first level in the data is the subject level, within each subject level we have a response level, and then within each response level, we have the occasions at which the response was recorded. The multilevel nature of the data is illustrated in Figure 10 where the levels in data are clearly elucidated.

Rearranging the data as described above allows us to include interaction terms between the various other covariates and the response covariate if needed. As such, if it is needed, we can have unique coefficients for each immune response and effectively have separate models for each of the different response profiles. This allows for a significant amount of flexibility in the models that we consider.

The above structure also allows us to include correlations between the within-group errors across the different responses; i.e. the matrix $\mathbf{\Lambda}_i$ will consist of seven submatrices $\mathbf{\Lambda}_{i(j)}$ along the diagonal each relating to a specific response j which is the covariance structure for that response. The remaining elements of $\mathbf{\Lambda}_i$ are the within-subject between response covariance structures and so the model can allow for within-subject between response correlations. Estimating the complete general covariance structure can require estimating a large number of parameters; we struggled to fit models with overly complex within-subject covariance structures and as such kept them relatively simple.

This extension to multiple responses also applies to NLMEMs and GLMEMs.

2.1.3 *Two Levels of Grouping*

The above model can be extended to the case where we have two levels of grouping. Suppose now we observe y_{ijk} which is the kth observation for the ith subject, the first-level grouping, at the jth level

of the second-level grouping, for $i = 1, \ldots, M$, $j = 1, \ldots, M_i$, and $k = 1, \ldots, n_{ij}$. The model with two levels of grouping is then given by

$$\boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{i,j}\boldsymbol{b}_i + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad \text{for } i = 1, \ldots, M, \quad j = 1, \ldots, M_i$$

where $\boldsymbol{X}_{ij} : n_{ij} \times p$ is the fixed effect model matrix, $\boldsymbol{b}_i : q_1 \times 1$ is the first-level random effect with associated model matrix $\boldsymbol{Z}_{i,j}$, and $\boldsymbol{b}_{ij} : q_2 \times 1$ is the second-level random effect with associated model matrix $\boldsymbol{Z}_{ij}$. We assume that

$$\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}_2),$$

and

$$\boldsymbol{\epsilon}_{ij} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{\Lambda}_i).$$

The random effect $\boldsymbol{b}_i$ is assumed to be independent for different i, $\boldsymbol{b}_{ij}$ is independent of the first-level random effect and independent for different i or j, and the within-group errors are assumed to be independent for different i or j and independent of the random effects.

In the data set we are considering, we can have at most two groupings. Specifically, the first-level grouping is at the subject level and the second-level grouping is at the response within subject level. The reason for including the second level of grouping is to account for the subject specific variation at the response level.

While we need only two levels of grouping, LMEMs can be extended to include further levels of grouping; we do not pursue this further here, but extensions to higher levels of grouping can be seen in the work of Pinheiro and Bates (2000), for example.

2.2 NONLINEAR MIXED EFFECT MODELS

The theory presented here is a summary of that presented in Chapters 6–8 of Pinheiro and Bates (2000) and that presented by Davidian and Giltinan (2003).

LMEMs have fixed and random effects which enter the model linearly, but can be used to model nonlinear profiles by the inclusion of polynomial and spline terms, for example. Nonlinear mixed effect models (NLMEMs), on the other hand, have at least one of the fixed or random effects which enters the model in a nonlinear manner. As such NLMEMs, typically attempt to explicitly model the nonlinear profiles of the subjects by using a nonlinear function $f(\cdot)$. The primary advantages of using NLMEMs as opposed to LMEMs with polynomial or spline terms are:

- NLMEMs tend to be more efficient in the sense that we can use fewer parameters to model the nonlinear profile; i.e. they are parsimonious.

- the form of $f(\cdot)$ can often be mechanistic or semi-mechanistic in the sense that it is intended to mimic the underlying mechanisms that generate the nonlinear profile and as a result the parameters can often have explicit interpretation and theoretical meaning. LMEMs with polynomials and splines are purely empirical models relating the response to the covariates.
- the NLMEMs are more likely to be valid beyond the range for which we have data when compared to using a LMEM with a polynomial or spline, for example.

There are disadvantages to NLMEMs. The main one is that their estimation can be quite difficult and very dependent on the starting values provided. There are several strategies suggested in the literature for arriving at possibly useful starting values, e.g. see the work of Mauff (2011) where several strategies are outlined.

In the following subsections we define NLMEMs with one and two levels of grouping.

2.2.1 *Single Level of Grouping*

The stage one specification of the NLMEM for the jth observation of subject i is given by

$$y_{ij} = f(\boldsymbol{x}_{ij}, \boldsymbol{\phi}_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i$$

where $f(\cdot)$ is a general real-valued nonlinear differentiable function of the group-specific parameter $\boldsymbol{\phi}_{ij}$ and the covariate vector $\boldsymbol{x}_{ij}$. The group-specific parameter is modelled as follows

$$\boldsymbol{\phi}_{ij} = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i$$

where $\boldsymbol{\beta} : p \times 1$ is the vector of fixed effects and $\boldsymbol{b}_i : q \times 1$ is the vector of random effects. The model matrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ contain values specific to the group and can include covariate values. The subscript j is included for these matrices as they can include time varying covariates. Alternatively, the subject specific parameters may be modelled as

$$\boldsymbol{\phi}_{ij} = \mathrm{d}(\boldsymbol{A}_{ij}, \boldsymbol{\beta}, \boldsymbol{b}_i)$$

where $\mathrm{d}(\cdot)$ is a p dimensional function depending on a fixed effects vector $\boldsymbol{\beta}$ and the random effects vector $\boldsymbol{b}_i$. This is the stage two specification of the NLMEM and captures both the systematic inter-individual variation due to different covariates and the random or unexplained variation due to the subjects being a sample from the population of potential subjects.

Again we assume that

$$\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_i),$$

and, as outlined above for the LMEMs, $\boldsymbol{\Lambda}_i = \boldsymbol{V}_i \boldsymbol{C}_i \boldsymbol{V}_i$ with various possible options for the variance and correlation structures.

The function $f(\cdot)$ determines the profile for the ith individual via the individual specific parameters $\boldsymbol{\phi}_{ij}$. It is in this way that the function determines the inter-subject variability by relating the subject covariates $\boldsymbol{x}_{ij}$ to their response via the subject specific parameters. The form of $f(\cdot)$ can be chosen for (1.) theoretical reasons or (2.) empirical reasons:

1. the form of $f(\cdot)$ may be chosen for theoretical mechanistic reasons whereby it mimics the underlying theoretical mechanisms of the processes which produce the nonlinear profiles of the responses. In this case the parameters are likely to have scientific meaning and be of interest.

2. the form of $f(\cdot)$ may also be chosen as empirically it provides a good representation of the nonlinear profiles of the responses. In this case the parameters are less likely to have explicit scientific meaning, but may still be of interest.

In our applications in Chapter 3, the forms for $f(\cdot)$ that we consider are primarily motivated from an empirical point of view, but also have mechanistic interpretations.

EXAMPLE To make the ideas more tangible, we consider a simple example here initially with only the time covariate t_j. Suppose we have observations y_{ij}, for $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$, and suppose we wish to model these observations by using a modified version of the transit-compartment model of Savic *et al.* (2007). Specifically the models is given by

$$y_{ij} = \phi_{0,i} + \phi_{1,i}(\phi_{2,i} t_j) \exp(-\phi_{3,i} t_j) + \epsilon_{ij}$$

where

$$\underbrace{\begin{bmatrix} \phi_{0,i} \\ \phi_{1,i} \\ \phi_{2,i} \\ \phi_{3,i} \end{bmatrix}}_{\boldsymbol{\phi}_{ij}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{A}_{ij}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{B}_{ij}} \underbrace{\begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix}}_{\boldsymbol{b}_i}$$

and

$$\boldsymbol{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} \sim \mathrm{N}\left(\boldsymbol{0}, \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} \\ \psi_{21} & \psi_{22} & \psi_{23} & \psi_{24} \\ \psi_{31} & \psi_{32} & \psi_{33} & \psi_{34} \\ \psi_{41} & \psi_{42} & \psi_{43} & \psi_{44} \end{bmatrix} \right).$$

We can extend the above model to allow for a vaccine fixed effect. Suppose for this example that we have two vaccines and each subject can be assigned to one of the two vaccines. We want to allow the

baseline response parameter $\phi_{0,i}$ to vary by the vaccine fixed effect. We can do this by extending the above model as follows

$$\underbrace{\begin{bmatrix} \phi_{0,i} \\ \phi_{1,i} \\ \phi_{2,i} \\ \phi_{3,i} \end{bmatrix}}_{\boldsymbol{\phi}_{ij}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & \text{vaccine}_i \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}}_{\boldsymbol{A}_{ij}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{B}_{ij}} \underbrace{\begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix}}_{\boldsymbol{b}_i}$$

and

$$\boldsymbol{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} \sim \mathrm{N}\left(\boldsymbol{0}, \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} \\ \psi_{21} & \psi_{22} & \psi_{23} & \psi_{24} \\ \psi_{31} & \psi_{32} & \psi_{33} & \psi_{34} \\ \psi_{41} & \psi_{42} & \psi_{43} & \psi_{44} \end{bmatrix}\right),$$

where vaccine_i is an indicator variable taking a value of one when the subject has been given Vaccine B and zero when the subject has been given Vaccine A. The baseline response parameter $\phi_{0,i}$ now varies according to the vaccine that was administered to subject i.

The extension to multiple responses outlined earlier for LMEMs in Subsection 2.1.2 applies to NLMEMs as well in a similar manner. In the next subsection we will use this extension when we discuss the extension to two levels of grouping.

2.2.2 *Two Levels of Grouping*

Extensions to two or more levels of grouping are relatively straightforward. Again suppose that we have the observation y_{ijk} which is for subject i, the first-level of grouping, at the jth level of the second-level grouping and is the kth such observation, where $i = 1, \ldots, M$, $j = 1, \ldots, M_i$ and $k = 1, \ldots, n_{ij}$. Then the stage one specification of the model with two levels of grouping is given by

$$y_{ijk} = f(\boldsymbol{x}_{ijk}, \boldsymbol{\phi}_{ijk}) + \epsilon_{ijk}$$

where $f(\cdot)$ is as before, $\boldsymbol{x}_{ijk}$ is the vector of covariates, $\boldsymbol{\phi}_{ijk}$ is the vector of group-specific covariates, and

$$\boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_i).$$

The stage two specification is such that the group-specific parameter $\boldsymbol{\phi}_{ijk}$ is now modelled as

$$\boldsymbol{\phi}_{ijk} = \boldsymbol{A}_{ijk}\boldsymbol{\beta} + \boldsymbol{B}_{i,jk}\boldsymbol{b}_i + \boldsymbol{B}_{ij,k}\boldsymbol{b}_{ij}$$

where

$$\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}_2).$$

The vector $\boldsymbol{b}_i : q_1 \times 1$ is the first-level random effect vector which is independent of the second-level random effect vector $\boldsymbol{b}_{ij} : q_2 \times 1$.

Both of the random effect vectors are independent of the within-group errors $\boldsymbol{\epsilon}_i$. The random effect model matrices $\boldsymbol{B}_{i,jk}$ and $\boldsymbol{B}_{ij,k}$ depend on the first and second-level groups and possibly values of the covariates which may be time varying. The vector $\boldsymbol{\beta} : p \times 1$ is the vector of fixed effects with the model matrix $\boldsymbol{A}_{ijk}$ that incorporates the covariates which may also be time varying.

2.3 GENERALISED LINEAR MIXED EFFECT MODELS

The theory that we present for generalised linear mixed effect models (GLMEMs) here is a summary of the general theory presented by Fitzmaurice *et al.* (2008, ch. 4) and the preceding review of generalised linear models (GLMs) and the more general exponential dispersion family models is a summary of that presented by McCullagh and Nelder (1989) and Jørgensen (1987). The details related to the univariate Tweedie Compound Poisson Linear Mixed Effect Models (Tweedie GLMEMs) are a summary of that presented by Zhang (2013) and the details of the multivariate Tweedie Compound Poisson Linear Mixed Effect Models (Tweedie M-GLMEMs) are a summary of that presented by Hui (2016) and Warton *et al.* (2015).

GLMEMs are an extension of generalised linear models where we now include random effects. We review the high level theory of exponential dispersion models, which GLMs are an example of, before discussing GLMEMs. We then discuss univariate Tweedie GLMEMs before moving on to the multivariate case.

2.3.1 *Review of Exponential Dispersion Models*

Exponential dispersion models (Jørgensen, 1987) are a general class of models that encompass a wide range of practically useful models including GLMs. The two-parameter formulation of the univariate model that Jørgensen (1987) considers is given by

$$p(y|\theta,\lambda) = a(y,\zeta)\exp\left(\lambda\{y\theta - \kappa(\theta)\}\right), \quad y \in \mathbb{R} \tag{2}$$

where a and κ are known functions, θ is the natural parameter, and $\lambda \in \mathbb{R}^+$ is the inverse of dispersion parameter ζ, i.e. $\zeta = \frac{1}{\lambda}$. The above model is a GLM if the $y_1, \ldots, y_n$ are independent and has parameters λ and $\theta_i = g(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})$ where $g(\cdot)$ is a monotonic link function, $\boldsymbol{x}_i$ is a vector of covariates associated with y_i and $\boldsymbol{\beta}$ is a parameter vector.

McCullagh and Nelder (1989) and Fitzmaurice *et al.* (2008) give the following properties of the above model

$$\mu = \mathrm{E}(y) = \kappa'(\theta) \equiv \frac{\partial \kappa(\theta)}{\partial \theta}$$

and

$$\mathrm{Var}(y) = \zeta\kappa''(\theta) = \zeta V(\mu)$$

where, for the variance of y, we replace $\kappa''(\theta)$ by $V(\mu)$, the variance function. The focus of the work here are the models with the power variance functions with

$$V(\mu) = \mu^p.$$

In general, for this class of models

$$p \in \{(-\infty, 0] \cup [1, \infty)\};$$

the case where $p \in (0,1)$ does not correspond to an exponential dispersion model. This class of models, the exponential dispersion models with power variance functions, includes several common models:

- $p = 0$: the normal distribution,
- $p = 1$: the Poisson distribution,
- $p = 2$: the Gamma distribution,
- $p = 3$: inverse Gaussian distribution, and
- $p \in (1,2)$: compound Poisson distributions;

see Jørgensen (1987) for details of the distributions for the remaining values of p. Our particular focus is on the models with $p \in (1,2)$, as the compound Poisson distributions are continuous for $y > 0$ and have a positive probability of $y = 0$. As seen in Section 1.3, the data we are considering have a nonneglible proportion of zeros which we need to account for. Following convention, we refer to these models as Tweedie models given the work of Tweedie (1984).

The compound Poisson model can be written as a combination of two exponential dispersion models; i.e. as the sum of Poisson Gammas. Specifically, it is the combination of a Poisson and Gamma model with

$$Y = \sum_{i=1}^{T} D_i \quad \text{with } T \sim \text{Poisson}(\eta) \text{ and } D_i \sim \text{Gamma}(\alpha, \gamma) \tag{3}$$

where T and D_i are independent. This representation can be related to that in (2) as follows

$$\begin{aligned} \mu &= \eta\alpha\gamma, \\ p &= \frac{\alpha+2}{\alpha+1}, \text{ and} \\ \zeta &= \frac{\eta^{1-p}(\alpha\gamma)^{2-p}}{2-p}. \end{aligned}$$

These equations are derived by equating the cumulant generating functions for (2) and (3) (Zhang, 2013).

For (3), we can write down the joint density of T and Y which is given by

$$\begin{aligned} p(y,t|\eta,\alpha,\gamma) &= p(y|t,\eta,\alpha,\gamma)p(t|\eta) \\ &= \begin{cases} \exp(-\eta) & \text{for } y=0, t=0 \\ \frac{y^{t\alpha-1}\exp(-y/\gamma)}{\Gamma(t\alpha)\gamma^{t\alpha}}\frac{\eta^t\exp(-\eta)}{t!} & \text{for } y\in\mathbb{R}^+, t\in\mathbb{Z}^+ \end{cases}. \end{aligned}$$

To retrieve the marginal density as in Equation (2) we need to sum over t

$$p(y|\eta,\alpha,\gamma) = \sum_{t=0}^{\infty} p(y,t|\eta,\alpha,\gamma).$$

From here the function $a(y,\zeta,p)$ can be found; see the details provided by Zhang (2013). It turns out that $a(y,\zeta,p)$ is an infinite sum, and to evaluate it the methods proposed by Dunn and Smyth (2005) are used in the `R` package `cplm` (Zhang, 2013; R Core Team, 2019). As the summands are positive, the methods of Dunn and Smyth (2005) involve identifying the index of the summand which is likely to be the maximum, and then summing over the summands with indices below and above that of the likely maximum for summands larger than some small threshold. In this way the infinite sum is reduced to a finite sum over the terms which are likely to contribute the most.

For the Tweedie model we use a log-link to relate the mean μ to the covariates. That is for subject i we have

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{x}_i^\intercal\boldsymbol{\beta}$$

and in matrix notation for all subjects,

$$g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$$

where the ith row of $\boldsymbol{X}$ equals $\boldsymbol{x}_i^\intercal$. The power parameter p and dispersion parameter ζ are kept constant across the different subjects i. In the next subsection we extend this model to include random effects.

2.3.2 *Univariate Mixed Model Description*

Single Level of Grouping

To extend the above models to include random effects is in principle straightforward. We first do this for a single-level of grouping.

Let us consider $\boldsymbol{y}_i$, which is the response vector for subject i where the jth element is given by y_{ij}, the jth observation for subject i, for $i = 1,\ldots,M$ and $j = 1,\ldots,n_i$. Then we can include random effects in the linear predictor as follows

$$g(\boldsymbol{\mu}_i) = \log(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \quad \text{for } i = 1,\ldots,M$$

where $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{y}_i)$, $\boldsymbol{X}_i$ is the model matrix associated with the fixed effects $\boldsymbol{\beta}$, and $\boldsymbol{Z}_i$ is the model matrix associated with the random effects $\boldsymbol{b}_i$. As before we have

$$\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}).$$

For the GLMEMs, we assume that the observations are conditionally independent given the covariates and the random effects and that the conditional distributions are given by (2); i.e. the conditional density for y_{ij} is given by

$$f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = a(y_{ij}, \zeta) \exp\left(\zeta^{-1}\{y_{ij}\theta_{ij} - \kappa(\theta_{ij})\}\right), \tag{4}$$

where θ_{ij} is a function of the linear predictor $\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i$.

The approach to modelling the multiple immune responses in the data that we have is again to stack the multiple responses into a single column vector and the particular response is indicated by including an additional factor covariate that identifies the response. This means that we can use the above GLMEMs to model all of the immune responses at once in a similar manner to what we described for the LMEMs in Subsection 2.1.2. We make use of this extension in the next subsection when we discuss two levels of grouping.

Two Levels of Grouping

To extend the above to two levels of grouping can be done as follows. Let us consider $\boldsymbol{y}_{ij}$ which is the response vector for subject i at the second-level of grouping j where the kth element is given by y_{ijk}, for $i = 1, \ldots, M$ and $j = 1, \ldots, M_i$, and $k = 1, \ldots, n_{ij}$. We can include random effects at both the subject-level and the grouping-level j in the linear predictor as follows

$$g(\boldsymbol{\mu}_{ij}) = \log(\boldsymbol{\mu}_{ij}) = \\ \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{i,j}\boldsymbol{b}_i + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij} \quad \text{for } i = 1, \ldots, M, \quad j = 1, \ldots, M_i$$

where $\boldsymbol{\mu}_{ij} = \mathrm{E}(\boldsymbol{y}_{ij})$, $\boldsymbol{X}_{ij}$ is the model matrix associated with the fixed effects $\boldsymbol{\beta}$, and $\boldsymbol{Z}_{i,j}$ and $\boldsymbol{Z}_{ij}$ are the model matrices associated with the random effects $\boldsymbol{b}_i$ and $\boldsymbol{b}_{ij}$ respectively. As before we have

$$\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}_1) \quad \text{and} \quad \boldsymbol{b}_{ij} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\psi}_2)$$

where the random effect at the subject-level grouping and that at the second-level grouping are independent.

The inclusion of random effects is in principle straightforward, but it does make the estimation of parameters (discussed in Section 2.4) and the calculation of marginal predictions more challenging. Marginal

prediction is more challenging because to calculate $\mathrm{E}(\boldsymbol{y}_{ij})$ we need to integrate out the random effects, i.e. for y_{ijk} we need to calculate

$$\mathrm{E}\left(y_{ijk}|\boldsymbol{x}_{ijk}, \boldsymbol{z}_{i,jk}, \boldsymbol{z}_{ijk}\right) = \int\int g^{-1}(\boldsymbol{x}_{ijk}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{i,jk}^{\mathsf{T}}\boldsymbol{b}_i + \boldsymbol{z}_{ijk}^{\mathsf{T}}\boldsymbol{b}_{ij})\xi(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\psi}_1)\,\mathrm{d}\boldsymbol{b}_i\,\xi(\boldsymbol{b}_{ij}; \boldsymbol{0}, \boldsymbol{\psi}_2)\,\mathrm{d}\boldsymbol{b}_{ij}$$

where $\xi(\cdot; \boldsymbol{a}, \boldsymbol{b})$ is the density of a multivariate normal with mean $\boldsymbol{a}$ and covariance $\boldsymbol{b}$. This marginal mean is often referred to as the population averaged mean as it gives the mean for an entire subpopulation or population strata with the specific covariate values. The conditional mean given the covariate values and the random effects is much easier to calculate, and is referred to as the subject specific mean.

2.3.3 *Multivariate Mixed and Latent Variable Model Description*

In this section we describe a multivariate mixed and latent variable model extension of the above univariate GLMEMs; we will refer to these models as M-GLMEMs. By a multivariate model we mean that we model the response matrix $\boldsymbol{Y}$ which has $N = \sum_i n_{ij}$ rows (where now we assume $n_{ij} = n_{ij'} = n_i$ for all $j \neq j'$) corresponding to each occasion k for subject i and the columns respond to the second-level of grouping j, the immune responses in our case. In particular we consider the models implemented by Hui (2016) in the `R` package `boral` (Hui, 2018). These models were developed for ecological applications where they are used to model multivariate abundance data; see for example the work of Warton *et al.* (2015), Letten *et al.* (2015), Hui *et al.* (2015), Hui (2016), and Jamil *et al.* (2013). In the multivariate abundance applications the rows of the response matrix correspond to the different sites where abundance data are collected and the columns correspond to the different species for which abundance is measured. Abundance can mean several different things depending on the application; for example, abundance may mean the counts of organisms, the presence or absence of organisms, or the relative abundance (composition) for a site.

Warton *et al.* (2015) make the distinction between multivariate GLMEMs and multivariate latent variable (LVM) models. Both of these models are extensions to exponential dispersion models. In the definitions given by Warton *et al.* (2015), the multivariate GLMEMs model the correlation across the columns of $\boldsymbol{Y}$ by using a multivariate normal random effect whereas the LVMs model the correlation by using loadings on a latent low dimension multivariate normal random variable. Warton *et al.* (2015) argue that the LVMs have several advantages over the GLMEMs; the most important being that the correlation across the columns of $\boldsymbol{Y}$ can be modelled with fewer parameters which makes the model estimation more likely to converge. The LVMs also allow for model based ordination to be performed where by ordina-

tion we mean the visualisation of the main structures in a multivariate response matrix $\boldsymbol{Y}$ in a low dimension space (e.g. one, two or three dimensions).

Model Structure

The form of the model that we consider here is given by

$$g(\mu_{ijk}) = b_i + \beta_{0j} + \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\beta}_j + \boldsymbol{z}_{ik}^{T}\boldsymbol{\theta}_j \quad \text{for } i = 1,\ldots,M,\ j = 1,\ldots,J, \text{ and } k = 1,\ldots,n_{i,} \quad (5)$$

where $\mu_{ijk} = \mathrm{E}(y_{ijk})$, y_{ijk} is the jth immune response for subject i at occasion k (t_k), $\boldsymbol{x}_{ik}$ is a vector of occasion specific covariates for subject i at occasion k (t_k), $b_i \sim \mathrm{N}(0, \sigma^2)$ is the subject specific random effect, and $\boldsymbol{z}_{ik} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ is the latent variable which, along with the factor loadings, induces correlation between the responses j of subject i. The parameter β_{0j} is the intercept for immune response j, $\boldsymbol{\beta}_j$ is the vector of parameters multiplying the subject and occasion specific covariates $\boldsymbol{x}_{ik}$, and $\boldsymbol{\theta}_j$ is the vector of parameters (factor loadings) multiplying the latent variables $\boldsymbol{z}_{ik}$ and which determines the correlation between the responses. These models are strictly an extension to the LVM models considered by Warton *et al.* (2015) as b_i is a subject specific random effect here as opposed to being a subject specific fixed effect.

The latent variables and factor loadings in the model given in Equation (5) allow us to model the correlation across the different immune responses in an efficient manner. Including the $\boldsymbol{z}_{ik}^{T}\boldsymbol{\theta}_j$ term in the model induces correlation across the different immune responses because the latent variable $\boldsymbol{z}_{ik}$ is common to all immune responses for subject i at occasion k. The degree of correlation between the immune responses is controlled by the factor loadings $\boldsymbol{\theta}_j$. An important point to note is that if we use a ω-dimensional latent variable, we will need to estimate ωJ parameters to model the correlation across the J immune responses. This can be contrasted against the number of parameters that need to be estimated when including random effects in a multivariate GLMEM to model the correlation between the immune responses. Such a multivariate GLMEM would typically have the form

$$g(\mu_{ijk}) = b_i + \beta_{0j} + \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\beta}_j + u_{ij}$$

where the corresponding terms are the same as those in Equation (5) and u_{ij} is an element of a multivariate normal distribution such that

$$\boldsymbol{U}_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{ij} \\ \vdots \\ u_{iJ} \end{pmatrix} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

In such a case, the correlation across the immune responses is modelled by using a multivariate normal random variable and will involve estimating $\mathbf{\Sigma}$ which has $\frac{1}{2}J(J-1)$ parameters. As such, the use of latent variables as in Equation (5) is an efficient manner of modelling the correlation across immune responses relative to the use of random effects in multivariate GLMEMs. As mentioned, this is one of the main advantages of the models that we consider.

In the applications in Chapter 3, the models that we consider will use a Tweedie distribution for each of the responses with a log link function. A separate estimate of the dispersion parameter ζ_j will be found for each of the response columns and one overall estimate for the power parameter p will be found.

Including Traits in the Model

The above model can be extended to include traits for the responses j. In the species abundance setting, where these models originated and where the columns correspond to specific species, the traits are species specific characteristics which we can include in the model. For example in our setting where the columns correspond to different combinations of cytokine responses, the traits are the specific characteristics of the combinations of cytokine responses.

The inclusion of traits can simplify the overall model by reducing the number of parameters to be estimated as the coefficients are now modelled as random effects. Specifically if each of the responses has a set of traits given by trait_j, then we can model the coefficients β_{0j} and $\boldsymbol{\beta}_j$ as follows

$$\beta_{0j} \sim \text{N}(\kappa_{0j} + \text{trait}_j^\intercal \boldsymbol{\kappa}_0, \sigma_0^2) \tag{6}$$

and

$$\beta_{jk} \sim \text{N}(\kappa_{0k} + \text{trait}_j^\intercal \boldsymbol{\kappa}_k, \sigma_k^2) \tag{7}$$

where β_{jk} is the kth element of $\boldsymbol{\beta}_j$. The above relate the coefficients β_{0j} and $\boldsymbol{\beta}_j$ linearly to the trait vectors via the coefficients $\boldsymbol{\kappa}_k$ $(k = 0, 1, \ldots)$ which will be estimated. There will be deviations from the above linear relationships and these will be accounted for as the β coefficients are normally distributed with standard errors σ_k $(k = 0, 1, \ldots)$.

Modelling Correlation

The model allows us to calculate the residual correlation / covariation between the responses j after accounting for the predictors. The residual covariance matrix is calculated as $\mathbf{\Theta}\mathbf{\Theta}^\intercal$ where

$$\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)^\intercal.$$

As mentioned, the residual correlation is modelled by the latent variables and their loadings, and may be attributed to several different factors. For example, the residual correlation may be due to

- joint response to unmeasured covariates, and
- biotic reactions, such as competition and facilitation (Warton *et al.*, 2015).

By modelling the residual correlation / covariation, we can identify what proportion of covariation between the responses j is due to the measured covariates. This can be done by comparing the trace of the residual covariance matrix of a model without the covariates to the trace of the residual covariance matrix of a model with the covariates (Legendre *et al.*, 2005). The reduction in the trace when we include the covariates tells us what proportion of the covariation the covariates are responsible for.

Model-Based Ordination

The inclusion of the latent variables allows us to produce ordination plots. These ordinations are model-based as they have an explicit statistical model underlying them.

Model-based ordination is an alternative to algorithmic approaches such as nonmetric multidimensional scaling and correspondence analysis. There are several advantages to model-based ordination relative to algorithmic methods. These advantages are discussed by Hui *et al.* (2015) and include:

- residual analysis is available for model checking,
- model selection tools are available,
- methods for formal statistical inference are available, and
- improved efficiency.

The number of axes in the ordination plots is determined by the dimension of the latent variable. Typically a latent variable with two or three dimensions is chosen so that a low dimension representation of the data can be produced. In the case of the model we have introduced above, the ordination plot involves constructing a scatter plot by plotting the estimated latent variable values (there is one for each row of the response matrix $\boldsymbol{Y}$) and a biplot is created by adding the points corresponding to the factor loadings (there is one for each column of $\boldsymbol{Y}$). The interpretation of these plots is similar to that of other unconstrained ordination biplots: clustering of the row points indicates similarity of those observations and the biplot axes give the trends in the data (Greenacre, 2010).

Comparison to Univariate GLMEMS for Multiples Responses

The univariate GLMEMs that we introduced and extended for multiple responses are relatively simple when compared to the M-GLMEMs

we have outlined. The main difference between these two models is that the univariate GLMEMs for multiple responses do not model the residual covariation across the multiple responses. That is the univariate GLMEMs allow for covariation in the responses via the covariates and shared coefficient estimates, but any residual covariation that is not captured by the covariates is left unmodelled. The M-GLMEMs account for this covariation explicitly by including the latent variables and their loadings.

There are other less significant differences, for example,

- the univariate GLMEMs for multiple responses do not require us to estimate separate coefficients for each of the responses where as the M-GLMEM without traits does.
- the univariate GLMEMs we have introduced do not consider the response traits. These could easily be included, but may not simplify the model.

2.4 MODEL ESTIMATION

In this section we outline the estimation methods that we use for the models that we fit in Chapter 3. The inclusion of the random effects and latent variables mean that estimation of the model parameters is not straightforward and requires some ingenuity.

2.4.1 *Linear Mixed Effect Models*

The methods presented here outline those given by Pinheiro and Bates (2000, ch. 2). The estimation methods that we use for the LMEMs are likelihood based methods. For the single-level LMEM, the marginal likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi}|\boldsymbol{y}) &= \prod_{i=1}^{M} p(\boldsymbol{y}_i|\boldsymbol{\beta}, \sigma, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{M} \int p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \sigma) p(\boldsymbol{b}_i|\boldsymbol{\psi}) \, \mathrm{d}\boldsymbol{b}_i \end{aligned}$$

where the conditional density of $\boldsymbol{y}_i$ is

$$p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \sigma) = \frac{\exp\left(-||\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i||^2/2\sigma^2\right)}{(2\pi\sigma^2)^{n_i/2}}$$

and the marginal density of $\boldsymbol{b}_i$ is

$$p(\boldsymbol{b}_i|\boldsymbol{\psi}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{b}_i^{\mathsf{T}}\boldsymbol{\psi}^{-1}\boldsymbol{b}_i\right)}{(2\pi)^{q/2}\sqrt{|\boldsymbol{\psi}|}}.$$

In the above we have assumed that the within-group errors are independent and identically distributed, $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. To arrive at

parameter estimates from the likelihood the following general steps are carried out:

- first the marginal likelihood function above is restated by using, what Pinheiro and Bates (2000, p. 63) refer to as, augmented data and model matrices,
- then the augmented model matrices are decomposed by using the $\boldsymbol{QR}$ matrix decomposition, which simplifies the marginal likelihood significantly as the integrals can be evaluated. The $\boldsymbol{QR}$ matrix decomposition also provides several advantages including reduced round-off error, efficient use of memory, and increased speed of execution (Lindstrom and Bates, 1988).
- Estimates of $\boldsymbol{\beta}$ and σ are then found in terms of the parameter vector $\boldsymbol{\theta}$ which parameterises $\boldsymbol{\Delta}$, the relative precision factor, where
$$\boldsymbol{\Delta}^{\intercal}\boldsymbol{\Delta} = \sigma^2\boldsymbol{\psi}^{-1}.$$
- Substituting back the estimates of $\boldsymbol{\beta}$ and σ in terms of $\boldsymbol{\theta}$, we get the profiled likelihood function. This profiled marginal likelihood can then be maximised with respect to $\boldsymbol{\theta}$ by using a Newton-Raphson routine or the EM-algorithm (Lindstrom and Bates, 1988). The implementation in the `R` function `lme` of the package `nlme` (Pinheiro and Bates, 2000; Pinheiro *et al.*, 2019), which we make use of in our applications, uses a hybrid approach. It starts by using an EM-algorithm and then switches to using a Newton-Raphson routine.
- Once we have $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}$ can be found.

See Pinheiro and Bates (2000) for a detailed description of the above steps and the extension to multiple levels of grouping.

The above outlines how maximum-likelihood estimates (MLEs) may be found and these are what we use in our applications. However, a criticism of maximum-likelihood estimation is that the variance components are biased downwards. Restricted maximum-likelihood estimates are typically preferred, but do not allow for model comparisons by using information criteria such as Akaike's information criterion (Akaike, 1974, AIC) or Bayesian information criterion (Schwarz *et al.*, 1978, BIC) when the fixed effects change.

2.4.2 *Nonlinear Mixed Effect Models*

The methods we outline here are those described by Lindstrom and Bates (1990) and Pinheiro and Bates (2000, ch. 7). These methods are implemented in the `nlme` function of the `R` package of the same name (Pinheiro *et al.*, 2019) which are the function and package we use in our applications in Chapter 3.

The marginal likelihood function for a nonlinear mixed effect model with one level of grouping is given by

$$L(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi}|\boldsymbol{y}) = \prod_{i=1}^{M} \int p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \sigma) p(\boldsymbol{b}_i|\boldsymbol{\psi}) \, \mathrm{d}\boldsymbol{b}_i \tag{8}$$

where the conditional density of $\boldsymbol{y}_i$ is

$$p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \sigma) = \frac{\exp\left(-||\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)||^2/2\sigma^2\right)}{(2\pi\sigma^2)^{n_i/2}},$$

where

$$\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) = [f(\boldsymbol{x}_{i1}, \boldsymbol{\phi}_{i1}), \ldots, f(\boldsymbol{x}_{in_i}, \boldsymbol{\phi}_{in_i})]^\intercal,$$

and the marginal density of $\boldsymbol{b}_i$ is

$$p(\boldsymbol{b}_i|\boldsymbol{\psi}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{b}_i^\intercal \boldsymbol{\psi}^{-1} \boldsymbol{b}_i\right)}{(2\pi)^{q/2}\sqrt{|\boldsymbol{\psi}|}}.$$

The key difficulty with evaluating the above marginal likelihood function is that $f(\cdot)$ can be nonlinear in the random effects. This means that the likelihood function may not have a closed form expression and approximations have to be used to make the maximisation of the marginal likelihood a tractable problem. The approximation that we outline here is that of Lindstrom and Bates (1990) which involves approximating $f(\cdot)$ by a first-order Taylor expansion around the conditional modes of the random effects where the conditioning is on the current estimate of the precision factor $\boldsymbol{\Delta}$. Pinheiro and Bates (2000, ch. 7) discuss this approximation as well as a Laplacian approximation to the marginal likelihood and an adaptive Gaussian quadrature (AGQ) rule which has improved accuracy relative to the other methods. Pinheiro and Bates (1995) provide a comparison of all three of these methods.

The method of Lindstrom and Bates (1990) consists of the following two steps.

PNLS Step: this step involves minimising the following penalised nonlinear least squares (PNLS) function with respect to $\boldsymbol{b}_i$ and $\boldsymbol{\beta}$, holding $\boldsymbol{\Delta}$ constant at its current estimate:

$$\sum_{i=1}^{M} \left[||\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)||^2 + ||\boldsymbol{\Delta}\boldsymbol{b}_i||^2\right].$$

This provides conditional estimates of the mode of $\boldsymbol{b}_i$ and an estimate of $\boldsymbol{\beta}$.

LME Step: this step involves linearising the log-likelihood function corresponding to the likelihood function in Equation 8. The resulting approximate log-likelihood is identical to that for a LMEM and so the methods outlined for LMEMs can then be used. This then provides us with an updated estimate of $\boldsymbol{\Delta}$.

The algorithm involves iterating between the PNLS Step and the LME Step until some convergence criterion is met. The work of Pinheiro and Bates (1995) demonstrates that the estimates found by using this method, referred to as the LME method in their work, are accurate, reliable, and computationally efficient to find. However, it should be noted that Pinheiro and Bates (1995) do recommend a hybrid approach where by the LME method is used to find starting values for the Laplacian approximation method.

2.4.3 *Univariate Generalised Linear Mixed Effect Models*

Zhang (2013) provides several different means of estimating the Tweedie GLMEMs. We outline the likelihood based method presented by Zhang (2013) which uses a Laplace approximation; this is the estimation method that we use in our applications. The alternative methods Zhang (2013) provide are likelihood based methods which make use of AGQ and Bayesian methods. The Laplace approximation is a special case of the AGQ method with only one knot.

The marginal likelihood function for the univariate GLMEMs with one level of grouping is given by

$$L(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi}, \zeta, p|\boldsymbol{y}) = \prod_{i=1}^{M} \int \boldsymbol{f}(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \sigma, \boldsymbol{\psi}, p) p(\boldsymbol{b}_i|\boldsymbol{\psi}) \, \mathrm{d}\boldsymbol{b}_i \tag{9}$$

where $\boldsymbol{f}(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \sigma, \zeta, p) = [f(y_{i1}|\boldsymbol{b}_i, \boldsymbol{x}_{i1}, \boldsymbol{z}_{i1}), \ldots, f(y_{in_i}|\boldsymbol{b}_i, \boldsymbol{x}_{in_i}, \boldsymbol{z}_{in_i})]^\intercal$ with $f(\cdot)$ given in (4), and

$$p(\boldsymbol{b}_i|\boldsymbol{\psi}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{b}_i^\intercal \boldsymbol{\psi}^{-1} \boldsymbol{b}_i\right)}{(2\pi)^{q/2}\sqrt{|\boldsymbol{\psi}|}}.$$

The integral in (9) is typically intractable and as such to evaluate it approximations have to be used. The Laplace approximation does this by approximating the integrand in (9) by its second order Taylor expansion at the conditional mode of the random effects $\boldsymbol{b}_i$, given the current estimate of $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, ζ, and p. The resulting approximate likelihood function can then be maximised using numerical optimisation methods to find updated estimates of $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, ζ, and p, from where an update estimate of the conditional mode of the random effects $\boldsymbol{b}_i$ can be found by using a penalized iteratively reweighted least squares routine. These two steps, estimating the parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, ζ, and p, given the conditional mode of the random effects, and estimating the conditional mode of the random effects, are iterated between until the estimates converge. See (Zhang, 2013) for the details and the exact implementation.

2.4.4 *Multivariate Generalised Linear Mixed Effect and Latent Variable Models*

The estimation of the M-GLMEMs is done by using Bayesian Markov Chain Monte-Carlo (MCMC) estimation methods which are implemented in `JAGS` (Plummer *et al.*, 2003) in the `boral` package in `R` (Hui, 2018). The work of Gelman *et al.* (2013, ch. 11) provides details of MCMC estimation methods.

The marginal likelihood function for the M-GLMEMs, assuming $\boldsymbol{\beta}$ has K elements, is given by

$$L(\boldsymbol{\Theta}, \sigma, \eta_1, \ldots, \eta_J, p, \boldsymbol{\kappa}_0, \boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K, \sigma_0, \sigma_1, \ldots, \sigma_K | \boldsymbol{Y}) =$$
$$\prod_{i=1}^{M} \int \prod_{j=1}^{J} \int \cdots \int \prod_{k=1}^{n_i} \int p(y_{ijk} | b_i, \boldsymbol{z}_{ik}, \beta_{0j}, \beta_{j1}, \ldots, \beta_{jK}, \ldots) p(\boldsymbol{z}_{ik}) \mathrm{d}z_{ik} \times$$
$$p(\beta_{0j} | \boldsymbol{\kappa}_0, \sigma_0) \mathrm{d}\beta_{0j} \ldots p(\beta_{jK} | \boldsymbol{\kappa}_K, \sigma_K) \, \mathrm{d}\beta_{Kj} \, p(b_i | \sigma) \, \mathrm{d}b_i \quad (10)$$

where $p(y_{ijk} | b_i, \boldsymbol{z}_{ik}, \beta_{0j}, \beta_{j1}, \ldots, \beta_{jK}, \ldots)$ is the conditional density function for y_{ijk} given the random effects, latent variables, and βs, $p(\boldsymbol{z}_{ik})$ is the density of the latent variables which is multivariate standard normal, $p(\beta_{0j} | \boldsymbol{\kappa}_0, \sigma_0)$ is normal with parameters given in (6), $p(\beta_{jk} | \boldsymbol{\kappa}_k, \sigma_k)$ is normal with parameters given in (7), and, as for the other models, $p(b_i | \sigma)$ is the density function for the random effects which is normal with mean of zero and variance of σ^2. To estimate the posterior densities for the parameters, prior distributions need to be specified. The prior distributions that we use for the parameters, which are assumed to be independent, are as follows:

- a uniform prior for σ with maximum of 30,
- normal priors are used for $\boldsymbol{\theta}_j$ with means of zero and variances of 10,
- normal priors are used for the $\boldsymbol{\kappa}$s with means of zero and variances of 10, and for the σ_k a uniform prior is used with maximum of 30,
- uniform priors are used for the dispersion parameters η_j with maximum equal to 30, and
- a uniform prior with range $(1, 2)$ is used for p.

Given the priors and the likelihood, sampling from the posterior distribution is done by using MCMC via `JAGS`.

For parameter estimation, the MCMC is run for 80 000 iterations, the first 20 000 of which are discarded for burnin, and a thinning rate of 30 is used. Only one MCMC chain is run by the `boral` package as the latent variable component of the model is invariant to sign switching; i.e. $\boldsymbol{z}_{ik}\boldsymbol{\theta}_j = -\boldsymbol{z}_{ik}(-\boldsymbol{\theta}_j)$ (Hui, 2018). Convergence of the MCMC chain is

checked by inspecting the trace plots of the MCMC chains for each parameter.

For the parameter estimates that we report, we use the medians of the sampled posterior distributions, and where we report credibility intervals, we use 95% highest posterior density intervals.

2.5 MODEL BUILDING AND SELECTION

The model building process that we follow for the LMEMs and the NLMEMs follows the approach of Pinheiro and Bates (2000) which relies extensively on plots of the standardised residuals and the estimated random effects. The model building process involves the following stages:

Stage 1, specifying the base or starting model: we start by specifying a basic model that only includes the time point covariate and no other covariates. Including the time point covariate allows us to model the non-linear profile of the vaccine responses.

Stage 2, fitting individual-specific base models: in this stage we fit base models specific to the data for each individual. The output from which are individual-specific parameter estimates. We then plot these parameter estimates along with their 95% confidence intervals to gauge whether any of the parameters require random effects.

Stage 3, adding (or removing) random effects: given the results from Stage 2, we then include random effects in the base model. We typically add random effects for all structural parameters at all levels of grouping where the random effects have a simple covariance structure. To assess whether a random effect is needed, we examine the resulting estimates of the covariance matrices. Random effects with variances close to zero are removed.

Stage 4, examining model diagnostic plots: to assess the fit of the model we examine model diagnostic plots. We examine plots of the residuals versus the fitted values and time points to identify whether the within-group variance structure assumptions are appropriate and to identify whether the non-linear profile by time is appropriate, respectively. We also examine plots of the random effect estimates by various covariates to assess whether they vary by those covariates and so whether fixed effects for the covariates need to be incorporated into the models. The plots of the standardised residuals and random effects also allow us to assess whether the underlying model assumptions are met. These assumptions are:

- the within-group errors are normally distributed with zero mean and covariance of $\sigma^2\mathbf{\Lambda}_i$, and are independent across subjects and independent of the random effects.
- the random effects are normally distributed with zero mean and covariance $\boldsymbol{\psi}$, and are independent for different groups.

We iterate through Stage 3, adding or removing random effects, and Stage 4 several times attempting to improve the fit of the model.

Stage 5, refinement and reassessment: the above process is a stepwise process whereby complexity is added to the base model as we see is necessary, but as we build the models, complexity added in earlier iterations may no longer be required and so the models can be simplified. In Stage 5, the objective is to refine earlier models and to simplify them by removing unnecessary complexity.

For the GLMEMs and M-GLMEMs we do not work through a full model development process, i.e. we do not start from a base model and increase the complexity. Instead, we rely on the form of the best LMEM to inform the structure of the linear predictors we consider for the GLMEMs and M-GLMEMs.

For the GLMEMs, we start with the linear predictor from the best LMEMs, and then try to improve the fit of the model from there. We use various residual plots for the fitted GLMEMs to try identify areas where the models may be deficient. Plots of the random effects are also used to try identify which random effects are needed in the models and whether any fixed effects are missing.

The form of the linear predictor for the most general M-GLMEMs that we consider has a form similar to that of the best LMEMs. In addition, we also consider a M-GLMEM without covariates so as to be able to construct unconstrained ordination plots which we can compare to the residual ordination plots when we do include covariates. The model diagnostic plots that we use for the M-GLMEMs are plots of the Randomised Quantile or Dunn-Smyth residuals (Dunn and Smyth, 1996). If the fitted model is correct, then the Dunn-Smyth residuals have a standard normal distribution, apart from sampling variability of the model parameters. The Dunn-Smyth residuals are calculated for each response by inverting the estimated distribution function and then finding the corresponding standard normal quantiles. Random variation is introduced into the calculation of the Dunn-Smyth residuals when the distribution function has discrete components to ensure the resulting residuals are continuous; see the definition given by Dunn and Smyth (1996).

Model selection is done by using AIC, but we also report the BIC values, where

$$\text{AIC} = 2n_{\text{param}} - 2\log\left(\widehat{L}\right) \quad \text{and} \quad \text{BIC} = \log(n_{\text{obs}})n_{\text{param}} - 2\log\left(\widehat{L}\right)$$

and where n_{param} is the number of estimated parameters in the model, $\widehat{L}$ is the likelihood evaluated at the MLEs, and n_{obs} is the number of observations. As we are using maximum-likelihood estimation for the LMEMs, NLMEMs, and GLMEMs, their AIC and BIC values are comparable. For the M-GLMEMs models, we use the AIC and BIC values which are calculated by using the marginal likelihood (10) evaluated at the medians of the posterior distributions, and where the AIC and BIC are not returned by `boral`, which happens when traits are included in the model, we make use of the expected AIC (EAIC) and expected BIC (EBIC) values (Carlin and Louis, 2008). As such the AIC and BIC values for the M-GLMEMs are comparable to those of the other models. The EAIC and EBIC values are not comparable to the AIC and BIC values. It must be noted that Hui (2018) indicates that these information criteria for the M-GLMEMs are calculated by heuristic methods and are experimental, and are not being actively maintained as of Version 1.6 of `boral`.

2.6 SOFTWARE USED

We have mentioned the software that we use in this work and summarise it here. The statistical software package `R` (version 3.6.0) was used for the analysis carried out here (R Core Team, 2019). The packages that we use for the model estimation and analysis are:

- for the LMEM and NLMEMs, the `nlme` package (version 3.1-139) was used (Pinheiro *et al.*, 2019),
- for the Tweedie GLMEMs, the `cplm` package (0.7-8) was used (Zhang, 2013), and
- for the Tweedie M-GLMEMs, the `boral` package (1.7) was used (Hui, 2018).

3

APPLICATIONS

In this chapter we use the methods outlined in Chapter 2 with the immune response data described in Chapter 1. The objective is to fit suitable models which capture the features of the data that we outlined in Chapter 1. Specifically, the models should allow for

1. the repeated measures for each of the subjects which is likely to induce correlation between observations for a subject,
2. the zero-inflated skew distributions of the immune responses,
3. the multiple immune responses in the data, and
4. the nonlinear immune response profiles.

The models should also be capable of interpretation and be suitable for inference so as to allow us to identify the vaccine candidates that should be advanced in the vaccine trial process.

In the applications presented in this chapter, we initially consider LMEMs with orthogonal polynomials and cubic B-splines, before considering NLMEMs. Following the applications of the LMEMs and NLMEMs, we consider Tweedie GLMEMs and M-GLMEMs with cubic B-splines. Given that the immune responses have zero-inflated skewed distributions, we do expect that the LMEMs and the NLMEMs, which both have normally distributed within-subject errors, will struggle to fit the data adequately. We include these models as they are the classic models considered in this context. We do expect that the Tweedie GLMEMs and M-GLMEMs will fit the data better than the LMEMs and NLMEMs as the Tweedie distribution has a mass at zero and has a skewed distribution.

At the end of the chapter we discuss the results from some of the fitted models and attempt to draw conclusions about which vaccine candidates should be advanced.

Example R code for several of the models we implement is provided in Appendix E

3.1 LINEAR MIXED EFFECT MODELS

In this section we consider the application of LMEMs to the vaccine immune response data. We consider LMEMs with orthogonal poly-

nomials and cubic B-splines to capture the nonlinear profile of the immune responses. The orthogonal polynomials we consider are of three degrees and the splines have three degrees of freedom. In building these models we follow the framework outlined in Section 2.5.

We first present the models with orthogonal polynomials before presenting the models with cubic B-splines. At the end of this section we provide a brief discussion of the fitted LMEMs.

3.1.1 *LMEMS with Orthogonal Polynomials of Three Degrees*

The general structure of the models that we consider is given by

$$y_{ijk} = \phi_{0,ij} P_0(t_k) + \sum_{w=1}^{3} \phi_{w,ij} P_w(t_k) + \epsilon_{ijk}$$

where y_{ijk} is the immune response j (= `CD4Gp2pTp`, `CD4Gn2pTp`, `CD4Gp2nTp`, `CD4Gp2pTn`, `CD4Gn2nTp`, `CD4Gn2pTn`, and `CD4Gp2nTn`) for subject i at time k (t_k), $P_w(t)$ is the wth order orthogonal polynomial such that

$$P_0(t_k) = 1, \qquad \sum_{k=1}^{n} P_r(t_k) P_r(t_k) = 1, \quad \text{and}$$
$$\sum_{k=1}^{n} P_r(t_k) P_s(t_k) = 0 \quad \text{for } r \neq s,\ r, s = 1, 2, 3,$$

and $\epsilon_{ijk} \sim N(0, \sigma^2)$. The parameters to be estimated are the $\phi_{w,ij}$s and σ. The nonlinear profiles of the immune responses are to be modelled by the polynomial terms in the summand in the model outlined above.

Individual Specific Models

Initially we fit separate models for each of the subjects including the response covariate as a fixed effect; i.e. we fit models of the form

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} P_w(t_k) + \epsilon_{ijk}$$

with

$$\phi_{0,ij} = \beta_0 + \beta_j \times \text{Response}_j \quad j = \texttt{CD4Gn2pTp}, \ldots, \texttt{CD4Gp2nTn},$$

where β_0 is the intercept and corresponds to response `CD4Gn2nTp`, and

$$\phi_{w,ij} = \beta_w \quad w = 1, 2, 3.$$

Fitting individual specific models involves estimating models which are specific to each subject's immune response profiles. The estimated parameters along with their approximate 95% confidence intervals can

then be used to decide which parameters we need to include random effects for.

Figure 11 contains a plot of the approximate 95% confidence intervals for each of the parameters and subjects. The panel for (Intercept) corresponds to β_0, each of the response panels corresponds to a β_j for $j =$ `CD4Gn2pTp`, ..., `CD4Gp2nTn`, and the pl1, pl2, and pl3 panels correspond to a β_w for $w = 1, 2, 3$ respectively. We are concerned with non-overlapping confidence intervals as these suggest that we may need to include random effects for the corresponding parameter.

From the figure, we can see that many of the confidence intervals do overlap, but not in all cases. Most notably we see that there are non-overlapping confidence intervals for β_0, $\beta_{\text{CD4Gn2pTp}}$, $\beta_{\text{CD4Gp2nTn}}$, $\beta_{\text{CD4Gp2pTp}}$, β_1, β_2, and β_3. For the parameters not listed, there are some confidence intervals that do not overlap, but these are less obvious. In the first model that we consider, we choose to include random effects for all of the parameters and then to remove those that are not needed.

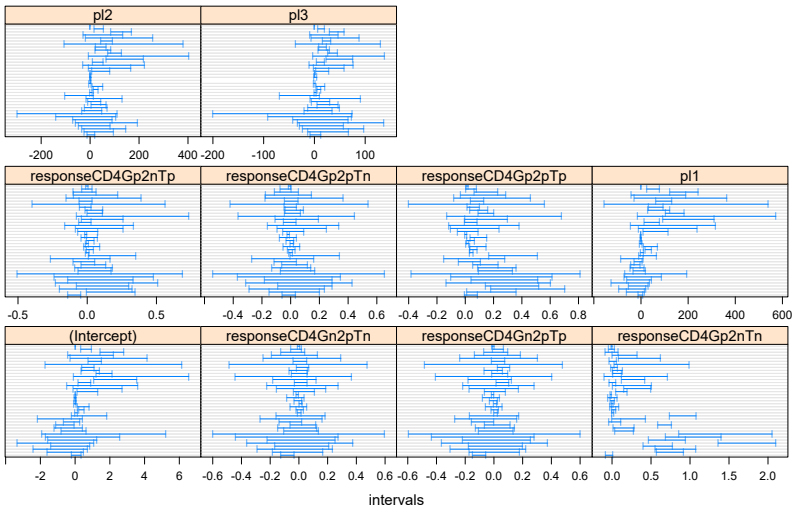


Figure 11: Approximate 95% confidence intervals for the parameter estimates for the subject specific polynomial models.

Multi-Level Models

In this section we fit models which include random effects and we add fixed effects as needed. We consider random effects at the subject level and the response within subject level. We include the random effects at the subject level because:

- each subject is a random selection from a population of potential subjects and as such we need to control for the variability introduced by that subject as we are concerned with population level effects, and

- the inclusion of random effects induces correlation between the subject's observations in the LMEMs.

The random effects at the response within subject level are included to control for the subject specific variability at the response level.

The models in this section are fitted to all of the vaccine immune response data at once; i.e. the reported parameter estimates are not subject specific.

MODEL 1 The first LMEM that we fit has the form

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} P_w(t_k) + \epsilon_{ijk}$$

with

$$\phi_{0,ij} = \beta_0 + b_{0i} + b_{0i,j}$$

and

$$\phi_{w,ij} = \beta_w + b_{wi} + b_{wi,j} \quad \text{for } w = 1, 2, 3,$$

where

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_1) \quad \text{and} \quad \boldsymbol{b}_{i,j} = \begin{pmatrix} b_{0i,j} \\ b_{1i,j} \\ b_{2i,j} \\ b_{3i,j} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_2),$$

and $\boldsymbol{b}_i$ is independent of $\boldsymbol{b}_{i,j}$. We consider diagonal structures for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ as more complex structures proved challenging to fit. The random effects $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i,j}$ are also independent of ϵ_{ijk}.

The above model was estimated by using the `nlme` package in `R` (Pinheiro *et al.*, 2019). Table 2 summarises the results from the fitted model. The table shows the parameter estimates for the fixed effects together with their standard errors, the degrees of freedom, and associated significance. The AIC and BIC values for this model are -730.1 and -663.0 respectively.

Table 2: Output for Model 1, a LMEM with a 3 degree polynomial.

Parameter	Estimate	Std. Error	DF	t-value	p-value
β_0	0.12	0.015	1026	7.921	<0.001
β_1	-0.664	0.1738	2825	-3.824	0.236
β_2	-0.135	0.2018	2825	-0.667	0.393
β_3	1.237	0.1622	2825	7.6276	<0.001

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.055^2 & 0 & 0 & 0 \\ 0 & 0.608^2 & 0 & 0 \\ 0 & 0 & 0.234^2 & 0 \\ 0 & 0 & 0 & (1.239 \times 10^{-5})^2 \end{pmatrix}$$

and

$$\widehat{\psi}_2 = \begin{pmatrix} 0.184^2 & 0 & 0 & 0 \\ 0 & 1.220^2 & 0 & 0 \\ 0 & 0 & (3.768 \times 10^{-5})^2 & 0 \\ 0 & 0 & 0 & (2.194 \times 10^{-5})^2 \end{pmatrix},$$

and the parameter estimate for σ, the standard deviation for the within-group error, is 0.138. We see that the estimates of the standard deviations for several of the random effects are close to zero suggesting that they are not needed. We will remove these in the next iteration of this model.

To assess the fit of the model, we present several diagnostic plots in Figures 12 to 18. Important features that can be noted about the residuals are:

- from Figure 12, which plots the residuals versus fitted values, that for several of the vaccine and response combinations there are clear fan shaped patterns with the variance in the residuals increasing as the fitted values increase. This indicates that the assumption of constant variance was invalid.

- from Figure 13, which plots the residuals versus the time points, we see that there are slight humped patterns in the residuals for some of the responses for vaccines 3 and 7. These patterns are slight, but suggest that the model may not be capturing the nonlinear structure for some of the vaccine and response combinations adequately, and that we may need to possibly include an interaction with the polynomial, vaccine, and response terms.

- from Figure 14, which gives the QQ-plots of the standardised residuals by vaccine and response combination, that the residuals are heavily skewed and are clearly non-normal. The assumption that the within group errors are normally distributed is clearly violated.

Important features that can be noted about the random effect estimates are:

- from Figure 15, which gives box and whisker plots for the subject specific random effects by vaccine, that the random effect b_{3i} is not needed as it has very little variability. For the b_{0i} and b_{1i} random effects, we see that they are not centered at zero and that there may be a vaccine effect particularly for the intercept term. The possible need for a vaccine fixed effect may indicate that there is a systematic difference in the immune responses by vaccine, which is what we are trying to investigate.

- from Figure 16, which gives box and whisker plots for the response within subject random effects by vaccine and response,

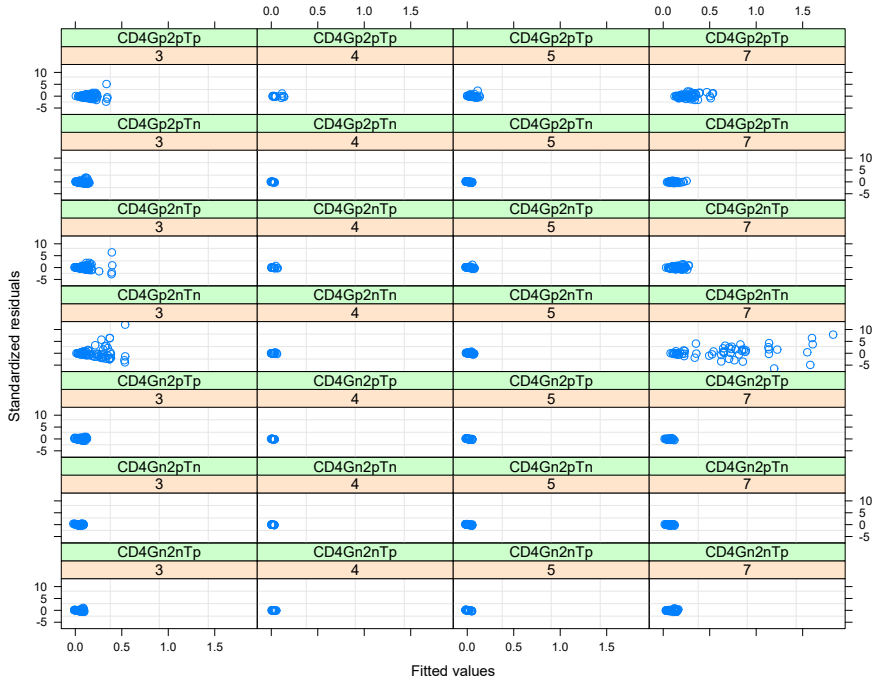


Figure 12: Standardised residuals versus fitted values by vaccine and response for Model 1, a LMEM with a 3 degree polynomial.

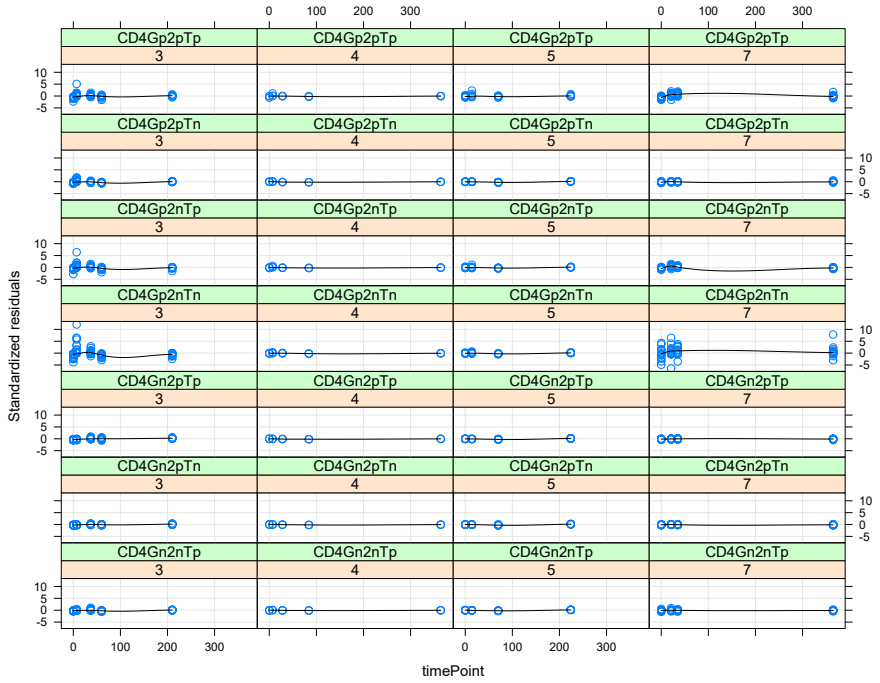


Figure 13: Standardised residuals versus time point by vaccine and response for Model 1, a LMEM with a 3 degree polynomial.

that the random effects $b_{2i,j}$ and $b_{3i,j}$ are not needed as they have very little variability. For $b_{0i,j}$, we see that the random effects are

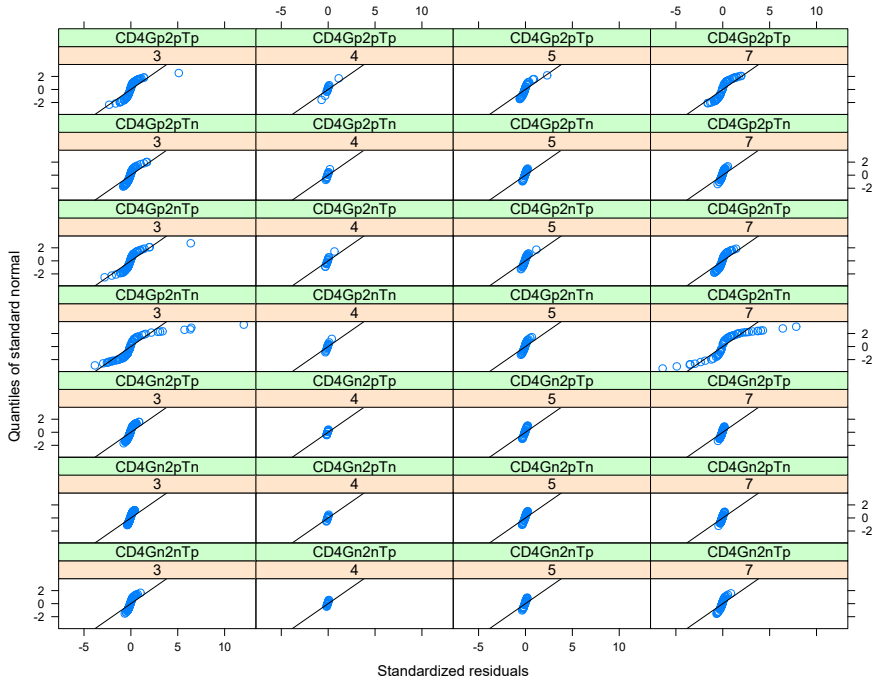


Figure 14: QQ-plot for the standardised residuals by vaccine and response for Model 1, a LMEM with a 3 degree polynomial.

not all centered around zero; it looks like we may need vaccine and response fixed effects.

- from Figures 17 and 18, which give the QQ-plots of the random effects at the subject and response within subject levels, that several of the random effects are clearly non-normal as the points do not lie on a straight line. The assumption that the random effects are multivariate normal is clearly violated.

MODEL 2 To try and improve the fit of the Model 1, we make the following changes:

- we include vaccine and response fixed effects for the intercept term, and
- we drop the random effects b_{3i}, $b_{2i,j}$, and $b_{3i,j}$.

Explicitly, the form of Model 2 is given by

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} P_w(t_k) + \epsilon_{ijk}$$

with

$$\phi_{0,ij} = (\beta_{0,1} + b_{0i} + b_{0i,j}) + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j$$

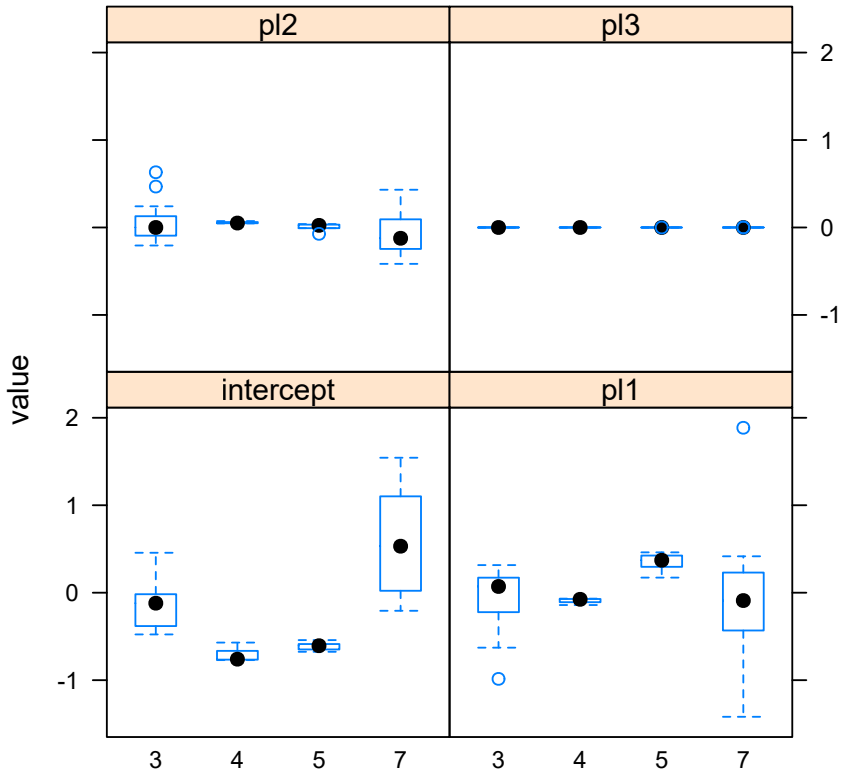


Figure 15: Subject specific random effect estimates $\widehat{b_{0i}}, \widehat{b_{1i}}, \widehat{b_{2i}}$ and $\widehat{b_{3i}}$ for Model 1, a LMEM with a 3 degree polynomial, by vaccine.

$$\phi_{1,ij} = \beta_1 + b_{1i} + b_{1i,j},$$
$$\phi_{2,ij} = \beta_2 + b_{2i},$$

and

$$\phi_{3,ij} = \beta_3.$$

The distribution assumptions for ϵ_{ijk} and the random effects remain the same with the relevant terms removed.

Table 3 gives the parameter estimates for the fixed effects in Model 2, along with their standard errors and significance levels. The AIC and BIC values for the model are -847.6 and -749.6 respectively. These are significantly better than Model 1.

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} (9.646 \times 10^{-5})^2 & 0 & 0 \\ 0 & 0.626^2 & 0 \\ 0 & 0 & 0.154^2 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{\psi}}_2 = \begin{pmatrix} 0.143^2 & 0 \\ 0 & 1.208^2 \end{pmatrix}$$

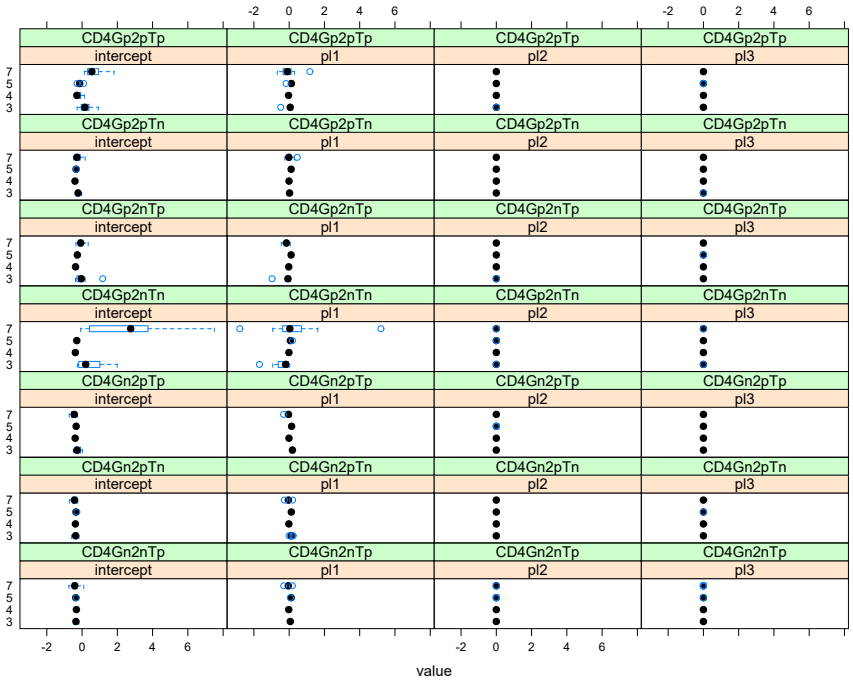


Figure 16: Response within subject random effect estimates $\widehat{b_{0i,j}}$, $\widehat{b_{1i,j}}$, $\widehat{b_{2i,j}}$ and $\widehat{b_{3i,j}}$ for Model 1, a LMEM with a 3 degree polynomial, by vaccine and response.

and the parameter estimate for σ, the standard deviation for the within-group error, is 0.139. Interestingly, it would appear that we do not need the random effect b_{0i} as it has a standard deviation close to zero; the inclusion of the vaccine and response fixed effects appear to have removed the need for b_{0i}.

Figures 70 to 76, which are in Appendix B, present diagnostic plots for the model. We use these plots to check the fit of the model as well as its underlying assumptions. The observations that we can make regarding the residuals are similar to those of Model 1; most importantly, the residuals are still heavily skewed and non-normal (Figure 72). For the subject level random effect estimates, we see that b_{0i} is not needed, and that there is less of a pattern by vaccine. For the response within subject level random effect estimates, we see that there is still a pattern by vaccine and response for $b_{0i,j}$. Importantly, from the QQ-plots, we still see that several of the random effects are not normal.

MODEL 3 To try and improve the fit of the Model 2, we make the following changes:

- we include vaccine and response fixed effects and their interaction for the intercept,

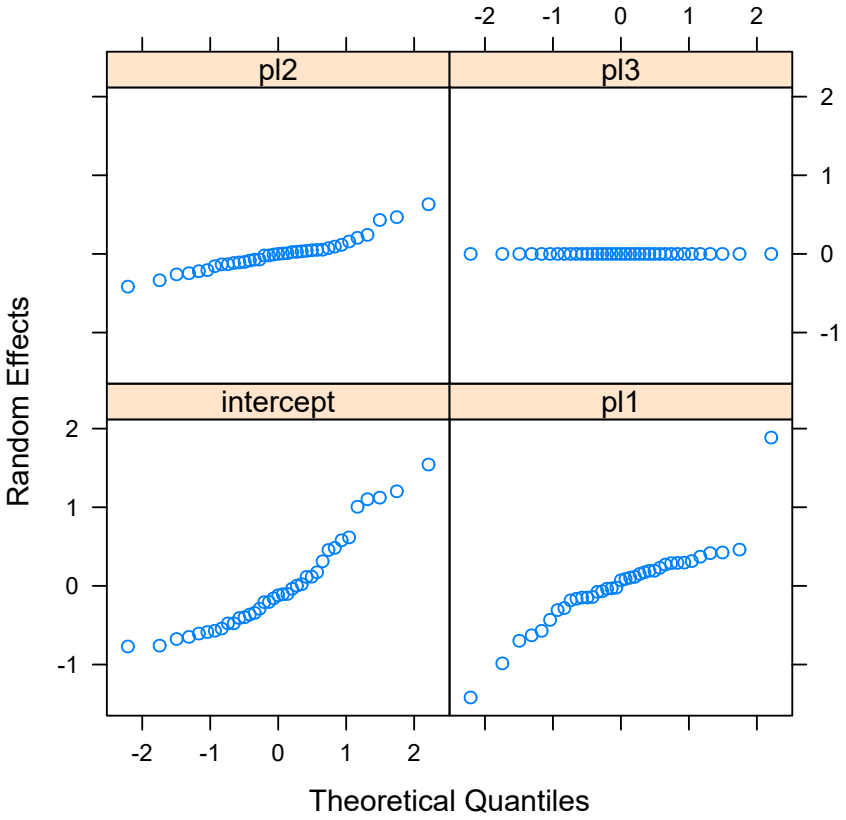


Figure 17: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i}}$, $\widehat{b_{1i}}$, $\widehat{b_{2i}}$ and $\widehat{b_{3i}}$ for Model 1, a LMEM with a 3 degree polynomial.

- we include the vaccine fixed effect for the polynomial terms to try remove the remaining pattern by time point seen in the residuals (Figure 71), and
- we simplify the random effects structure by dropping b_{2i}. The results from Model 2 suggest removing b_{0i}, but we found that removing b_{2i} instead of b_{0i} provided a significant improvement to the fit of the model.

The second change above is intended to ensure that the model captures the different non-linear profiles of the immune responses for each vaccine.

The form of Model 3 is given by

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} P_w(t_k) + \epsilon_{ijk}$$

with

$$\begin{aligned} \phi_{0,ij} &= (\beta_{0,1} + b_{0i} + b_{0i,j}) + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j + \\ &\quad \beta_{0,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \end{aligned}$$

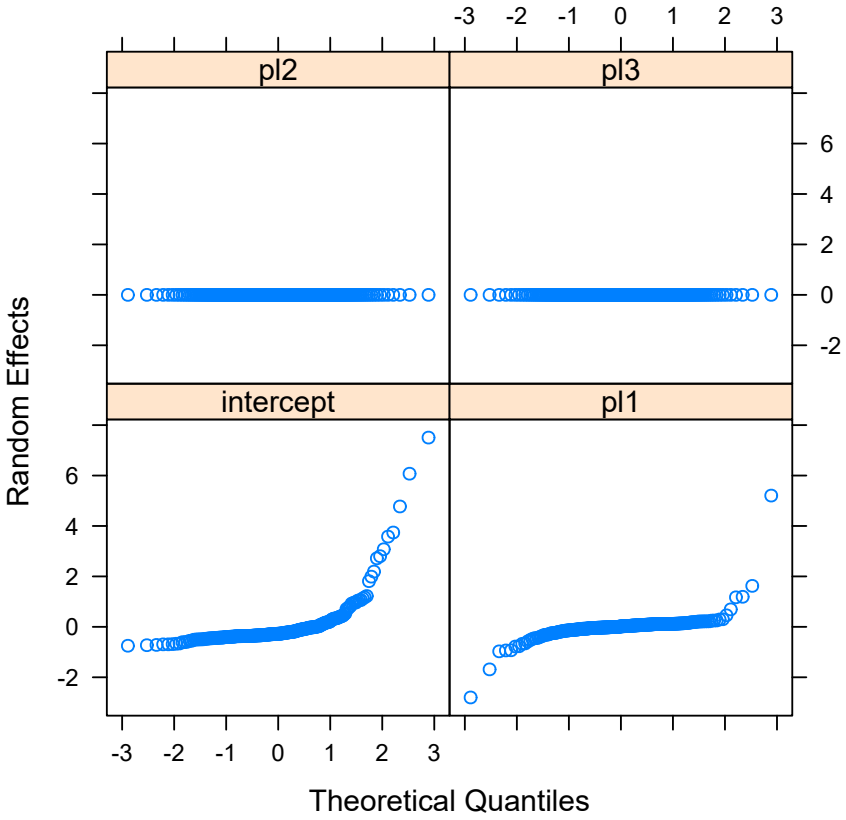


Figure 18: QQ-plot for the response within subject random effect estimates $\widehat{b_{0i,j}}$, $\widehat{b_{1i,j}}$, $\widehat{b_{2i,j}}$ and $\widehat{b_{3i,j}}$ for Model 1, a LMEM with a 3 degree polynomial.

$$\phi_{1,ij} = (\beta_{1,1} + b_{1i} + b_{1i,j}) + \beta_{1,\text{vaccine}_i}\text{vaccine}_i$$

and

$$\phi_{w,ij} = \beta_{w,1} + \beta_{w,\text{vaccine}_i}\text{vaccine}_i \quad \text{for } w = 2, 3.$$

The distribution assumptions for ϵ_{ijk} and the random effects remain the same with the relevant terms removed.

Table 4 presents the results from Model 3. The table presents the parameter estimates, their standard deviation, and their significance. The AIC and BIC values for this model are -1004.4 and -772.1 respectively. This is a significant improvement on Model 1 and Model 2.

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.037^2 & 0 \\ 0 & 0.571^2 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{\psi}}_2 = \begin{pmatrix} 0.101^2 & 0 \\ 0 & 1.360^2 \end{pmatrix}$$

and the parameter estimate for σ, the standard deviation for the within-group error, is 0.132.

Table 3: Output for Model 2, a LMEM with a 3 degree polynomial.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	0.021	0.028	1026	0.753	0.451
$\beta_{0,\text{vaccine4}}$	-0.08	0.038	33	-2.109	0.043
$\beta_{0,\text{vaccine5}}$	-0.067	0.03	33	-2.252	0.031
$\beta_{0,\text{vaccine7}}$	0.112	0.023	33	4.856	<0.001
$\beta_{0,\text{responseCD4Gn2pTn}}$	-0.007	0.036	216	-0.188	0.851
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.001	0.036	216	0.032	0.975
$\beta_{0,\text{responseCD4Gp2nTn}}$	0.308	0.036	216	8.438	<0.001
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.052	0.036	216	1.435	0.153
$\beta_{0,\text{responseCD4Gp2pTn}}$	0.02	0.036	216	0.542	0.589
$\beta_{0,\text{responseCD4Gp2pTp}}$	0.133	0.036	216	3.655	<0.001
$\beta_{1,1}$	-0.29	0.203	1026	-1.429	0.153
$\beta_{2,1}$	0.045	0.2	1026	0.226	0.821
$\beta_{3,1}$	0.642	0.154	1026	4.173	<0.001

Figures 77 to 83, which are in Appendix B, present diagnostic plots for the model which we can use to check the fit of the model and the underlying assumptions. The graphs are broadly similar to those presented for Model 1. Most notably, for the residuals we see that :

- there is still a fan pattern when we plot the residuals versus the fitted values, with the variance in the residuals increasing as the fitted values increase (Figure 77).
- there is still some pattern in the residuals when plotted by time point. Relative to the plots for Model 1 and Model 2, much of the pattern has been removed (Figure 78).
- importantly, the residuals are still skewed and non-normal (Figure 79).

For the random effect estimates, we see that:

- $\widehat{b_{0i}}$ and $\widehat{b_{1i}}$ are relatively well centered around zero with no strong pattern by vaccine (Figure 80).
- the estimates of the random effect $\widehat{b_{0i,j}}$ and $\widehat{b_{1i,j}}$ are now centered around zero which is an improvement from Model 2 (Figure 81).
- the QQ-plots of the random effect estimates show that there are clear departures from normality for $\widehat{b_{0i,j}}$ and $\widehat{b_{1i,j}}$. The QQ-plots for $\widehat{b_{0i}}$ and $\widehat{b_{1i}}$ show some deviations from normality with some curvature in the plots (Figures 82 and 83).

MODEL 4 In this model we make the following changes to Model 3:

Table 4: Output for Model 3, a LMEM with a 3 degree polynomial.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	1.463	0.182	1017	8.048	<0.001
$\beta_{0,\text{vaccine4}}$	-1.457	0.199	33	-7.338	<0.001
$\beta_{0,\text{vaccine5}}$	-1.236	0.27	33	-4.57	<0.001
$\beta_{0,\text{vaccine7}}$	-1.615	0.238	33	-6.773	<0.001
$\beta_{0,\text{responseCD4Gn2pTn}}$	-0.001	0.043	198	-0.016	0.987
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.021	0.043	198	0.482	0.63
$\beta_{0,\text{responseCD4Gp2nTn}}$	0.162	0.043	198	3.725	<0.001
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.061	0.043	198	1.396	0.164
$\beta_{0,\text{responseCD4Gp2pTn}}$	0.025	0.043	198	0.573	0.567
$\beta_{0,\text{responseCD4Gp2pTp}}$	0.11	0.043	198	2.528	0.012
$\beta_{1,1}$	125.795	15.892	1017	7.916	<0.001
$\beta_{2,1}$	88.47	11.105	1017	7.967	<0.001
$\beta_{3,1}$	30.997	3.827	1017	8.1	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-0.012	0.106	198	-0.108	0.914
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	0.003	0.084	198	0.035	0.972
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	-0.015	0.063	198	-0.241	0.81
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-0.032	0.106	198	-0.303	0.762
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	-0.016	0.084	198	-0.188	0.851
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.04	0.063	198	-0.637	0.525
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-0.165	0.106	198	-1.555	0.122
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	-0.147	0.084	198	-1.756	0.081
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	0.492	0.063	198	7.761	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-0.059	0.106	198	-0.555	0.579
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	-0.046	0.084	198	-0.547	0.585
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	0.009	0.063	198	0.135	0.893
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-0.039	0.106	198	-0.368	0.713
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	-0.022	0.084	198	-0.263	0.793
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	0.003	0.063	198	0.05	0.96
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-0.072	0.106	198	-0.682	0.496
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	-0.061	0.084	198	-0.728	0.467
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	0.105	0.063	198	1.65	0.1
$\beta_{1,\text{vaccine4}}$	-126.551	15.992	1017	-7.913	<0.001
$\beta_{1,\text{vaccine5}}$	-106.736	23.224	1017	-4.596	<0.001
$\beta_{1,\text{vaccine7}}$	-134.462	16.842	1017	-7.984	<0.001
$\beta_{2,\text{vaccine4}}$	-87.61	11.352	1017	-7.718	<0.001
$\beta_{2,\text{vaccine5}}$	-74.385	16.713	1017	-4.451	<0.001
$\beta_{2,\text{vaccine7}}$	-75.885	14.095	1017	-5.384	<0.001
$\beta_{3,\text{vaccine4}}$	-30.034	4.529	1017	-6.631	<0.001
$\beta_{3,\text{vaccine5}}$	-25.423	6.16	1017	-4.127	<0.001
$\beta_{3,\text{vaccine7}}$	-20.411	6.977	1017	-2.926	0.004

- Firstly we now include a more complex variance structure for the within-subject errors. Specifically, we model the variance of the within-subject errors as a function of the fitted values as follows

$$\text{Var}(\epsilon_{ijk}) = \sigma^2 \widehat{y}_{ijk}.$$

  We use this structure to capture the fan shape observed for the residuals versus the fitted values; we did try to fit more complex variance structures, but the models failed to fit in these cases.

- Secondly, we remove all random effects except b_{0i} as the other random effects are not needed; they had very little to no variability in this model .

This is the final iteration of the LMEMs with orthogonal polynomials that we consider. Additional complexity is unlikely to be supported by the data, particularly for Vaccines 4 and 5 where we have limited data.

Table 5 presents the fixed effect parameter estimates for Model 4. The AIC and BIC values for this model are -3008.9 and -2792.1. These values are significantly better than Model 3. The parameter estimate for ψ_1 is $(6.786 \times 10^{-7})^2$ and the parameter estimate for σ, the standard deviation for the within-group error, is 0.342.

Figures 19 to 23 present diagnostic plots for the model which we can use to check the fit of the model and the underlying assumptions. We see that for the residuals:

- that we have removed most of the fan pattern seen when the residuals were plotted against the fitted values for previous models. This is an improvement on the previous models (Figure 19).

- there is no marked pattern in the residuals when plotted against the time points. This is an improvement over Models 1 and 2, and similar to Model 3 (Figure 20).

- the residuals are still skewed and non-normal, but are closer to normal than seen for the earlier models (Figure 21).

For the random effect estimates, we see that:

- the estimates $\widehat{b_{0i}}$ are mostly centered around zero with some deviation seen for Vaccine 4 (Figure 22).

- the QQ-plot of the random effect estimates shows no clear sign that $\widehat{b_{0i}}$ is non-normal (Figure 23).

Table 5: Output for Model 4, a LMEM with a 3 degree polynomial.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	1.488	0.144	1215	10.316	<0.001
$\beta_{0,\text{vaccine4}}$	-1.484	0.145	33	-10.24	<0.001
$\beta_{0,\text{vaccine5}}$	-1.322	0.163	33	-8.131	<0.001
$\beta_{0,\text{vaccine7}}$	-1.563	0.202	33	-7.754	<0.001
$\beta_{0,\text{responseCD4Gn2pTn}}$	0.002	0.005	1215	0.47	0.638
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.009	0.007	1215	1.275	0.203
$\beta_{0,\text{responseCD4Gp2nTn}}$	0.123	0.017	1215	7.129	<0.001
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.02	0.008	1215	2.398	0.017
$\beta_{0,\text{responseCD4Gp2pTn}}$	0.008	0.007	1215	1.009	0.313
$\beta_{0,\text{responseCD4Gp2pTp}}$	0.084	0.015	1215	5.653	<0.001
$\beta_{1,1}$	126.679	12.515	1215	10.122	<0.001
$\beta_{2,1}$	88.639	8.794	1215	10.08	<0.001
$\beta_{3,1}$	31.47	3.038	1215	10.358	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-0.013	0.012	1215	-1.057	0.291
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	-0.003	0.008	1215	-0.341	0.734
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	-0.017	0.017	1215	-0.999	0.318
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-0.018	0.015	1215	-1.195	0.232
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	-0.007	0.01	1215	-0.692	0.489
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.027	0.018	1215	-1.552	0.121
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-0.131	0.021	1215	-6.312	<0.001
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	-0.114	0.019	1215	-5.911	<0.001
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	0.53	0.045	1215	11.823	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-0.02	0.017	1215	-1.191	0.234
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	-0.011	0.011	1215	-0.972	0.331
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	0.038	0.022	1215	1.737	0.083
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-0.019	0.014	1215	-1.392	0.164
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	-0.008	0.009	1215	-0.917	0.359
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	0.021	0.02	1215	1.027	0.305
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-0.047	0.027	1215	-1.748	0.081
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	-0.042	0.02	1215	-2.081	0.038
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	0.122	0.031	1215	3.914	<0.001
$\beta_{1,\text{vaccine4}}$	-127.561	12.527	1215	-10.18	<0.001
$\beta_{1,\text{vaccine5}}$	-113.329	14.07	1215	-8.055	<0.001
$\beta_{1,\text{vaccine7}}$	-132.704	13.563	1215	-9.784	<0.001
$\beta_{2,\text{vaccine4}}$	-87.614	8.823	1215	-9.931	<0.001
$\beta_{2,\text{vaccine5}}$	-78.766	10.011	1215	-7.868	<0.001
$\beta_{2,\text{vaccine7}}$	-80.059	11.984	1215	-6.681	<0.001
$\beta_{3,\text{vaccine4}}$	-30.318	3.124	1215	-9.705	<0.001
$\beta_{3,\text{vaccine5}}$	-27.548	3.546	1215	-7.769	<0.001
$\beta_{3,\text{vaccine7}}$	-24.384	6.22	1215	-3.92	<0.001

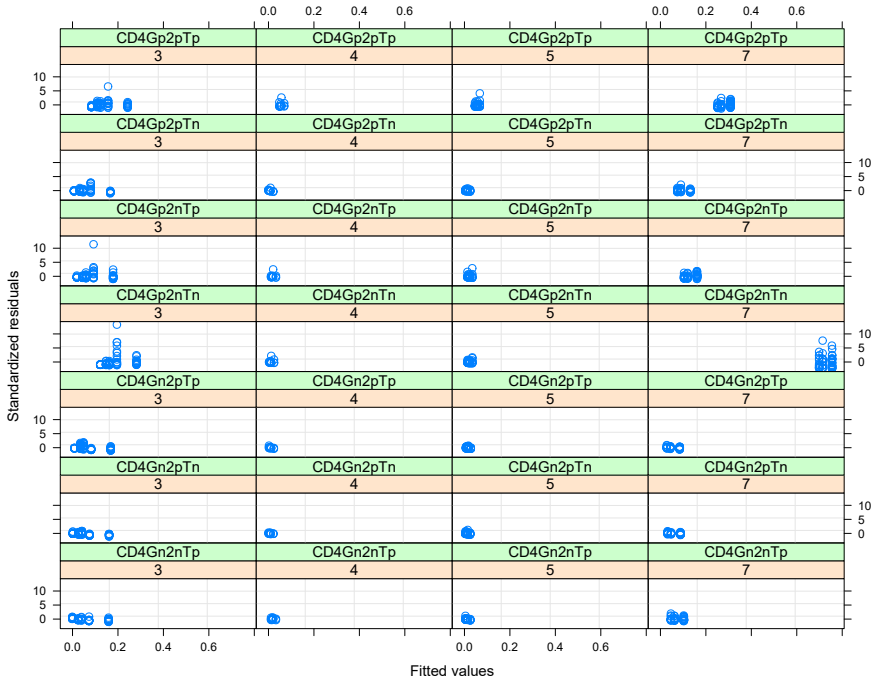


Figure 19: Standardised residuals versus fitted values by vaccine and response for Model 4, a LMEM with a 3 degree polynomial.

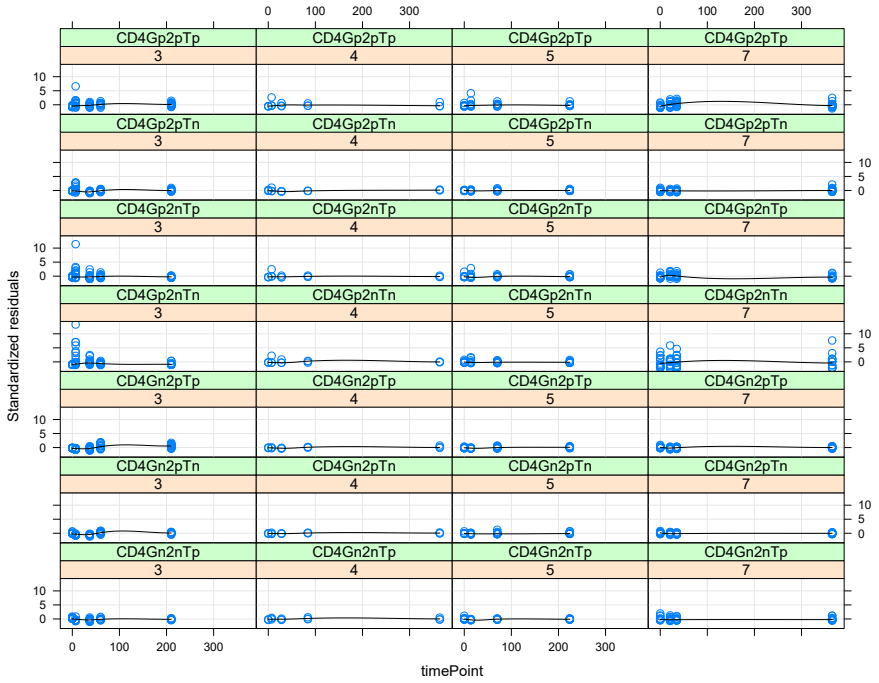


Figure 20: Standardised residuals versus time point by vaccine and response for Model 4, a LMEM with a 3 degree polynomial.

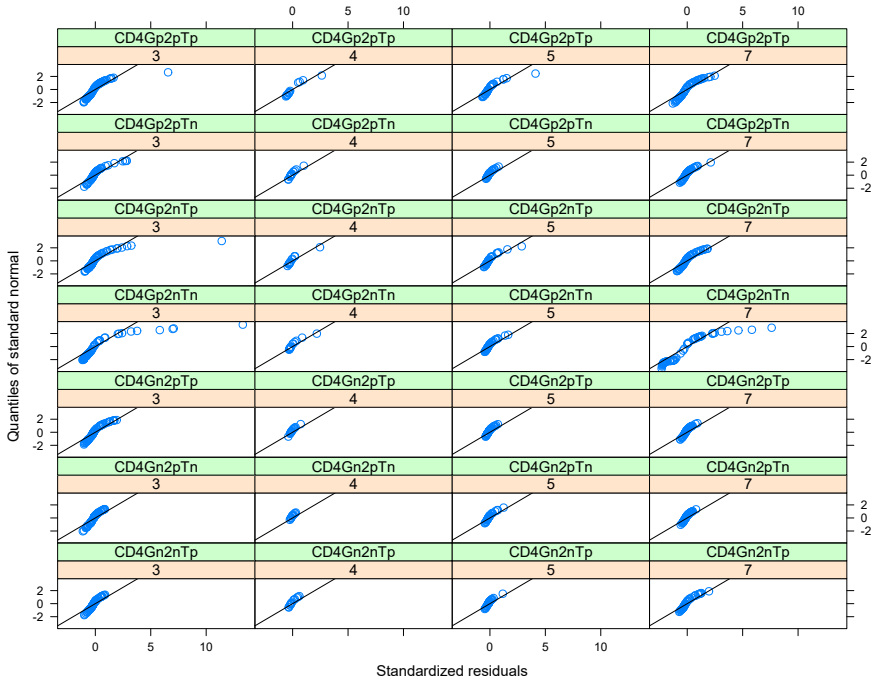


Figure 21: QQ-plot for the standardised residuals by vaccine and response for Model 4, a LMEM with a 3 degree polynomial.

3.1.2 *Comments on the LMEMs with Orthogonal Polynomials*

In this subsection we considered several LMEMs each with orthogonal polynomials of three degrees. The polynomials were used to capture the non-linear profiles for the immune responses.

We have seen that these models do not fit the data well and that the assumptions underlying the LMEMs are not satisfied; i.e. the within-group errors are consistently non-normal and several of the random effect estimates tend to be non-normal. As the fitted models do not satisfy the underlying assumptions we cannot reliably perform inference with them, as such with these models we cannot make comments about which vaccine candidates are the most promising and should be advanced. However, some important points which may be useful for the other models that we consider are:

- that we needed to include vaccine and response fixed effects, and their interaction, for the intercept parameter, and
- that we needed to include the vaccine fixed effect for the polynomial parameters to capture the different immune response profiles for the different vaccines.

We did consider models with higher order orthogonal polynomials, but there is insufficient data to support the additional flexibility they provide; the shapes of the curves in many cases were unreasonable

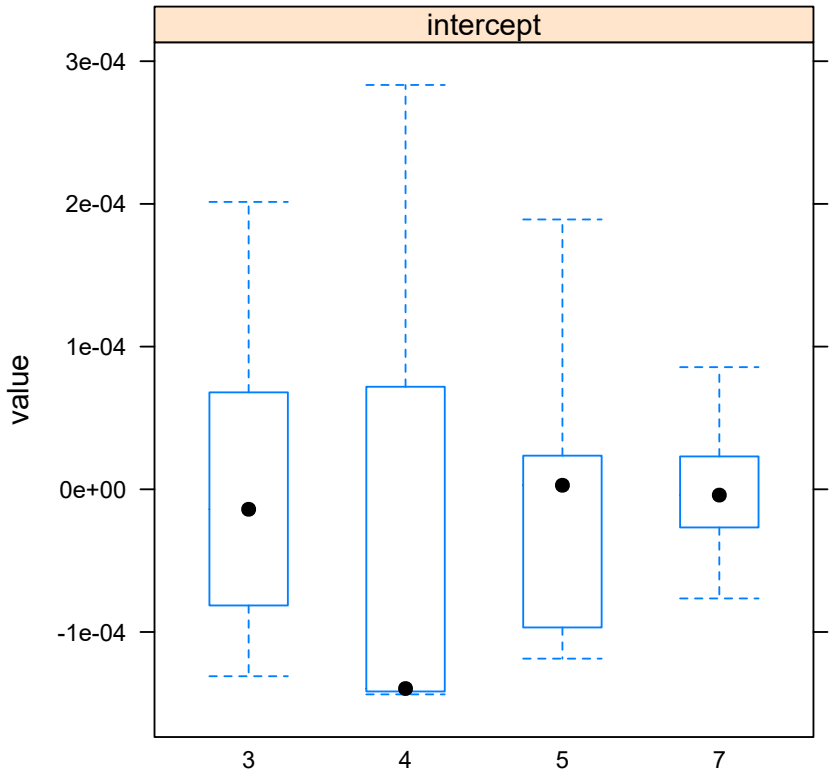


Figure 22: Subject specific random effect estimates $\widehat{b_{0i}}$ for Model 4, a LMEM with a 3 degree polynomial.

given the profiles discussed in Section 1.2. In the next subsection, we explore LMEMs which use cubic B-spline terms in place of the orthogonal polynomial terms.

3.1.3 *LMEMS with Cubic B-Splines of Three Degrees of Freedom*

In this section we consider models with cubic B-splines in place of orthogonal polynomials. The main advantage of cubic B-splines over orthogonal polynomials is that they are more flexible. However, as we are using a limited number of degrees of freedom for the cubic B-splines given the limited number of time points for which we have data, we are not taking full advantage of their increased flexibility.

The general structure of the models that we consider in this subsection is given by

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \epsilon_{ijk}$$

where y_{ijk} is the immune response j for subject i at time k (t_k), $B_{w,m}(t)$ is the w^{th} B-spline basis function of order m and $\epsilon_{ijk} \sim N(0, \sigma^2)$. The

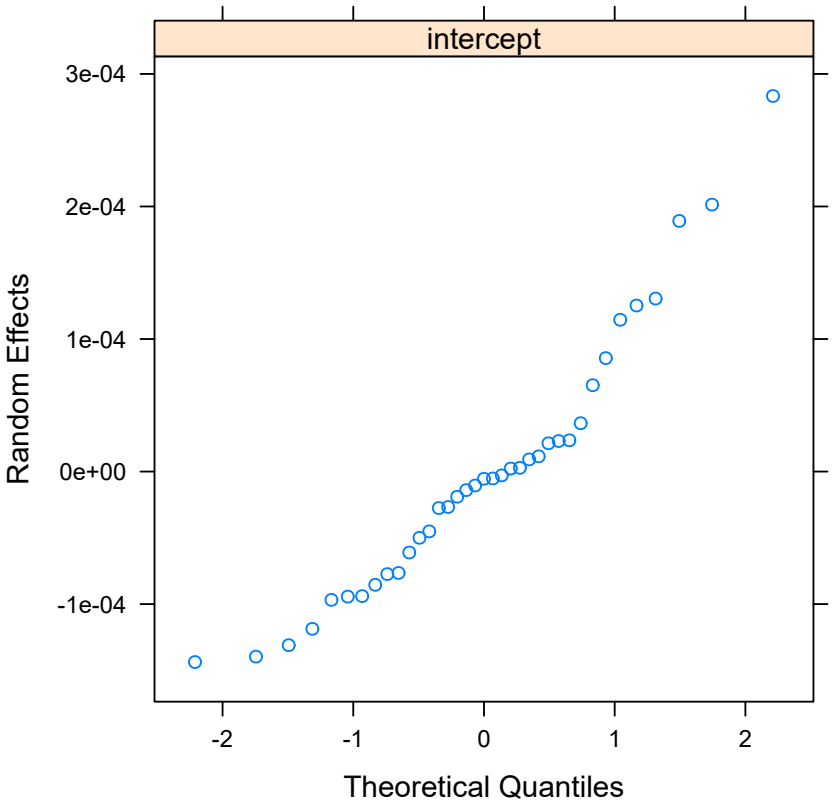


Figure 23: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i}}$ for Model 4, a LMEM with a 3 degree polynomial.

parameters to be estimated are the $\phi_{w,ij}$s and σ. See Hastie *et al.* (2009, p. 186) for a definition of B-splines. For the B-splines, we use knots placed at the percentiles, but as we use only three degrees of freedom, this means we only have knots at the boundaries, $t = 0$ and $t = 365$. Increasing the degrees of freedom beyond three provides too much flexibility to the spline and results in profiles which are not consistent with those discussed and shown in Section 1.2. Ultimately, we have data at too few time points to reasonably estimate coefficients for a more flexible spline function without the estimated immune response profiles having unreasonable shapes.

Individual Specific Models

Before fitting multi-level models to all of the data, we first consider models fitted for each subject's data. In these subject specific models, we treat the response as a fixed effect for the intercept in a similar

manner to what we did for the LMEMs with orthogonal polynomials. The explicit form of these subject specific models is given by

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \epsilon_{ijk}$$

where

$$\phi_{0,ij} = \beta_{0,1} + \beta_{0,\text{Response}_j} \times \text{Response}_j \quad j = \texttt{CD4Gn2pTp}, \ldots, \texttt{CD4Gp2nTn},$$

where $\beta_{0,1}$ is the intercept and corresponds to response `CD4Gn2nTp`, and

$$\phi_{w,ij} = \beta_{w,1} \quad w = 1, 2, 3.$$

This model is fitted for each individual subject separately and the result is a set of parameters which are subject specific.

Figure 24 presents the approximate 95% confidence intervals for the subject specific parameters for each subject and parameter. The panel for (Intercept) corresponds to $\beta_{0,1}$, each of the response panels corresponds to a $\beta_{0,\text{Response}_j}$ for $j = \texttt{CD4Gn2pTn}, \ldots, \texttt{CD4Gp2pTp}$, and the bs1, bs2, and bs3 panels correspond to a $\beta_{w,1}$ for $w = 1, 2, 3$ respectively. As before, we are concerned with non-overlapping confidence intervals as these suggest that we may need to include random effects for the corresponding parameter. We can see that many of the confidence intervals do overlap, but not in all cases. For each parameter, there are confidence intervals that do not overlap. This is particularly clear for $\beta_{0,\text{CD4Gp2nTn}}$, $\beta_{0,\text{CD4Gp2pTp}}$, $\beta_{1,1}$ and $\beta_{3,1}$ where we see that several of the confidence intervals do not overlap. For the other parameters there are confidence intervals that do not overlap, but they are not as pronounced. As a starting point in our multi-level models, we include random effects for each of the parameters at the subject and response within subject levels, and then remove those that are not needed.

Multi-Level Models

In this section we fit models which include random effects and add fixed effects as needed. We consider random effects at the subject level and the response within subject level.

MODEL 5 The first model with a cubic B-spline that we fit has the form

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \epsilon_{ijk}$$

with

$$\phi_{0,ij} = \beta_{0,1} + b_{0i} + b_{0i,j}$$

and

$$\phi_{w,ij} = \beta_{w,1} + b_{wi} + b_{wi,j} \quad w = 1, 2, 3,$$

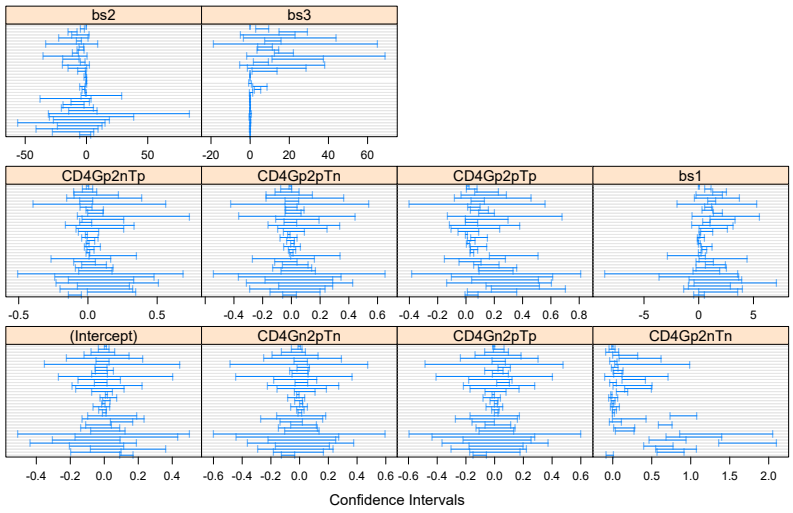


Figure 24: Approximate 95% confidence intervals for the parameter estimates from the subject specific cubic B-spline models fitted to the vaccine data.

where

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_1) \quad \text{and} \quad \boldsymbol{b}_{i,j} = \begin{pmatrix} b_{0i,j} \\ b_{1i,j} \\ b_{2i,j} \\ b_{3i,j} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_2),$$

and $\boldsymbol{b}_i$ is independent of $\boldsymbol{b}_{i,j}$, and both random effects are independent of ϵ_{ijk}. We consider diagonal structures for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ as more complex structures proved challenging to fit.

Table 6 summarises the results from this model. It shows the parameter estimates for the fixed effects together with their standard errors, the degrees of freedom, and associated significance. The AIC and BIC values for this model are -754.4 and -687.3 respectively. This is better than Model 1, the comparably polynomial model, but not better than Model 4, the best polynomial model that we considered.

Table 6: Output for Model 5, a cubic B-spline model with three degrees of freedom.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	0.106	0.016	1026	6.779	<0.001
$\beta_{1,1}$	-0.664	0.173753	2825	-3.8243	<0.001
$\beta_{2,1}$	-0.135	0.201827	2825	-0.66652	<0.001
$\beta_{3,1}$	1.237	0.162191	2825	7.627624	0.299

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\psi}_1 = \begin{pmatrix} 0.053^2 & 0 & 0 & 0 \\ 0 & 0.177^2 & 0 & 0 \\ 0 & 0 & (1.104 \times 10^{-5})^2 & 0 \\ 0 & 0 & 0 & 0.044^2 \end{pmatrix}$$

and

$$\widehat{\psi}_2 = \begin{pmatrix} 0.180^2 & 0 & 0 & 0 \\ 0 & (1.699 \times 10^{-5})^2 & 0 & 0 \\ 0 & 0 & (1.400 \times 10^{-5})^2 & 0 \\ 0 & 0 & 0 & 0.151^2 \end{pmatrix},$$

and the parameter estimate for σ, the standard deviation for the within group error, is 0.135. We see that the estimates of the standard deviations for several of the random effects are close to zero suggesting that they are not needed. We will remove these in the next iteration of this model.

We present several diagnostic plots for this model in Figures 84 to 90 which are in Appendix B. The plots are not too dissimilar to those for Model 1. The important features to note about the residual plots are:

- the variance in the residuals increases as the fitted values increase for several of the vaccine and immune response combinations (Figure 84). This indicates that the assumption of constant variance is inappropriate.
- for most of the vaccines, there is little evidence of patterns in the residuals over time, the exception is vaccine 3 and 7, where we see a minor pattern over time for some of the immune responses (Figure 85).
- the residuals are heavily skewed and are clearly non-normal. The assumption that the within-group errors are normally distributed is clearly not satisfied.

From the plots of the random effect estimates, we can note:

- for the estimates of the b_{0i}, b_{1i}, and b_{3i} random effects, there is a pattern by vaccine and so we should include the vaccine fixed effect (Figure 87). We expect there to be a vaccine effect as we expect the vaccines to elicit different immune responses.
- for $b_{0i,j}$ and $b_{3i,j}$, the random effect estimates are not all centered around zero, and that vaccine and response fixed effects may be required (Figure 88).
- from the QQ-plots of the random effects estimates at the subject and response within subject levels, several of the random effects appear to be non-normal (Figures 89 and 90). The assumption that the random effects are normally distributed is not satisfied.

MODEL 6 We try to improve on the previous model by:

- including fixed effects for vaccine and response, and their interaction, for the intercept parameter and
- dropping the random effects b_{2i}, $b_{1i,j}$, and $b_{2i,j}$. We also drop the random effect b_{1i} as it turned out not to be needed.

The specific form of the model that we fit is given by

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \epsilon_{ijk}$$

with

$$\begin{aligned} \phi_{0,ij} &= (\beta_{0,1} + b_{0i} + b_{0i,j}) + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j + \\ &\quad \beta_{0,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\ \phi_{1,ij} &= \beta_{1,0}, \\ \phi_{2,ij} &= \beta_{2,0}, \end{aligned}$$

and

$$\phi_{3,ij} = \beta_{3,0} + b_{3i} + b_{3i,j}.$$

The distribution assumptions for ϵ_{ijk} and the random effects remain the same with the relevant terms removed.

Table 7 summarises the results from this model. It shows the parameter estimates for the fixed effects together with their standard errors, the degrees of freedom, and associated significance. The AIC and BIC values for this model are -958.5 and -772.7 respectively. This is better than Model 5, the previous model, but not better than Model 4, the best polynomial model that we considered.

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.040^2 & 0 \\ 0 & 0.049^2 \end{pmatrix},$$

and

$$\widehat{\boldsymbol{\psi}}_2 = \begin{pmatrix} 0.093^2 & 0 \\ 0 & 0.150^2 \end{pmatrix},$$

and the parameter estimate for σ, the standard deviation for the within group error, is 0.138.

We present several diagnostic plots for this model in Figures 91 to 97 in Appendix B. Very little has changed in the residual plots relative to the previous model that we considered. There is still a fan shape to some of the residuals when plotted against the fitted values and there is some minor evidence of a time pattern in the residuals for some of the vaccine and response combinations. Importantly, the residuals are also still clearly non-normal. For the random effects, we see that they are mostly well centered around zero indicating that the inclusion of the fixed effects for vaccine and response were needed in the model. The QQ-plots of the random effect estimates still indicate that several of them are non-normal, similar to what we saw for Model 5.

Table 7: Output for Model 6, a cubic B-spline model with three degrees of freedom.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	0.03	0.031	1026	0.957	0.339
$\beta_{0,\text{vaccine4}}$	-0.026	0.077	33	-0.339	0.737
$\beta_{0,\text{vaccine5}}$	-0.028	0.06	33	-0.464	0.646
$\beta_{0,\text{vaccine7}}$	0.033	0.046	33	0.724	0.474
$\beta_{0,\text{responseCD4Gn2pTn}}$	-0.001	0.041	198	-0.028	0.977
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.02	0.041	198	0.481	0.631
$\beta_{0,\text{responseCD4Gp2nTn}}$	0.166	0.041	198	4.029	<0.001
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.062	0.041	198	1.513	0.132
$\beta_{0,\text{responseCD4Gp2pTn}}$	0.025	0.041	198	0.616	0.538
$\beta_{0,\text{responseCD4Gp2pTp}}$	0.11	0.041	198	2.665	0.008
$\beta_{1,1}$	0.136	0.036	1026	3.758	<0.001
$\beta_{2,1}$	-0.161	0.039	1026	-4.138	<0.001
$\beta_{3,1}$	0.008	0.024	1026	0.318	0.75
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-0.011	0.102	198	-0.108	0.914
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	0.003	0.079	198	0.043	0.965
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	-0.015	0.061	198	-0.245	0.807
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-0.032	0.102	198	-0.309	0.758
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	-0.015	0.079	198	-0.186	0.853
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.04	0.061	198	-0.651	0.516
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-0.169	0.102	198	-1.649	0.101
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	-0.151	0.079	198	-1.906	0.058
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	0.481	0.061	198	7.847	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-0.06	0.102	198	-0.587	0.558
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	-0.047	0.079	198	-0.598	0.551
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	0.011	0.061	198	0.186	0.853
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-0.04	0.102	198	-0.39	0.697
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	-0.022	0.079	198	-0.283	0.777
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	0	0.061	198	0.001	0.999
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-0.072	0.102	198	-0.706	0.481
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	-0.061	0.079	198	-0.764	0.446
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	0.107	0.061	198	1.753	0.081

MODEL 7 This is the last iteration of the cubic B-spline LMEMs that we consider. To try and improve the fit of the previous model, we make the following changes:

- we include a vaccine fixed effect for the spline terms,
- we drop the random effect b_{3i} as doing so improves the information criteria significantly, and
- we now model the variance of the within-group errors as a function of the fitted values; i.e.

$$\text{Var}(\epsilon_{ijk}) = \sigma^2 \widehat{y}_{ijk}.$$

The explicit form of this model is given by

$$y_{ijk} = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \epsilon_{ijk}$$

with

$$\begin{aligned}
\phi_{0,ij} &= (\beta_{0,1} + b_{0i} + b_{0i,j}) + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j + \\
&\quad \beta_{0,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\
\phi_{1,ij} &= \beta_{1,0} + \beta_{1,\text{vaccine}_i}\text{vaccine}_i, \\
\phi_{2,ij} &= \beta_{2,0} + \beta_{2,\text{vaccine}_i}\text{vaccine}_i,
\end{aligned}$$

and

$$\phi_{3,ij} = (\beta_{3,0} + \beta_{3,\text{vaccine}_i}\text{vaccine}_i + b_{3i,j}).$$

The distribution assumption for the random effects remains the same.

Table 8 summarises the results from this model. It shows the parameter estimates for the fixed effects together with their standard errors, the degrees of freedom, and associated significance. The AIC and BIC values for this model are -3045.3 and -2818.2 respectively. This is better than Model 6, the previous model, and better than Model 4, the best polynomial model that we considered.

The parameter estimates for ψ_1 and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = (9.935 \times 10^{-7})^2$$

and

$$\widehat{\boldsymbol{\psi}}_2 = \begin{pmatrix} (6.516 \times 10^{-7})^2 & 0 \\ 0 & (7.242 \times 10^{-10})^2 \end{pmatrix},$$

and the parameter estimate for σ, the standard deviation for the within group error, is 0.333.

We present several diagnostic plots for this model in Figures 25 to 31. For the within-group errors, we see that the fan pattern is removed and the time pattern is removed. The within-group errors are still skewed and non-normal; the assumptions regarding the within-group errors are clearly not satisfied. For the random effect estimates, we see that they are all very close to zero and that the assumption regarding normality is not satisfied with deviations in the tails for the random effects on the intercept parameter.

Table 8: Output for Model 7, a cubic B-spline model with three degrees of freedom.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	0.001	0.003	1017	0.414	0.679
$\beta_{0,\text{vaccine4}}$	0.012	0.012	33	0.996	0.327
$\beta_{0,\text{vaccine5}}$	0.003	0.005	33	0.613	0.544
$\beta_{0,\text{vaccine7}}$	0.046	0.014	33	3.349	0.002
$\beta_{0,\text{responseCD4Gn2pTn}}$	0	0.004	198	-0.019	0.985
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.013	0.007	198	1.712	0.088
$\beta_{0,\text{responseCD4Gp2nTn}}$	0.111	0.016	198	7.041	<0.001
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.021	0.009	198	2.235	0.027
$\beta_{0,\text{responseCD4Gp2pTn}}$	0.004	0.005	198	0.795	0.428
$\beta_{0,\text{responseCD4Gp2pTp}}$	0.082	0.014	198	5.744	<0.001
$\beta_{1,1}$	1.491	0.134	1017	11.157	<0.001
$\beta_{2,1}$	-8.004	0.799	1017	-10.014	<0.001
$\beta_{3,1}$	15.419	1.557	1017	9.902	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-0.012	0.012	198	-0.939	0.349
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	0	0.007	198	-0.028	0.977
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	-0.014	0.016	198	-0.881	0.379
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-0.021	0.015	198	-1.423	0.156
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	-0.01	0.01	198	-0.999	0.319
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.03	0.017	198	-1.779	0.077
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-0.118	0.021	198	-5.746	<0.001
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	-0.102	0.018	198	-5.714	<0.001
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	0.54	0.043	198	12.451	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-0.021	0.018	198	-1.16	0.247
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	-0.012	0.012	198	-1.004	0.317
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	0.036	0.022	198	1.623	0.106
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-0.015	0.013	198	-1.219	0.224
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	-0.005	0.008	198	-0.591	0.555
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	0.024	0.019	198	1.239	0.217
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-0.045	0.027	198	-1.675	0.095
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	-0.04	0.019	198	-2.036	0.043
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	0.123	0.03	198	4.056	<0.001
$\beta_{1,\text{vaccine4}}$	-1.365	0.159	1017	-8.567	<0.001
$\beta_{1,\text{vaccine5}}$	-1.248	0.166	1017	-7.52	<0.001
$\beta_{1,\text{vaccine7}}$	-0.961	0.263	1017	-3.655	<0.001
$\beta_{2,\text{vaccine4}}$	7.623	0.848	1017	8.984	<0.001
$\beta_{2,\text{vaccine5}}$	6.964	0.929	1017	7.493	<0.001
$\beta_{2,\text{vaccine7}}$	5.282	2.412	1017	2.189	0.029
$\beta_{3,\text{vaccine4}}$	-15.417	1.557	1017	-9.9	<0.001
$\beta_{3,\text{vaccine5}}$	-13.733	1.744	1017	-7.876	<0.001
$\beta_{3,\text{vaccine7}}$	-15.404	1.557	1017	-9.892	<0.001

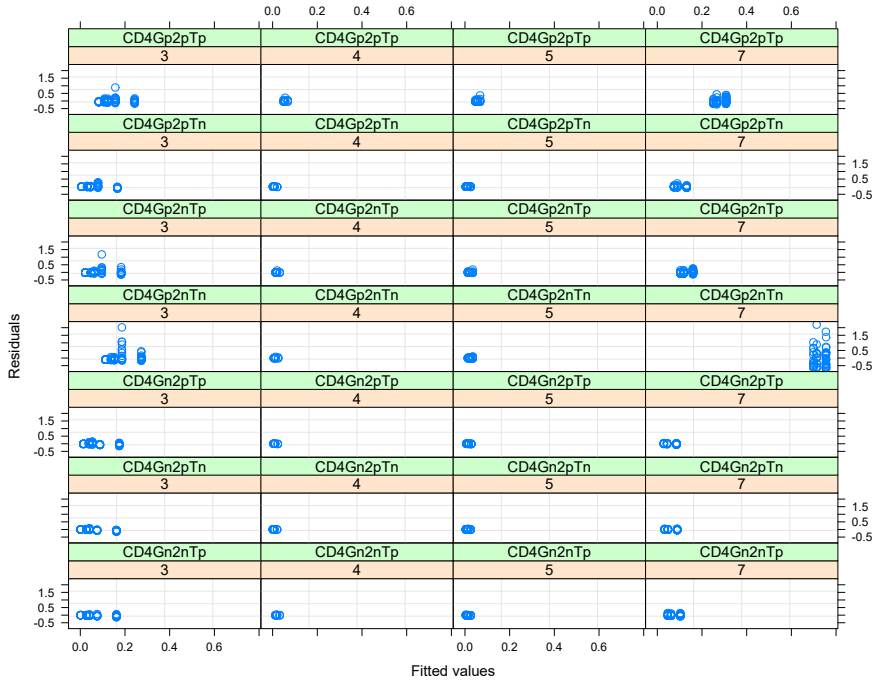


Figure 25: Standardised residuals versus fitted values by vaccine and response for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

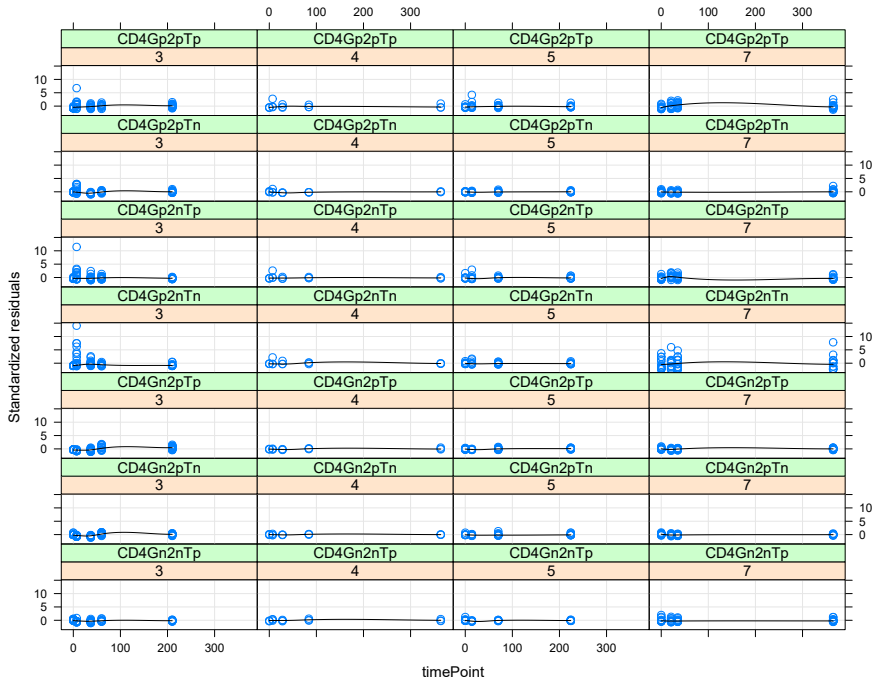


Figure 26: Standardised residuals versus time point by vaccine and response for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

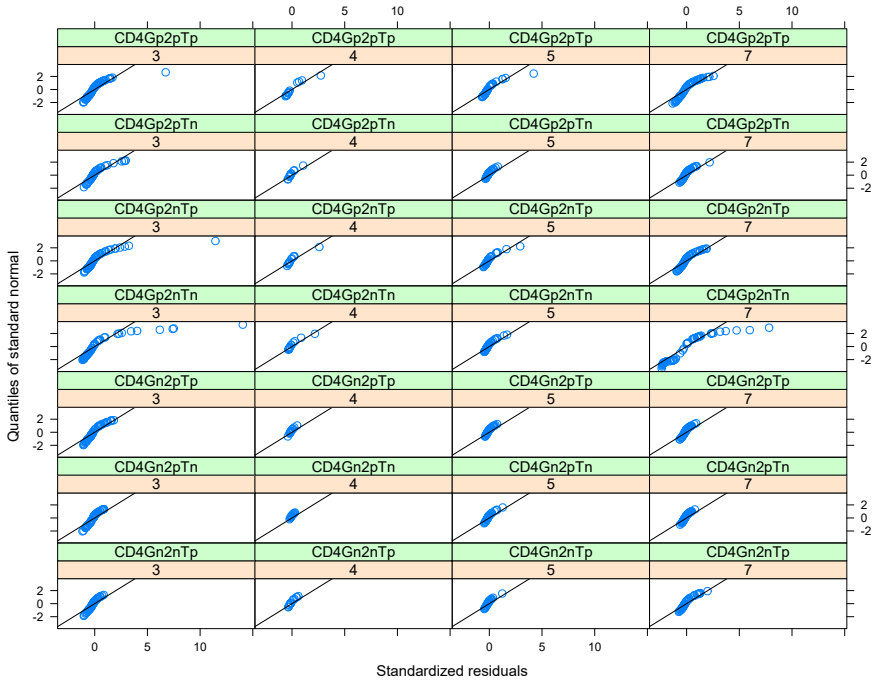


Figure 27: QQ-plot for the standardised residuals by vaccine and response for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

3.1.4 *Comments on the LMEMs with cubic B-splines*

In this subsection we have explored several LMEMs with cubic B-splines with three degrees of freedom. We have seen that, similar to the LMEMs with orthogonal polynomials, these models do not fit the data well and that the assumptions underlying the LMEMs are not satisfied. Specifically, the within-group errors are non-normal and the random effect estimates in some cases are also non-normal. Again, as the fitted models do not satisfy the assumptions underlying them, we cannot reliably use them to perform inference to decide which vaccine candidates are the most promising. In the next subsection, we explore several NLMEMs.

3.2 NONLINEAR MIXED EFFECT MODELS

In this section we explore NLMEMs which have the following general form

$$y_{ijk} = f(\boldsymbol{\phi}_{ij}, t_k) + \epsilon_{ijk}$$

where $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$, and the function $f(\cdot)$ is nonlinear in one of the elements of $\boldsymbol{\phi}_{ij}$. The parameters can contain fixed and random effects in the same way as the parameters of the LMEMs can.

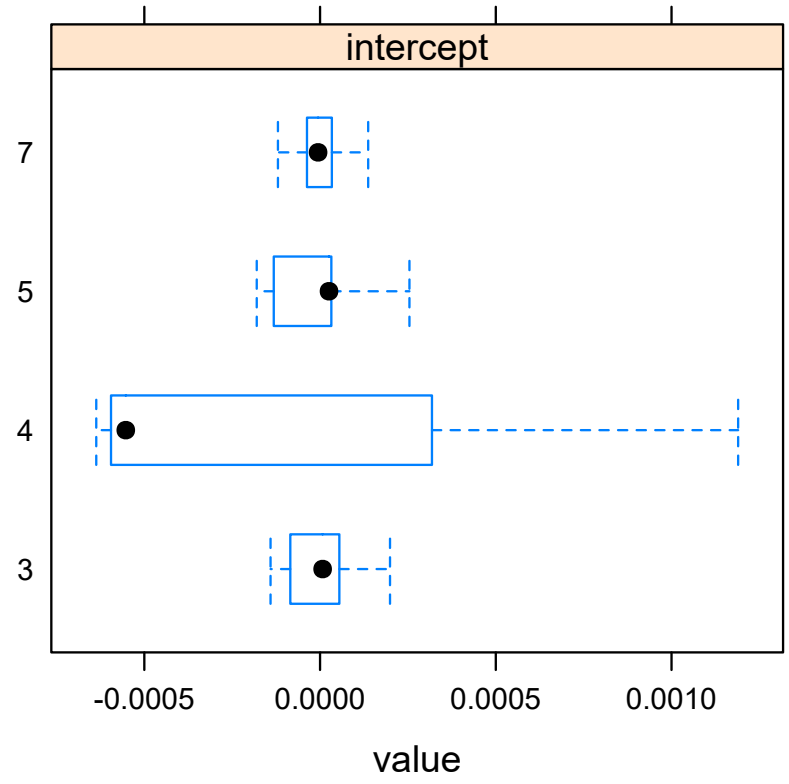


Figure 28: Subject specific random effect estimates $\widehat{b_{0i}}$ for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

There are several possible options for $f(\cdot)$ from the literature which broadly match the profile seen for the immune responses in Section 1.2. We consider the following possible options in this section (we suppress $\boldsymbol{\phi}_{ij}$ for convenience):

- the bi-exponential model:

$$f(x) = \phi_1 \exp[-\exp(\phi_2)x] - \phi_3 \exp[-\exp(\phi_4)x]$$

  as presented by Pinheiro and Bates (2000, p. 514),

- the first-order compartment model:

$$f(x) = \frac{D \exp(\phi_1) \exp(\phi_2)}{\exp(\phi_3)[\exp(\phi_2) - \exp(\phi_1)]} \times (\exp[-\exp(\phi_1)x] - \exp[-\exp(\phi_2)x])$$

  as presented by Pinheiro and Bates (2000, p. 516),

- the model

$$f(x) = \frac{A(1+d)\exp(-bx)}{1 + d\exp(-cx)}$$

  as considered by Weigend *et al.* (1997), and

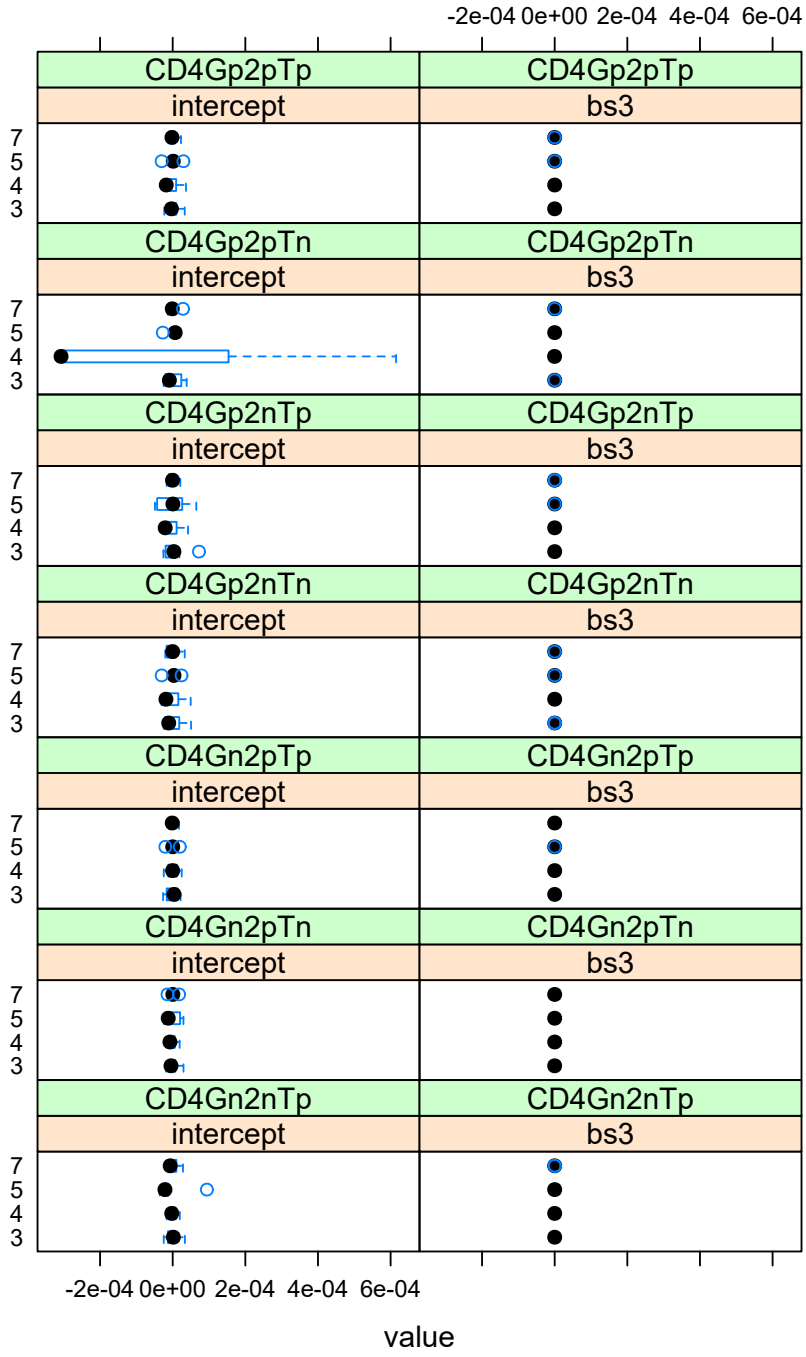


Figure 29: Response within subject random effect estimates $\widehat{b_{0i,j}}$ and $\widehat{b_{3i,j}}$ for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

- and a modified version of the transit compartment model:

$$f(x) = B + D\frac{(kx)^n}{n!}\exp(-kx) \tag{11}$$

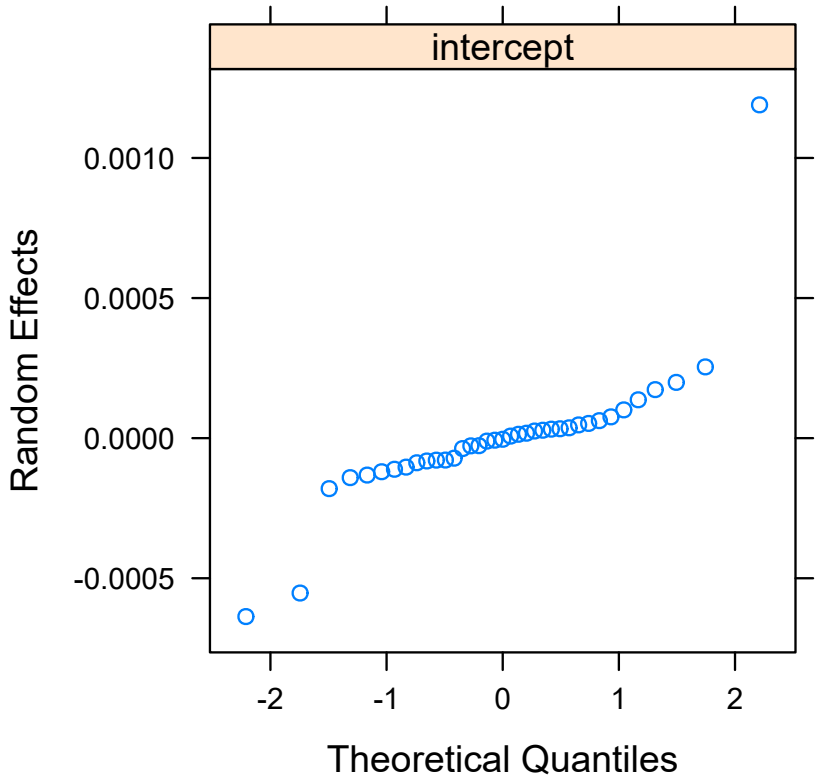


Figure 30: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i}}$ for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

as considered by Savic *et al.* (2007). The change that we made was to add the baseline term B which allows for a non-zero starting value.

We struggled to fit most of these models to the vaccine data. The model which we had the most success with was the modified transit compartment model with $n = 1$ and this is the one that we present here.

In the subsections that follow, we first consider subject specific models and then move on to consider multilevel models.

3.2.1 *Individual Specific Models*

Subject specific fits of the modified transit compartment model are considered here. The specific model we consider is given by

$$y_{ijk} = \phi_{0,ij} + \phi_{1,ij}(\phi_{2,ij}t_k)\exp\left(-\phi_{2,ij}t_k\right) + \epsilon_{ijk} \tag{12}$$

where the $\phi_{w,ij}$s are the parameters to be estimated for each subject.

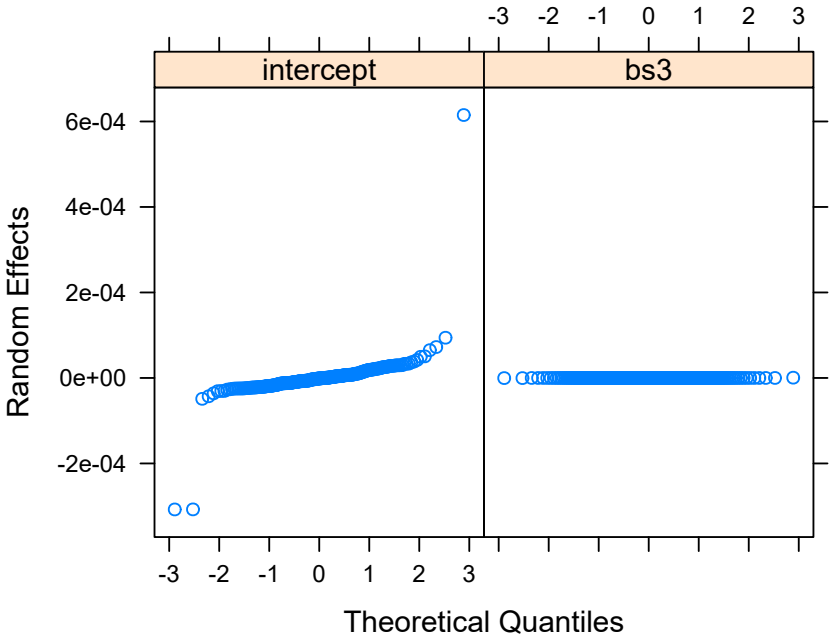


Figure 31: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i,j}}$ and $\widehat{b_{3i,j}}$ for Model 7, a LMEM with a cubic B-spline with three degrees of freedom.

The parameters given in Equation (11) correspond to those in Equation (12); the graphs in the figures that we present will make use of the labelling given in Equation (11).

Figure 32 presents a plot of the parameter estimates and their approximate 95% confidence intervals for the subject specific models. We can use this graph to try and identify the parameters that may require random effects; Figure 33 contains the same graph but excludes the individuals with large confidence intervals. We can see that there are non-overlapping confidence intervals for $\phi_{0,ij}$ (B), $\phi_{1,ij}$ (D), and $\phi_{2,ij}$ (k) suggesting that we may need random effects for all of the parameters. As such, the first multilevel NLMEM that we consider has random effects at the subject and response within subject levels for all of the parameters.

3.2.2 *Multi-level NLMEMs*

In this subsection we investigate multi-level versions of the above model which include random and fixed effects as needed.

MODEL 8 The first multi-level NLMEM that we consider includes random effects for all three parameters. Specifically it has the form:

$$y_{ijk} = \phi_{0,ij} + \phi_{1,ij}(\phi_{2,ij}t_k)\exp\left(-\phi_{2,ij}t_k\right) + \epsilon_{ijk}$$

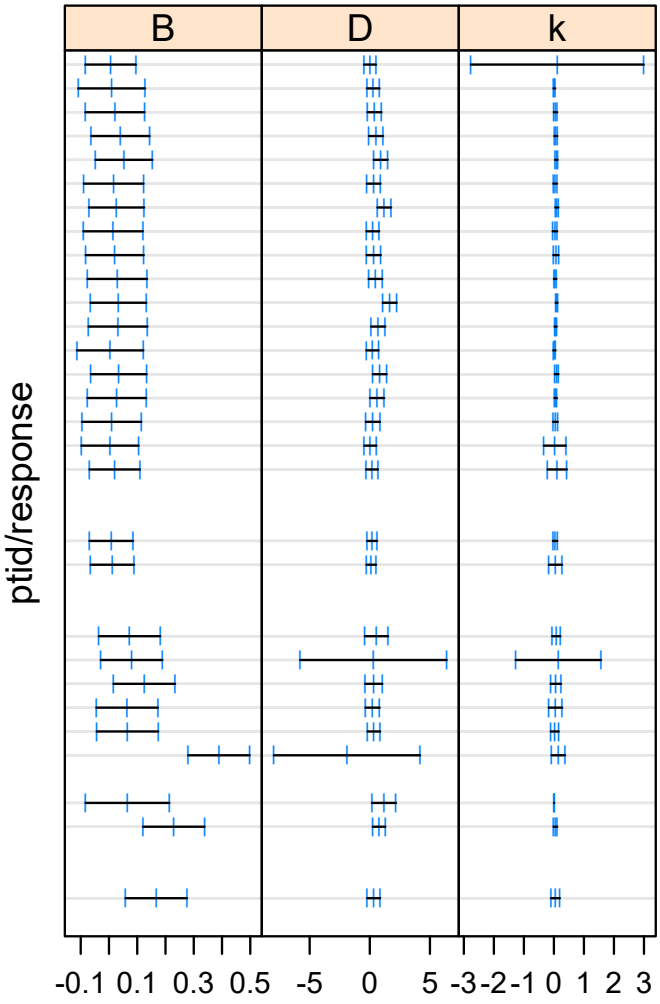


Figure 32: Approximate 95% confidence intervals for the parameter estimates from the subject specific nonlinear models fitted to the vaccine data (I).

where

$$\begin{aligned} \phi_{0,ij} &= \beta_{0,1} + b_{0i} + b_{0i,j} \\ \phi_{1,ij} &= \beta_{1,1} + b_{1i} + b_{1i,j} \\ \phi_{2,ij} &= \beta_{2,1} + b_{2i} + b_{2i,j}, \end{aligned}$$

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_1) \quad \boldsymbol{b}_{i,j} = \begin{pmatrix} b_{0i,j} \\ b_{1i,j} \\ b_{2i,j} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_2),$$

and $\boldsymbol{b}_i$ is independent of $\boldsymbol{b}_{i,j}$, and both are independent of ϵ_{ijk}. We consider diagonal structures for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ as more complex structures proved challenging to fit. We assume that $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ and are independent.

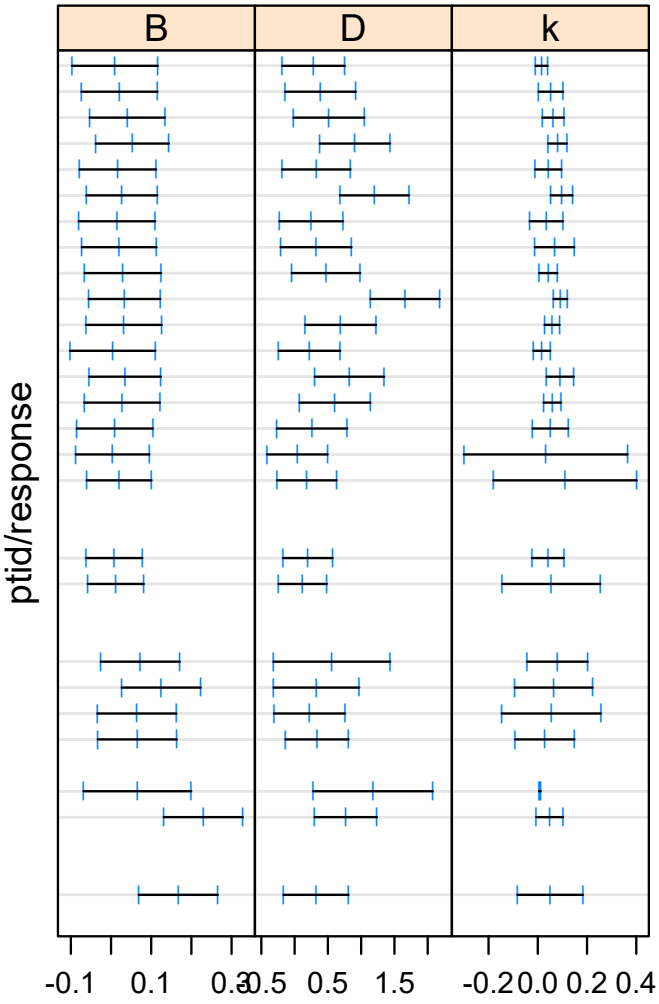


Figure 33: Approximate 95% confidence intervals for the parameter estimates from the subject specific nonlinear models fitted to the vaccine data (II).

The estimates for the fixed effects $\beta_{0,1}$, $\beta_{1,1}$, and $\beta_{2,1}$ are given in Table 9 along with their significance. The AIC and BIC for this model are -1313.2 and -1261.6 which are significantly better than many of the models we have considered so far, but not the best.

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.063^2 & 0 & 0 \\ 0 & 0.291^2 & 0 \\ 0 & 0 & 0.031^2 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{\psi}}_2 = \begin{pmatrix} 0.185^2 & 0 & 0 \\ 0 & 0.743^2 & 0 \\ 0 & 0 & 0.002^2 \end{pmatrix},$$

Table 9: Output from Model 8, a NLME model.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	0.079	0.016	1027	5.037	<0.001
$\beta_{1,1}$	0.336	0.069	1027	4.87	<0.001
$\beta_{2,1}$	0.058	0.006	1027	9.398	<0.001

and the parameter estimate for σ, the standard deviation for the within group error, is 0.079. We see that the random effect $b_{2i,j}$ may be unnecessary as it has a relatively small variance.

We present some diagnostic plots for Model 8 in Figures 34 to 40. In these figures we consider the residuals and random effect estimates.

Looking at Figures 34 to 36 which consider the residuals from the model, we make the following observations:

- there is no clear pattern in the residuals when plotted against the fitted values; the fan pattern we saw previously for the LMEMs is not as apparent.
- there is no clear pattern in the residuals when plotted against the time points.
- the residuals are clearly non-normal as the points in the QQ-plots do not lie on straight lines. As such, the assumption that the within-group residuals are normally distributed is not satisfied.

Figures 37 to 40 consider the random effect estimates. We make the following observations about these figures:

- there appears to be clear vaccine effects for b_{0i} (B) and b_{1i} (D) (Figure 37).
- there also appears to be a slight vaccine and response effect for $b_{0i,j}$ (B) and $b_{1i,j}$ (D) for some of the responses (Figure 38).
- several of the random effect estimates arise from distributions that are clearly non-normal and as such do not satisfy the assumptions of the model (Figures 39 to 40).

MODEL 9 To try and improve the above NLMEM we make the following changes to Model 8:

- we include vaccine and response fixed effects for $\phi_{0,ij}$ and $\phi_{1,ij}$. We also include the interactions between response and vaccine as these proved to be useful.
- we dropped the $b_{2i,j}$ random effect as it was not needed.

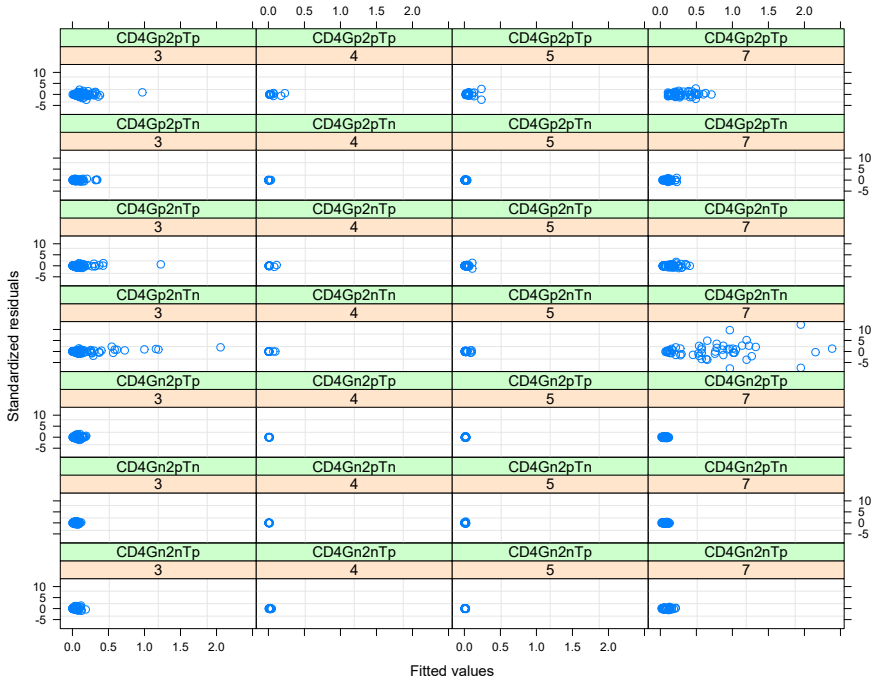


Figure 34: Standardised residuals versus fitted values by vaccine and response for Model 8, a NLMEM.

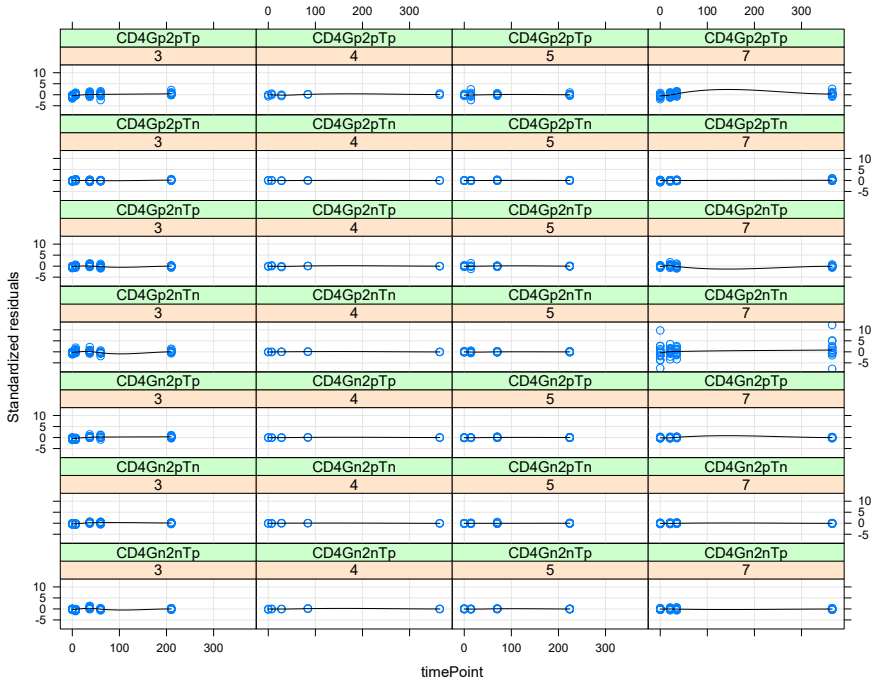


Figure 35: Standardised residuals versus time point by vaccine and response for Model 8, a NLMEM.

Specifically, the model we consider is given by

$$y_{ijk} = \phi_{0,ij} + \phi_{1,ij}(\phi_{2,ij}t_k)\exp\left(-\phi_{2,ij}t_k\right) + \epsilon_{ijk}$$

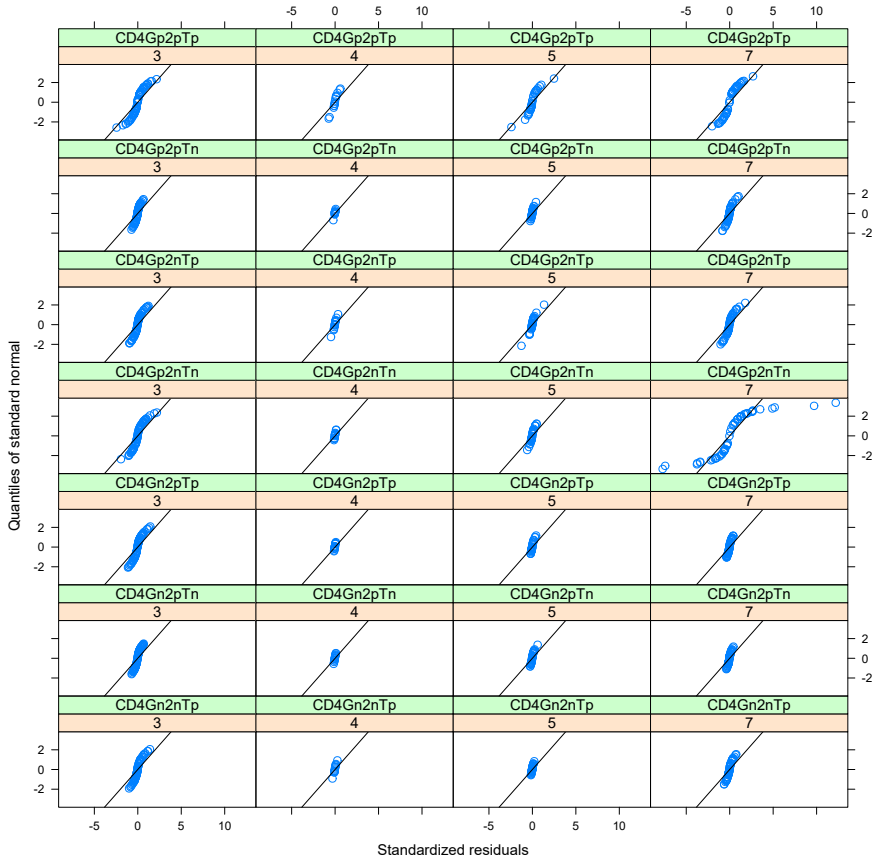


Figure 36: QQ-plots for the standardised residuals by vaccine and response for Model 8, a NLMEM.

where

$$
\begin{aligned}
\phi_{0,ij} &= \beta_{0,1} + b_{0i} + b_{0i,j} + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j + \\
&\quad \beta_{0,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\
\phi_{1,ij} &= \beta_{1,1} + b_{1i} + b_{1i,j} + \beta_{1,\text{vaccine}_i}\text{vaccine}_i + \beta_{1,\text{response}_j}\text{response}_j + \\
&\quad \beta_{1,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\
\phi_{2,ij} &= \beta_{2,1} + b_{2i},
\end{aligned}
$$

$$
\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\psi}_1) \quad \boldsymbol{b}_{i,j} = \begin{pmatrix} b_{0i,j} \\ b_{1i,j} \end{pmatrix} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\psi}_2),
$$

and $\boldsymbol{b}_i$ is independent of $\boldsymbol{b}_{i,j}$, and both random effects are independent of ϵ_{ijk}. We still consider diagonal structures for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ as more complex structures proved challenging to fit.

The estimates for the fixed effects are given in Table 10 along with their significance. The AIC and BIC for this model are -1492.4 and -1167.3; which is significantly better than the previous NLMEM, but not better than Model 7, the best cubic B-spline LMEM we considered.

Table 10: Output from Model 9, a NLME model.

Parameter	Estimate	Std. Error	DF	t-value	p-value
$\beta_{0,1}$	0.016	0.035	1000	0.464	0.643
$\beta_{0,\text{vaccine4}}$	-0.002	0.084	33	-0.024	0.981
$\beta_{0,\text{vaccine5}}$	-0.009	0.067	33	-0.127	0.9
$\beta_{0,\text{vaccine7}}$	0.043	0.05	33	0.856	0.398
$\beta_{0,\text{responseCD4Gn2pTn}}$	0.01	0.047	198	0.208	0.835
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.028	0.047	198	0.6	0.549
$\beta_{0,\text{responseCD4Gp2nTn}}$	-0.002	0.047	198	-0.033	0.974
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.004	0.047	198	0.091	0.928
$\beta_{0,\text{responseCD4Gp2pTn}}$	0	0.047	198	-0.01	0.992
$\beta_{0,\text{responseCD4Gp2pTp}}$	0.063	0.047	198	1.348	0.179
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-0.02	0.113	198	-0.176	0.86
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	-0.009	0.09	198	-0.098	0.922
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	-0.026	0.068	198	-0.381	0.703
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-0.035	0.113	198	-0.312	0.755
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	-0.023	0.09	198	-0.253	0.801
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.046	0.068	198	-0.679	0.498
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-0.008	0.113	198	-0.072	0.943
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	0.007	0.09	198	0.076	0.939
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	0.621	0.068	198	9.122	<0.001
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-0.012	0.113	198	-0.108	0.914
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	0.002	0.09	198	0.026	0.98
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	0.02	0.068	198	0.296	0.767
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-0.013	0.113	198	-0.11	0.912
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	0.001	0.09	198	0.008	0.994
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	0.031	0.068	198	0.452	0.652
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-0.044	0.113	198	-0.392	0.696
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	-0.033	0.09	198	-0.362	0.718
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	0.091	0.068	198	1.336	0.183
$\beta_{1,1}$	0.187	0.181	1000	1.032	0.302
$\beta_{1,\text{vaccine4}}$	-0.145	0.439	1000	-0.33	0.742
$\beta_{1,\text{vaccine5}}$	-0.191	0.364	1000	-0.527	0.599
$\beta_{1,\text{vaccine7}}$	-0.087	0.261	1000	-0.332	0.74
$\beta_{1,\text{responseCD4Gn2pTn}}$	-0.093	0.239	1000	-0.388	0.698
$\beta_{1,\text{responseCD4Gn2pTp}}$	-0.081	0.239	1000	-0.341	0.733
$\beta_{1,\text{responseCD4Gp2nTn}}$	1.612	0.239	1000	6.738	<0.001
$\beta_{1,\text{responseCD4Gp2nTp}}$	0.568	0.239	1000	2.378	0.018
$\beta_{1,\text{responseCD4Gp2pTn}}$	0.266	0.239	1000	1.114	0.266
$\beta_{1,\text{responseCD4Gp2pTp}}$	0.433	0.239	1000	1.815	0.07
$\beta_{1,\text{vaccine4:responseCD4Gn2pTn}}$	0.075	0.576	1000	0.131	0.896
$\beta_{1,\text{vaccine5:responseCD4Gn2pTn}}$	0.113	0.477	1000	0.236	0.814
$\beta_{1,\text{vaccine7:responseCD4Gn2pTn}}$	0.106	0.343	1000	0.309	0.758
$\beta_{1,\text{vaccine4:responseCD4Gn2pTp}}$	0.05	0.576	1000	0.087	0.931
$\beta_{1,\text{vaccine5:responseCD4Gn2pTp}}$	0.081	0.477	1000	0.17	0.865
$\beta_{1,\text{vaccine7:responseCD4Gn2pTp}}$	0.082	0.343	1000	0.241	0.81
$\beta_{1,\text{vaccine4:responseCD4Gp2nTn}}$	-1.559	0.576	1000	-2.704	0.007
$\beta_{1,\text{vaccine5:responseCD4Gp2nTn}}$	-1.528	0.477	1000	-3.2	0.001
$\beta_{1,\text{vaccine7:responseCD4Gp2nTn}}$	-1.404	0.346	1000	-4.057	<0.001
$\beta_{1,\text{vaccine4:responseCD4Gp2nTp}}$	-0.485	0.576	1000	-0.841	0.401
$\beta_{1,\text{vaccine5:responseCD4Gp2nTp}}$	-0.497	0.477	1000	-1.041	0.298
$\beta_{1,\text{vaccine7:responseCD4Gp2nTp}}$	-0.275	0.343	1000	-0.801	0.424
$\beta_{1,\text{vaccine4:responseCD4Gp2pTn}}$	-0.275	0.576	1000	-0.477	0.633
$\beta_{1,\text{vaccine5:responseCD4Gp2pTn}}$	-0.239	0.477	1000	-0.5	0.617
$\beta_{1,\text{vaccine7:responseCD4Gp2pTn}}$	-0.266	0.343	1000	-0.777	0.437
$\beta_{1,\text{vaccine4:responseCD4Gp2pTp}}$	-0.27	0.576	1000	-0.468	0.64
$\beta_{1,\text{vaccine5:responseCD4Gp2pTp}}$	-0.282	0.478	1000	-0.59	0.555
$\beta_{1,\text{vaccine7:responseCD4Gp2pTp}}$	-0.025	0.344	1000	-0.073	0.942
$\beta_{2,1}$	0.058	0.006	1000	9.657	<0.001

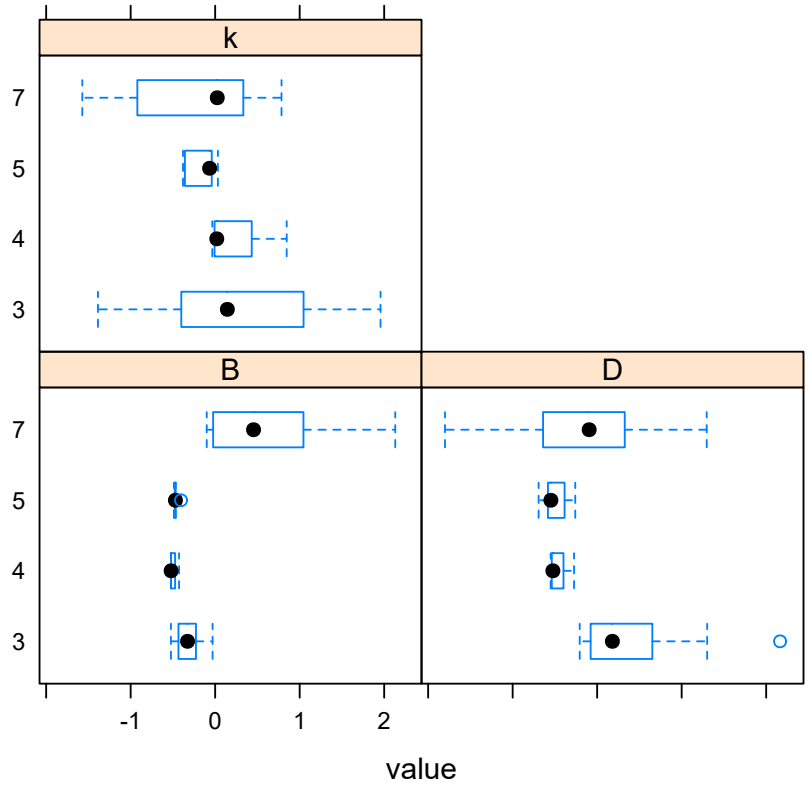


Figure 37: subject specific random effect estimates $\widehat{b}_{0i}$, $\widehat{b}_{1i}$, and $\widehat{b}_{2i}$ for Model 8, a NLMEM.

The parameter estimates for $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.040^2 & 0 & 0 \\ 0 & 0.251^2 & 0 \\ 0 & 0 & 0.029^2 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{\psi}}_2 = \begin{pmatrix} 0.114^2 & 0 \\ 0 & 0.570^2 \end{pmatrix},$$

and the parameter estimate for σ, the standard deviation for the within group error, is 0.082.

In Figures 98 to 104 in Appendix C, we present some diagnostic plots where we consider the residuals and random effect estimates for Model 9. These are broadly similar to those of Model 8. Importantly, the residuals are still non-normal as the points in the QQ-plots do not lie on straight lines. As such, the assumption that the within-group residuals are normally distributed is still not satisfied and the fitted model cannot reliably be used for inference.

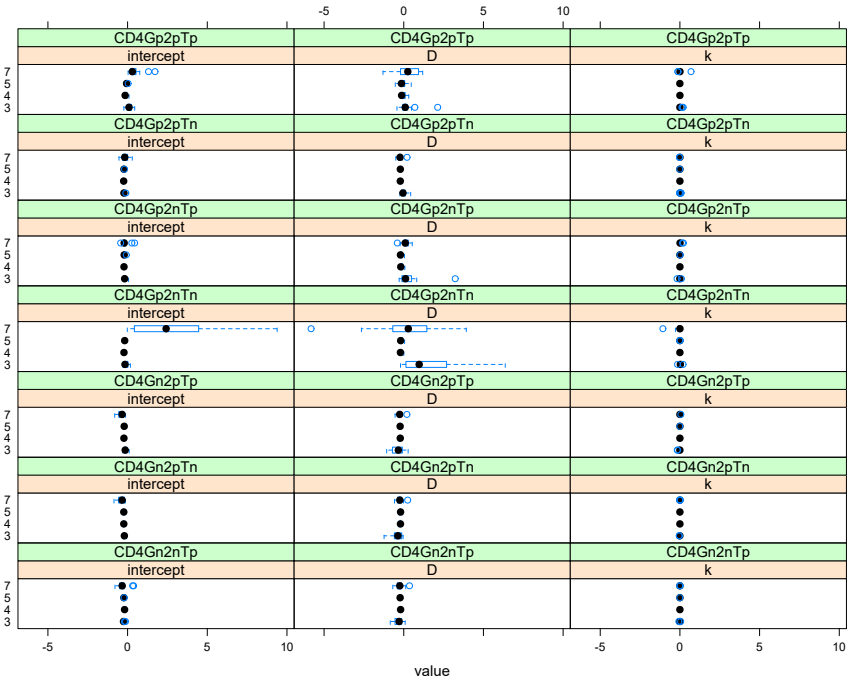


Figure 38: Response within subject random effect estimates $\widehat{b}_{0i,j}$, $\widehat{b}_{1i,j}$, and $\widehat{b}_{2i,j}$ for Model 8, a NLMEM.

3.2.3 *Comments on the NLMEMs*

We have investigated two NLMEMs in this subsection. These models do relatively well when compared to the LMEMs, especially given their relatively simple structure, but similar to the LMEMs, the assumptions underlying the models are not satisfied. This means that we cannot draw meaningful inference from the fitted models.

In the next section we investigate applications of the univariate Tweedie GLMEM to the vaccine data. These models are expected to perform significantly better than the LMEMs and NLMEMs as the Tweedie GLMEMs explicitly allow for the zero-inflated skew distribution of the immune responses in the data which the LMEMs and NLMEMs didn't allow for.

3.3 UNIVARIATE GENERALISED LINEAR MIXED EFFECT MODELS

In this section we investigate applications of univariate Tweedie GLMEMs to the vaccine data. The form of the Tweedie GLMEMs that we consider is described in Section 2.3. These models allow for an explicit mass at zero and a skew distribution, and so should perform better than the previous models that we have considered. Similar to the LMEMs that we considered, to capture the nonlinear profile of the

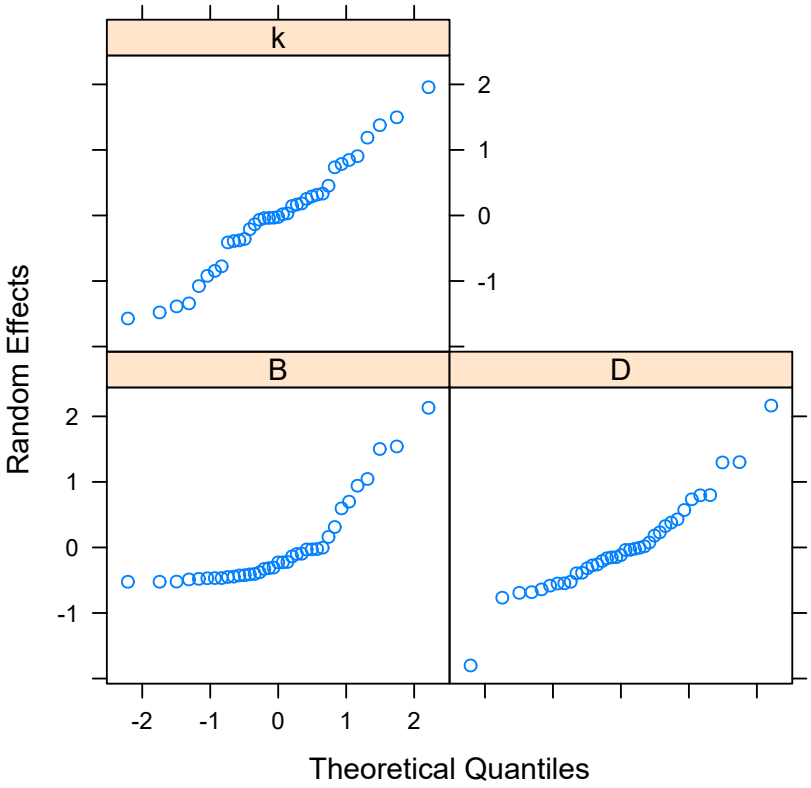


Figure 39: QQ-plots for the subject specific random effect estimates $\widehat{b}_{0i}$, $\widehat{b}_{1i}$, and $\widehat{b}_{2i}$ for Model 8, a NLMEM.

immune responses, we use a cubic B-spline as described in Subsection 3.1.3.

MODEL 10 For the first model that we consider in this section, we assume a similar form for the linear predictor to that of Model 7. We do remove some of the random effects in the linear predictor here as they were not needed; they had variances close to zero.

Specifically, the log of the expected immune response value μ_{ijk} under the Tweedie GLMEM that we consider is given by

$$\log(\mu_{ijk}) = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k)$$

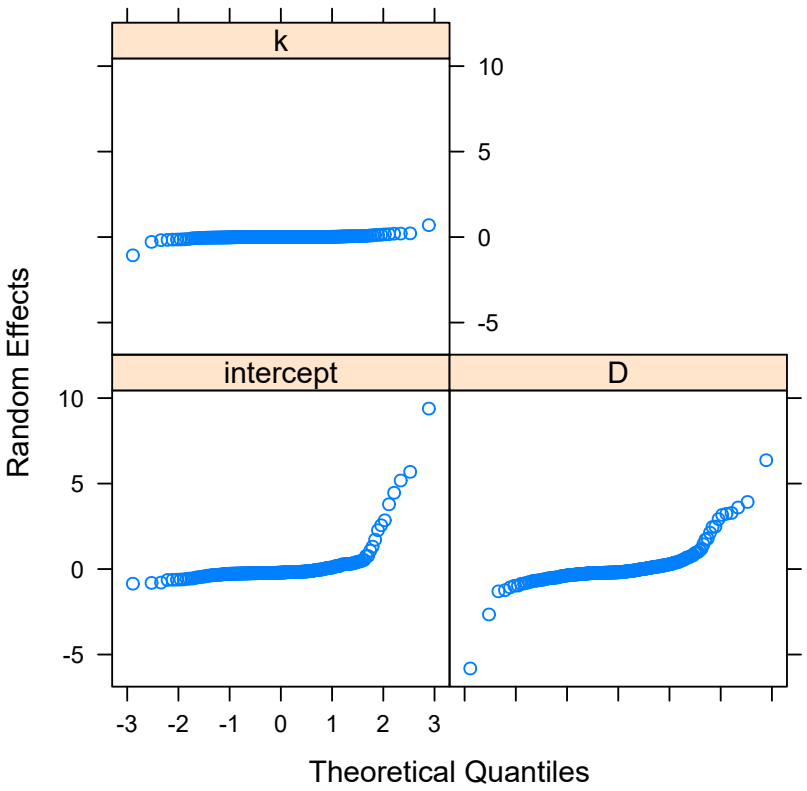


Figure 40: QQ-plots for response within subject random effect estimates $\widehat{b}_{0i,j}$, $\widehat{b}_{1i,j}$, and $\widehat{b}_{2i,j}$ for Model 8, a NLMEM.

with

$$\begin{aligned}
\phi_{0,ij} &= (\beta_{0,1} + b_{0i} + b_{0i,j}) + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j + \\
&\quad \beta_{0,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\
\phi_{1,ij} &= \beta_{1,0} + b_{1i} + \beta_{1,\text{vaccine}_i}\text{vaccine}_i, \\
\phi_{2,ij} &= \beta_{2,0} + b_{2i} + \beta_{2,\text{vaccine}_i}\text{vaccine}_i,
\end{aligned}$$

and

$$\phi_{3,ij} = \beta_{3,0} + b_{3i} + \beta_{3,\text{vaccine}_i}\text{vaccine}_i,$$

where $\mu_{ijk} = \text{E}(y_{ijk})$ and y_{ijk} is the immune response j for subject i at occasion k (t_k). For this model we assume that

$$\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{\psi}_1) \quad \text{and} \quad b_{i,j} = b_{0i,j} \sim \text{N}(\boldsymbol{0}, \psi_2)$$

where $\boldsymbol{\psi}_1$ is diagonal, and $\boldsymbol{b}_i$ and $b_{i,j}$ are independent. This structure allows for different mean immune response profiles for each vaccine and a different level for each vaccine and response combination. As such it could be used to identify which vaccines induce a greater immune response for each specific combination of cytokines.

Table 11 gives the output of Model 10 for the fixed effects. The output also includes the estimates of the dispersion parameter ζ and the index parameter p. The AIC and BIC for this model are -4249.9 and -4017.6 respectively. This is the best model that we have fitted so far as measured by AIC and BIC.

The parameter estimates for $\boldsymbol{\psi}_1$ and ψ_2 are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.533^2 & 0 & 0 \\ 0 & 1.627^2 & 0 \\ 0 & 0 & 0.541^2 \end{pmatrix}$$

and

$$\widehat{\psi}_2 = 0.342.$$

Figures 41 to 46 present diagnostic plots for the model that allow us to check the model assumptions.

Figures 41 and 42 present plots of the residuals versus the fitted values and the time points respectively. From these two figures we can see

- that there is no clear pattern for the residuals and the fitted values (larger plots for each vaccine do not reveal any pattern in the residuals when plotted against the fitted values).
- there is some minor pattern in the residuals when plotted against time, most prominently for Vaccines 3 and 7. This suggests that the model is not adequately capturing the nonlinear profile of the immune responses. To try and address this we could include response fixed effects for the cubic B-spline parameters in the next iteration of this model.

Figure 43 to 46 present plots of the random effect estimates. The box-and-whisker plots for the subject level random effect estimates are fairly well centered around zero. The box-and-whisker plots for the response within subject level random effects show that there is some variability around zero suggesting that there may still be a vaccine and response effect. This also suggests that we may need to include the response fixed effects for the cubic B-spline parameters. The QQ-plots show that the assumption of normality for the random effects appears to be reasonable.

MODEL 11 For the second model that we consider in this section, we include vaccine and response fixed effects, and their interaction, for the intercept and cubic B-spline parameters. We do this so as to try and capture the time effects seen in Figure 42 for Model 10 and to reduce the vaccine and response variation seen for the response within subject random effects estimates. We keep the random effect structure the same as the previous model.

Table 11: Output from Model 10, a GLMEM with a cubic B-spline.

Parameter	Estimate	Std. Error	t-value
$\beta_{0,1}$	-4.505	0.202	-22.272
$\beta_{0,\text{vaccine4}}$	0.045	0.516	0.086
$\beta_{0,\text{vaccine5}}$	-1.222	0.412	-2.969
$\beta_{0,\text{vaccine7}}$	1.419	0.289	4.909
$\beta_{0,\text{responseCD4Gn2pTn}}$	0.128	0.189	0.678
$\beta_{0,\text{responseCD4Gn2pTp}}$	0.542	0.185	2.932
$\beta_{0,\text{responseCD4Gp2nTn}}$	1.54	0.178	8.649
$\beta_{0,\text{responseCD4Gp2nTp}}$	0.831	0.183	4.55
$\beta_{0,\text{responseCD4Gp2pTn}}$	0.585	0.185	3.167
$\beta_{0,\text{responseCD4Gp2pTp}}$	1.472	0.178	8.252
$\beta_{1,1}$	17.668	0.977	18.085
$\beta_{2,1}$	-89.481	5.129	-17.446
$\beta_{3,1}$	172.294	10.178	16.928
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-1.171	0.518	-2.262
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	0.117	0.383	0.305
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	-0.323	0.27	-1.195
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-1.415	0.511	-2.771
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	-0.029	0.377	-0.078
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.805	0.268	-3
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-2.217	0.502	-4.414
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	-0.584	0.368	-1.586
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	0.599	0.251	2.383
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-1.008	0.491	-2.055
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	-0.038	0.373	-0.103
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	-0.239	0.261	-0.915
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-2.293	0.539	-4.257
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	-0.345	0.381	-0.906
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	-0.229	0.264	-0.866
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-0.557	0.467	-1.193
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	0.382	0.36	1.063
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	-0.127	0.254	-0.498
$\beta_{1,\text{vaccine4}}$	-10.731	2.396	-4.478
$\beta_{1,\text{vaccine5}}$	-0.762	2.426	-0.314
$\beta_{1,\text{vaccine7}}$	-11.604	1.643	-7.061
$\beta_{2,\text{vaccine4}}$	70.189	7.93	8.851
$\beta_{2,\text{vaccine5}}$	21.516	10.49	2.051
$\beta_{2,\text{vaccine7}}$	56.886	12.659	4.494
$\beta_{3,\text{vaccine4}}$	-172.66	10.184	-16.954
$\beta_{3,\text{vaccine5}}$	-61.608	18.306	-3.365
$\beta_{3,\text{vaccine7}}$	-172.035	10.178	-16.902
ζ	0.212		
p	1.637		

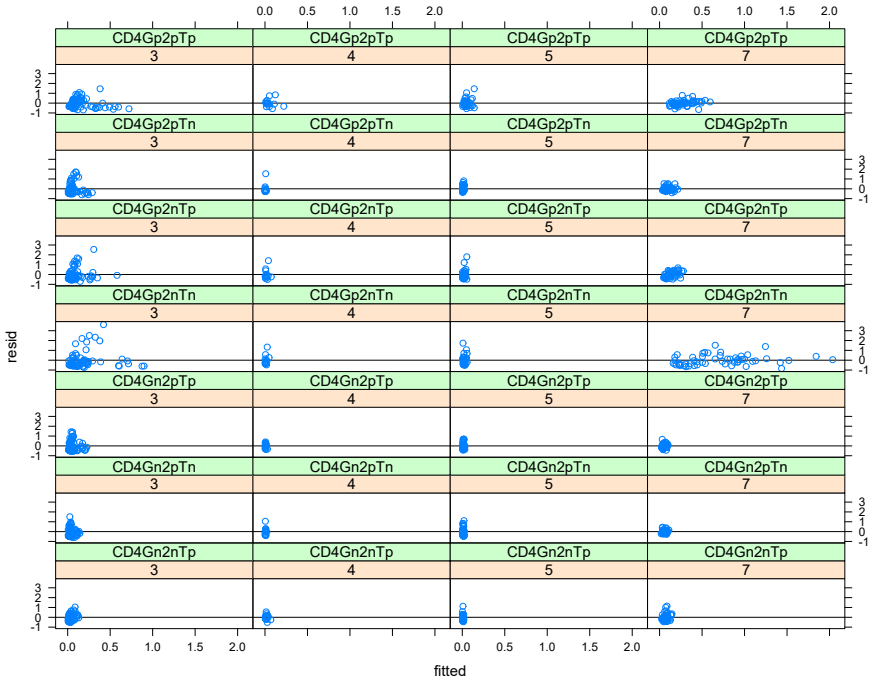


Figure 41: Residuals versus fitted values by vaccine and response for Model 10, a GLMEM.

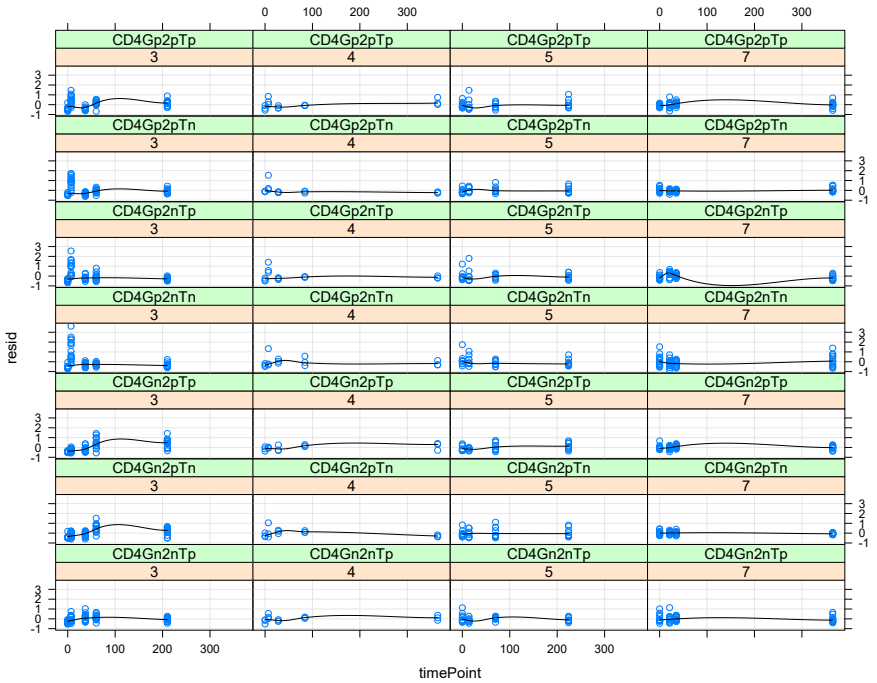


Figure 42: Residuals versus time point by vaccine and response for Model 10, a GLMEM.

Specifically, the log of the expected immune response value under the Tweedie GLMEM that we consider here is given by

$$g(\mu_{ijk}) = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k)$$

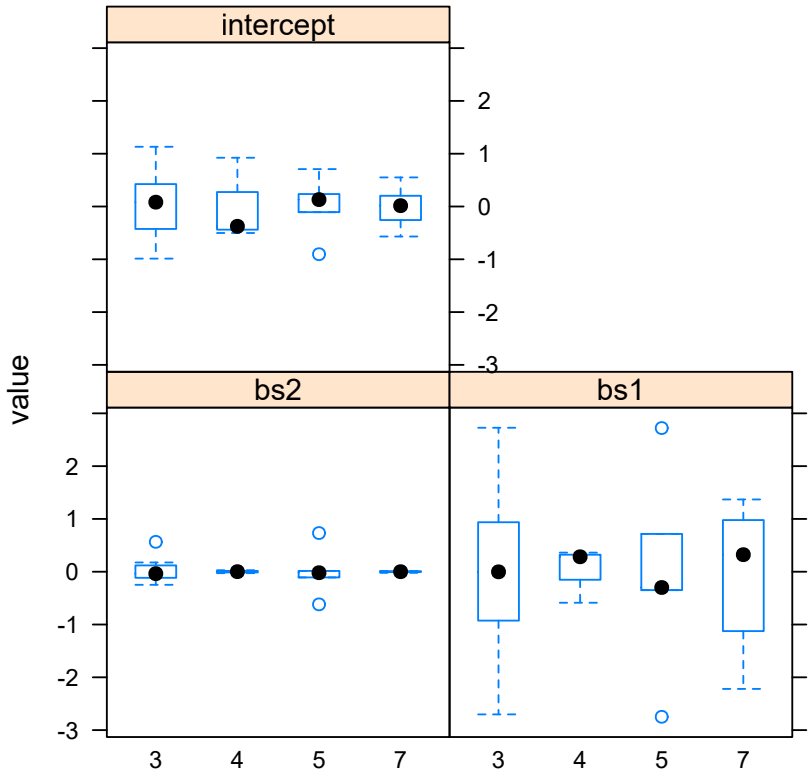


Figure 43: Subject specific random effect estimates $\widehat{b}_{0i}$ (intercept), $\widehat{b}_{1i}$ (bs1), and $\widehat{b}_{2i}$ (bs2) for Model 10, a GLMEM.

with

$$\begin{aligned}
\phi_{0,ij} &= (\beta_{0,1} + b_{0i} + b_{0i,j}) + \beta_{0,\text{vaccine}_i}\text{vaccine}_i + \beta_{0,\text{response}_j}\text{response}_j + \\
&\quad \beta_{0,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\
\phi_{1,ij} &= (\beta_{1,1} + b_{1i}) + \beta_{1,\text{vaccine}_i}\text{vaccine}_i + \beta_{1,\text{response}_j}\text{response}_j + \\
&\quad \beta_{1,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j \\
\phi_{2,ij} &= (\beta_{2,1} + b_{2i}) + \beta_{2,\text{vaccine}_i}\text{vaccine}_i + \beta_{2,\text{response}_j}\text{response}_j + \\
&\quad \beta_{2,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j
\end{aligned}$$

and

$$\begin{aligned}
\phi_{3,ij} &= (\beta_{3,1} + b_{3i}) + \beta_{3,\text{vaccine}_i}\text{vaccine}_i + \beta_{3,\text{response}_j}\text{response}_j + \\
&\quad \beta_{3,\text{vaccine}_i \times \text{response}_j}\text{vaccine}_i \times \text{response}_j.
\end{aligned}$$

The assumptions for $\boldsymbol{b}_i$ and $b_{i,j}$ remain the same. As mentioned, in effect this model allows for a separate immune response curve to be estimated for each vaccine and response combination. As such it is a relatively complicated model.

Tables 12 and 13 give the output of Model 11 for the fixed effects. The output also includes the estimates of the dispersion parameter

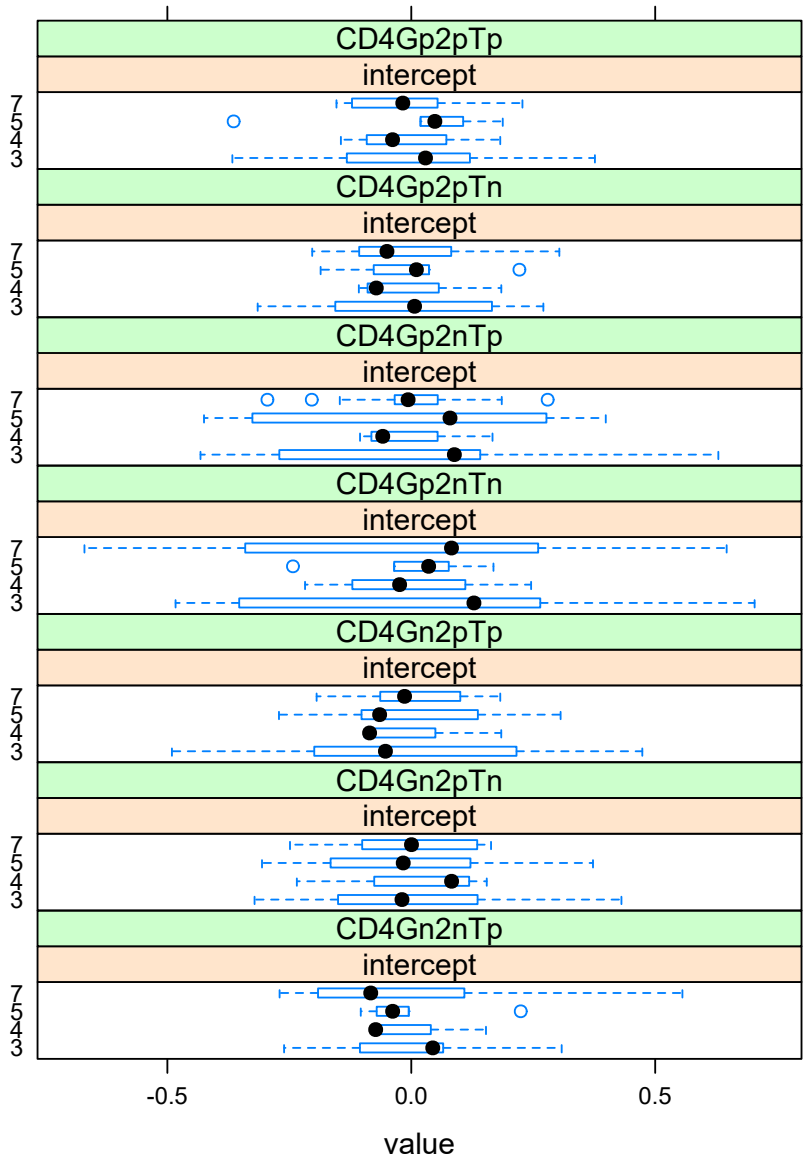


Figure 44: Response within subject random effect estimates $\widehat{b}_{0i,j}$ (intercept) for Model 10, a GLMEM.

ζ and the index parameter p. The AIC and BIC for this model are -4355.7 and -3751.9 respectively. This is the best model that we have fit so far as measured by AIC, but as the number of parameters has increased significantly, the BIC has increased relative to that of Model 10.

The parameter estimates for $\boldsymbol{\psi}_1$ and ψ_2 are

$$\widehat{\boldsymbol{\psi}}_1 = \begin{pmatrix} 0.250^2 & 0 & 0 \\ 0 & 2.484^2 & 0 \\ 0 & 0 & 0.410^2 \end{pmatrix}$$

Table 12: Output from Model 11, a GLMEM with a B-spline.

Parameter	Estimate	Std. Error	t-value
$\beta_{0,1}$	-5.157	0.27	-19.096
$\beta_{0,\text{vaccine4}}$	0.556	0.644	0.863
$\beta_{0,\text{vaccine5}}$	-0.211	0.526	-0.401
$\beta_{0,\text{vaccine7}}$	2.12	0.37	5.73
$\beta_{0,\text{responseCD4Gn2pTn}}$	-0.308	0.346	-0.89
$\beta_{0,\text{responseCD4Gn2pTp}}$	-0.125	0.34	-0.368
$\beta_{0,\text{responseCD4Gp2nTn}}$	2.62	0.288	9.105
$\beta_{0,\text{responseCD4Gp2nTp}}$	1.778	0.299	5.937
$\beta_{0,\text{responseCD4Gp2pTn}}$	1.657	0.303	5.469
$\beta_{0,\text{responseCD4Gp2pTp}}$	2.317	0.293	7.912
$\beta_{1,1}$	24.22	2.48	9.765
$\beta_{2,1}$	-113.8	13.91	-8.181
$\beta_{3,1}$	217.9	27.63	7.886
$\beta_{0,\text{vaccine4:responseCD4Gn2pTn}}$	-0.909	0.87	-1.045
$\beta_{0,\text{vaccine5:responseCD4Gn2pTn}}$	0.028	0.666	0.043
$\beta_{0,\text{vaccine7:responseCD4Gn2pTn}}$	0.096	0.464	0.208
$\beta_{0,\text{vaccine4:responseCD4Gn2pTp}}$	-0.951	0.861	-1.105
$\beta_{0,\text{vaccine5:responseCD4Gn2pTp}}$	0.008	0.656	0.012
$\beta_{0,\text{vaccine7:responseCD4Gn2pTp}}$	-0.239	0.462	-0.516
$\beta_{0,\text{vaccine4:responseCD4Gp2nTn}}$	-3.427	0.813	-4.217
$\beta_{0,\text{vaccine5:responseCD4Gp2nTn}}$	-1.721	0.594	-2.899
$\beta_{0,\text{vaccine7:responseCD4Gp2nTn}}$	-0.366	0.391	-0.936
$\beta_{0,\text{vaccine4:responseCD4Gp2nTp}}$	-1.433	0.761	-1.884
$\beta_{0,\text{vaccine5:responseCD4Gp2nTp}}$	-1.374	0.613	-2.24
$\beta_{0,\text{vaccine7:responseCD4Gp2nTp}}$	-1.51	0.422	-3.577
$\beta_{0,\text{vaccine4:responseCD4Gp2pTn}}$	-2.524	0.837	-3.013
$\beta_{0,\text{vaccine5:responseCD4Gp2pTn}}$	-2.164	0.655	-3.303
$\beta_{0,\text{vaccine7:responseCD4Gp2pTn}}$	-1.305	0.424	-3.081
$\beta_{0,\text{vaccine4:responseCD4Gp2pTp}}$	-1.26	0.726	-1.735
$\beta_{0,\text{vaccine5:responseCD4Gp2pTp}}$	-0.871	0.581	-1.499
$\beta_{0,\text{vaccine7:responseCD4Gp2pTp}}$	-1.173	0.406	-2.891
$\beta_{1,\text{vaccine4}}$	-17.9	5.315	-3.368
$\beta_{1,\text{vaccine5}}$	-16.11	6.34	-2.54
$\beta_{1,\text{vaccine7}}$	-18.16	4.189	-4.335
$\beta_{1,\text{responseCD4Gn2pTn}}$	-3.699	3.519	-1.051
$\beta_{1,\text{responseCD4Gn2pTp}}$	-4.104	3.432	-1.196
$\beta_{1,\text{responseCD4Gp2nTn}}$	-5.228	3.031	-1.725
$\beta_{1,\text{responseCD4Gp2nTp}}$	-5.559	3.181	-1.747
$\beta_{1,\text{responseCD4Gp2pTn}}$	-9.073	3.28	-2.766
$\beta_{1,\text{responseCD4Gp2pTp}}$	-11.16	3.074	-3.631
$\beta_{2,\text{vaccine4}}$	99.4	19.7	5.045
$\beta_{2,\text{vaccine5}}$	86.18	29.12	2.959
$\beta_{2,\text{vaccine7}}$	79.22	34.05	2.326
$\beta_{2,\text{responseCD4Gn2pTn}}$	36.46	19.6	1.86
$\beta_{2,\text{responseCD4Gn2pTp}}$	42.54	19.05	2.233
$\beta_{2,\text{responseCD4Gp2nTn}}$	-4.093	17.71	-0.231
$\beta_{2,\text{responseCD4Gp2nTp}}$	6.426	18.51	0.347
$\beta_{2,\text{responseCD4Gp2pTn}}$	25.34	19.08	1.328
$\beta_{2,\text{responseCD4Gp2pTp}}$	48.57	17.62	2.757
$\beta_{3,\text{vaccine4}}$	-217.9	27.65	-7.882
$\beta_{3,\text{vaccine5}}$	-175.7	50.7	-3.465
$\beta_{3,\text{vaccine7}}$	-217.8	27.64	-7.883
$\beta_{3,\text{responseCD4Gn2pTn}}$	-69.86	38.86	-1.798
$\beta_{3,\text{responseCD4Gn2pTp}}$	-79.87	37.76	-2.115
$\beta_{3,\text{responseCD4Gp2nTn}}$	8.535	35.26	0.242
$\beta_{3,\text{responseCD4Gp2nTp}}$	-12.77	36.82	-0.347
$\beta_{3,\text{responseCD4Gp2pTn}}$	-46.8	37.93	-1.234
$\beta_{3,\text{responseCD4Gp2pTp}}$	-91.58	34.98	-2.618

Table 13: Cont.: Output from Model 11, a GLMEM with a B-spline.

Parameter	Estimate	Std. Error	t-value
$\beta_{1,\text{vaccine4:responseCD4Gn2pTn}}$	8.636	7.936	1.088
$\beta_{1,\text{vaccine5:responseCD4Gn2pTn}}$	13.11	8.772	1.495
$\beta_{1,\text{vaccine7:responseCD4Gn2pTn}}$	3.57	5.981	0.597
$\beta_{1,\text{vaccine4:responseCD4Gn2pTp}}$	5.015	8.015	0.626
$\beta_{1,\text{vaccine5:responseCD4Gn2pTp}}$	10.53	8.701	1.21
$\beta_{1,\text{vaccine7:responseCD4Gn2pTp}}$	2.908	5.988	0.486
$\beta_{1,\text{vaccine4:responseCD4Gp2nTn}}$	10.34	7.389	1.4
$\beta_{1,\text{vaccine5:responseCD4Gp2nTn}}$	13.49	7.923	1.702
$\beta_{1,\text{vaccine7:responseCD4Gp2nTn}}$	2.258	5.032	0.449
$\beta_{1,\text{vaccine4:responseCD4Gp2nTp}}$	3.253	7.198	0.452
$\beta_{1,\text{vaccine5:responseCD4Gp2nTp}}$	17.01	8.18	2.079
$\beta_{1,\text{vaccine7:responseCD4Gp2nTp}}$	12.15	5.479	2.218
$\beta_{1,\text{vaccine4:responseCD4Gp2pTn}}$	4.465	8.286	0.539
$\beta_{1,\text{vaccine5:responseCD4Gp2pTn}}$	22.9	8.732	2.622
$\beta_{1,\text{vaccine7:responseCD4Gp2pTn}}$	7.75	5.663	1.369
$\beta_{1,\text{vaccine4:responseCD4Gp2pTp}}$	10.43	6.712	1.553
$\beta_{1,\text{vaccine5:responseCD4Gp2pTp}}$	21.11	7.653	2.759
$\beta_{1,\text{vaccine7:responseCD4Gp2pTp}}$	11.79	5.241	2.25
$\beta_{2,\text{vaccine4:responseCD4Gn2pTn}}$	-50.83	28.99	-1.753
$\beta_{2,\text{vaccine5:responseCD4Gn2pTn}}$	-76.19	40.18	-1.896
$\beta_{2,\text{vaccine7:responseCD4Gn2pTn}}$	-33.03	48.92	-0.675
$\beta_{2,\text{vaccine4:responseCD4Gn2pTp}}$	-42.96	28.94	-1.485
$\beta_{2,\text{vaccine5:responseCD4Gn2pTp}}$	-65.83	39.84	-1.652
$\beta_{2,\text{vaccine7:responseCD4Gn2pTp}}$	-21.82	49.16	-0.444
$\beta_{2,\text{vaccine4:responseCD4Gp2nTn}}$	-13.6	26.99	-0.504
$\beta_{2,\text{vaccine5:responseCD4Gp2nTn}}$	-41.85	36.75	-1.139
$\beta_{2,\text{vaccine7:responseCD4Gp2nTn}}$	19.24	41.38	0.465
$\beta_{2,\text{vaccine4:responseCD4Gp2nTp}}$	-7.086	27.16	-0.261
$\beta_{2,\text{vaccine5:responseCD4Gp2nTp}}$	-60.09	37.86	-1.587
$\beta_{2,\text{vaccine7:responseCD4Gp2nTp}}$	-52.62	45.03	-1.169
$\beta_{2,\text{vaccine4:responseCD4Gp2pTn}}$	-23.76	30.6	-0.777
$\beta_{2,\text{vaccine5:responseCD4Gp2pTn}}$	-83.7	40	-2.092
$\beta_{2,\text{vaccine7:responseCD4Gp2pTn}}$	-16.24	46.95	-0.346
$\beta_{2,\text{vaccine4:responseCD4Gp2pTp}}$	-50.51	25.32	-1.995
$\beta_{2,\text{vaccine5:responseCD4Gp2pTp}}$	-95.95	35.45	-2.706
$\beta_{2,\text{vaccine7:responseCD4Gp2pTp}}$	-41.77	42.98	-0.972
$\beta_{3,\text{vaccine4:responseCD4Gn2pTn}}$	67.74	38.89	1.742
$\beta_{3,\text{vaccine5:responseCD4Gn2pTn}}$	137.2	70.08	1.958
$\beta_{3,\text{vaccine7:responseCD4Gn2pTn}}$	69.88	38.86	1.798
$\beta_{3,\text{vaccine4:responseCD4Gn2pTp}}$	80.2	37.78	2.123
$\beta_{3,\text{vaccine5:responseCD4Gn2pTp}}$	121.3	69.34	1.75
$\beta_{3,\text{vaccine7:responseCD4Gn2pTp}}$	80.04	37.76	2.119
$\beta_{3,\text{vaccine4:responseCD4Gp2nTn}}$	-9.238	35.28	-0.262
$\beta_{3,\text{vaccine5:responseCD4Gp2nTn}}$	68.55	64.12	1.069
$\beta_{3,\text{vaccine7:responseCD4Gp2nTn}}$	-8.367	35.26	-0.237
$\beta_{3,\text{vaccine4:responseCD4Gp2nTp}}$	11.65	36.84	0.316
$\beta_{3,\text{vaccine5:responseCD4Gp2nTp}}$	102.5	66.12	1.55
$\beta_{3,\text{vaccine7:responseCD4Gp2nTp}}$	12.93	36.82	0.351
$\beta_{3,\text{vaccine4:responseCD4Gp2pTn}}$	44.39	37.96	1.169
$\beta_{3,\text{vaccine5:responseCD4Gp2pTn}}$	145	69.7	2.08
$\beta_{3,\text{vaccine7:responseCD4Gp2pTn}}$	47.06	37.93	1.241
$\beta_{3,\text{vaccine4:responseCD4Gp2pTp}}$	91.7	34.99	2.621
$\beta_{3,\text{vaccine5:responseCD4Gp2pTp}}$	172.6	61.97	2.785
$\beta_{3,\text{vaccine7:responseCD4Gp2pTp}}$	91.9	34.98	2.627
ζ	0.157		
p	1.605		

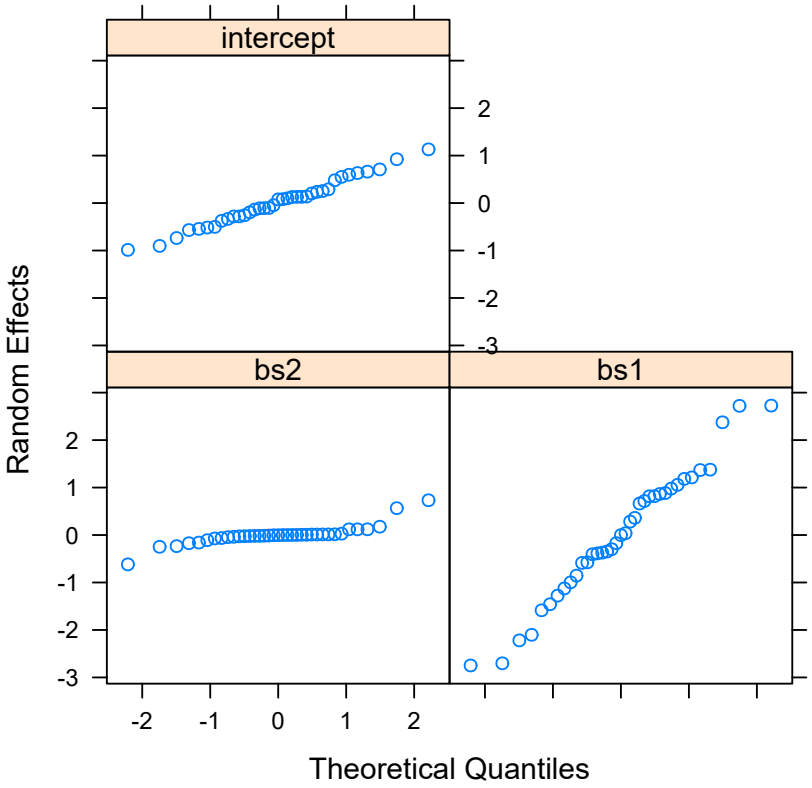


Figure 45: QQ-plots for the subject specific random effect estimates $\widehat{b}_{0i}$ (intercept), $\widehat{b}_{1i}$ (bs1), and $\widehat{b}_{2i}$ (bs2) for Model 10, a GLMEM.

and

$$\widehat{\psi}_2 = 0.130.$$

The Figures 47 to 52 present diagnostic plots for the model.

Figures 47 and 48 present plots of the residuals versus the fitted values and the time points. From these two figures we can see

- that there is no clear pattern in the residuals when plotted against the fitted values, which is good.
- that there is some time pattern in the residuals when plotted against the time points. This is mostly seen for Vaccine 3; the other vaccine candidates do not have such prominent patterns by time point. The visible pattern is fairly minor.

Figure 49 to 52 present plots of the random effect estimates. We see that

- in Figure 49, the random effect estimates are well centered around zero.

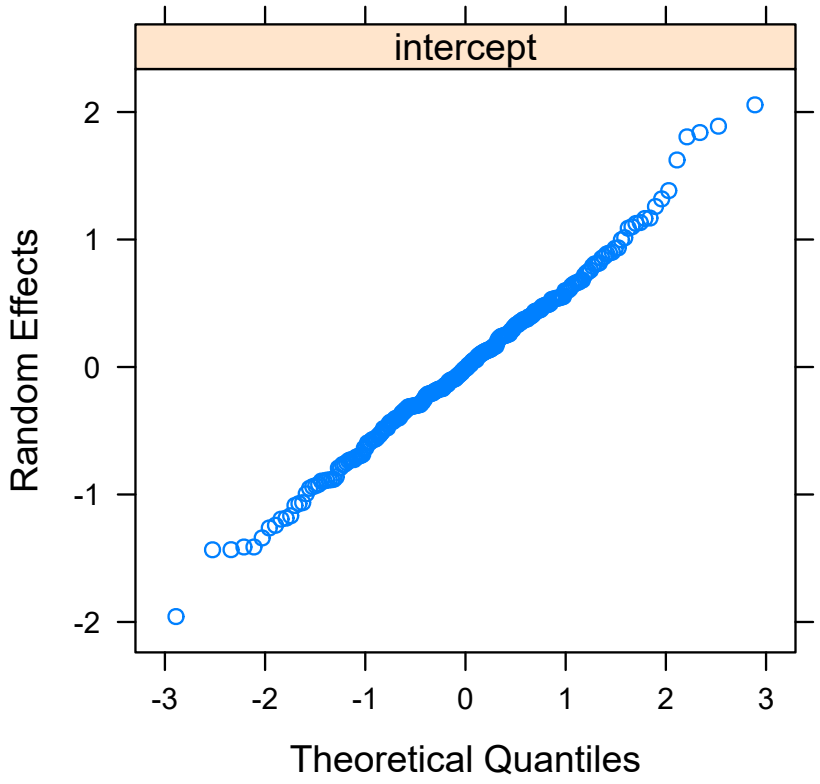


Figure 46: QQ-plot for response within subject random effect estimates $\widehat{b}_{0i,j}$ (intercept) for Model 10, a GLMEM.

- in Figure 50, not much has changed relative to the previous model. The random effect estimates do appear to be more closely clustered around zero, but this is a very minor change. This apparent response within vaccine effect may be overstated due to the limited data that we have available.
- in the QQ-plots, the random effect estimates mostly lie on straight lines suggesting that there is no significant deviation from normality. There are a few points in the tails that suggest a slight departure from normality, but these are a limited number of these points.

Figure 53 shows the fitted versus actual values for each of the vaccines and responses. We can see that the model fits well for some of the vaccine and response combinations, but there are cases where the fit looks poor. The most noticeable cases are for Vaccine 3 where we see poor fit for several of the immune responses, e.g. we see poor fit for CD4Gn2pTn and CD4Gn2pTp.

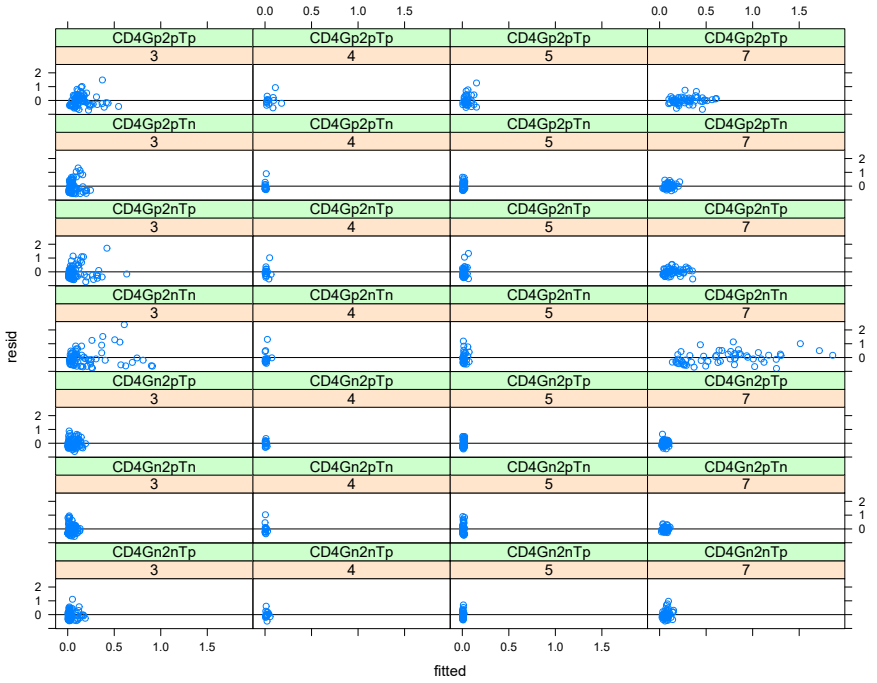


Figure 47: Residuals versus fitted values by vaccine and response for Model 11, a GLMEM.

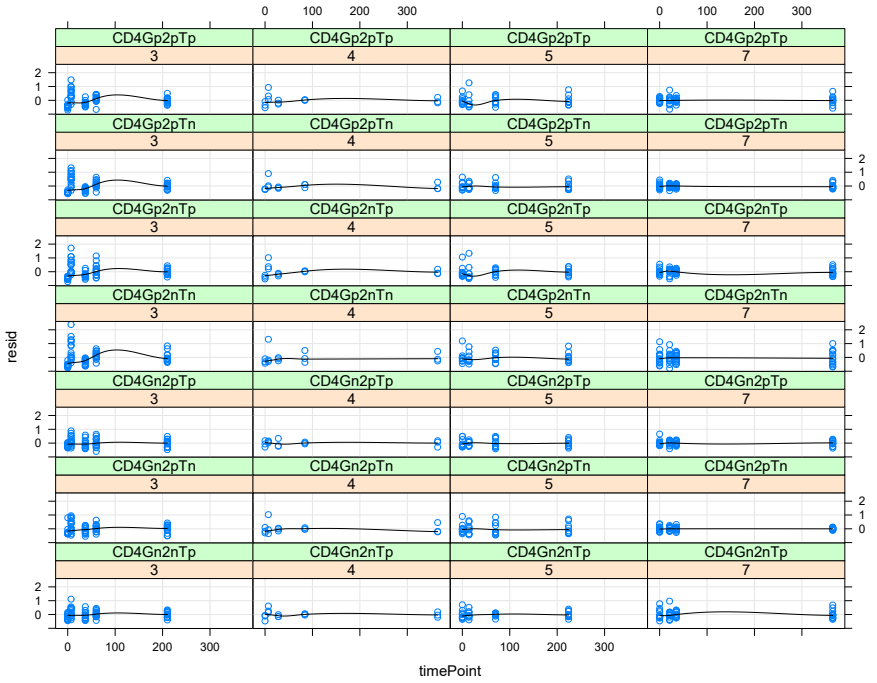


Figure 48: Residuals versus time point by vaccine and response for Model 11, a GLMEM.

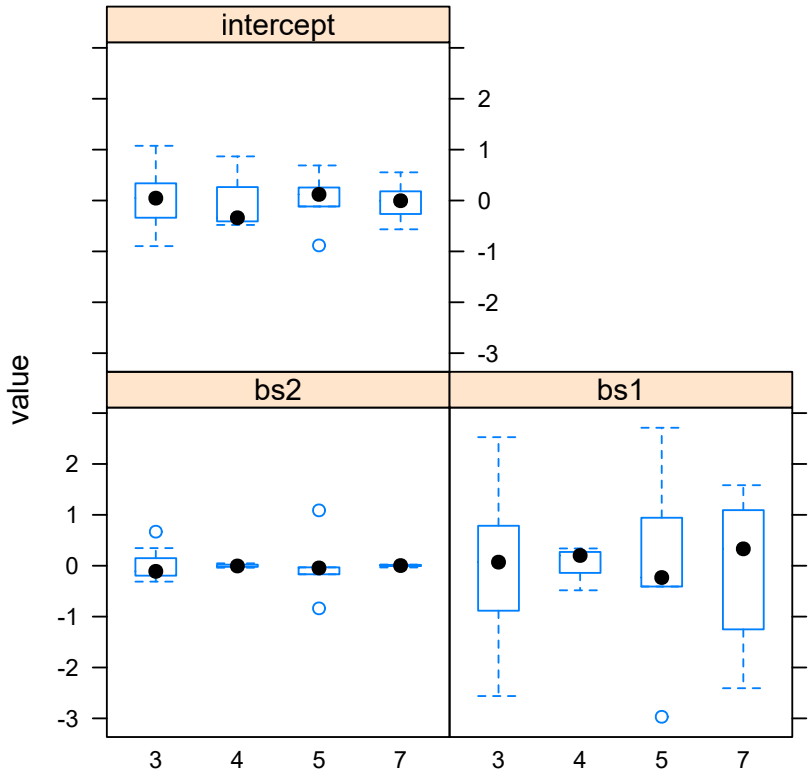


Figure 49: Subject specific random effect estimates $\widehat{b}_{0i}$ (intercept), $\widehat{b}_{1i}$ (bs1), and $\widehat{b}_{2i}$ (bs2) for Model 11, a GLMEM.

3.3.1 *Comment on the GLMEMs*

In this section we have investigated univariate Tweedie GLMEMs with cubic B-splines with three degrees of freedom. These models fit the data much better than the LMEMs and NLMEMs as they explicitly allow for the zero-inflated skew nature of the immune responses. In particular, we see that the assumptions underlying the models are mostly satisfied with the exception of the random effects not being consistently centered around zero. This is the case for both of the models that we considered in this section. The fit of Model 11 is good for some vaccine and response combinations, but not all. We suspect that this is due to the variable nature in the immune responses for different subjects.

Model 11 is a complicated model which attempted to address the vaccine and response patterns seen for the random effect estimates by including a large number of fixed effects. It failed to capture all of the pattern we were seeing, but this may be a function of the small sample sizes we have. Using Model 11 for inference is difficult given its complexity. One means of summarising the fitted models

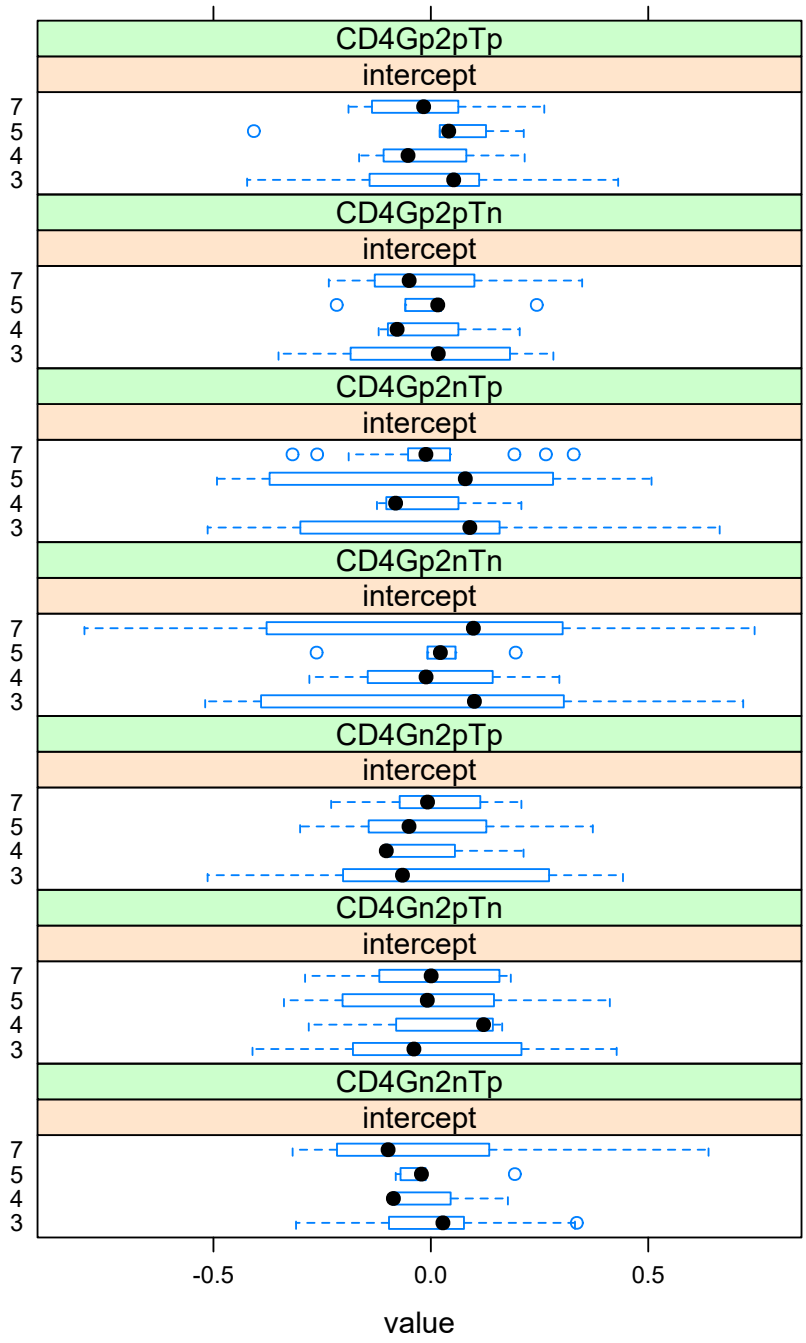


Figure 50: Response within subject random effect estimates $\widehat{b}_{0i,j}$ (intercept) for Model 11, a GLMEM.

is to consider plots of the marginal expected immune responses over time with confidence intervals. These plots would provide us with insight into the modelled immune response profiles. To construct such plots would involve integrating out the random effects in the model which could be done by using approximate methods or by Monte

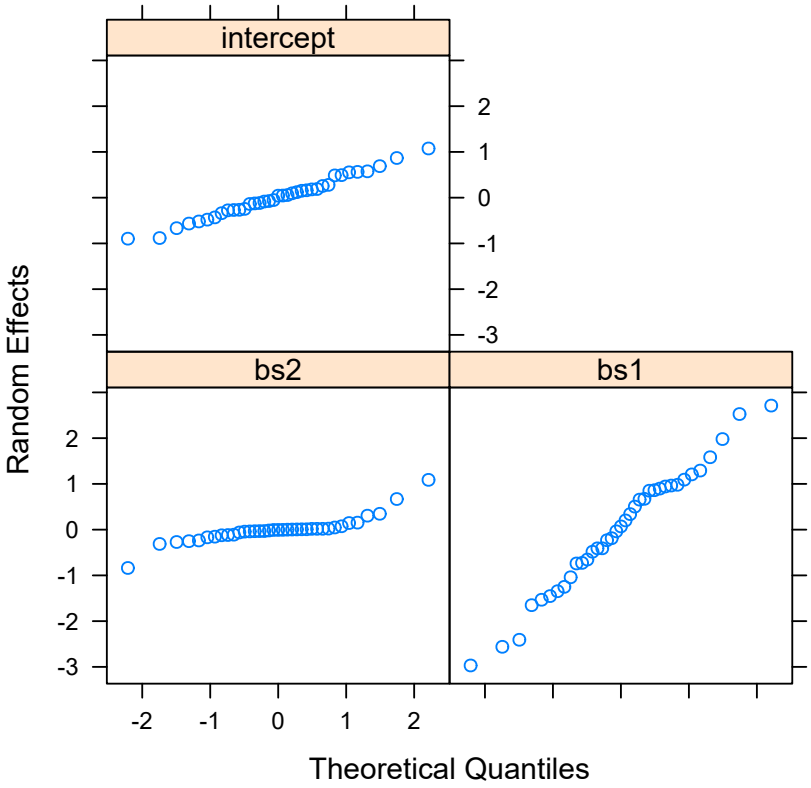


Figure 51: QQ-plots for the subject specific random effect estimates $\widehat{b}_{0i}$ (intercept), $\widehat{b}_{1i}$ (bs1), and $\widehat{b}_{2i}$ (bs2) for Model 11, a GLMEM.

Carlo integration. The plots would give us a relatively straightforward means of comparing the modelled immune response profiles across vaccines and would be useful in deciding which vaccines induced the largest immune responses over time. We could then provide recommendations about which vaccines to promote for further trials.

In the next section, where we consider multivariate generalised linear mixed effect and latent variable models, we make use of plots of the marginal expected immune responses over time to compare the modelled immune response profiles for the vaccines and to identify the vaccine candidates which are promising.

3.4 MULTIVARIATE GENERALISED LINEAR MIXED AND LATENT VARIABLE MODELS

In this section we apply the M-GLMEMs that we described in Section 2.3.3 to the vaccine data. The M-GLMEMs that we consider have a

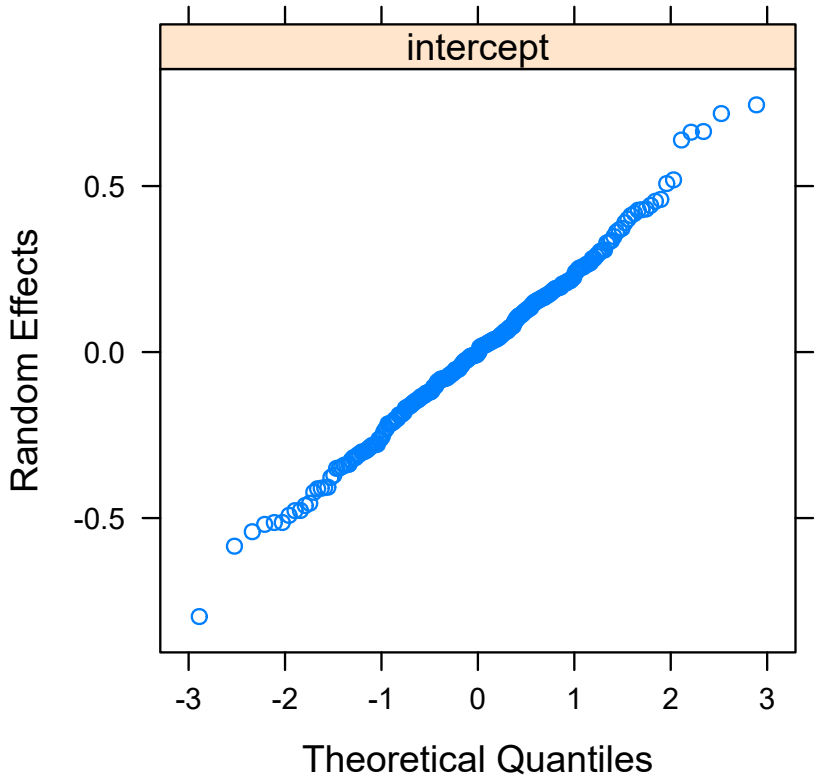


Figure 52: QQ-plot for response within subject random effect estimates $\widehat{b}_{0i,j}$ (intercept) for Model 11, a GLMEM.

Tweedie distribution and use a log link function. The log of the mean of the models that we consider is given by

$$\log(\mu_{ijk}) = b_i + \beta_{0j} + \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta}_j + \boldsymbol{z}_{ik}^{\mathsf{T}}\boldsymbol{\theta}_j$$

where $\mu_{ijk} = \mathrm{E}(y_{ijk})$ is the expected immune response j for subject i at occasion k (t_k), $\boldsymbol{x}_{ik}$ is the vector of occasion specific covariates for subject i at occasion k (t_k), $b_i \sim \mathrm{N}(0, \sigma^2)$ is the subject specific random effect, and $\boldsymbol{z}_{ik} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ is the latent variable which, along with the $\boldsymbol{\theta}_j$, induce correlation between the responses j of individual i. The parameter β_{0j} is the intercept value for response j, $\boldsymbol{\beta}_j$ is the vector of parameters multiplying the subject and occasion specific covariates $\boldsymbol{x}_{ik}$, and $\boldsymbol{\theta}_j$ is the vector of parameters multiplying the latent variables $\boldsymbol{z}_{ik}$ and which determine the correlation between the responses.

The main differences between the models considered in this section and those considered in Section 3.3 are:

- the models here are explicitly multivariate at the outset; i.e. the response variable in the models here is $\boldsymbol{Y}$, a matrix. In Section 3.3 to accommodate the multiple immune responses in

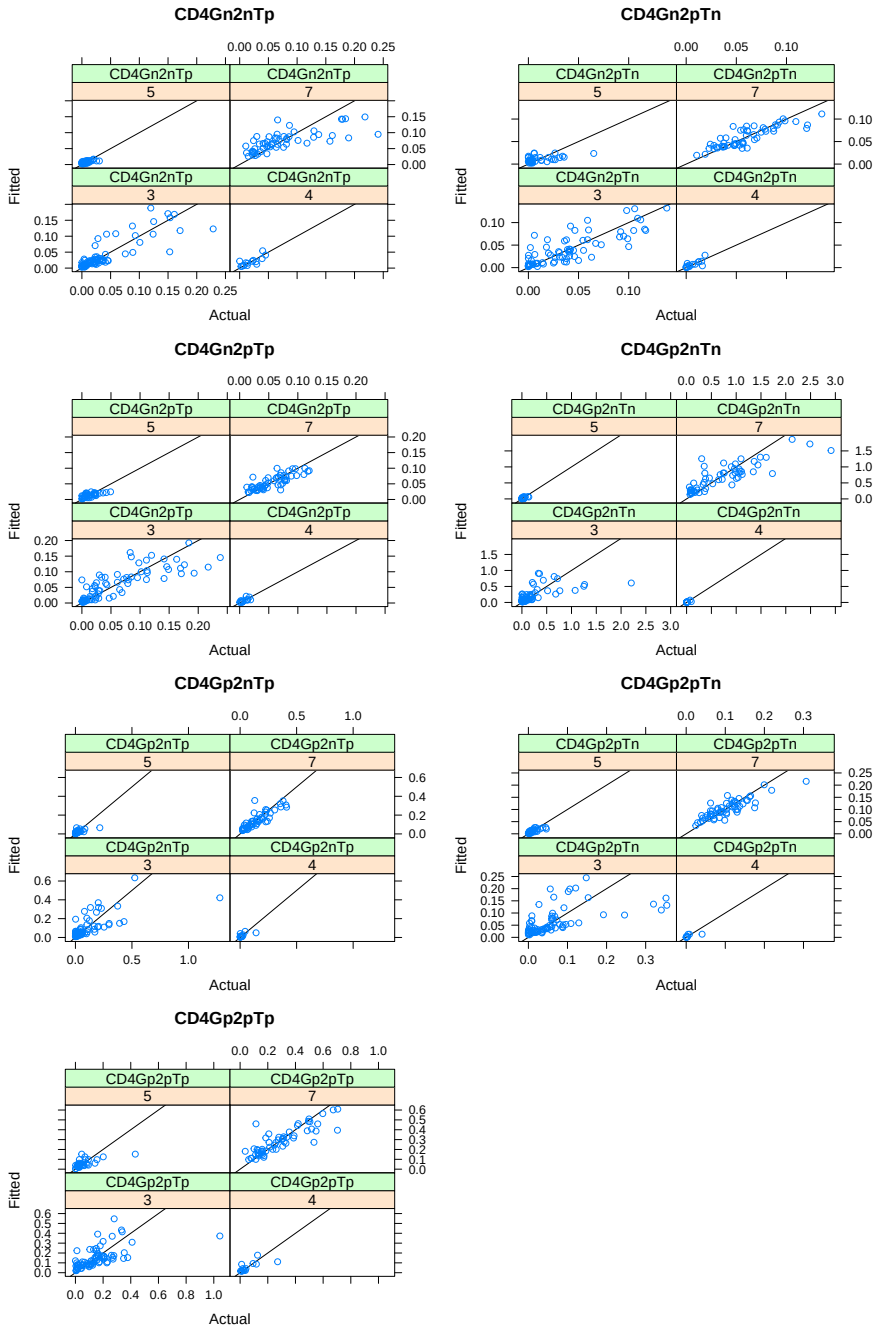


Figure 53: Plot of the fitted conditional response values versus the actual vaccine response values for each of the vaccines and responses for Model 11.

the univariate GLMEMs, we stacked the immune responses so as to form a long column vector and included labels for the responses as covariates. This allowed us to model the multiple immune responses by using univariate models.

- we include latent variables which model the residual correlation between the different responses of individual i. We didn't explic-

itly consider or model the residual correlation in the univariate GLMEMs that we considered in Section 3.3. The residual correlation captured by the latent variables in the M-GLMEMs can be thought of as capturing the correlation that is due to missing predictors (Warton *et al.*, 2015). The latent variables can also be thought of as representing the axes in a model-based ordination and represent the main axes of residual covariation between the vaccine responses (Warton *et al.*, 2015). The model-based ordination interpretation allows us to construct ordination plots to visualise this residual covariation of the responses.

In the subsections that follow, we consider several different M-GLMEMs. The first of these models does not include the occasion specific covariates and these types of models are referred to as a pure latent variable models by Hui (2016). The second and third models include the occasion specific covariates and these types of models are referred to as correlated response models by Hui (2016). The difference between the second and third model is that the third model is simplified relative to the second model by the inclusion of traits for the response variables. As discussed in Section 2.3.3, traits are covariates specific to the different responses j (the columns of $\boldsymbol{Y}$).

3.4.1 *Model without Occasion Specific Covariates*

MODEL 12 In this section we consider a M-GLMEM which does not include any of the time and vaccine covariates which, in the setting here, are the occasion specific covariates. Specifically the form of the model that we consider is given by

$$\log(\mu_{ijk}) = \phi_{0,ij} + \boldsymbol{z}_{ik}^{\mathsf{T}}\boldsymbol{\theta}_j$$

where

$$\phi_{0,ij} = b_i + \beta_{0j},$$

for $j =$ `CD4Gn2nTp`, `CD4Gn2pTn`, `CD4Gn2pTp`, `CD4Gp2nTn`, `CD4Gp2nTp`, `CD4Gp2pTn`, and `CD4Gp2pTp`,

$$b_i \sim \mathrm{N}(0, \sigma^2) \quad \text{and} \quad \boldsymbol{z}_{ik} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{2\times 2}).$$

The subject specific random effects b_i are independent of the latent random variables $\boldsymbol{z}_{ik}$. The parameters to be estimated in this model are the β_{0j}s, the $\boldsymbol{\theta}_j$s, the ζ_js, σ, and p. We estimated the model by using the `boral` package in `R` (Hui, 2016).

The estimated parameters for the model are presented in Table 14 and the estimate of σ is 0.998. These estimates are the median values from the posterior distributions of the parameters. The estimated marginal AIC and BIC at the median of the posterior distribution, which are calculated by using Equation (10), are -4503.8 and -4354.2 respectively. This is the best model in terms of AIC and BIC that we

have considered so far. In addition to AIC and BIC, it is also useful to report additional measures of fit, specifically the expected AIC (EAIC) and expected BIC (EBIC), which are -4741.1 and -2506.5 for this model. These measures are useful as the `boral` package in `R` does not provide the AIC and BIC in all cases.

Table 14: Output for Model 12, a M-GLMEM. Values in the brackets correspond to the 95% highest posterior density (credible) intervals.

	Parameter				
Response j	$\widehat{\beta}_{0j}$	$\widehat{\theta}_{j,1}$	$\widehat{\theta}_{j,2}$	$\widehat{\zeta}_j$	$\widehat{p}$
CD4Gn2nTp	-3.524 (-3.925,-3.149)	1.068 (0.86,1.278)	0 (0,0)	0.088 (0.063,0.111)	1.53 (1.496,1.566)
CD4Gn2pTn	-3.523 (-3.904,-3.173)	0.762 (0.569,0.946)	0.169 (0,0.322)	0.111 (0.083,0.145)	
CD4Gn2pTp	-3.459 (-3.843,-3.084)	1.04 (0.832,1.23)	0.379 (0.23,0.534)	0.053 (0.033,0.077)	
CD4Gp2nTn	-2.181 (-2.648,-1.813)	0.772 (0.432,1.142)	-1.079 (-1.293,-0.883)	0.136 (0.092,0.189)	
CD4Gp2nTp	-3.201 (-3.629,-2.812)	1.379 (1.13,1.636)	-0.694 (-0.876,-0.494)	0.052 (0.035,0.074)	
CD4Gp2pTn	-3.282 (-3.667,-2.918)	0.944 (0.75,1.161)	-0.484 (-0.645,-0.336)	0.083 (0.063,0.104)	
CD4Gp2pTp	-2.194 (-2.589,-1.852)	1.073 (0.915,1.247)	-0.223 (-0.354,-0.093)	0.055 (0.041,0.073)	

From Table 14, it is interesting to note how the estimates for the $\boldsymbol{\theta}_j$ vary across the different immune responses. This indicates that the different immune responses have differing levels of covariation. We explore this in more detail when we consider the ordination plots for this model.

Figure 54 presents several plots of the Dunn-Smyth residuals for Model 12 which can be used to assess the fit of the model. The four plots illustrate the following:

- the upper left panel shows a plot of the Dunn-Smyth residuals against the linear predictors. There appears to be an increase in the variance of the residuals as the linear predictor increases. This increase is mostly concentrated in the smaller values for the linear predictor and affects relatively few observations, and as such appears to be fairly insignificant.
- the upper right panel is a plot of the Dunn-Smyth residuals against the subject and occasion observations in the data; i.e. by row of the data. Again there is no clear pattern to the residuals indicating that the model appears to be suitable across the individuals and occasions.
- the lower left panel is a plot of the Dunn-Smyth residuals for the seven different responses. There is some difference in variability of the residuals across the different responses, but nothing which is too concerning. Each of the responses appears to be well modelled.
- the lower right panel is a normal QQ-plot of the Dunn-Smyth residuals. The residuals appear to be normally distributed with no obvious departures from normality.

It is difficult to see whether there is any time effect in the residuals plotted in Figure 54. Figure 55 plots the Dunn-Smyth residuals by

vaccine, response, and time point. As we have not allowed for the time covariate in the model and as we know there is a nonlinear pattern by time in the immune responses, we expect there to be patterns in the residuals by time, and this is what we see. However, these patterns have been dampened by including the latent variables which will capture some of the time effect. These patterns can be seen most prominently for response `CD4Gp2nTn` for Vaccines 3 and 4, response `CD4Gn2pTp` for Vaccine 3, and for response `CD4Gp2nTp` for Vaccine 7. We do expect that in the next iterations of this model where we include the time covariate, we will remove some of the pattern that we are seeing here.

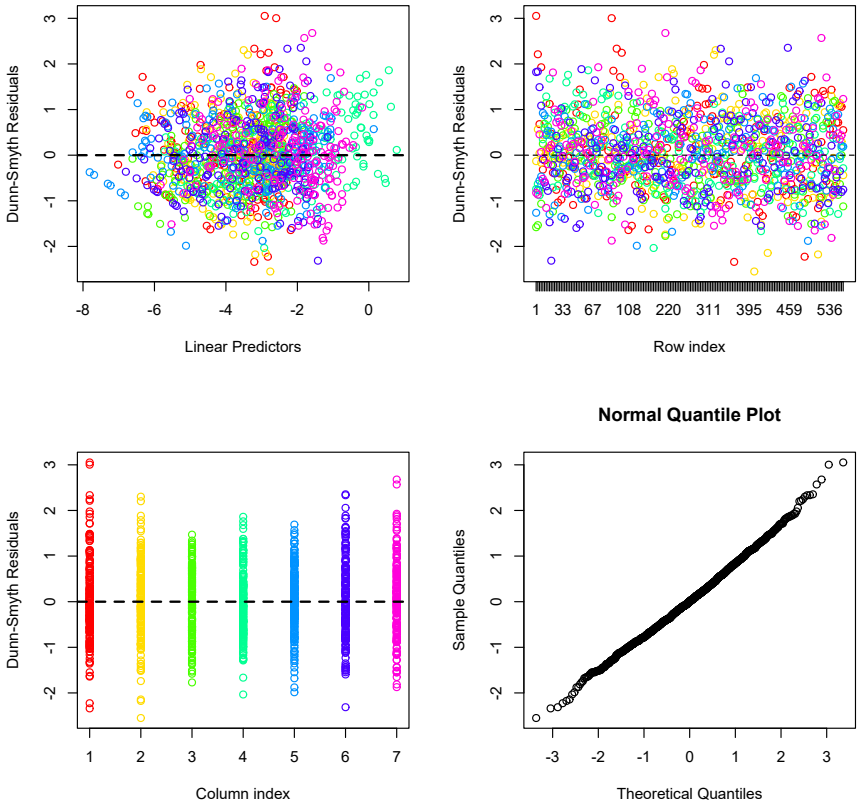


Figure 54: Dunn-Smyth residual plots for Model 12, a pure latent variable M-GLMEM. Each colour represents a different immune response.

We can use the fitted model to produce unconstrained ordination plots which are given in Figure 56. The unconstrained ordination plots involve plotting the scaled latent variable estimates for the observations and the latent variable loading estimates $\widehat{\theta}_j$ in biplots. The plots in Figure 56 are such biplots with the estimates of the latent variables shown as the numbered and coloured observations, and the corresponding coefficient estimates ($\widehat{\theta}_j$) labelled by the black response names. The numbering of the observations gives the time

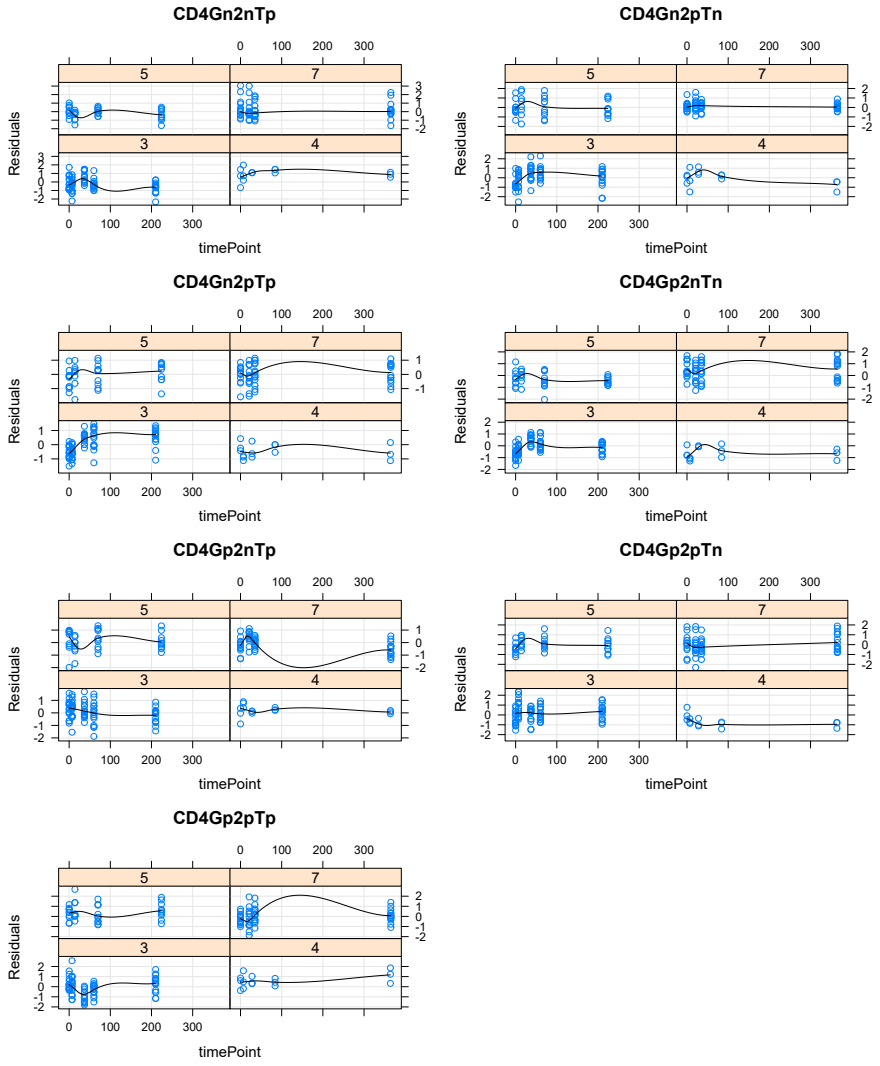


Figure 55: Dunn-Smyth residual plots by vaccine, response, and time point for Model 12, a pure latent variable M-GLMEM.

that they were recorded and the colouring gives the vaccine they pertain to. The plots are referred to as unconstrained as we have not accounted for the correlation between the responses that is due to the measured covariates and are to be interpreted in the same manner as plots constructed from nonmetric multi-dimensional scaling and correspondence analysis (Hui, 2016, 2018). From the plots we can see that:

- at a high-level, the observations from the Vaccines 4, 5, and 7 are relatively clustered indicating that the vaccines induce responses across similar cytokine combinations. Vaccine 3 has observations which overlap all of the other vaccines, but which are not as clustered indicating that it has a slightly wider range of induced responses.

- looking a bit closer, we see that Vaccines 4 and 5 have observations which lie in the same region while Vaccine 7 has observations which are in a slightly different region of the plot. This suggests that the observations for Vaccines 4 and 5 are similar, but marginally different to those of Vaccine 7.
- in particular, we see that Vaccines 3, 4, and 5 are more closely clustered to the `CD4Gp2pTp` response whereas Vaccine 7 is clustered between `CD4Gp2pTp` and `CD4Gp2pTn` responses. The proximity of the observations from a vaccine to specific responses indicates that the vaccines are more associated with those specific responses.
- As we have not allowed for the time covariate in the model we do expect there to be a time pattern in the plot as the latent variables will capture some of the time effect. To some extent this can be seen in the plots for the individual vaccines. We see that the observations at time zero tend to be to the left and later observations to the right of the plot; this suggests that latent variable one (the x-axis) can be interpreted as a time variable.
- the different response variables are more spread across latent variable two, this suggests that latent variable two (the y-axis) is accounting for the difference in the responses.

The large overlap in the observations for the different vaccines broadly indicates that the vaccine candidates induce immune responses across similar cytokine combinations. This lack of diversity in T cell responses for the different vaccine candidates is a point that is also noted by Rodo (2017) and Rodo *et al.* (2019a).

We can calculate the residual correlation / covariance matrix captured by the latent variables for Model 12. The residual covariance matrix is calculated as $\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Theta}}^{\mathsf{T}}$ where

$$\widehat{\mathbf{\Theta}} = (\widehat{\boldsymbol{\theta}}_{\texttt{CD4Gn2nTp}}, \ldots \widehat{\boldsymbol{\theta}}_{\texttt{CD4Gp2pTp}})^{\mathsf{T}}.$$

Substituting the estimates from the fitted model, we have the following estimate of the correlation matrix

	CD4Gn2nTp	CD4Gn2pTn	CD4Gn2pTp	CD4Gp2nTn	CD4Gp2nTp	CD4Gp2pTn	CD4Gp2pTp
CD4Gn2nTp	1	0.976	0.94	0.582	0.893	0.891	0.98
CD4Gn2pTn	0.976	1	0.991	0.388	0.772	0.769	0.911
CD4Gn2pTp	0.94	0.991	1	0.267	0.684	0.68	0.85
CD4Gp2nTn	0.582	0.388	0.267	1	0.887	0.89	0.736
CD4Gp2nTp	0.893	0.772	0.684	0.887	1	0.999	0.966
CD4Gp2pTn	0.891	0.769	0.68	0.89	0.999	1	0.964
CD4Gp2pTp	0.98	0.911	0.85	0.736	0.966	0.964	1

The correlation between two responses can be interpreted as the correlation due to interactions between the responses and due to the underlying covariates, measured and unmeasured. Figure 57 provides a plot of the correlation matrix showing the significant correlations. We can see that most of the correlations are positive and strong. The exceptions are the correlations between the immune response `CD4Gp2nTn`

and the two responses `CD4Gn2pTn` and `CD4Gn2pTp`. This reduced correlation may reflect an underlying immunological relationship between the production of the cytokines IFN-γ, IL-2, and TNF-α.

In the model that we have fitted here, we did not include any covariates and so the latent variables will be capturing any covariation in the immune responses that is due to the measured covariates that we have available. The trace of the estimated residual covariance matrix gives a measure of the total covariation captured by the latent variables. In this case it is 9.563. We can compare this value to the trace of the estimated residual covariance matrices of the subsequent models that we fit where we include covariates. This will give us a measure of the covariation that is captured by the measured covariates.

Model 12, which does not include occasion specific covariates, gives us insight into the covariation that exists between the immune responses for different vaccines and how they compare. This has allowed us to identify that there is limited diversity in the immune responses induced by the different vaccine candidates. In the subsequent models that we fit, we include covariates which should reduce the covariation between immune responses that is being captured by the latent variables. We also expect that the models with covariates to fit the data better and, in particular, to better capture the nonlinear profiles of the immune responses.

3.4.2 *Model with Occasion Specific Covariates*

MODEL 13 In this subsection we consider a model with vaccine and time point covariates. It is expected that these covariates will capture a large proportion of the covariation between the immune responses, but we also expect that there will be some residual covariation. This residual covariation is likely due to additional covariates that we have not measured or other biotic reactions not captured in the model. We include latent variables alongside the measured covariates to account for this remaining covariation. The time point covariates are included in the model via a cubic B-spline, as before, to capture the nonlinear profiles of the immune responses.

The model with covariates that we consider here has the form

$$\log(\mu_{ijk}) = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \boldsymbol{z}_{ik}^{\mathsf{T}} \boldsymbol{\theta}_j$$

with

$$\begin{aligned}
\phi_{0,ij} &= b_i + \beta_{0j} + \beta_{0j,\text{vaccine}_i}\text{vaccine}_i \\
\phi_{1,ij} &= \beta_{1j,0}, \\
\phi_{2,ij} &= \beta_{2j,0}, \\
\phi_{3,ij} &= \beta_{3j,0},
\end{aligned}$$

for $j =$ `CD4Gn2nTp`, ..., `CD4Gp2pTp`, and where $b_i \sim \mathrm{N}(0, \sigma^2)$ and $\boldsymbol{z}_{ik} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{2\times 2})$. We include the vaccine covariate in the intercept term of the model as a fixed effect. This allows for different levels in the immune responses for the different vaccines, but with the shapes of the immune response profiles for the different vaccines being the same.

Again we estimate this model by using the `boral` package in `R` (Hui, 2016). The estimated parameters for the model are presented in Tables 15 and 16, and the estimate of σ is 0.252. These are again the median values from the posterior distributions of the parameters. The estimated marginal AIC and BIC at the median of the posterior distribution, which are calculated by using Equation (10), are -4893.3 and -4526.8 respectively, and the estimated EAIC and EBIC are -4862.1 and -2410.7 respectively. These measures, apart from the EBIC, indicate that this model is a significant improvement over Model 12.

Table 15: Output for Model 13, a M-GLMEM. Values in the brackets correspond to the 95% highest posterior density (credible) intervals.

	Parameter				
Response j	$\widehat{\beta}_{0j}$	$\widehat{\theta}_{j,1}$	$\widehat{\theta}_{j,2}$	$\widehat{\zeta}_j$	$\widehat{p}$
CD4Gn2nTp	-4.211 (-4.628,-3.829)	0.814 (0.641,0.981)	0 (0,0)	0.063 (0.045,0.082)	1.499 (1.463,1.534)
CD4Gn2pTn	-4.264 (-4.607,-3.894)	0.357 (0.198,0.531)	0.131 (0.001,0.282)	0.093 (0.068,0.119)	
CD4Gn2pTp	-4.275 (-4.618,-3.945)	0.635 (0.482,0.807)	-0.054 (-0.184,0.09)	0.053 (0.039,0.069)	
CD4Gp2nTn	-2.832 (-3.364,-2.339)	0.898 (0.591,1.189)	1.088 (0.897,1.279)	0.069 (0.039,0.103)	
CD4Gp2nTp	-3.677 (-4.17,-3.206)	1.188 (0.946,1.432)	0.676 (0.454,0.847)	0.051 (0.036,0.069)	
CD4Gp2pTn	-3.478 (-3.878,-3.069)	0.785 (0.585,0.983)	0.53 (0.366,0.67)	0.064 (0.05,0.083)	
CD4Gp2pTp	-2.821 (-3.221,-2.469)	0.932 (0.777,1.085)	0.296 (0.143,0.443)	0.041 (0.028,0.057)	

Table 16: Continued output for Model 13, a M-GLMEM. Values in the brackets correspond to the 95% highest posterior density (credible) intervals.

	Parameter					
Response j	$\beta_{0j,\text{vaccine4}}$	$\beta_{0j,\text{vaccine5}}$	$\beta_{0j,\text{vaccine7}}$	$\beta_{1j,0}$	$\beta_{2j,0}$	$\beta_{3j,0}$
CD4Gn2nTp	-0.43 (-1.076,0.321)	-1.282 (-1.794,-0.743)	1.116 (0.687,1.559)	3.707 (2.403,4.999)	-2.268 (-3.629,-0.885)	0.085 (-0.514,0.648)
CD4Gn2pTn	-1.473 (-2.119,-0.771)	-1.076 (-1.572,-0.626)	0.881 (0.473,1.231)	4.548 (3.23,5.664)	-1.534 (-2.572,-0.244)	0.193 (-0.285,0.666)
CD4Gn2pTp	-1.548 (-2.166,-0.859)	-1.214 (-1.708,-0.804)	0.662 (0.269,1.027)	5.195 (4.072,6.327)	-0.469 (-1.528,0.584)	0.402 (-0.036,0.932)
CD4Gp2nTn	-2.411 (-3.404,-1.49)	-1.528 (-2.145,-0.84)	1.993 (1.453,2.604)	1.722 (-0.293,3.511)	-2.033 (-3.867,-0.33)	-0.001 (-0.769,0.849)
CD4Gp2nTp	-1.3 (-2.234,-0.447)	-1.098 (-1.678,-0.433)	1.279 (0.728,1.854)	2.939 (1.036,4.679)	-1.724 (-3.367,0.024)	-0.344 (-1.113,0.522)
CD4Gp2pTn	-2.539 (-3.347,-1.72)	-1.443 (-1.967,-0.89)	0.986 (0.537,1.414)	1.54 (0.044,2.927)	-0.908 (-2.191,0.443)	0.066 (-0.515,0.704)
CD4Gp2pTp	-0.809 (-1.438,-0.078)	-0.867 (-1.373,-0.403)	1.181 (0.772,1.629)	2.392 (0.99,3.627)	-0.227 (-1.45,0.919)	0.065 (-0.554,0.615)

Figure 58 presents several plots of the Dunn-Smyth residuals for Model 13 which we can use to assess the underlying assumptions of the model. The plots are very similar to those for Model 12 shown in Figure 54 and broadly the same comments apply. Most notably, we see that there is possibly increasing variance in the residuals as the linear predictor increases, but this is again fairly minor.

Figure 59 plots the Dunn-Smyth residuals by vaccine, response, and time point. As we have now allowed for the time point covariate via the cubic B-spline in the model, we expect that the patterns in the

residuals over time, that we saw in Figure 55, should be dampened. To some extent this is the case; some of the pattern in the residuals by time has been removed, but not all of it. The remaining pattern can be clearly seen for response `CD4Gp2pTp` for Vaccines 3 and 7 and response `CD4Gn2pTn` for all of the Vaccines, as well as for other response and vaccine combinations. This pattern does appear to be relatively minor as most of the residuals are still fairly well centered around zero, but the fit is not as good as previous models we have considered. This does suggest that these models struggle to fully capture the nonlinear profiles of the immune responses.

Plots of the conditional fitted immune response values from Model 13 against the actual immune response values are shown in Figure 60. This plot also provides us with a means of assessing the fit of the model; if the model fits well, we expect the points to lie close to the diagonal line. We see that for most of the vaccine and response combinations, the model appears to fit the data well. There are a few vaccine and response combinations where the model does not fit well. This can most notably be seen for Vaccine 7 for responses `CD4Gn2nTp` and `CD4Gn2pTn`, Vaccine 3 for response `CD4Gn2pTn`, and Vaccine 5 for response `CD4Gn2pTn`. It can also be seen that the model underestimates the larger immune responses in some cases, e.g. for Vaccine 3 for the responses `CD4Gn2pTp` and `CD4Gn2pTn`.

There are some deficiencies to the model that we have fitted here. Specifically it appears not to fully capture the nonlinear nature of the immune responses and is a poor fit for some vaccine and response combinations. Given these deficiencies, we can still attempt to interpret the fitted model and this is what we do below.

From the estimated parameters in Table 16, it is interesting to see that the vaccine effects are mostly significant as their highest posterior density intervals do not include zero. These vaccine effects are relative to Vaccine 3 which is included in the intercept. This indicates that the magnitude of the immune responses for the different vaccines are significantly different to that of Vaccine 3, with Vaccine 7 having the largest immune response.

From the fitted model we can calculate the marginal expected immune response profiles for each vaccine and compare them. By comparing the profiles, we can identify the vaccines that induced the greatest immune response and as a result which should be advanced for further trials. These fitted expected marginal profiles are calculated by numerically integrating over the random effects and latent variables for the fitted model. These fitted expected marginal immune response profiles are shown in Figure 61. The lack of smoothness is due to the numerical integration; increasing the number of evaluations at each time point would smooth out the profiles, but the shapes and levels are unlikely to change. We see that the shapes of the immune response profiles broadly conform to the description given in Section

1.2 with an initial increase in the responses and then a tapering off to a memory response level. The increase in the immune responses after 300 days is an artifact of models we have fitted and is due to the lack of flexibility in the cubic B-splines we used as well as having data at only a limited number of time points. As discussed, increasing the flexibility of the splines resulted in models which had immune response profiles which did not conform to the description given in Section 1.2.

We can interpret these marginal expected immune response profiles and use them to visually identify the vaccines that should be advanced. From the profiles, it is clear that Vaccine 7, the current TB vaccine BCG, has the most significant effect on the immune responses with limited overlap in its 95% credible intervals with other vaccines' 95% credible intervals. Based on these results, it is clear that none of the vaccine candidates included here are competitive relative to the current BCG vaccine. Vaccine 3 is the only novel vaccine which looks to be promising relative to the other novel vaccines we have included; its induced immune responses are mostly greater than those of the other novel vaccines.

Model 13 can also be used to produce residual ordination plots which allow us to explore the residual covariation between the immune responses for the different vaccines. Figure 62 presents the residual ordination plots for Model 13. The plots are similar to those in Figure 56, except for Model 13 we have now accounted for the covariation in the immune responses due to the time point and vaccine covariates, and as such the ordination plots here are examining the residual covariation not explained by these covariates. From the plots we see that there is less pattern observable in the plots relative to those in Figure 56. Specifically:

- the observations from the Vaccines 3, 4, 5, and 7 are now more clustered and overlap in the same region. Vaccine 3 still has slightly more variation in its observations across the latent variables.
- the observations lie between the `CD4Gp2pTp`, `CD4Gp2nTp`, `CD4Gp2pTn`, and `CD4Gn2pTn` responses.
- there is now less pattern in the time points and vaccines; i.e. we do not see any clear ordering of the time points and vaccines along either of the latent variables.

The interpretation of the latent variables as capturing the residual covariation due to missing covariates allows us to interpret the values of the z_{ik} as the values of the missing covariates and the θ_j as the corresponding coefficients. To this point it is interesting to note that the responses `CD4Gp2nTn`, `CD4Gn2nTp` and `CD4Gn2pTp` lie further away from the origin indicating that their coefficients are larger than the other

responses. This may indicate that there are important covariates or biotic reactions not captured in the current set of measured covariates for these particular responses. Identifying and including these missing covariates may help to explain some of the residual covariation in the immune responses.

For Model 13, we can calculate the residual correlation / covariance between the responses that is captured by the latent variables. The residual correlation matrix for Model 13 is given by

	CD4Gn2nTp	CD4Gn2pTn	CD4Gn2pTp	CD4Gp2nTn	CD4Gp2nTp	CD4Gp2pTn	CD4Gp2pTp
CD4Gn2nTp	1	0.937	0.995	0.63	0.865	0.82	0.952
CD4Gn2pTn	0.937	1	0.905	0.862	0.974	0.963	0.985
CD4Gn2pTp	0.995	0.905	1	0.562	0.818	0.767	0.921
CD4Gp2nTn	0.63	0.862	0.562	1	0.936	0.961	0.841
CD4Gp2nTp	0.865	0.974	0.818	0.936	1	0.997	0.977
CD4Gp2pTn	0.82	0.963	0.767	0.961	0.997	1	0.957
CD4Gp2pTp	0.952	0.985	0.921	0.841	0.977	0.957	1

Figure 63 plots the significant residual correlations for Model 13. We can see that most of the residual correlations between the immune responses are positive and strong which is probably expected.

Including the occasion specific covariates in the model should reduce the covariation between the immune responses that is explained by the latent variables. The trace of the residual covariance matrix is 6.958 for Model 13. This is a reduction relative to that of 9.563 for Model 12 which means that the occasion specific covariates explained 27.2% of the covariation in the immune responses. There is a still a large proportion of covariation being explained by the latent variables. This may indicate that there are significant covariates or biotic reactions not currently represented in the measured covariate set which could improve the fit of the model.

In addition to calculating the residual correlation matrix as captured by the latent variables, we can also calculate the correlations in immune responses due to the occasion specific covariates. These correlations between the immune responses for Model 13 due to the occasion specific covariates are given by

CD4Gn2nTp	CD4Gn2pTn	CD4Gn2pTp	CD4Gp2nTn	CD4Gp2nTp	CD4Gp2pTn	CD4Gp2pTp	
CD4Gn2nTp	1	0.89	0.797	0.869	0.937	0.827	0.929
CD4Gn2pTn	0.89	1	0.95	0.837	0.92	0.881	0.924
CD4Gn2pTp	0.797	0.95	1	0.699	0.817	0.793	0.863
CD4Gp2nTn	0.869	0.837	0.699	1	0.951	0.959	0.926
CD4Gp2nTp	0.937	0.92	0.817	0.951	1	0.934	0.96
CD4Gp2pTn	0.827	0.881	0.793	0.959	0.934	1	0.912
CD4Gp2pTp	0.929	0.924	0.863	0.926	0.96	0.912	1

and Figure 64 plots the significant correlations. We see that all of the correlations in the immune responses due to the occasion specific covariates are positive and strong. This is what we expect as all of the immune response profiles for the different vaccines and responses have similar shapes over time.

It should be noted that we did fit a more complex model to try and address the deficiencies identified for Model 13. This more complex model included vaccine fixed effects for the cubic B-spline terms, but it did not improve the information criteria and the deficiencies remained.

The model that we have fitted here is fairly complicated and we can try to simplify it by including traits for the responses. This is what we do in the next model that we investigate.

3.4.3 *Model with Occasion Specific Covariates and Response Traits*

MODEL 14 In this subsection we consider a model that has both occasion specific covariates and traits for the responses `CD4Gn2nTp`,..., `CD4Gp2pTp`. The traits give the characteristics of the cytokines for the responses and we include them to try explain the differences in the immune response profiles to the occasion specific covariates.

For each response, the trait variables are encapsulated in 3×1 vectors of indicator variables. The indicator variables identify the presence or absence of the specific cytokines for the associated response. For example, the trait vector associated with response `CD4Gn2pTn` is given by

$$\text{traits}_{\texttt{CD4Gn2pTn}} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

where the first zero indicates the absence of IFN-γ, the one indicates the presence of IL-2, and the last zero indicates the absence of TNF-α. Similar vectors can be constructed for the other responses.

The model that we consider in this subsection has the following form

$$\log(\mu_{ijk}) = \phi_{0,ij} + \sum_{w=1}^{3} \phi_{w,ij} B_{w,3}(t_k) + \boldsymbol{z}_{ik}^{\mathsf{T}} \boldsymbol{\theta}_j$$

with

$$\begin{aligned} \phi_{0,ij} &= b_i + \beta_{0j} + \beta_{0j,\text{vaccine}_i} \text{vaccine}_i \\ \phi_{1,ij} &= \beta_{1j,0}, \\ \phi_{2,ij} &= \beta_{2j,0}, \\ \phi_{3,ij} &= \beta_{3j,0}, \end{aligned}$$

for $j =$ `CD4Gn2nTp`, ..., `CD4Gp2pTn`, and `CD4Gp2pTp`.

To include the traits in the model, we now model the above β parameters as random effects drawn from normal distributions. Specifically, we model the column specific intercepts, the β_{0j}s, as random effects drawn from the normal distribution

$$\beta_{0j} \sim \text{N}(\kappa_{0,0} + \text{traits}_j^{\mathsf{T}} \boldsymbol{\kappa}_0, \sigma_0^2)$$

where $(\kappa_{0,0}, \boldsymbol{\kappa}_0)$ are the regression parameters relating the traits to the column specific intercepts. The remaining β parameters $(\beta_{0j,\text{vaccine}_4}, \beta_{0j,\text{vaccine}_5}, \beta_{0j,\text{vaccine}_7}, \beta_{1j,0}, \beta_{2j,0}, \beta_{3j,0})$ are also drawn from normal distributions of the form

$$\beta_{0j,\text{vaccine}_i} \sim \text{N}(\kappa_{0,\text{vaccine}_i} + \text{traits}_j^{\mathsf{T}} \boldsymbol{\kappa}_{\text{vaccine}_i}, \sigma^2_{\text{vaccine}_i})$$

and

$$\beta_{wj,0} \sim \text{N}(\kappa_{0,w} + \text{traits}_j^{\mathsf{T}} \boldsymbol{\kappa}_w, \sigma_w^2).$$

The β parameters that appeared in Model 13 have now all been regressed against the response traits and as such the inclusion of traits can be seen as a simplification of the model. That is, including the traits is a simplification of Model 13 as we are now saying that the coefficients in Model 13 are linearly related to the traits, and we are using the traits to explain similarities and differences in the immune responses to the occasion specific covariates (Warton *et al.*, 2015; Hui, 2018). The coefficients in the model that need to be estimated are now the κs.

We fit this model by using the `boral` package in `R`. The parameter estimates are shown in Tables 17 and 18. The EAIC and EBIC for the model, which are calculated by using Equation (10), are -4852.5 and -2401.1 which are not an improvement over those of Model 13. This indicates that the simplification that arises from the inclusion of the traits does not significantly improve the model. Looking at the parameter estimates in Table 18, we can see that none of the trait coefficients are significantly different to zero as all of their 95% credible intervals include zero. This is consistent with the lack of improvement in the information criteria relative to Model 13.

Table 17: Parameter estimates for Model 14, a correlated response M-GLMEM with traits.

	Parameter			
Response j	$\widehat{\theta}_{j,1}$	$\widehat{\theta}_{j,2}$	$\widehat{\zeta}_j$	$\widehat{p}$
CD4Gn2nTp	0.795 (0.635,0.96)	0 (0,0)	0.064 (0.047,0.085)	1.502 (1.469,1.539)
CD4Gn2pTn	0.354 (0.204,0.527)	0.122 (0,0.266)	0.095 (0.073,0.125)	
CD4Gn2pTp	0.639 (0.499,0.812)	-0.055 (-0.183,0.083)	0.053 (0.039,0.07)	
CD4Gp2nTn	0.879 (0.532,1.18)	1.077 (0.896,1.266)	0.073 (0.044,0.106)	
CD4Gp2nTp	1.173 (0.902,1.421)	0.679 (0.474,0.887)	0.053 (0.038,0.07)	
CD4Gp2pTn	0.783 (0.587,0.99)	0.54 (0.383,0.694)	0.065 (0.049,0.081)	
CD4Gp2pTp	0.926 (0.769,1.099)	0.301 (0.135,0.449)	0.041 (0.027,0.058)	

Table 18: Continued.: parameter estimates for Model 14, a correlated response M-GLMEM with traits. Parameter estimates are for the coefficient models for Model 14.

	Parameter				
$(\kappa_{...}, \boldsymbol{\kappa}_{...})$	$\kappa_{0,...}$	$\boldsymbol{\kappa}_{...}(1)$	$\boldsymbol{\kappa}_{...}(2)$	$\boldsymbol{\kappa}_{...}(3)$	$\sigma_{...}$
$(\kappa_{0,0}, \boldsymbol{\kappa}_0, \sigma_0)$	-3.02 (-4.286,-1.573)	-0.548 (-1.627,0.792)	-0.852 (-1.941,0.258)	-0.203 (-1.83,1.273)	0.481 (0.103,1.436)
$(\kappa_{0,\text{vaccine}_4}, \boldsymbol{\kappa}_{\text{vaccine}_4}, \sigma_{\text{vaccine}_4})$	-2.073 (-4.248,0.324)	-0.175 (-2.344,1.989)	-0.371 (-2.344,1.743)	0.92 (-2.125,3.278)	1.034 (0.348,2.624)
$(\kappa_{0,\text{vaccine}_5}, \boldsymbol{\kappa}_{\text{vaccine}_5}, \sigma_{\text{vaccine}_5})$	-1.349 (-2.294,-0.593)	-0.061 (-0.636,0.561)	-0.369 (-0.971,0.338)	0.539 (-0.26,1.438)	0.158 (0,0.705)
$(\kappa_{0,\text{vaccine}_7}, \boldsymbol{\kappa}_{\text{vaccine}_7}, \sigma_{\text{vaccine}_7})$	1.852 (1.04,2.79)	-0.133 (-0.743,0.51)	-0.28 (-1.008,0.294)	-0.666 (-1.598,0.284)	0.187 (0.001,0.717)
$(\kappa_{0,1}, \boldsymbol{\kappa}_1, \sigma_1)$	3.034 (-0.056,5.536)	1.399 (-1.099,3.873)	0.71 (-1.639,3.151)	0.48 (-2.413,3.567)	1.156 (0.034,2.905)
$(\kappa_{0,2}, \boldsymbol{\kappa}_2, \sigma_2)$	-2.403 (-5.198,0.57)	0.016 (-2.623,2.809)	0.065 (-2.477,2.947)	0.401 (-3.148,3.332)	1.463 (0.434,3.418)
$(\kappa_{0,3}, \boldsymbol{\kappa}_3, \sigma_3)$	0.452 (-0.664,1.394)	-0.002 (-0.715,0.764)	0.15 (-0.631,0.849)	-0.207 (-1.241,0.798)	0.218 (0.001,0.804)

The additional analysis that we performed for Model 13 can be performed for Model 14, but the results are almost identical, with most of the same comments and insights applying, and so we do not repeat the analysis here. We do provide the figures and results relating to this further analysis for Model 14 in Appendix D.

3.4.4 *Comment on the M-GLMEMs*

The M-GLMEMs are the final set of models that we consider in our applications. They have several technical advantages over the other models that we have considered in the context of modelling multivariate nonlinear immune response profiles that have zero-inflated skew distributions. These advantages are

- the models are explicitly multivariate at the outset,
- they can accommodate the zero-inflated skew distributions of the immune responses,
- they allow for the residual correlation between the immune responses to be modelled explicitly and parsimoniously, and
- they also provide us with the means of exploring the modelled residual correlations via ordination plots.

In addition to the above advantages, we have seen that in terms of information criteria, the M-GLMEMs provided the best models, with Model 13 outperforming all of the other models that we considered in terms of AIC and BIC. The models also appear to fit reasonably well apart from some difficulty in capturing the nonlinear profiles of the immune responses over time and underestimating the larger immune responses in some cases.

We have also been able to interpret the models relatively easily. We have seen from the models that we fitted, that there is limited diversity in the immune responses induced by the different vaccine candidates. This was seen in the unconstrained ordination plots for Model 12 in Figure 57 where the observations for all vaccines have considerable overlap. This lack of diversity in the immune responses is also identified by Rodo (2017) and Rodo *et al.* (2019a) who argue that it is a concern as vaccine candidates should induce a diverse immune response so as to increase the likelihood of finding an effective alternative to BCG.

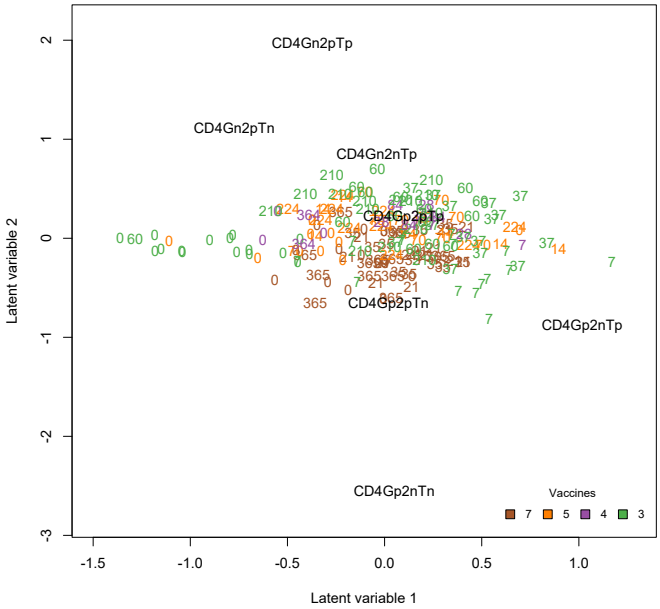
In addition to the above, we were able to see that the immune responses for the current TB vaccine, BCG, Vaccine 7 in the work here, were the largest with Vaccine 3 being the most promising from the novel vaccines we considered. This conclusion is consistent with that of Rodo (2017) and Rodo *et al.* (2019a) who identify that the M72/AS01$_{\text{E}}$ vaccine, Vaccine 3 in the work here, has the largest immune response for the novel vaccines that they considered.

The residual ordination plots for Model 13, which give us a means of examining the residual covariation between the immune responses for the different vaccines after accounting for the measured covariates, suggested that there may be unmeasured covariates which may be important for some of the responses. It would be interesting to identify whether there are additional covariates which have been measured and

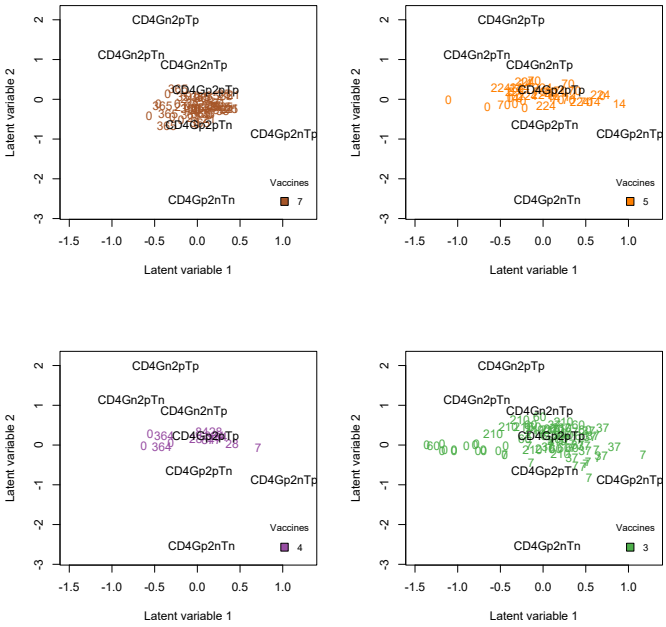
could be included in the models to try explain some of the residual covariation.

The inclusion of traits for the responses did not prove to be useful here. This may be due to the limited information carried by the trait variables that we used and or the limited data available. Including traits for the cytokines and cytokine combinations which carry information which is more immunologically relevant, for example, may still prove worthwhile. Such traits may be useful for understanding vaccine induced immune responses.

Ultimately, the M-GLMEMs are a very promising class of models which are able to capture the features of the multivariate immune response data that we considered here in a flexible and interpretable manner. The deficiencies in the fitted models here should not deter further applications to multivariate immune response data. It is likely with data collected at more regular time points and for a larger number of subjects these deficiencies would not arise.



(a) Model based unconstrained ordination plot for Model 12 for all of the vaccines.



(b) Model based unconstrained ordination plots for Model 12 for each of the vaccines.

Figure 56: Model based unconstrained ordination plots for Model 12, based on the median of the posterior distribution. The observations are given by the coloured and numbered points where the colouring indicates the vaccine the observation pertains to, and the numbering indicates the time that the observation was recorded. The immune responses, the biplot axes, are given by the points marked by black text.

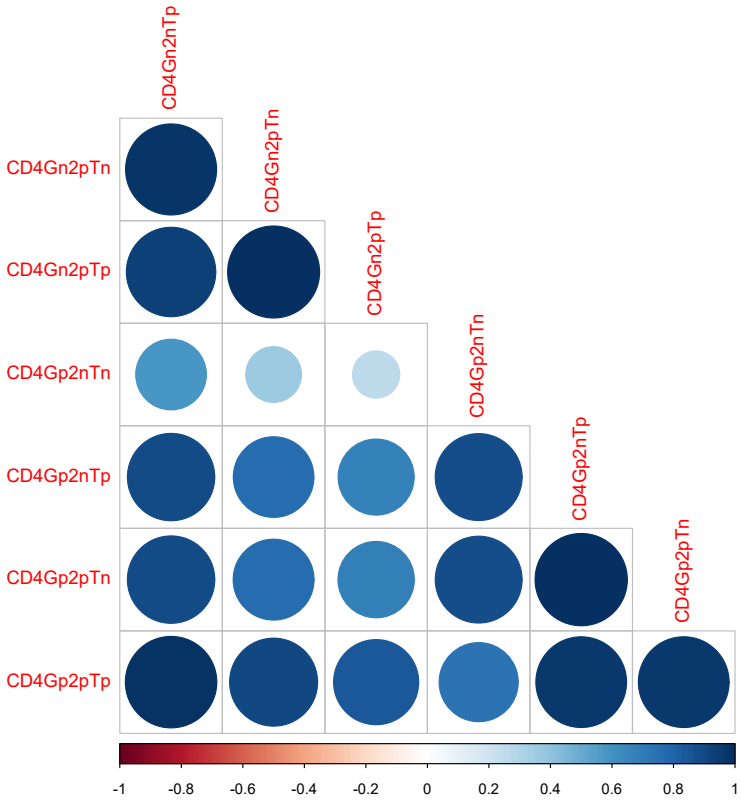


Figure 57: Plot of the significant correlations between the responses before controlling for time point and vaccine effects. Significance is determined from the 95% credible intervals; i.e. only the correlations with credible intervals excluding zero are plotted.

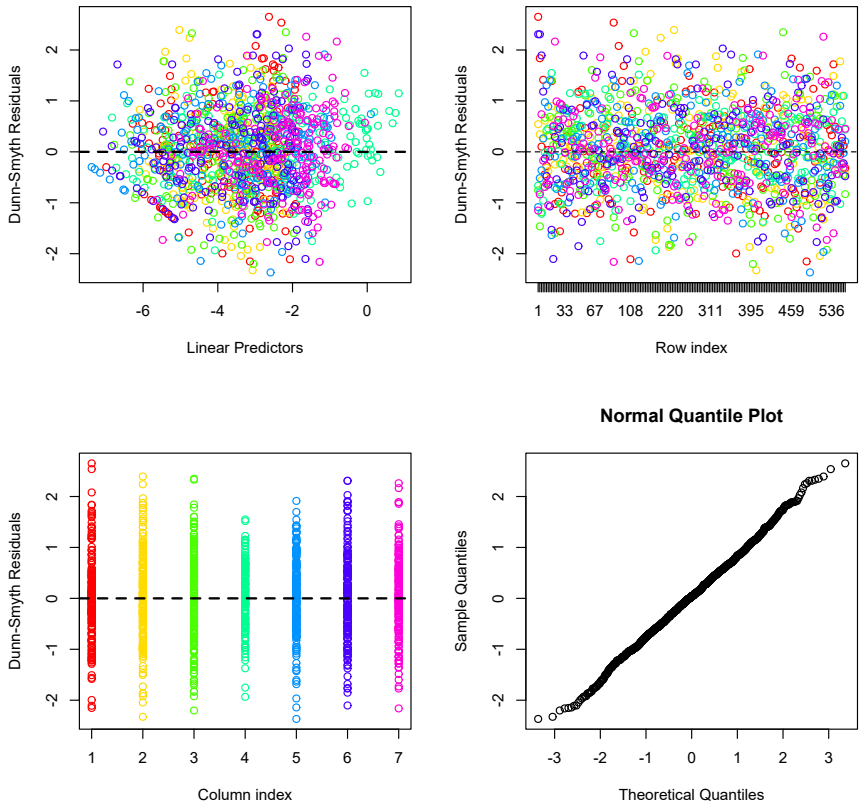


Figure 58: Dunn-Smyth residual plots for Model 13, a correlated response M-GLMEM. Each colour represents a different immune response.

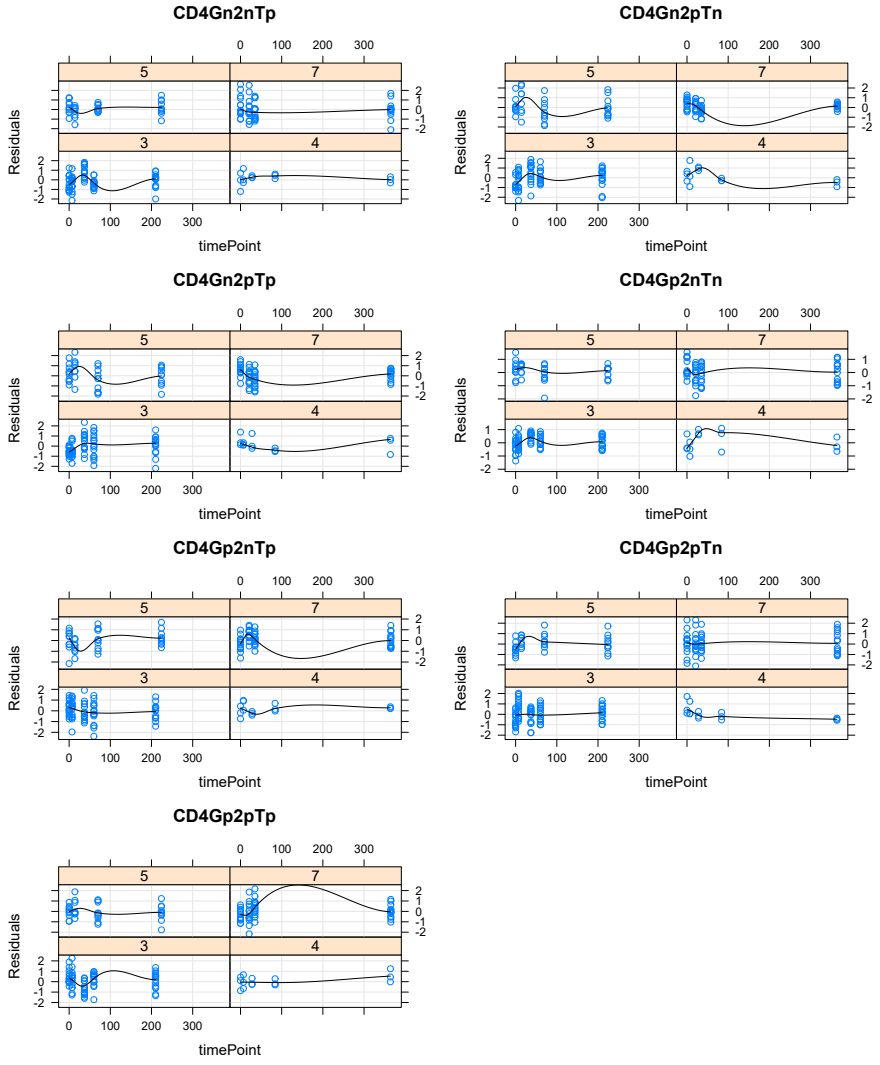


Figure 59: Dunn-Smyth residual plots by vaccine, response, and time point (days) for Model 13, a correlated response M-GLMEM.

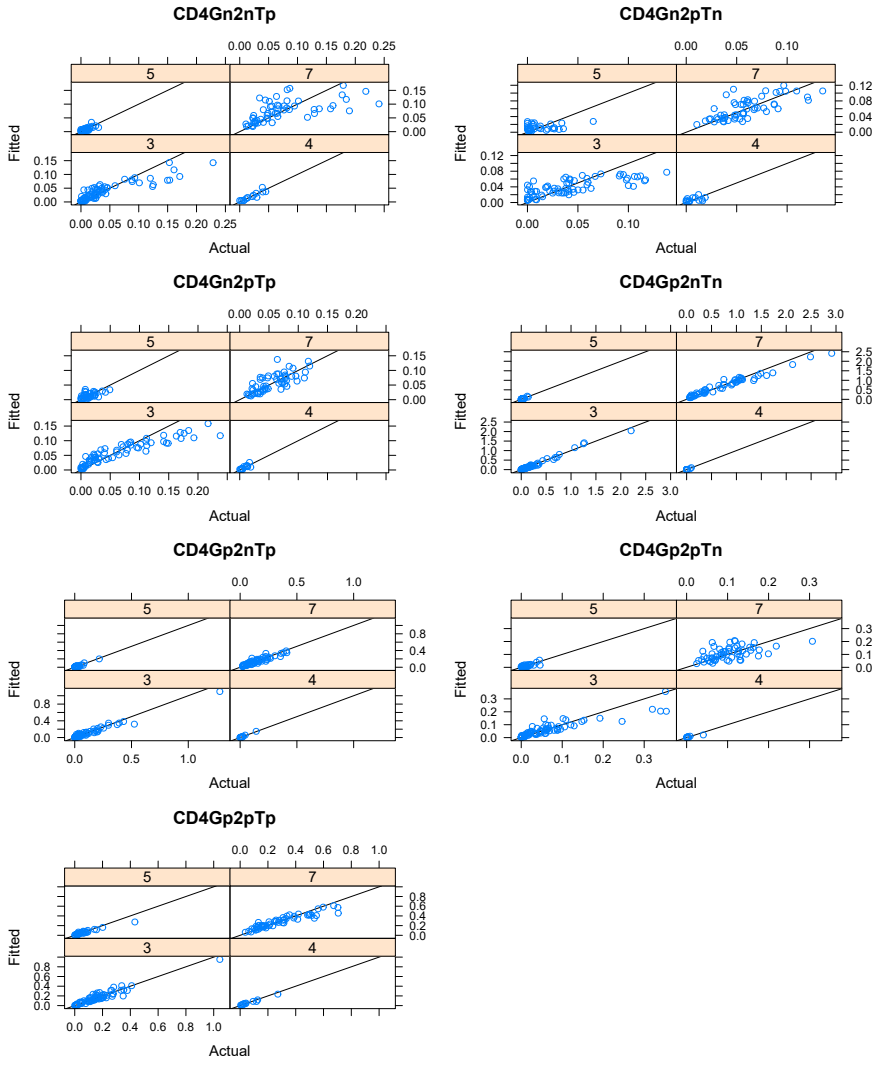


Figure 60: Plot of the fitted conditional response values versus the actual vaccine response values for each of the vaccines and responses for Model 13.

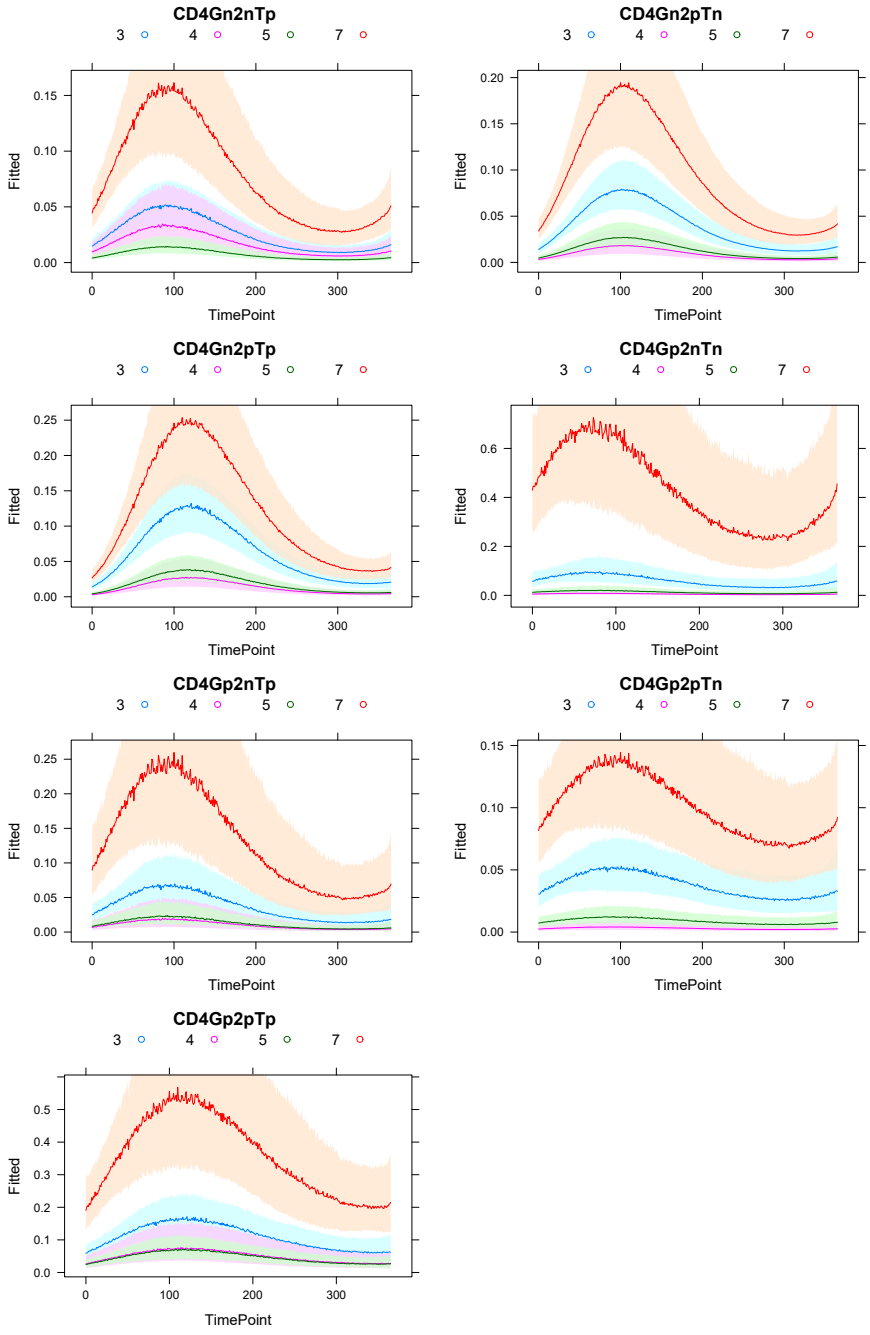
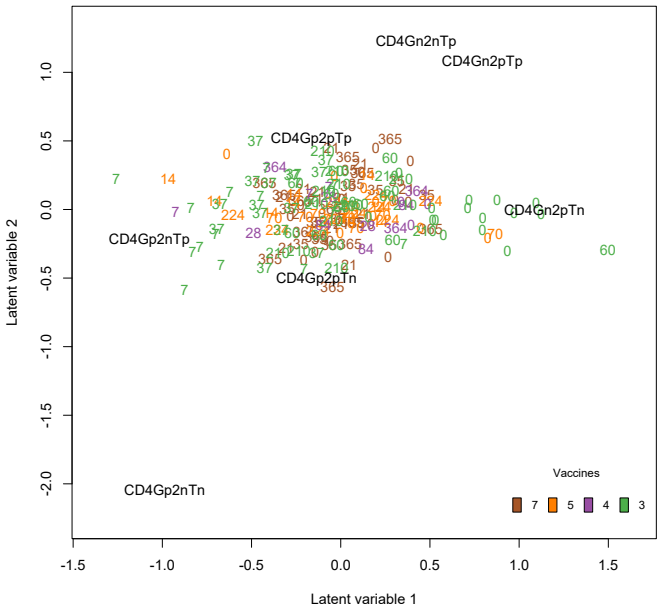
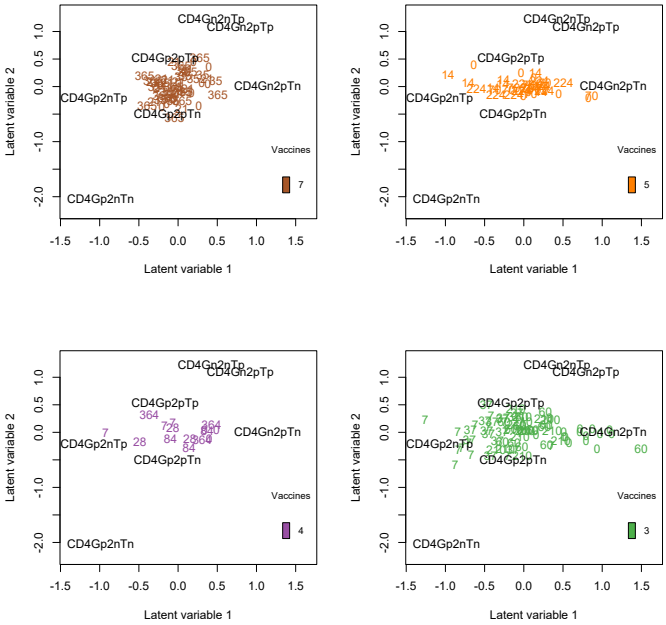


Figure 61: Plots of the fitted marginal expected immune response profiles for each of the vaccines and responses for Model 13. The 95% credible intervals are given by the shaded regions in the corresponding colours.



(a) Model based residual ordination plot for Model 13 for all of the vaccines.



(b) Model based residual constrained ordination plots for Model 13 for each of the vaccines.

Figure 62: Model based residual ordination plots for Model 13, a correlated response M-GLMEM. The observations are given by the coloured and numbered points where the colouring indicates the vaccine the observation pertains to, and the number the time that the observation was recorded. The immune responses, the biplot axes, are given by the points marked by black text.

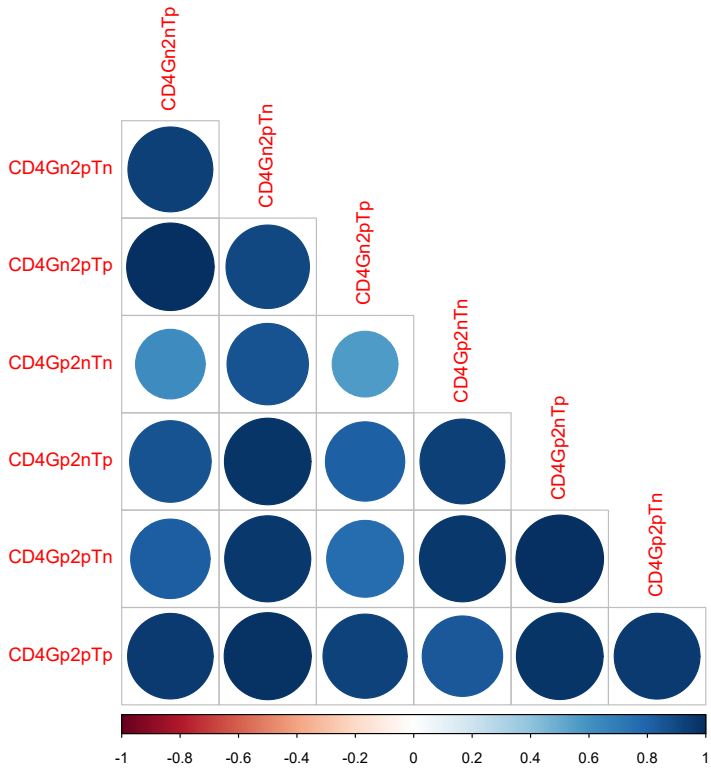


Figure 63: Plot of the significant residual correlations between the responses after controlling for time point and vaccine effects for Model 13. Significance is determined from the 95% credible intervals; i.e. only the correlations with credible intervals excluding zero are plotted.

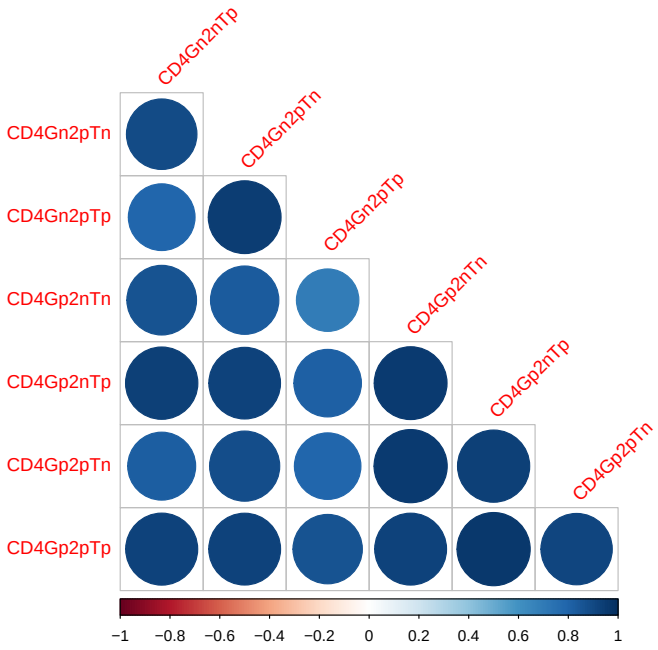


Figure 64: Plot of the significant correlations between the vaccine responses due to the occasion specific covariates for Model 13. Significance is determined from the 95% credible intervals; i.e. only the correlations with credible intervals excluding zero are plotted.

4

CONCLUDING REMARKS AND SUGGESTIONS FOR FURTHER WORK

There are currently no unanimously agreed criteria for identifying whether a candidate vaccine should be advanced through the clinical trials process or not (Voss *et al.*, 2018), and this can make it difficult for candidate vaccines to be promoted. The work of Rodo (2017) and Rodo *et al.* (2019a) aimed to provide a statistically rigorous means of comparing the immune responses for candidate vaccines to identify those which are the most promising and which should be advanced. Our aims were similar in that we wanted to demonstrate that statistical models could be used to capture the features of multivariate vaccine induced immune response data and identify the candidate vaccines that should be advanced in the trial process. The primary difference between our work and that of Rodo (2017) and Rodo *et al.* (2019a) was that we considered the longitudinal immune response profiles whereas Rodo (2017) and Rodo *et al.* (2019a) only considered the immune responses in the memory stage. A key advantage of our approach is that we are able to compare the modelled immune responses for the different vaccines over time and not just at the time points at which the immune responses were measured. This is useful when the immune responses are measured at different time points for different vaccines which is the case for the vaccine data we considered.

On a methodological front, which was our primary concern, in this dissertation we have shown that the M-GLMEMs are a promising class of models for modelling multivariate vaccine induced immune response data. We have shown that the M-GLMEMs perform considerably better than classical LMEMs and NLMEMs, and better than the univariate GLMEMs, which are often the preferred models in the context of modelling data with repeated measures and nonlinear profiles. We also demonstrated that the fitted M-GLMEMs are capable of capturing most of the features of the vaccine data that we considered. We showed that the models were able to account for

- the repeated measures for the subjects by the inclusion of subject specific random effects,
- the zero-inflated right skewed nature of the immune responses by using a Tweedie distribution, and

- the multivariate nature of the data.

The feature that the M-GLMEMs appeared to struggle with was the nonlinear profiles of the immune responses. The models were able to capture most of the nonlinear profile of the immune responses as seen by the relatively good fit of the models in Figures 60 and 107, and the clearly nonlinear profiles of the fitted marginal expected immune response profiles in Figures 61 and 108, but there were still patterns in the residuals when plotted against time. It is believed that this deficiency of the fitted M-GLMEMs is likely to be mediated by having data with more regular measurement time points and an increased number of subjects.

On a clinical front, from the fitted M-GLMEMs, we were also able to draw conclusions about the vaccines that were included in the data set that we modelled. The primary conclusions are:

- that there is limited diversity in the induced immune responses of the vaccines and
- that Vaccine 3, the M72/AS01$_E$ vaccine, is the most promising novel TB vaccine of the novel vaccines that we considered.

These conclusions are consistent with those of Rodo (2017) and Rodo *et al.* (2019a). Additional insights that we are able to draw from the M-GLMEMs that we fitted are

- that the covariates that we included in the M-GLMEMs, the vaccine and time point covariates, explain a small amount of the total covariation between the immune responses. They explain only 27.2% of the total covariation. This is less than we expected and may suggest there are other important unmeasured covariates.
- the residual correlation between the immune responses not explained by the occasion specific covariates is positive and strong for all responses. This is probably to be expected given the covariates captured a small proportion of the total covariation between the immune responses and that the vaccines are likely designed to induce responses for the cytokines included.

The ability of the M-GLMEMs to capture most of the features of the multivariate immune response data we considered and the ability for us to draw meaningful clinical conclusions from the fitted models, indicates that M-GLMEMs may be useful in applications to other vaccine data. However, there are several limitations to the work that we have carried out here. The main limitation is that we chose to only model the immune responses from the vaccines that were the most amenable to being modelled and we preprocessed the data significantly before attempting to model it. This may mean that the M-GLMEMs, which have been less widely investigated in the literature,

are not very robust and may struggle in further applications to other vaccine data. In addition, the preprocessing we carried out was not clinically informed and so may have significantly affected the results and conclusions that we have presented here. This must be borne in mind when considering the work here.

There are several areas for possible future work which are suggested by the work here.

- In future work, it would be important to involve a clinician in the data preprocessing. This would help to ensure that the results and conclusions drawn for the fitted models are clinically valid. In addition, involving a clinician may also help us to include more of the vaccine data and possibly include more of the vaccines in the applications.

- Clinical interpretation of the models and their parameter estimates may also be useful. Involving a clinician in future work may allow us to elicit such interpretations.

- It would also be interesting in future vaccine trials to identify, measure, and include additional covariates and immunologically relevant traits for the responses. The additional covariates may help to explain more of the covariation of the immune responses and the traits may help with our understanding of vaccine induced immune responses.

- We have used cubic B-splines to capture the nonlinear profiles in the M-GLMEMs. In future work it may be worthwhile to consider alternative means of capturing the nonlinear profiles, e.g. it may be worthwhile to consider alternative types of splines or possibly nonlinear functions.

- The NLMEMs that we fitted appeared to capture the nonlinear profiles of the immune responses well, but as they had normally distributed within-group errors, they did not perform well. In future work, we could also investigate multivariate NLMEMs with non-normal within-group errors. Such models may perform well relative to the M-GLMEMs we have considered here.

- In future work, more attention should be given to the Bayesian MCMC estimation of the M-GLMEMs. We have used the default prior distributions and thinning rates from the `boral` package. Sensitivity tests could be used to assess the impact of these choices. In addition, future work should use more formal testing, such as the Geweke convergence diagnostic (Geweke, 1991), to assess the convergence of the MCMC chains.

A

IMMUNE RESPONSE PROFILES

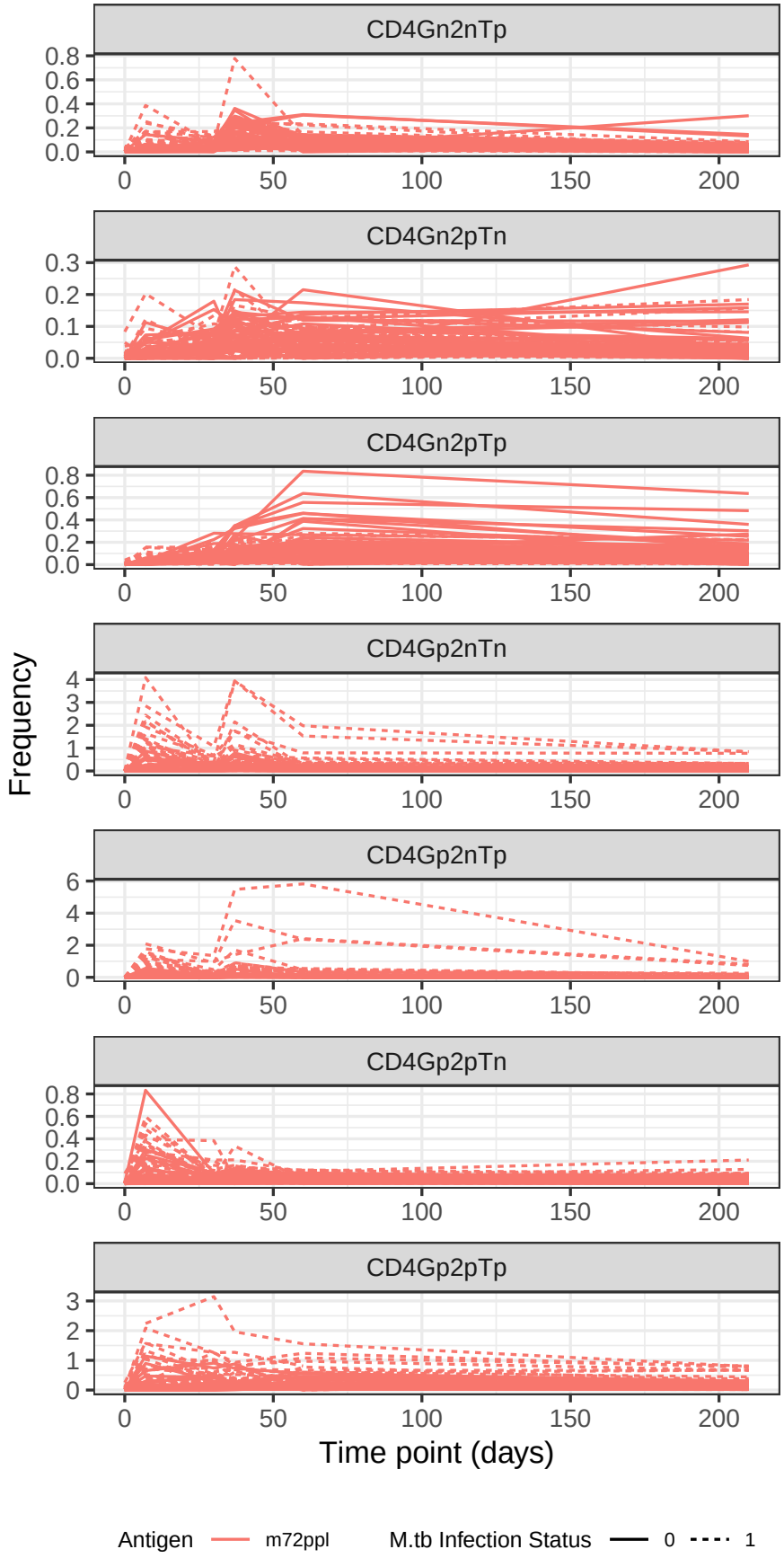


Figure 65: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 3.

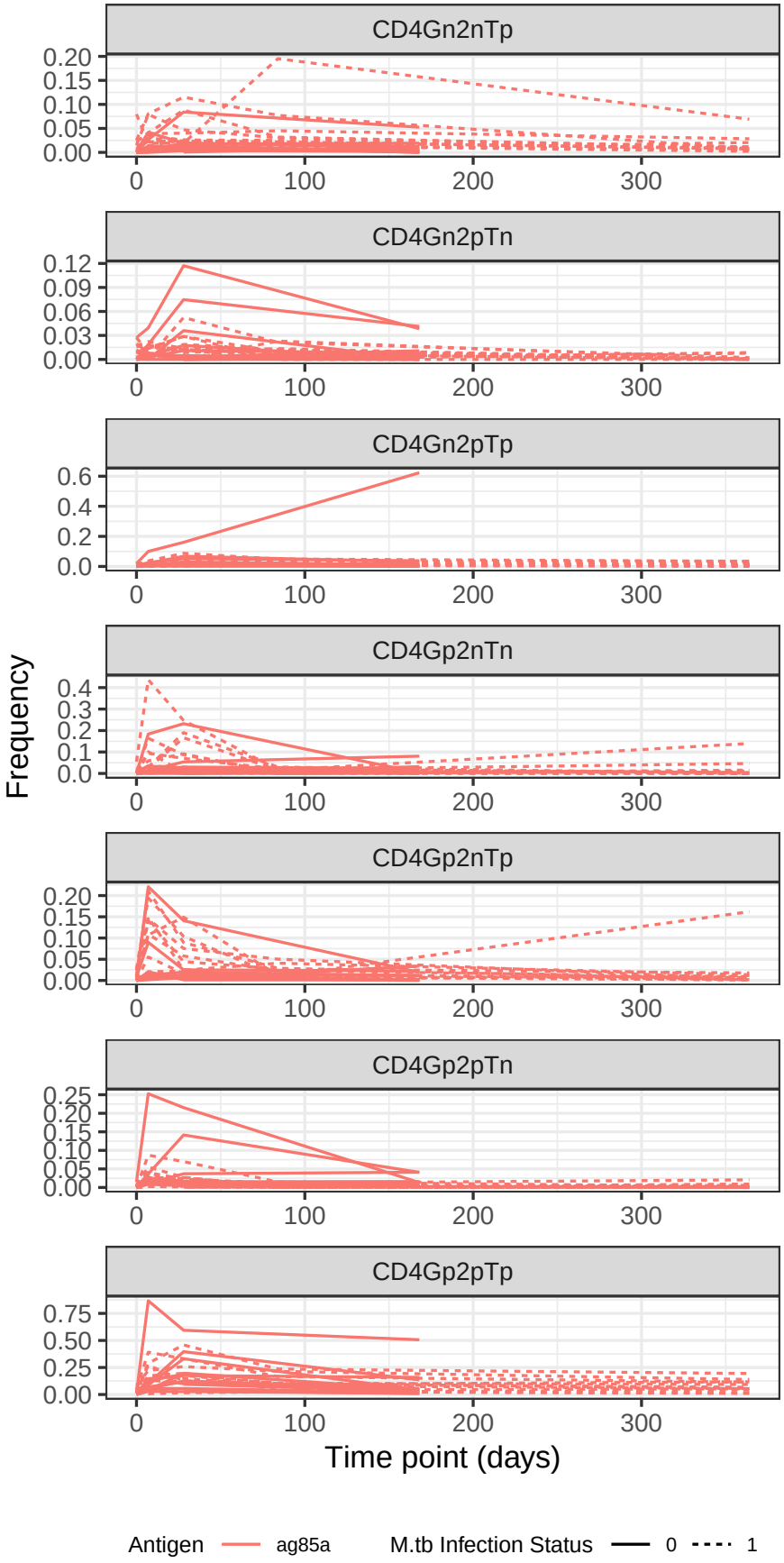


Figure 66: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 4.

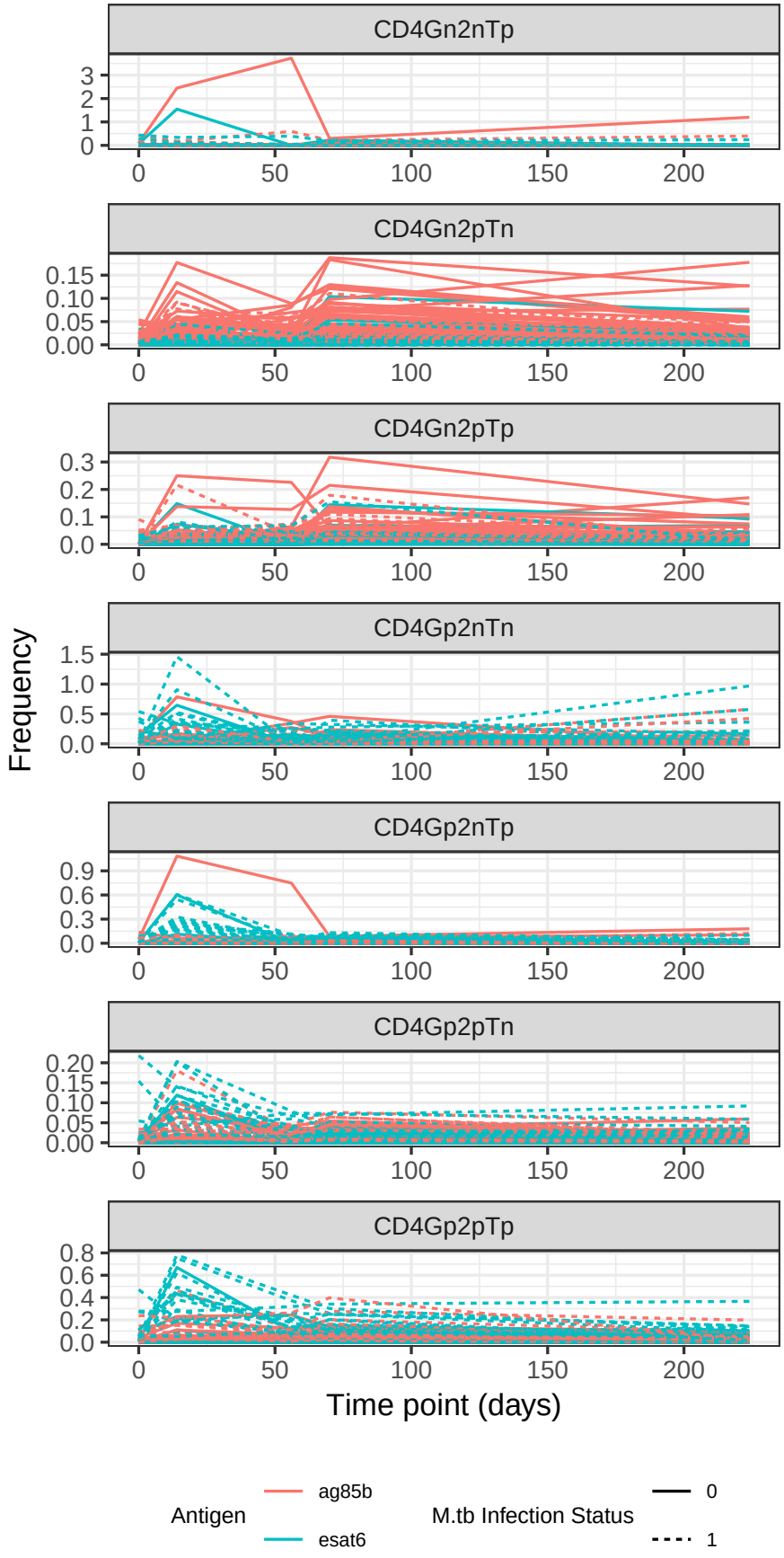


Figure 67: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 5.

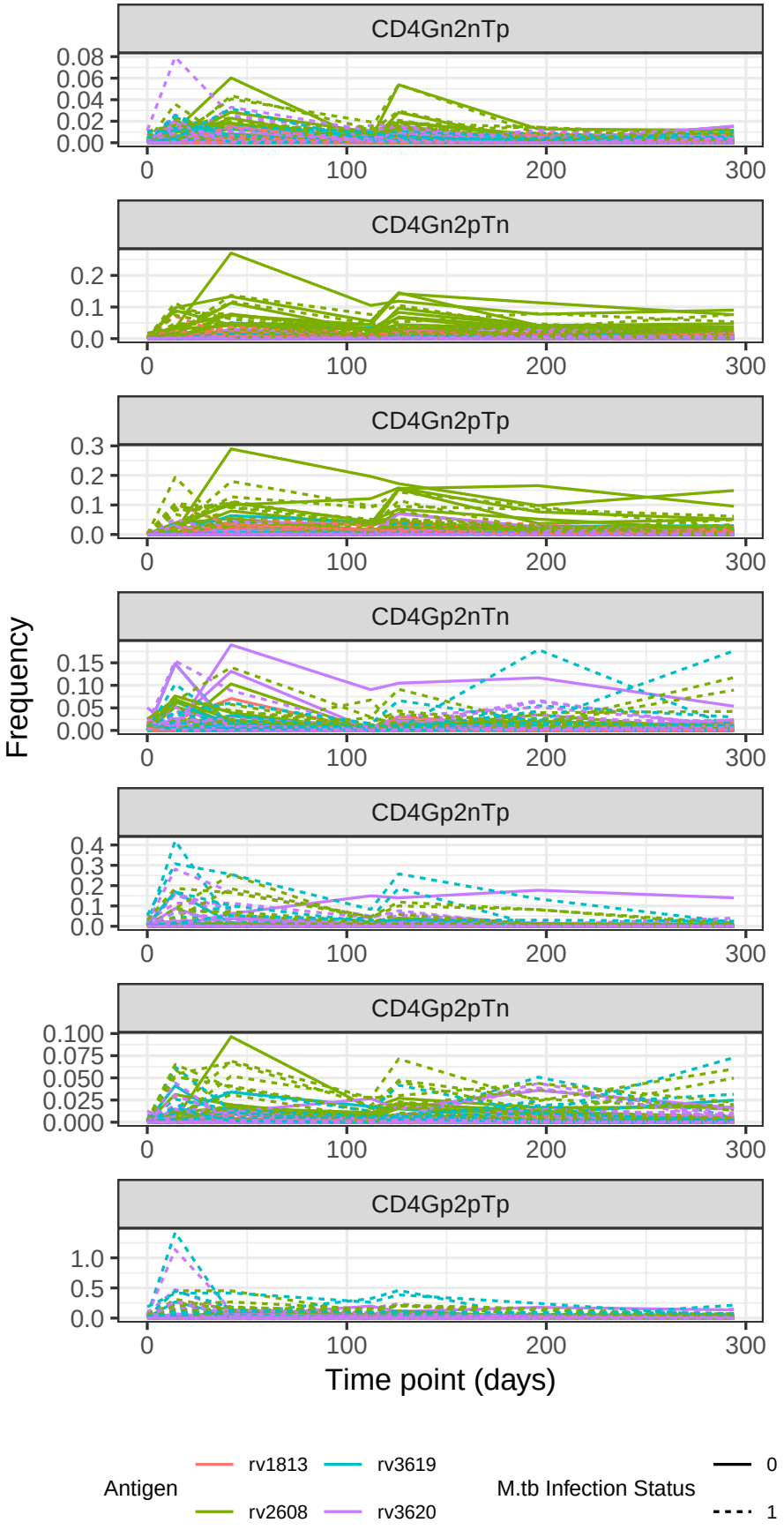


Figure 68: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 6.

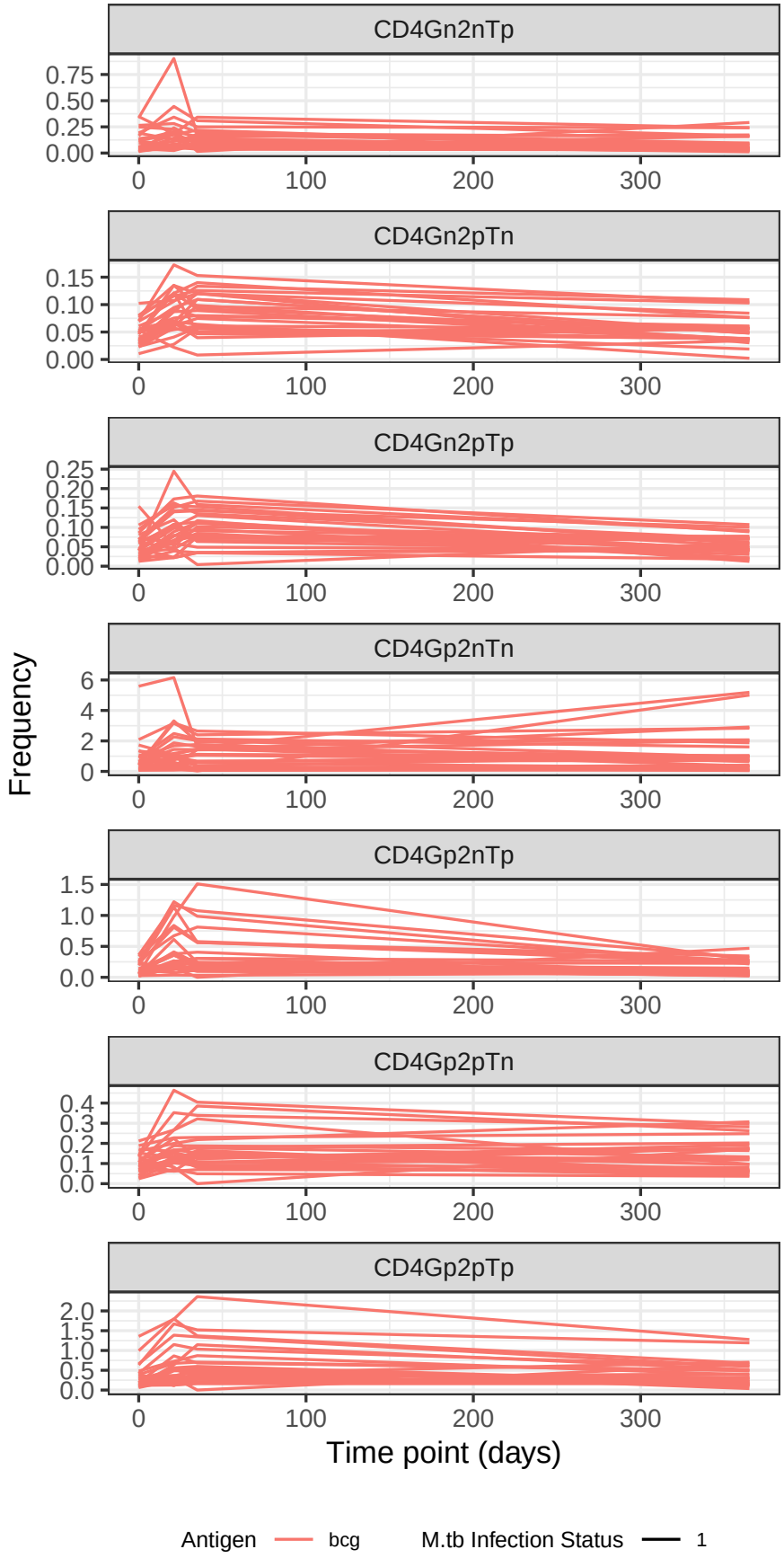


Figure 69: Plots of the antigen-specific frequencies for the CD4 T cells expressing the seven different cytokine combinations for Vaccine 7.

B

LINEAR MIXED EFFECT MODELS

B.1 LINEAR MIXED EFFECT MODELS WITH ORTHOGONAL POLYNOMIALS

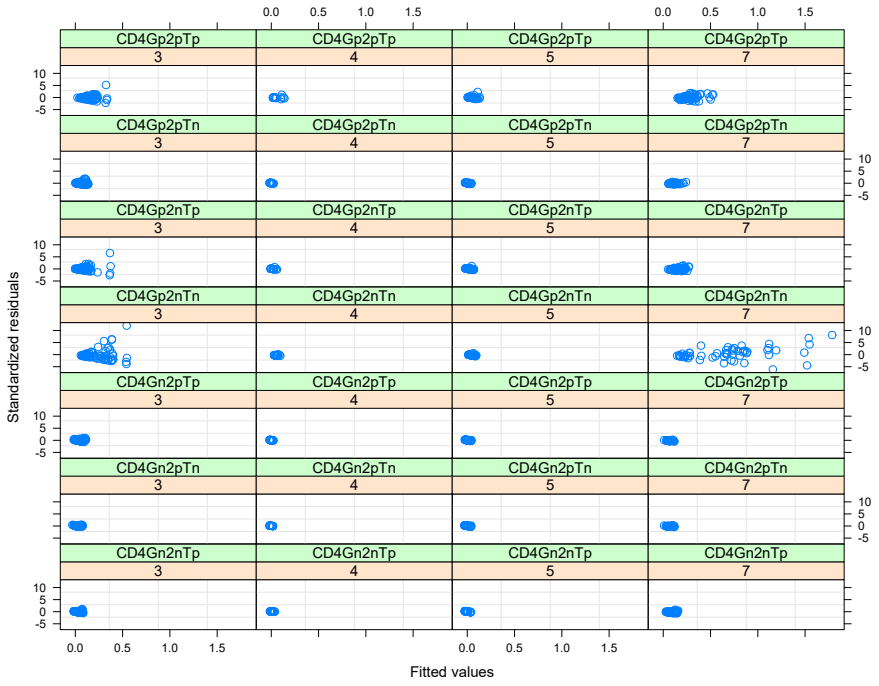


Figure 70: Standardised residuals versus fitted values by vaccine and response for Model 2, a LMEM with a 3 degree polynomial.

B.2 LINEAR MIXED EFFECT MODELS WITH CUBIC B-SPLINES

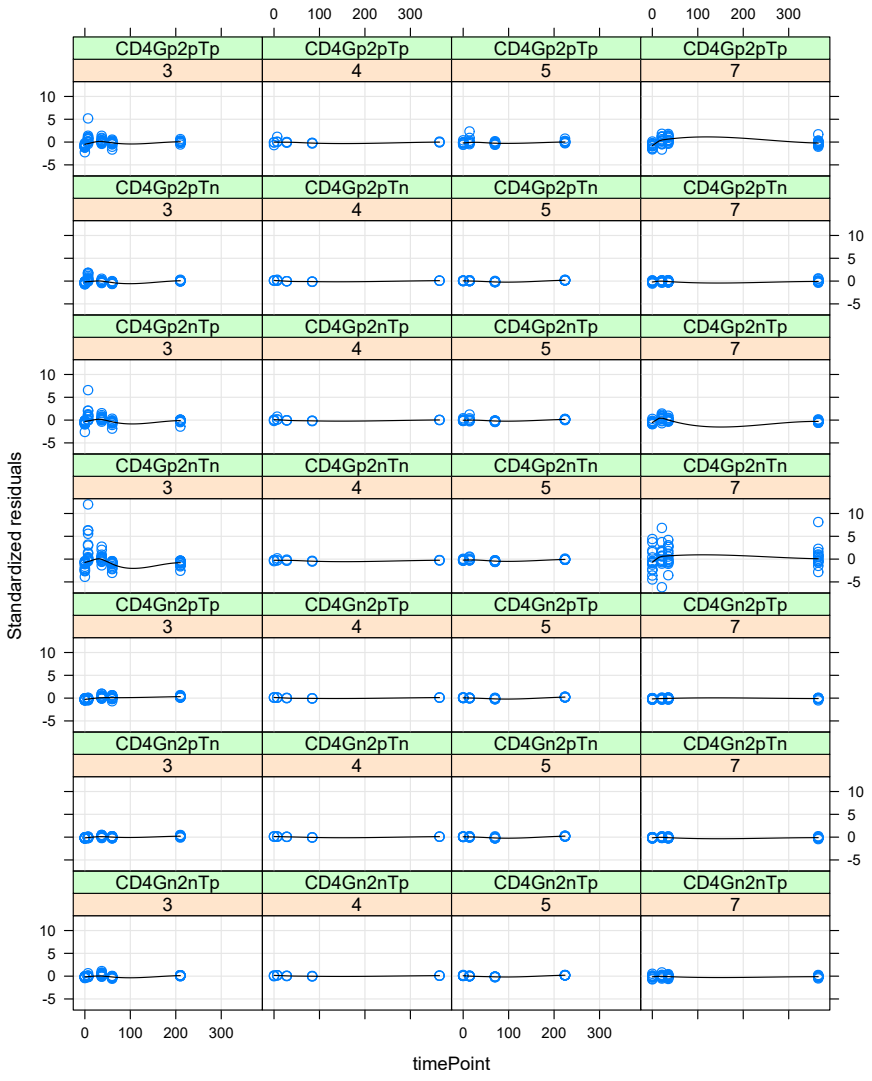


Figure 71: Standardised residuals versus time point by vaccine and response for Model 2, a LMEM with a 3 degree polynomial.

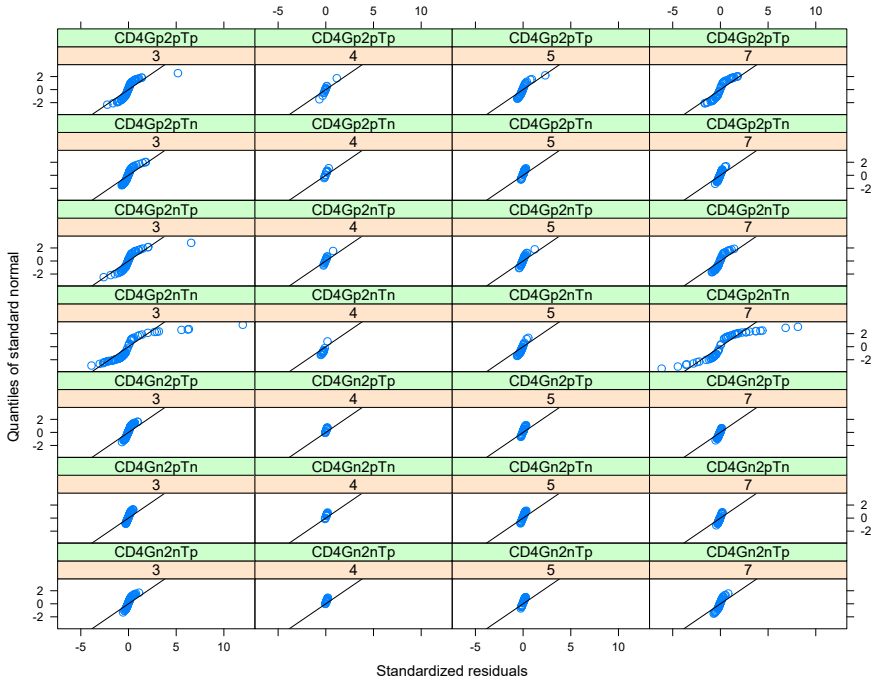


Figure 72: QQ-plot for the standardised residuals by vaccine and response for Model 2, a LMEM with a 3 degree polynomial.

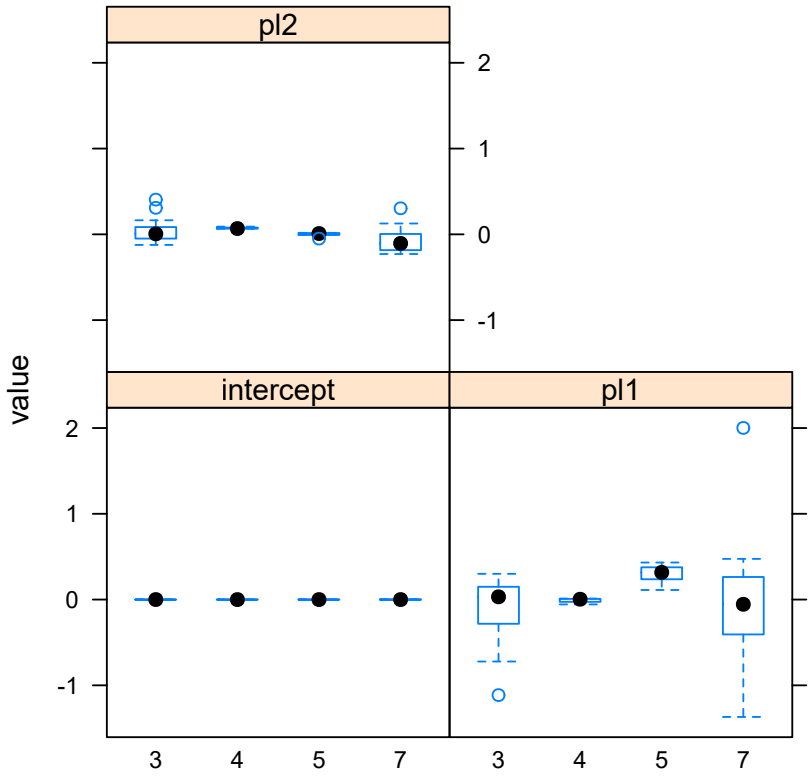


Figure 73: subject specific random effect estimates $\widehat{b_{0i}}$, $\widehat{b_{1i}}$ and $\widehat{b_{2i}}$ for Model 2, a LMEM with a 3 degree polynomial.

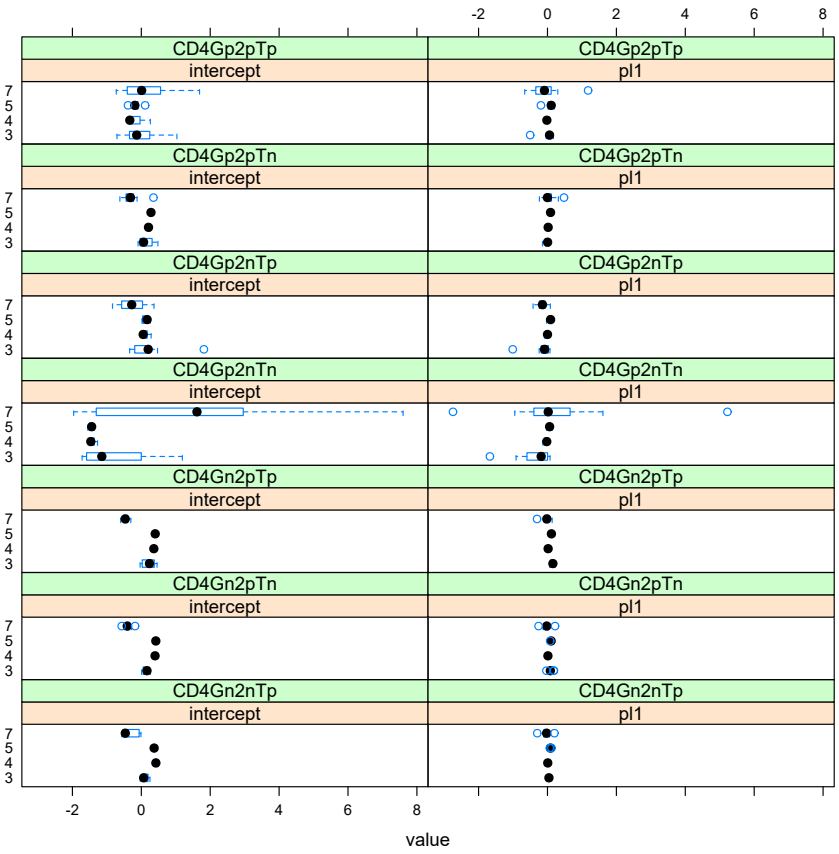


Figure 74: Response within subject random effect estimates $\widehat{b_{0i,j}}$ and $\widehat{b_{1i,j}}$ for Model 2, a LMEM with a 3 degree polynomial.

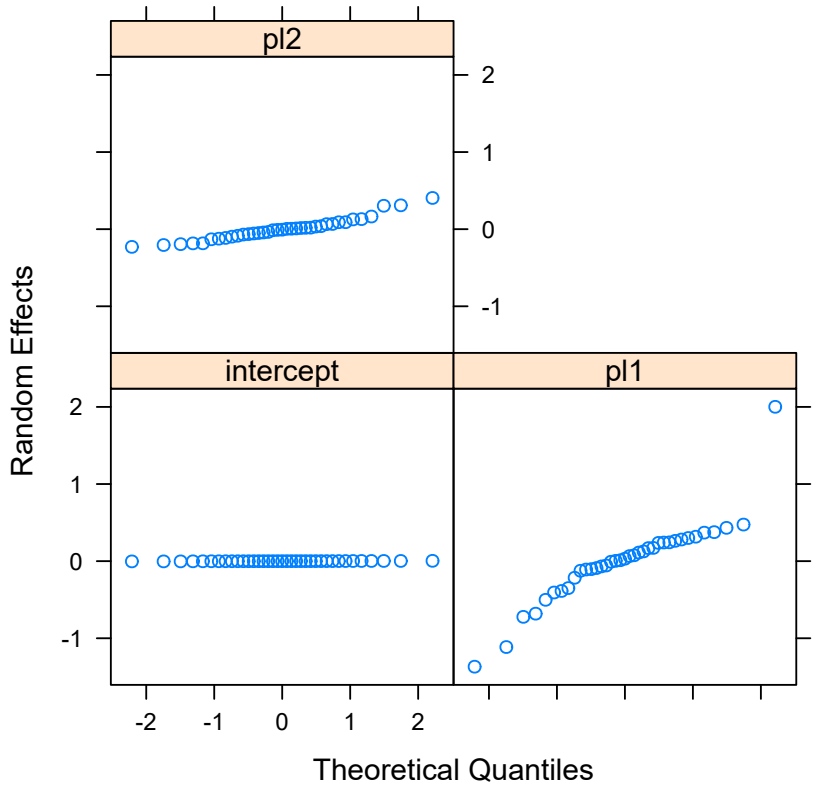


Figure 75: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i}}, \widehat{b_{1i}}$ and $\widehat{b_{2i}}$ for Model 2, a LMEM with a 3 degree polynomial.

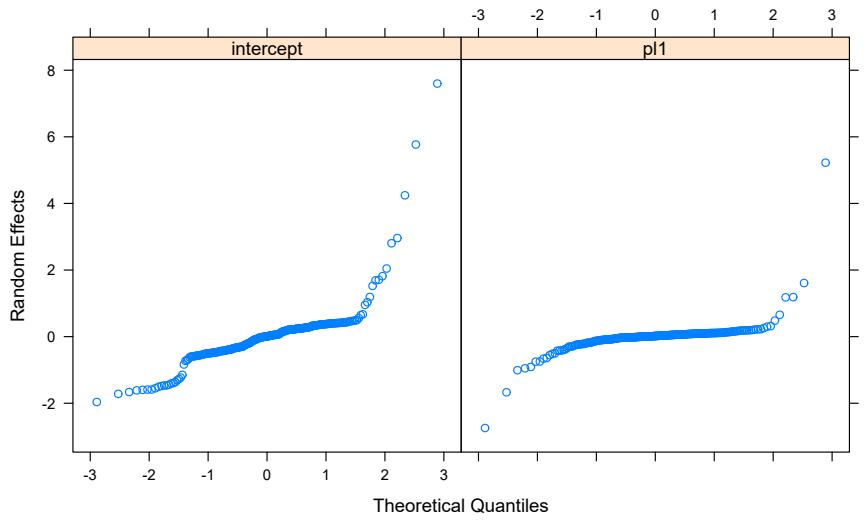


Figure 76: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i,j}}$ and $\widehat{b_{1i,j}}$ for Model 2, a LMEM with a 3 degree polynomial.

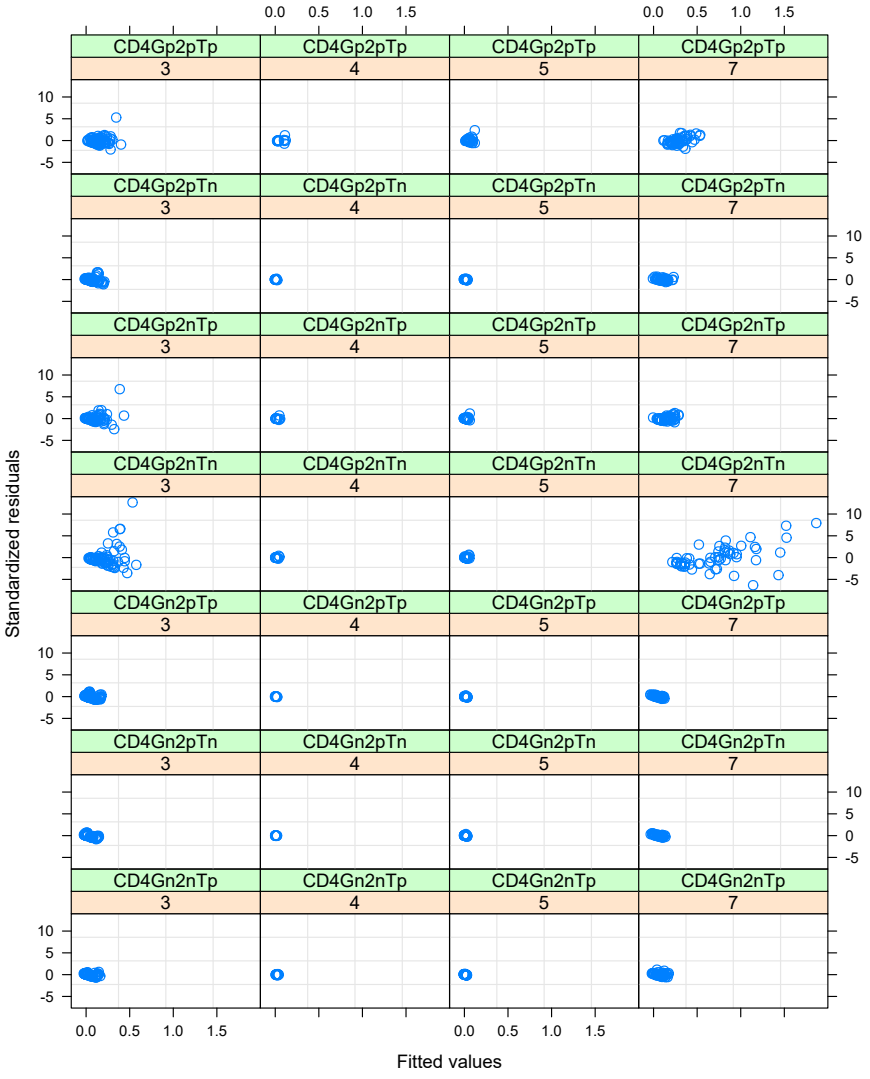


Figure 77: Standardised residuals versus fitted values by vaccine and response for Model 3, a LMEM with a 3 degree polynomial.

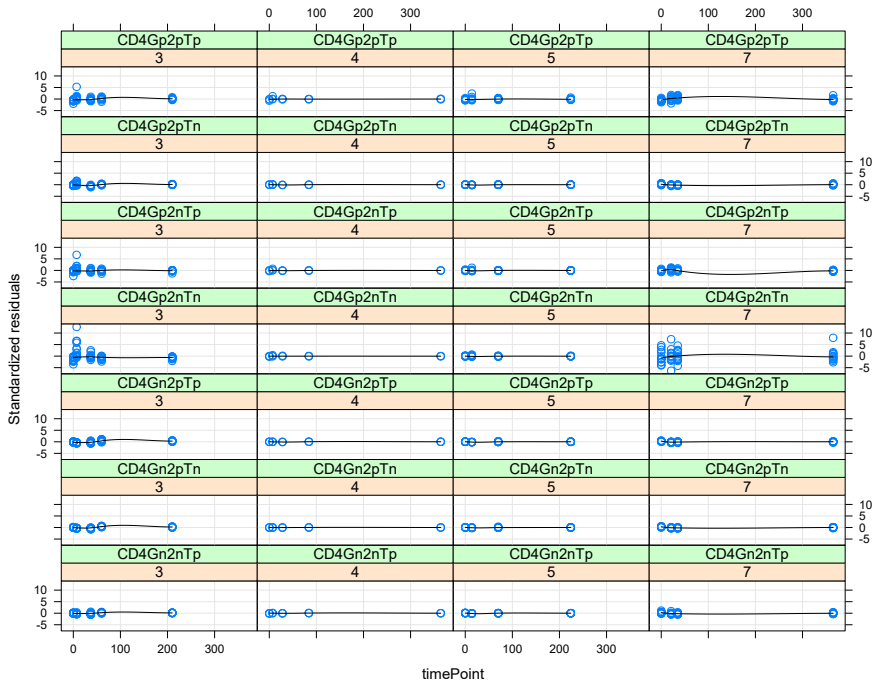


Figure 78: Standardised residuals versus time point by vaccine and response for Model 3, a LMEM with a 3 degree polynomial.

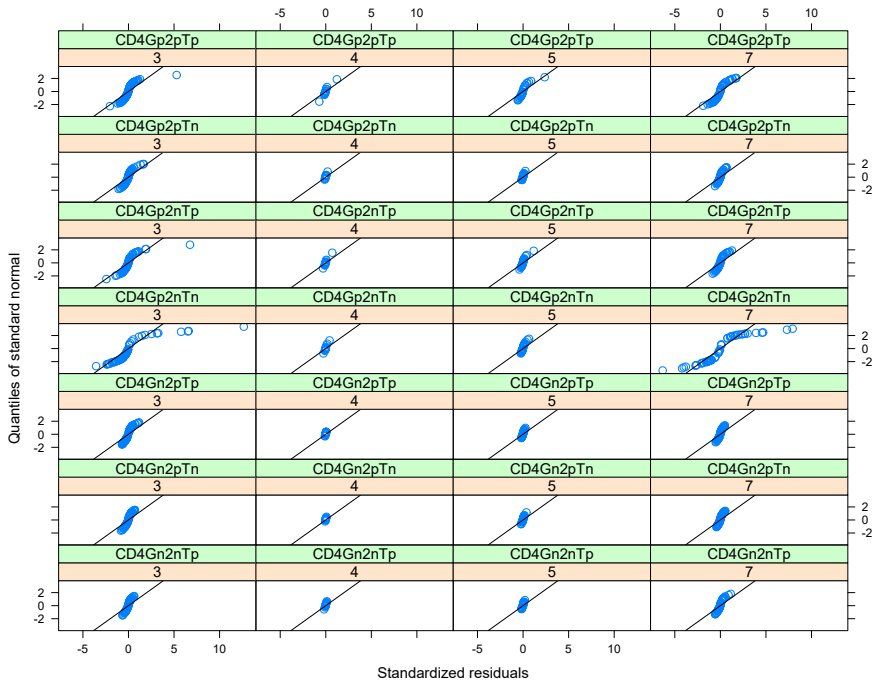


Figure 79: QQ-plot for the standardised residuals by vaccine and response for Model 3, a LMEM with a 3 degree polynomial.

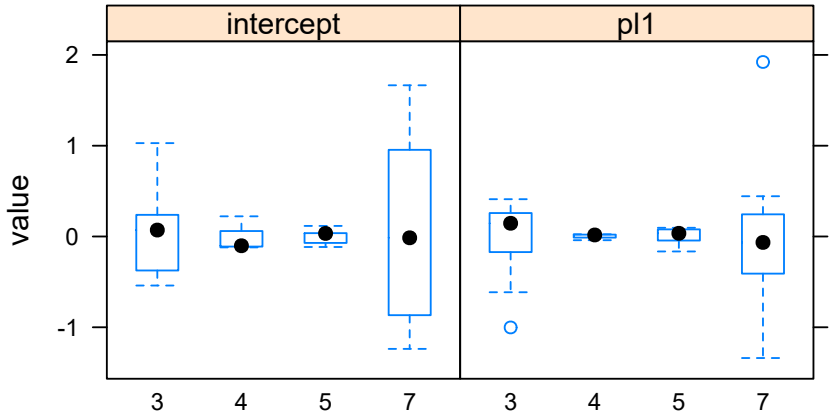


Figure 80: Subject specific random effect estimates $\widehat{b}_{0i}$ and $\widehat{b}_{1i}$ for Model 3, a LMEM with a 3 degree polynomial.

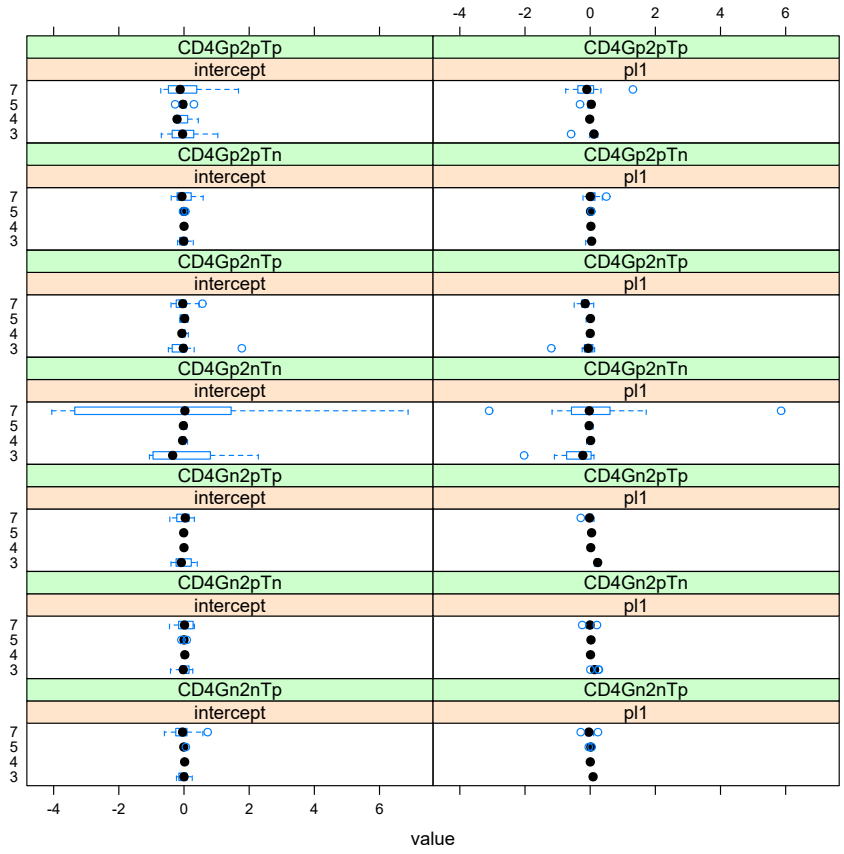


Figure 81: Response within subject random effect estimates $\widehat{b}_{0i,j}$ and $\widehat{b}_{1i,j}$ for Model 3, a LMEM with a 3 degree polynomial.

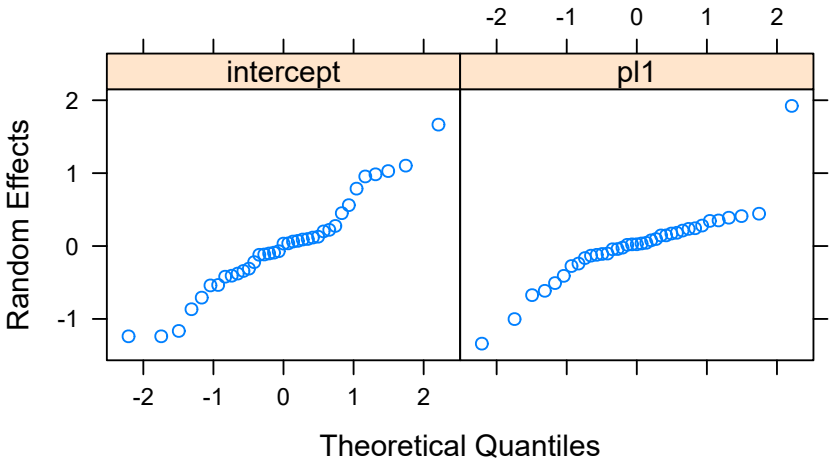


Figure 82: QQ-plot for the subject specific random effect estimates $\widehat{b}_{0i}$ and $\widehat{b}_{1i}$ for Model 3, a LMEM with a 3 degree polynomial.

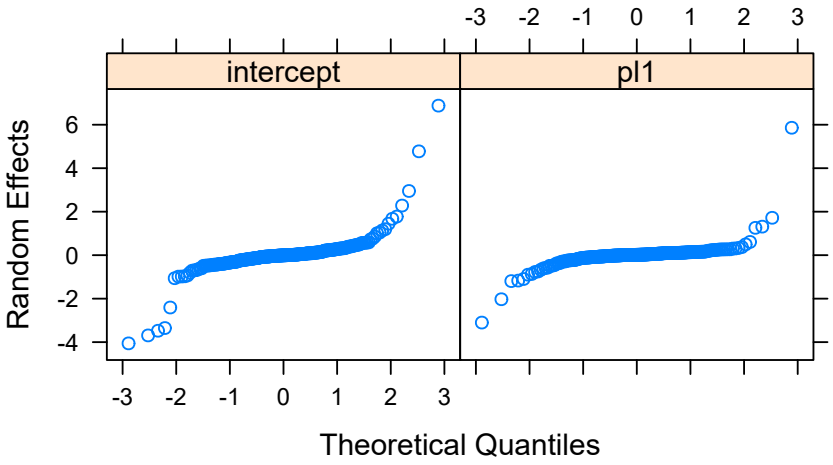


Figure 83: QQ-plot for the subject specific random effect estimates $\widehat{b}_{0i,j}$ and $\widehat{b}_{1i,j}$ for Model 3, a LMEM with a 3 degree polynomial.

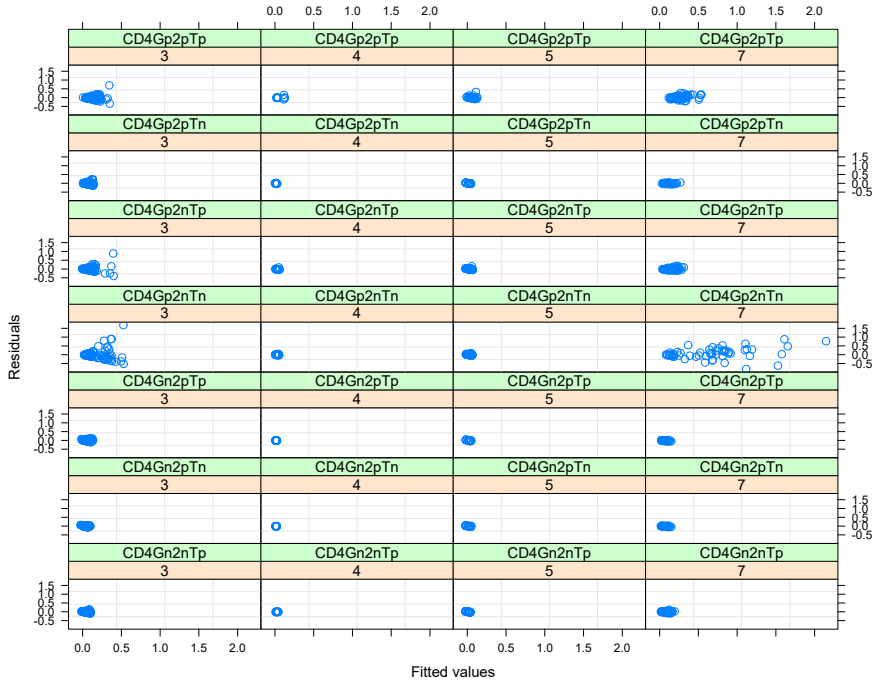


Figure 84: Standardised residuals versus fitted values by vaccine and response for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

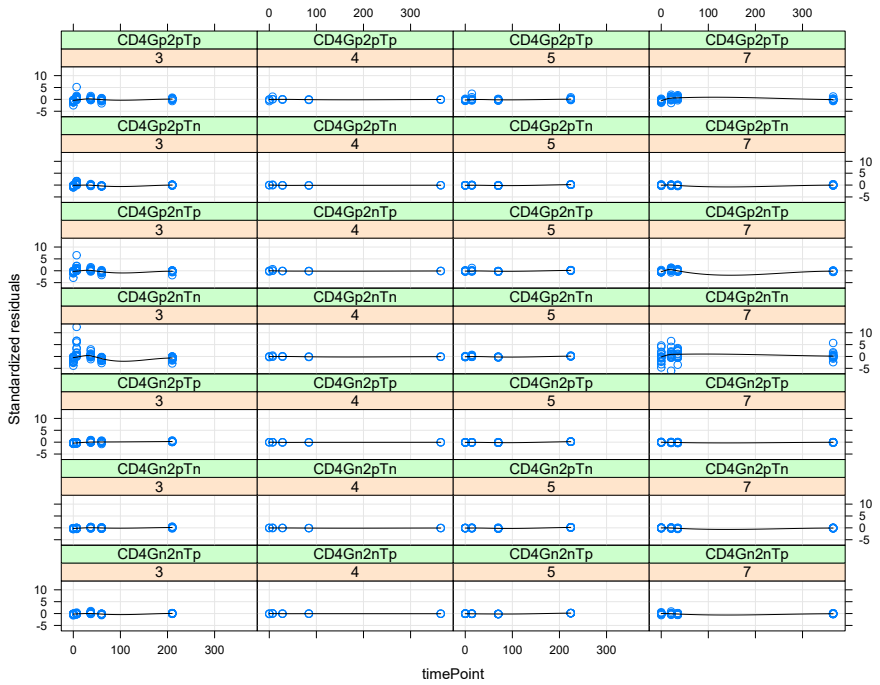


Figure 85: Standardised residuals versus time point by vaccine and response for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

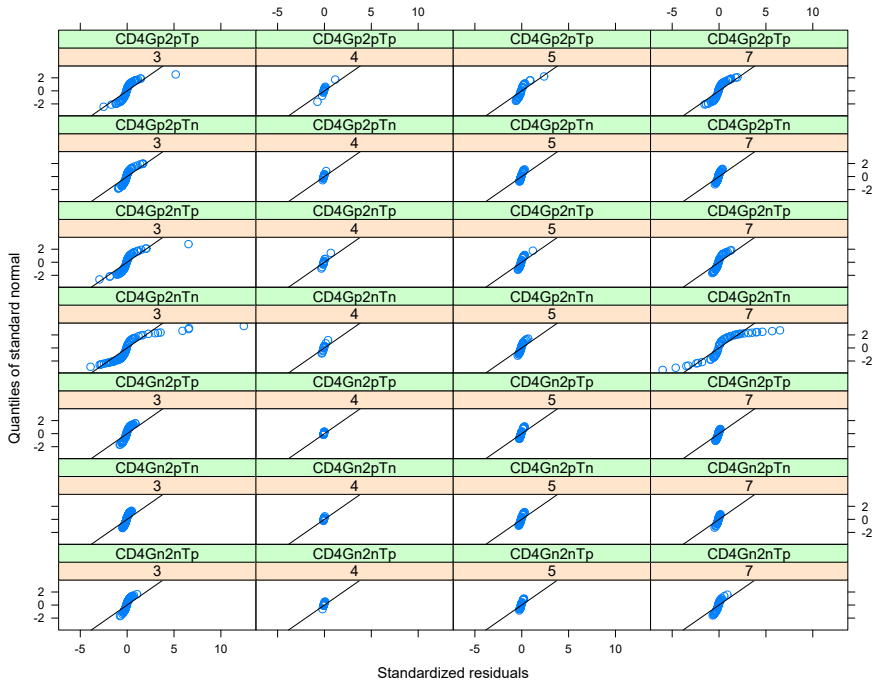


Figure 86: QQ-plot for the standardised residuals by vaccine and response for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

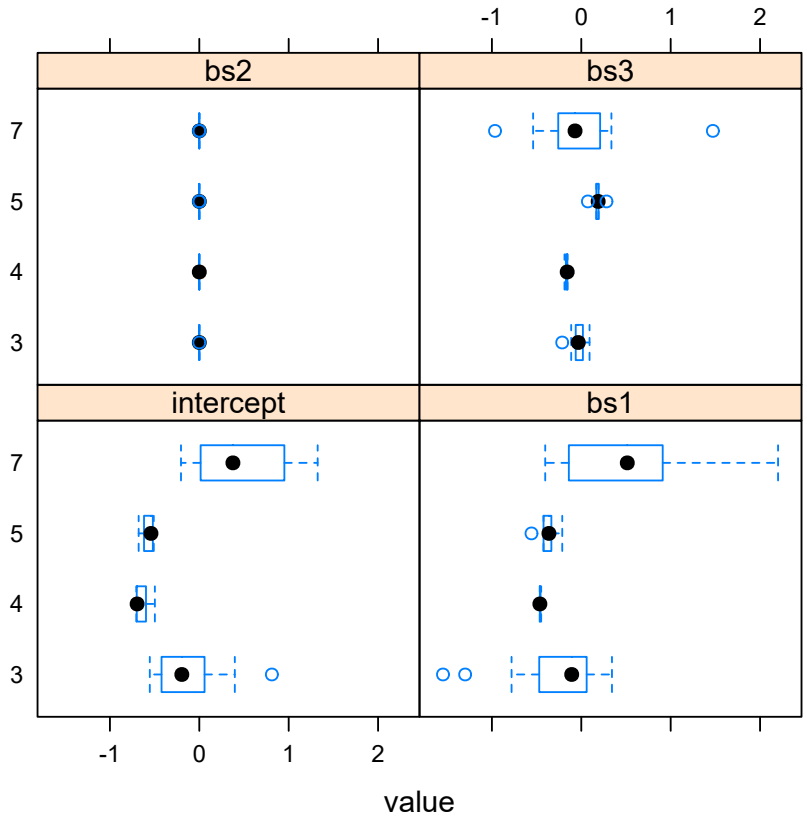


Figure 87: subject specific random effect estimates $\widehat{b_{0i}}, \widehat{b_{1i}}, \widehat{b_{2i}}$ and $\widehat{b_{3i}}$ for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

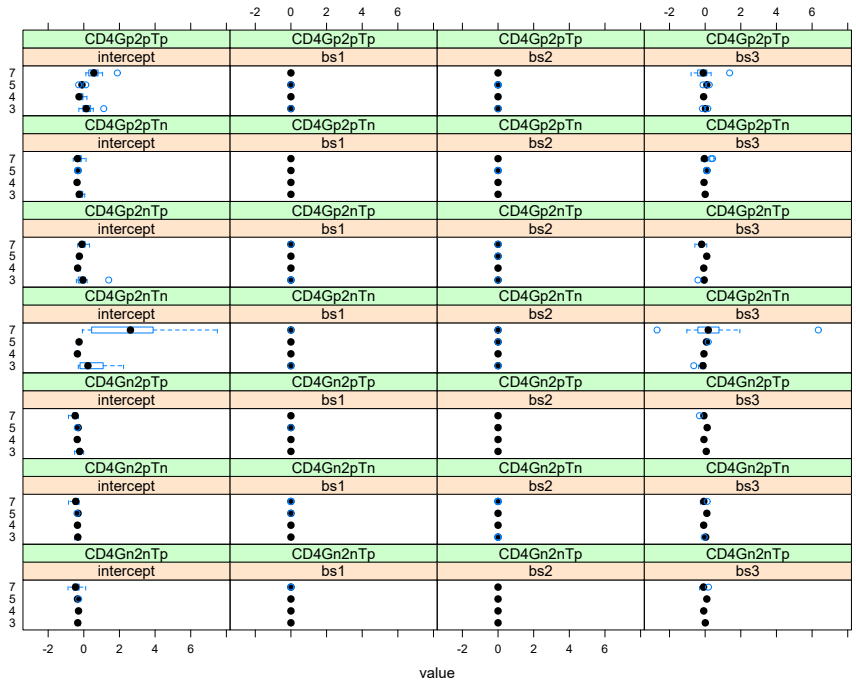


Figure 88: Response within subject random effect estimates $\widehat{b_{0i,j}}$, $\widehat{b_{1i,j}}$, $\widehat{b_{2i,j}}$ and $\widehat{b_{3i,j}}$ for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

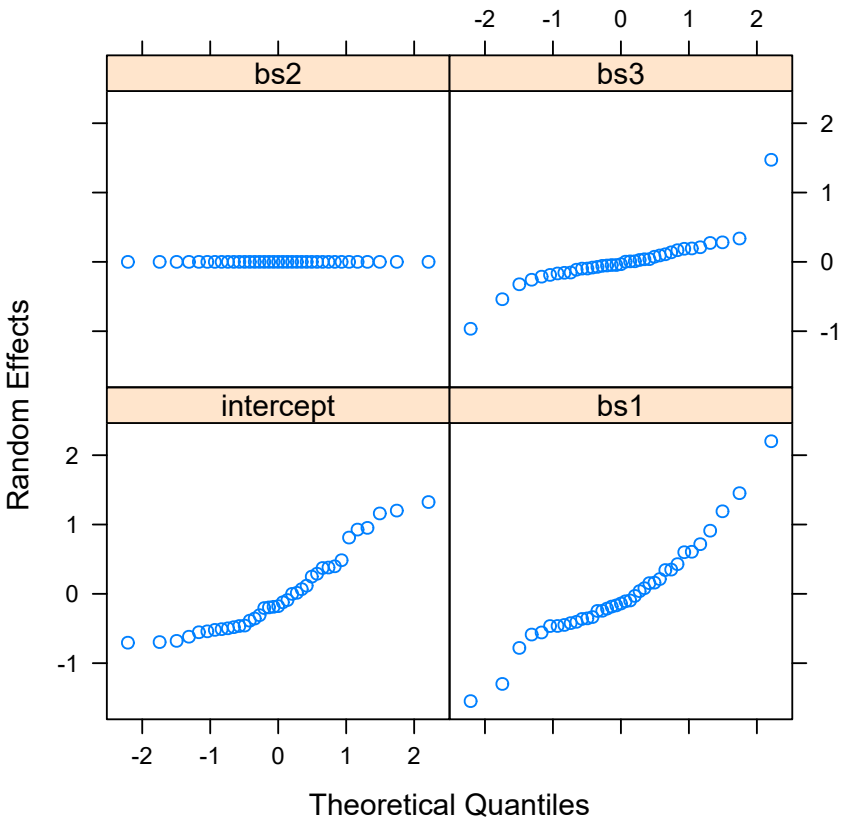


Figure 89: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i}}, \widehat{b_{1i}}, \widehat{b_{2i}}$ and $\widehat{b_{3i}}$ for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

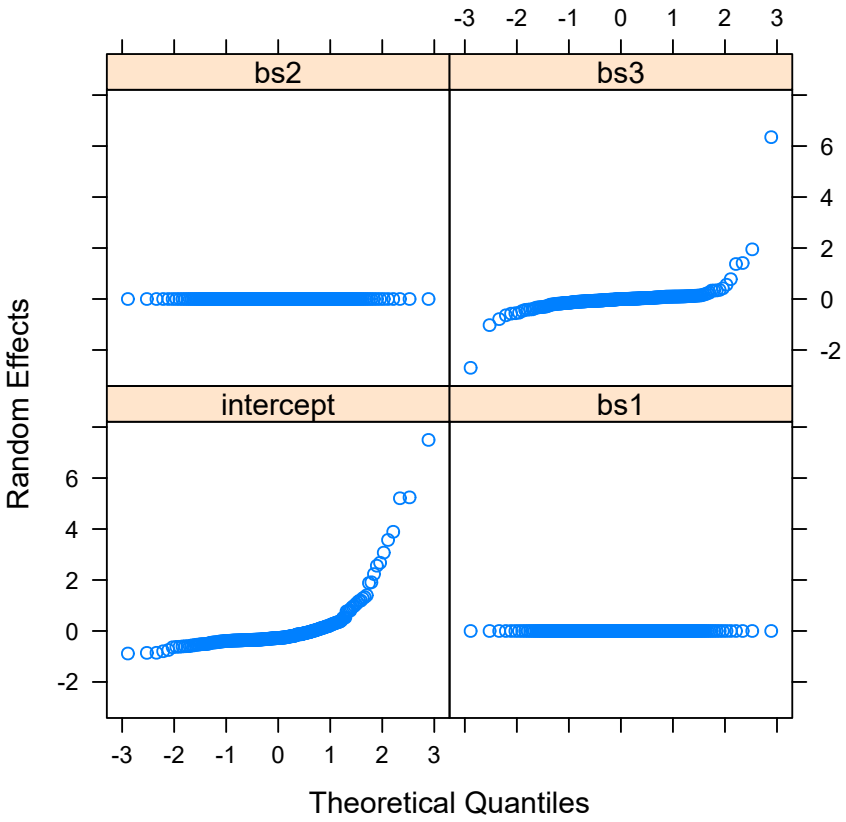


Figure 90: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i,j}}$, $\widehat{b_{1i,j}}$, $\widehat{b_{2i,j}}$ and $\widehat{b_{3i,j}}$ for Model 5, a LMEM with a cubic B-spline with 3 degrees of freedom.

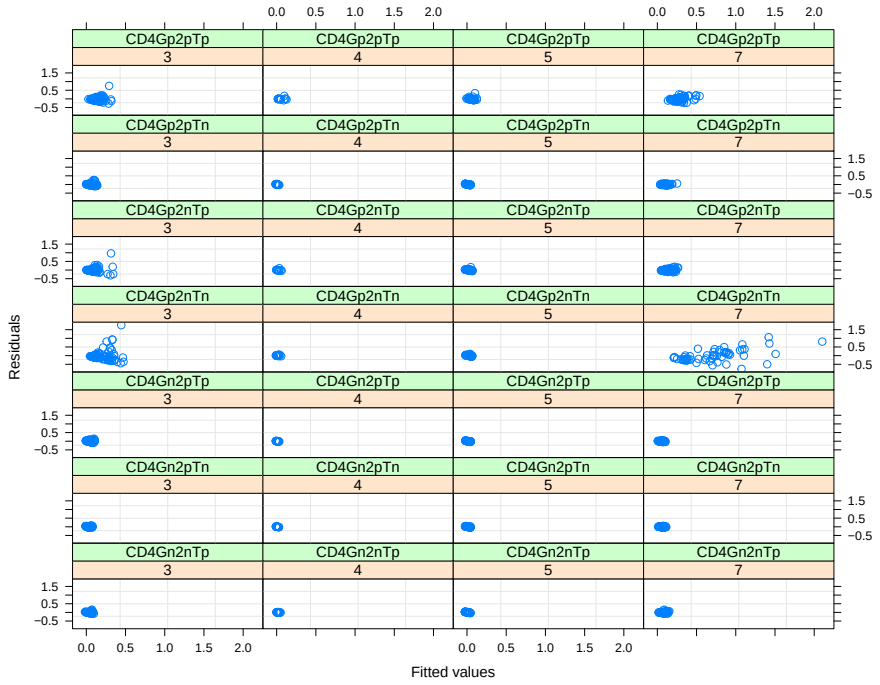


Figure 91: Standardised residuals versus fitted values by vaccine and response for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

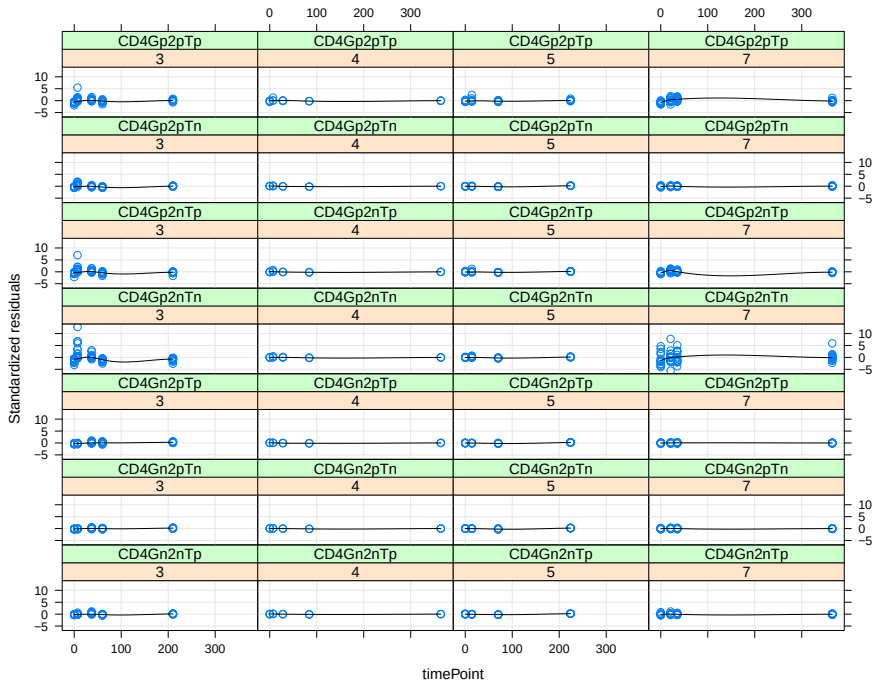


Figure 92: Standardised residuals versus time point by vaccine and response for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

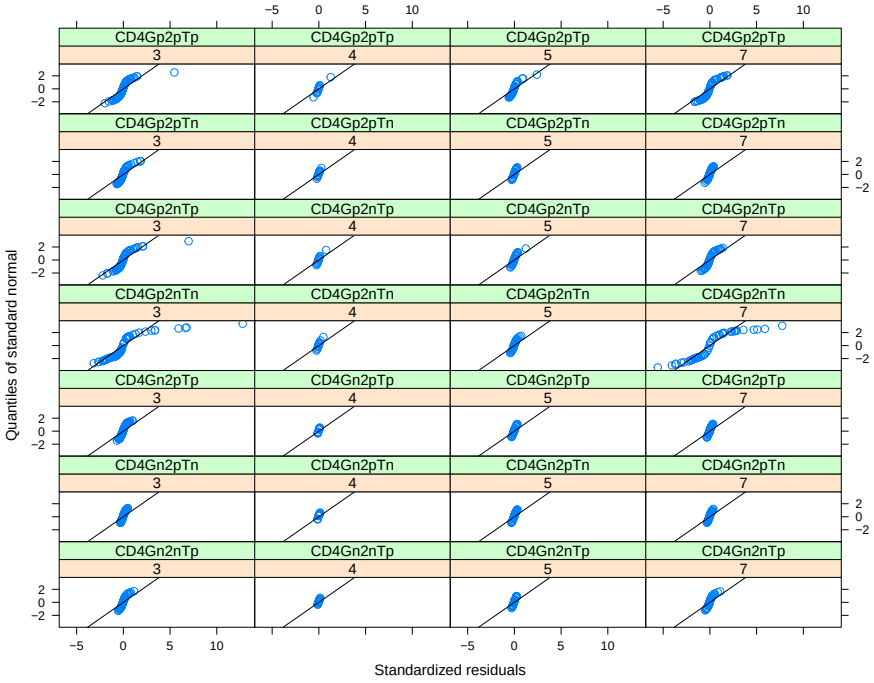


Figure 93: QQ-plot for the standardised residuals by vaccine and response for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

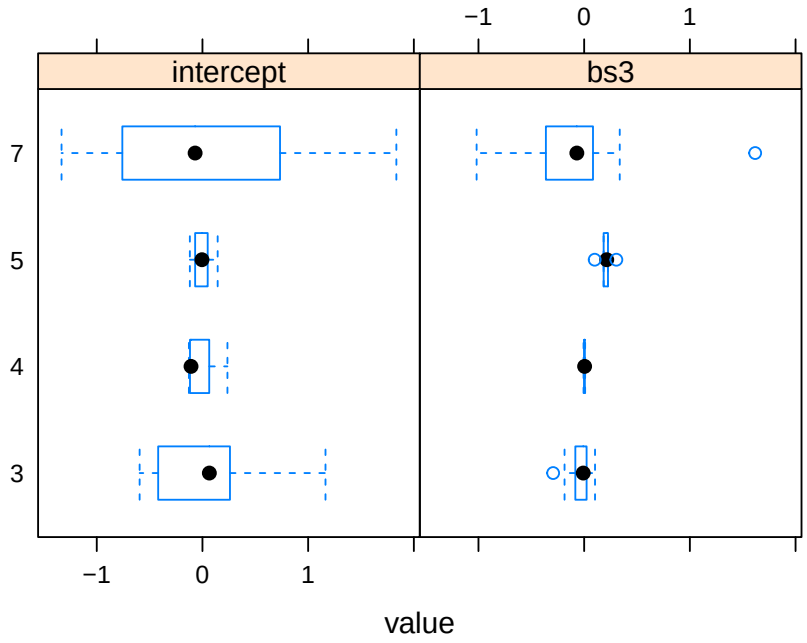


Figure 94: Subject specific random effect estimates $\widehat{b_{0i}}$ for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

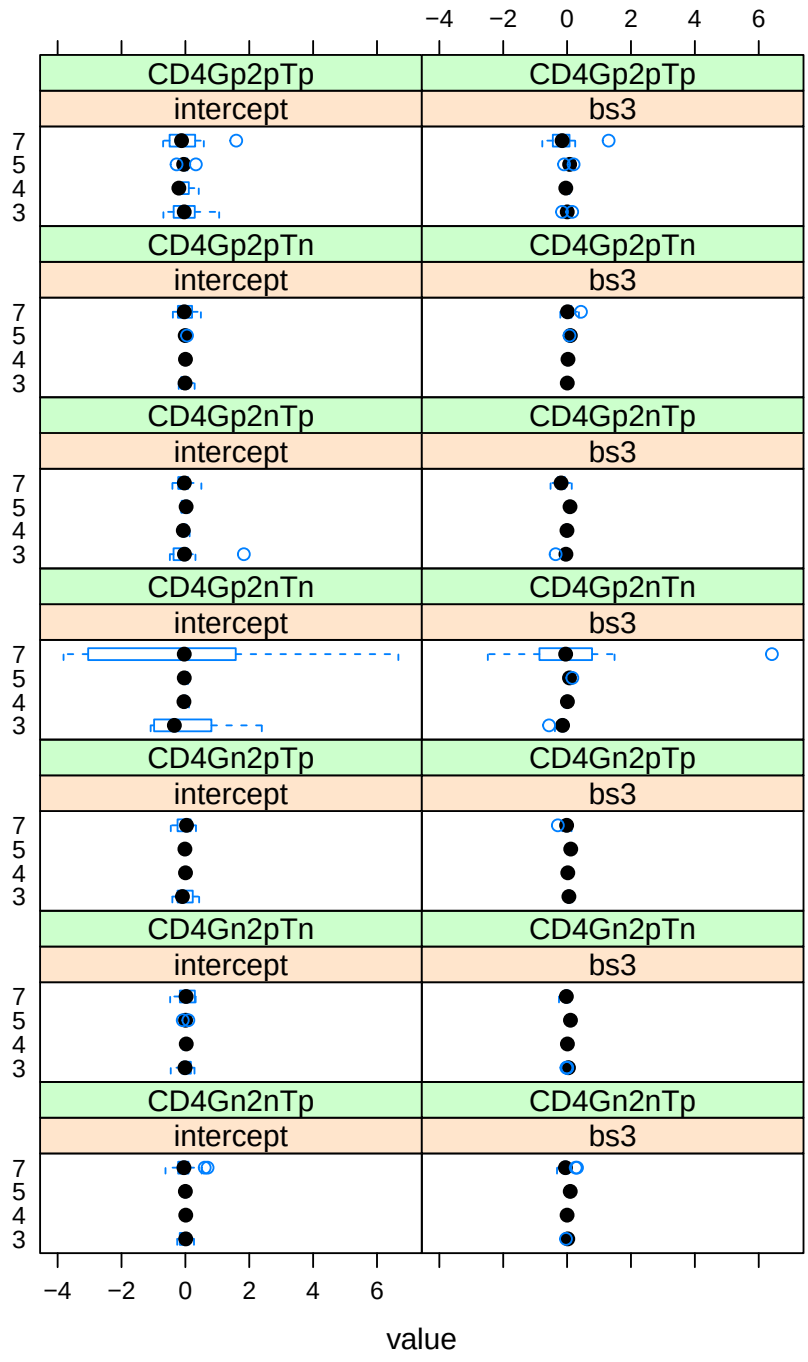


Figure 95: Response within subject random effect estimates $\widehat{b_{0i,j}}$ and $\widehat{b_{3i,j}}$ for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

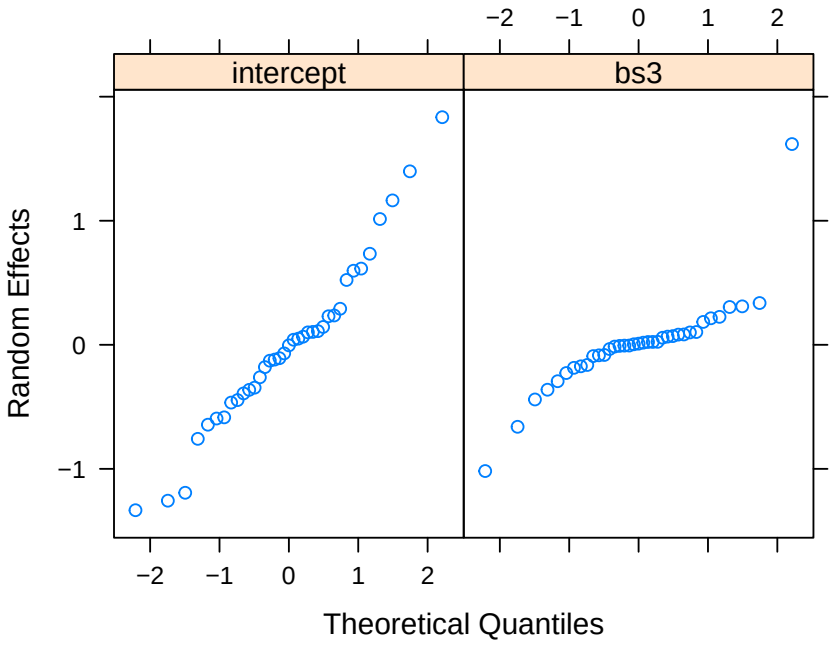


Figure 96: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i}}$ for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

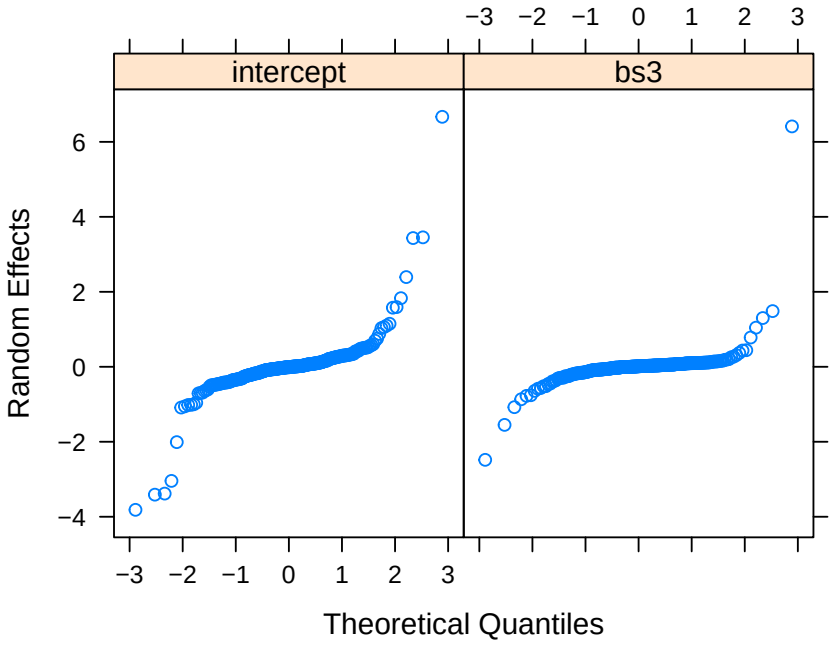


Figure 97: QQ-plot for the subject specific random effect estimates $\widehat{b_{0i,j}}$ and $\widehat{b_{3i,j}}$ for Model 6, a LMEM with a cubic B-spline with 3 degrees of freedom.

C

NON-LINEAR MIXED EFFECT MODELS

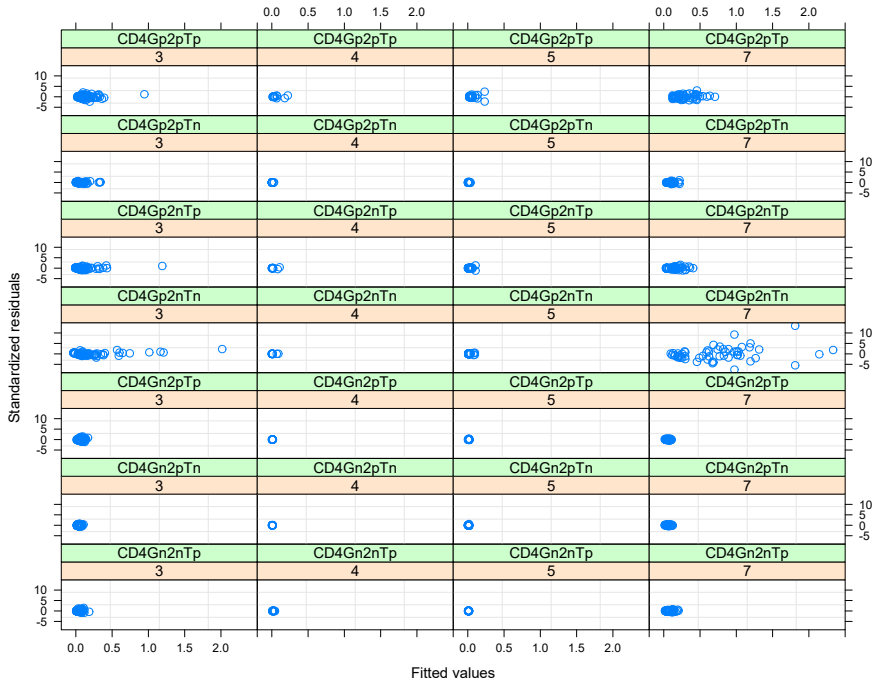


Figure 98: Standardised residuals versus fitted values by vaccine and response for Model 9, a NLMEM.

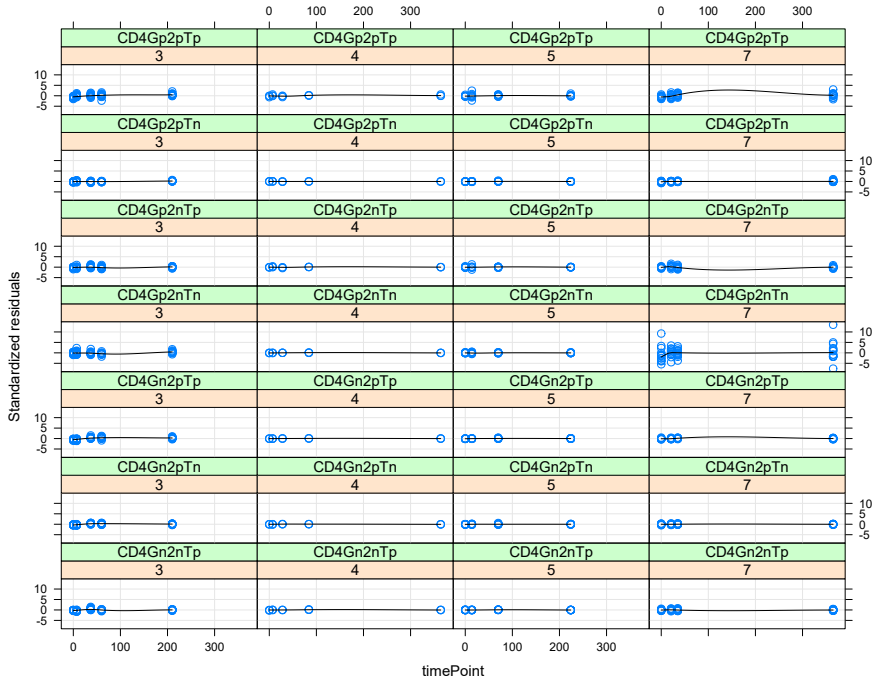


Figure 99: Standardised residuals versus time point by vaccine and response for Model 9, a NLMEM.

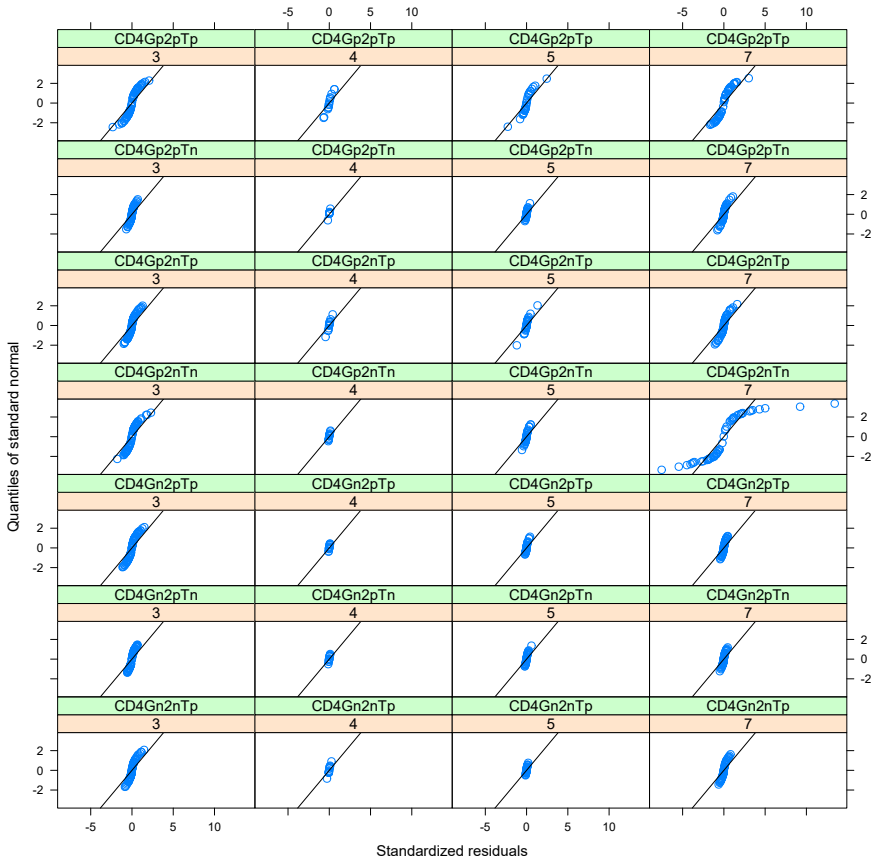


Figure 100: QQ-plots for the standardised residuals by vaccine and response for Model 9, a NLMEM.

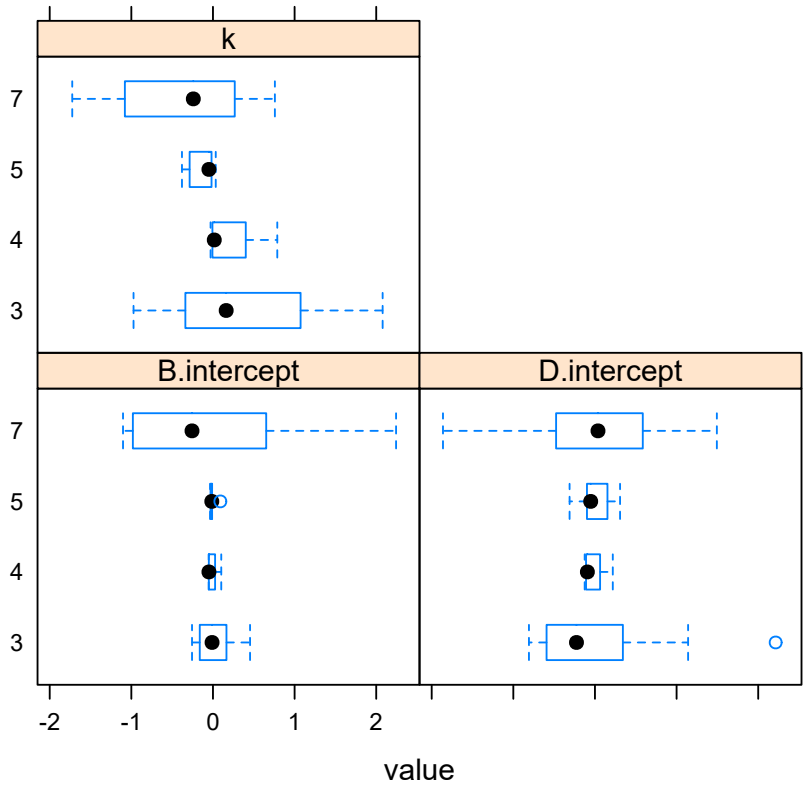


Figure 101: Subject specific random effect estimates $\widehat{b}_{0i}$ (B.intercept), $\widehat{b}_{1i}$ (D.intercept), and $\widehat{b}_{2i}$ (k) for Model 9, a NLMEM.

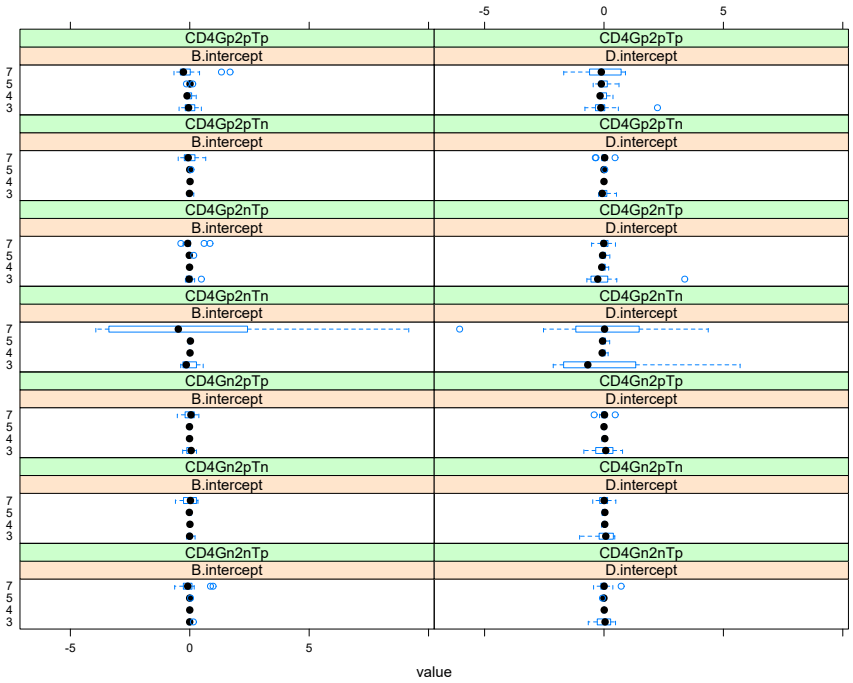


Figure 102: Response within subject random effect estimates $\widehat{b}_{0i,j}$ (B.intercept) and $\widehat{b}_{1i,j}$ (D.intercept) for Model 9, a NLMEM.

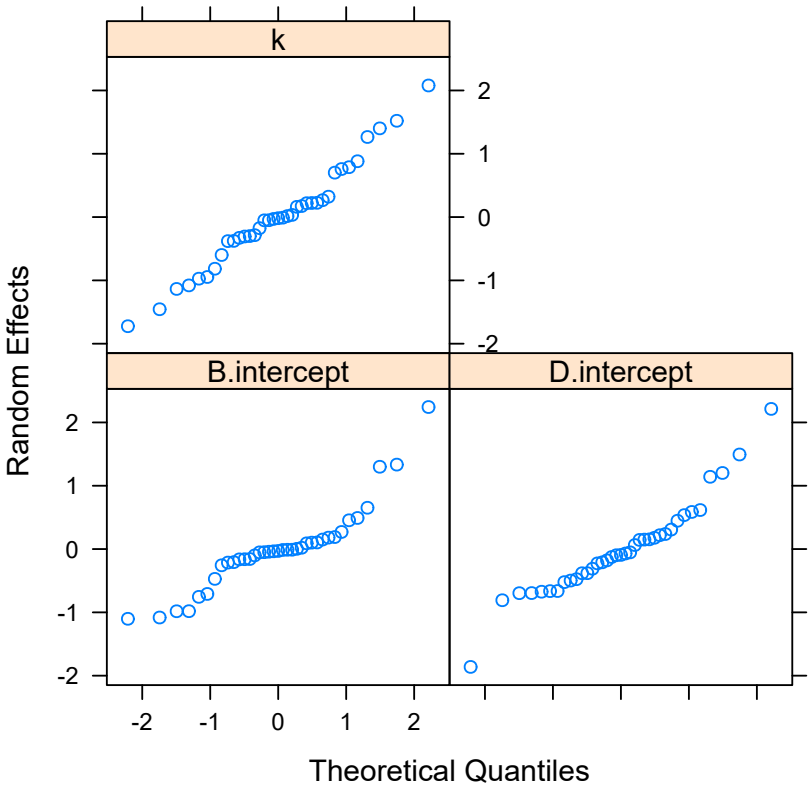


Figure 103: QQ-plots for the subject specific random effect estimates $\widehat{b}_{0i}$ (B.intercept), $\widehat{b}_{1i}$ (D.intercept), and $\widehat{b}_{2i}$ (k) for Model 9, a NLMEM.

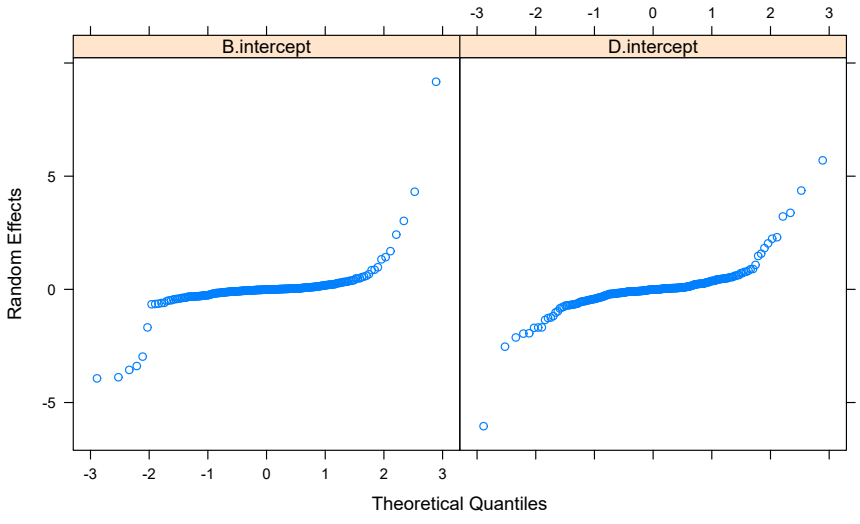


Figure 104: QQ-plots for response within subject random effect estimates $\widehat{b}_{0i,j}$ (B.intercept) and $\widehat{b}_{1i,j}$ (D.intercept) for Model 9, a NLMEM.

D

RESULTS FOR THE MULTIVARIATE GENERALISED LINEAR MIXED AND LATENT VARIABLE MODEL WITH OCCASION SPECIFIC COVARIATES AND RESPONSE TRAITS

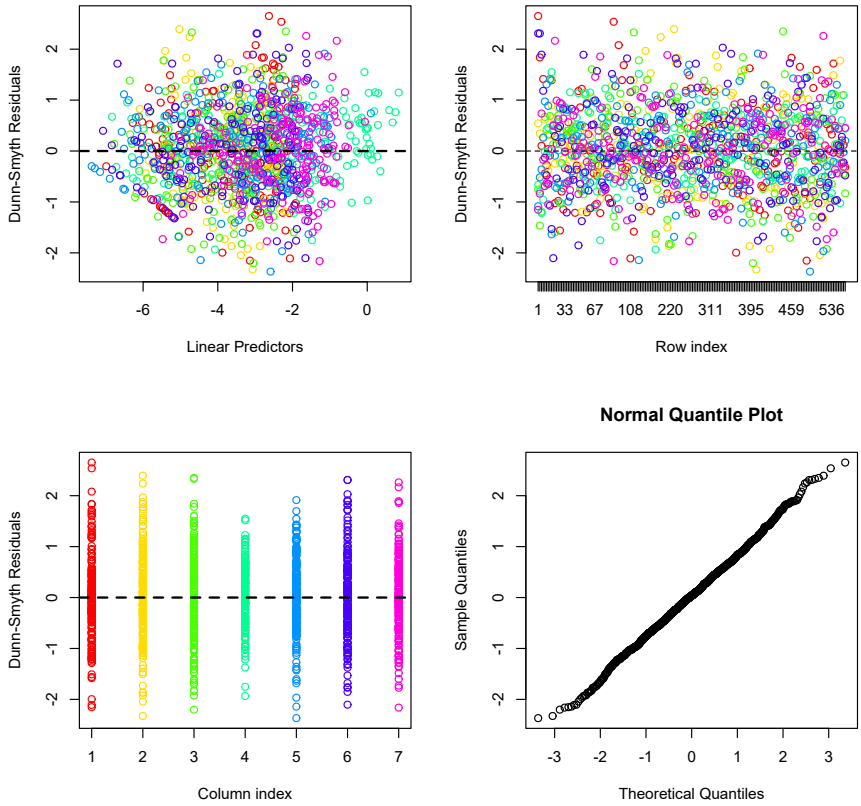


Figure 105: Dunn-Smyth residual plots for Model 14, a correlated response M-GLMEM with traits. Each colour represents a different response.

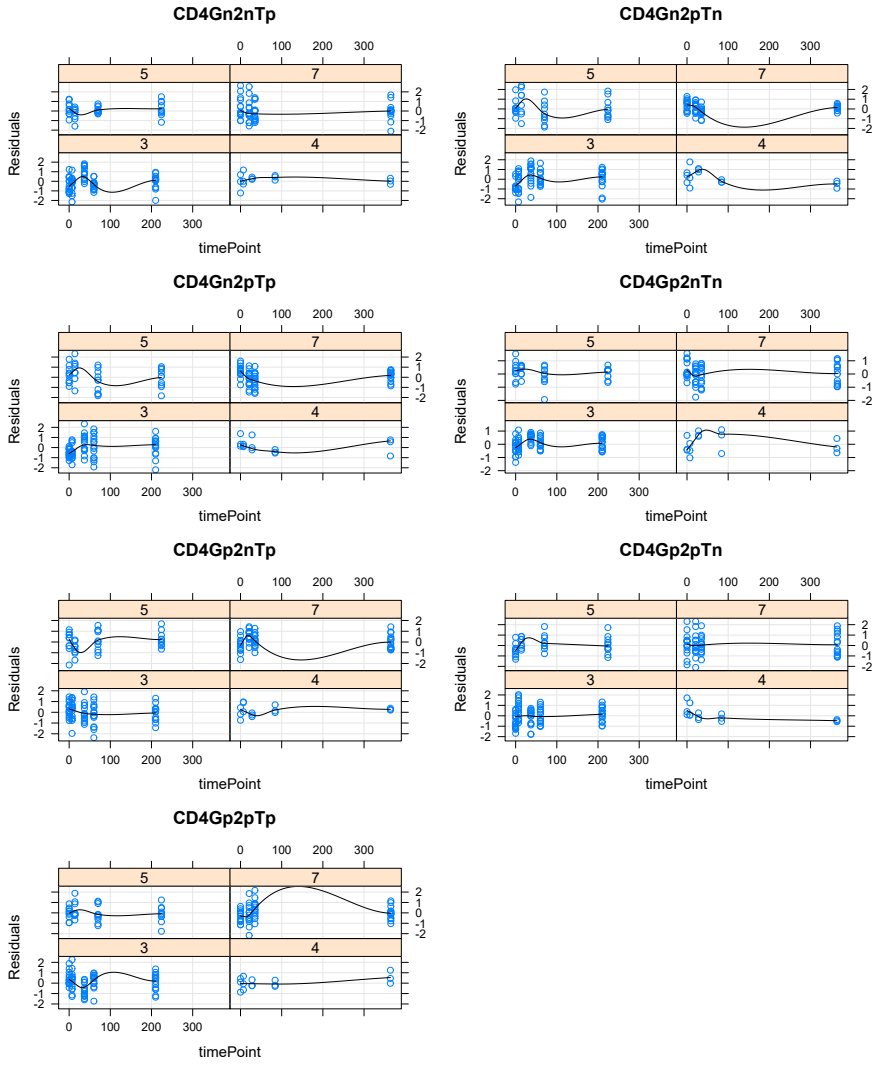


Figure 106: Dunn-Smyth residual plots by vaccine, response, and time point for Model 14, a correlated response M-GLMEM with traits.

For Model 14, the estimated residual correlation matrix is given by

	CD4Gn2nTp	CD4Gn2pTn	CD4Gn2pTp	CD4Gp2nTn	CD4Gp2nTp	CD4Gp2pTn	CD4Gp2pTp
CD4Gn2nTp	1	0.946	0.995	0.633	0.865	0.822	0.951
CD4Gn2pTn	0.946	1	0.911	0.851	0.974	0.961	0.988
CD4Gn2pTp	0.995	0.911	1	0.559	0.817	0.771	0.919
CD4Gp2nTn	0.633	0.851	0.559	1	0.936	0.961	0.843
CD4Gp2nTp	0.865	0.974	0.817	0.936	1	0.997	0.978
CD4Gp2pTn	0.822	0.961	0.771	0.961	0.997	1	0.959
CD4Gp2pTp	0.951	0.988	0.919	0.843	0.978	0.959	1

and Figure 110 provides a plot of this correlation matrix.

The correlation matrix giving the correlations between the immune responses for Model 14 due to the occasion specific covariates is given by

CD4Gn2nTp	CD4Gn2pTn	CD4Gn2pTp	CD4Gp2nTn	CD4Gp2nTp	CD4Gp2pTn	CD4Gp2pTp
CD4Gn2nTp1	0.918	0.827	0.884	0.95	0.861	0.94
CD4Gn2pTn0.918	1	0.956	0.87	0.951	0.926	0.954
CD4Gn2pTp0.827	0.956	1	0.735	0.865	0.838	0.897
CD4Gp2nTn0.884	0.87	0.735	1	0.938	0.962	0.921
CD4Gp2nTp0.95	0.951	0.865	0.938	1	0.938	0.965
CD4Gp2pTn0.861	0.926	0.838	0.962	0.938	1	0.922
CD4Gp2pTp0.94	0.954	0.897	0.921	0.965	0.922	1

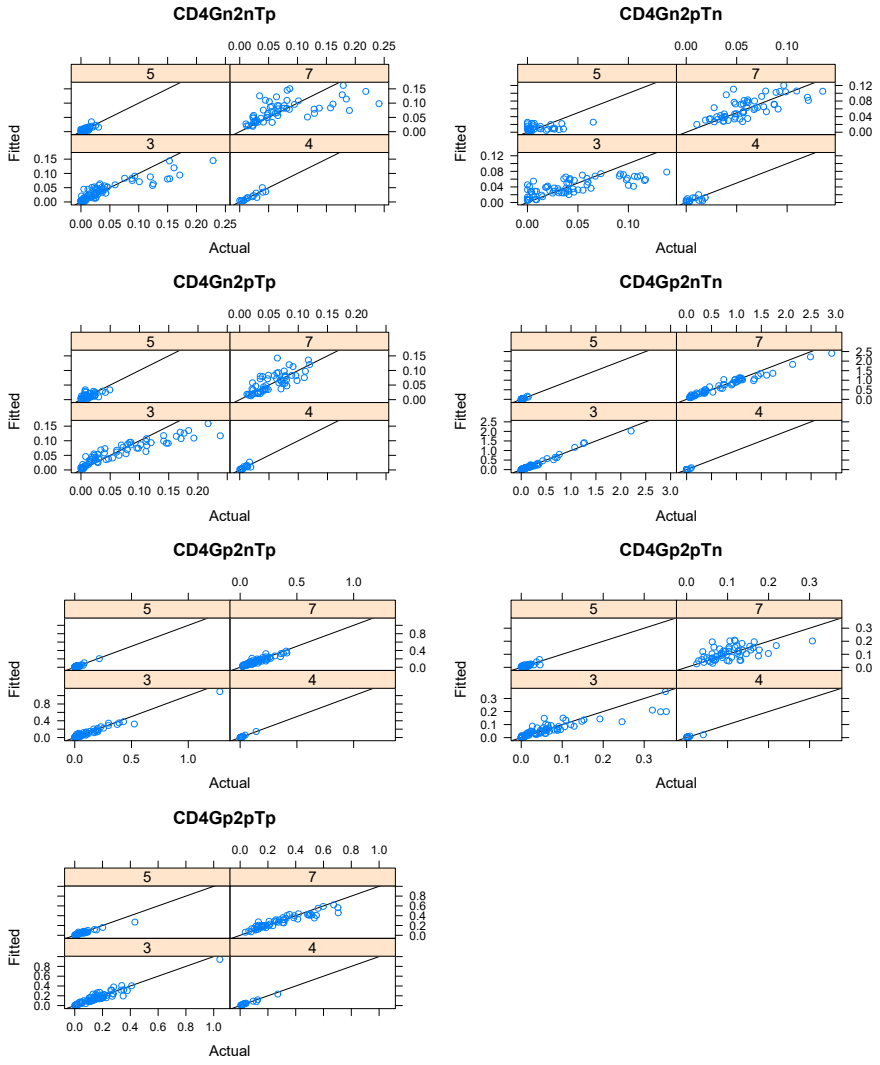


Figure 107: Plot of the fitted conditional response values versus the actual vaccine response values for each of the vaccines and responses for Model 14.

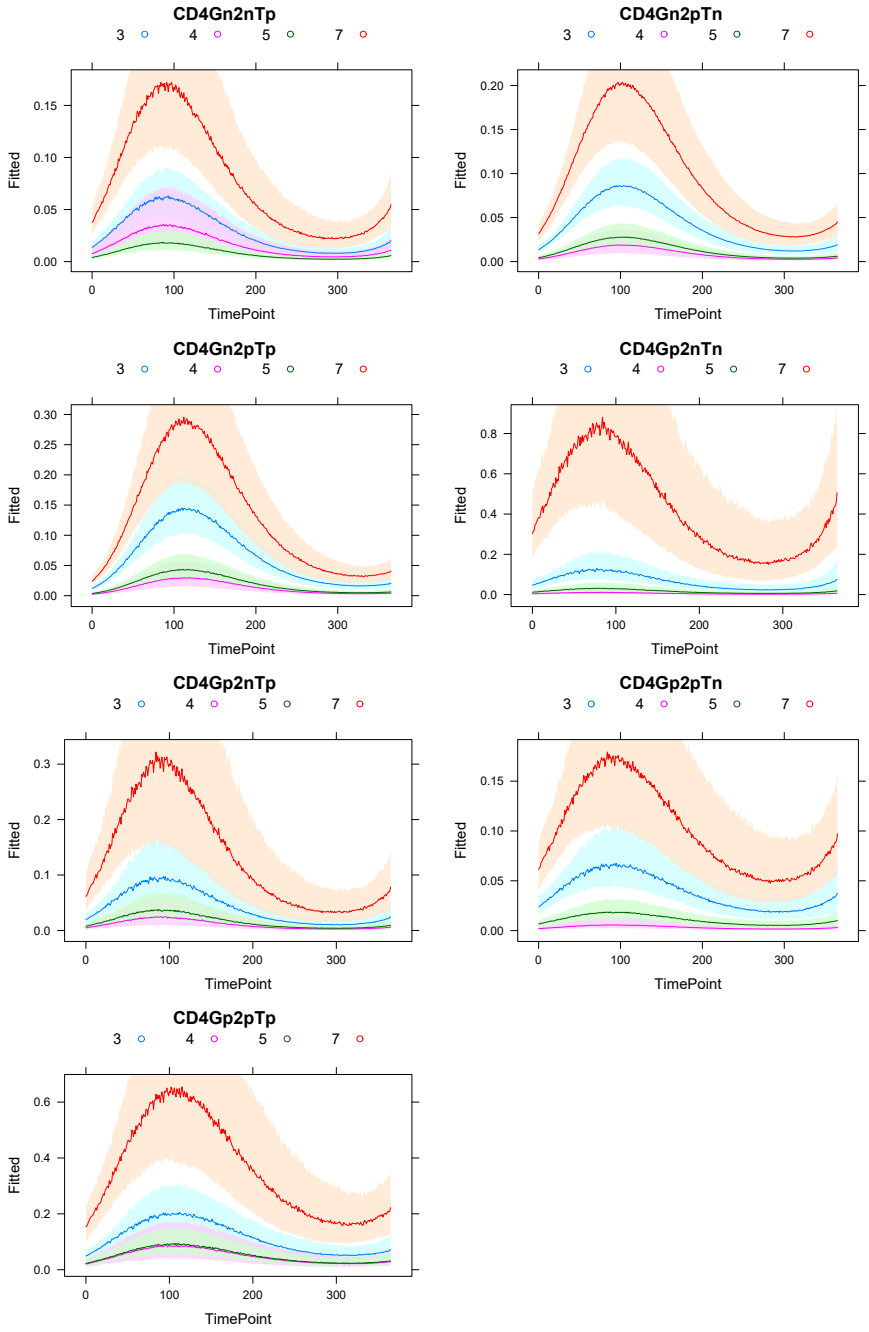
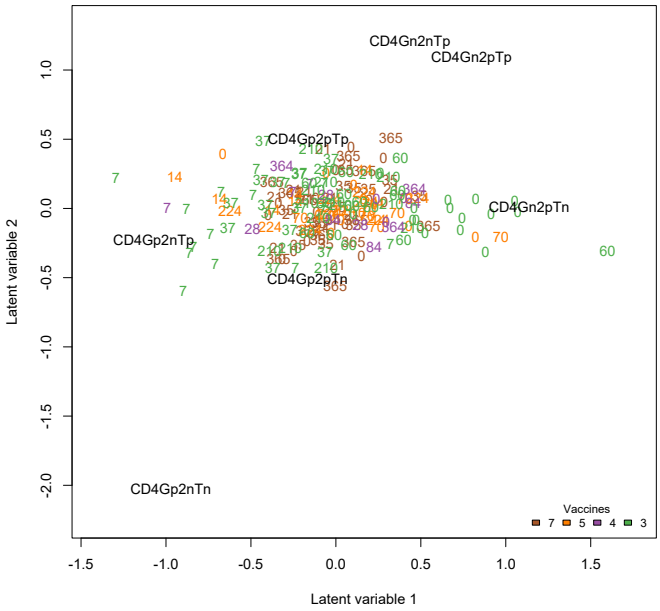
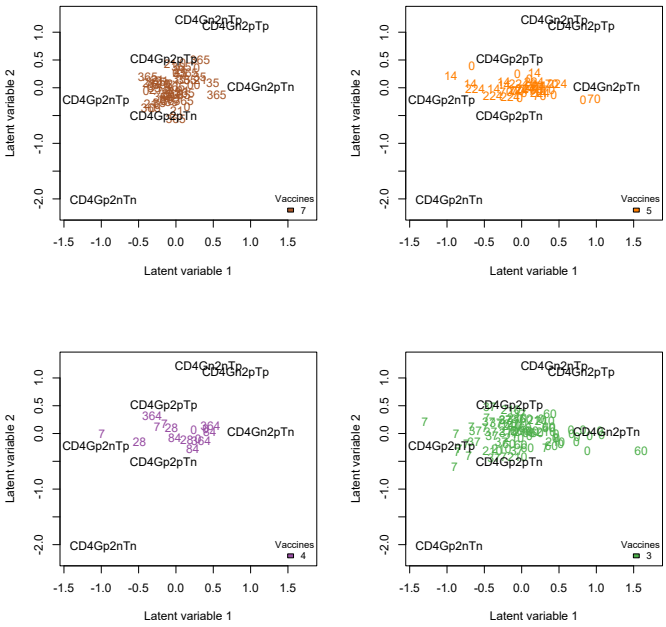


Figure 108: Plot of the fitted marginal immune response profiles for each of the vaccines and responses for Model 14.



(a) Model based constrained ordination plot for Model 14 for all of the vaccines.



(b) Model based constrained ordination plots for Model 14 for each of the vaccines.

Figure 109: Model based constrained ordination plots for Model 14, a correlated response M-GLMEM with traits. The observations are given by the coloured and numbered points where the colouring indicates the vaccine the response pertains to and the number the time that the observation was recorded. The responses, the biplot axes, are given by the black text.

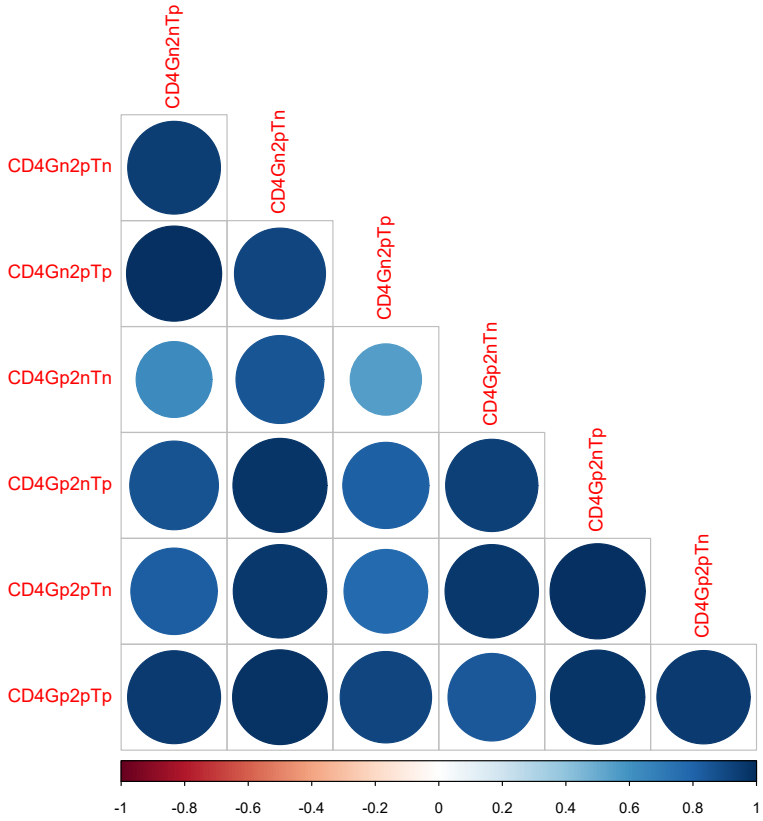


Figure 110: Plot of the residual correlation between the responses after controlling for time point and vaccine effects for Model 14.

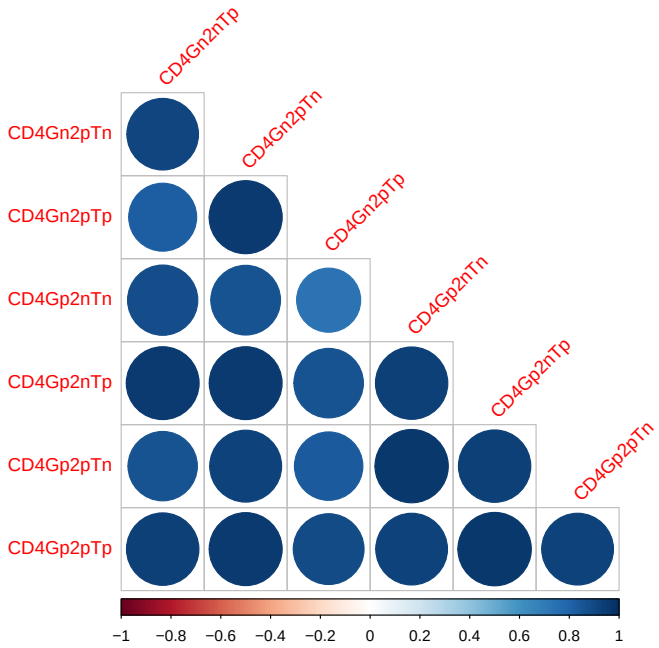


Figure 111: Plot of the significant correlations between the vaccine responses due to the occasion specific covariates for Model 14. Significance is determined from the 95% credible intervals; i.e. only the correlations with credible intervals excluding zero are plotted.

E

R CODE EXAMPLES

In the following Figures we provide snippets of R code which illustrate the implementations of some of the models that we considered in our applications.

```
#### Loading Packages ####################
library(readr)
library(dplyr)
library(reshape2)
library(nlme)
library(splines)
library(cplm)
library(boral)
#### Loading the data ####################
data <- read_delim("/Vaccines.csv",
                   "|", escape_double = FALSE, trim_ws = TRUE)
data$vaccine <- as.factor(data$vaccine)
data$infxn <- as.factor(data$infxn)
# Selecting the vaccines we want
data_upd <- data[data$vaccine %in% c(3,4,5,7),]
# Selecting infected lives
data_upd <- data_upd[data_upd$infxn==1,]
# Dropping unwanted time points
data_upd <- data_upd[!(data_upd$vaccine==3&data_upd$timePoint==30),]
data_upd <- data_upd[!(data_upd$vaccine==5&data_upd$timePoint==56),]
# Data manipulation
data_upd <- as.data.frame(data_upd)
responses <- c('CD4Gn2nTp',
               'CD4Gn2pTn',
               'CD4Gn2pTp',
               'CD4Gp2nTn',
               'CD4Gp2nTp',
               'CD4Gp2pTn',
               'CD4Gp2pTp')
vaccines <- unique(data_upd$vaccine)
# Removing outliers
out <- c()
for(i in 1:length(vaccines)){
  vacc_i <- vaccines[i]
  times <- unique(data_upd[data_upd$vaccine==vacc_i,]$timePoint)
  for(j in 1:length(times)){
    for(k in 1:length(responses))
    {
      idx_i <- data_upd[data_upd$vaccine==vacc_i&data_upd$timePoint==times[j],
                        responses[k]] %in%
        (boxplot.stats(
data_upd[data_upd$vaccine==vacc_i&data_upd$timePoint==times[j],
                                responses[k]])$out)
      out <- c(out,
        data_upd[
    data_upd$vaccine==vacc_i&data_upd$timePoint==times[j],'ptid'][idx_i])
    }
  }
}
out <- unique(out)
outliers <- out
data_upd <- data_upd[!data_upd$ptid %in% outliers,]
# Creating a long version of the data
data.long_upd <- melt(data_upd, id.vars=c("vaccine",
  "ptid","stim","infxn", "timePoint"), variable.name="response")
# Filtering the data
data.long_upd <- data.long_upd %>%
  filter(response %in% responses)
```

Figure 112: R code which performs the data preprocessing.

```
# Creating the spline
bspline_timePoint <-as.data.frame( bs(
  data_upd$timePoint,
  df = 3,
  degree = 3))
names(bspline_timePoint) <- c('bs1','bs2','bs3'
)
data.long_upd_bs <- cbind(data.long_upd,bspline_timePoint)
data.long_upd_bs_gr <- groupedData(value~time|vaccine/ptid/response,
data = data.long_upd_bs)
# Fitting the model
model_7 <- lme(
  value ~
    vaccine * response +
    vaccine * bs1 +
    vaccine * bs2 +
    vaccine * bs3,
  data = data.long_upd_bs_gr,
  random =  list(ptid = pdDiag( ~ 1),
response  = pdDiag( ~ 1 + bs3)),
  method = 'ML',
  weights = varPower(form = ~ fitted(.), fixed = 0.5),
  control = lmeControl(
    maxIter = 200,
    msMaxIter = 200,
    tolerance = 1e-3,
    niterEM = 200,
    msMaxEval = 500,
    opt = c("nlminb")
  )
)
```

Figure 113: R code which implements Model 7, a LMEM with a cubic B-spline.

```
# Setting up the nonlinear function
fun <- function(B, D, k, n, t) B + D * (k*t)^n *exp(-k*t) / factorial(n)
nform <- ~ B + D * (k*t)^1 *exp(-k*t) / factorial(1)
nfun <- deriv(nform,namevec=c("B","D","k"),
              function.arg=c("t","B","D","k"))
# Fitting the model
model_9 <- nlme(value ~ nfun(timePoint, B, D, k),
           start = c(
                     B = 0.05265518,
                     0,0,0,
                     0,0,0,0,0,0,
                     0,0,0,0,0,0,
                     0,0,0,0,0,0,
                     0,0,0,0,0,0,
                     D = 0.83466,
                     0,0,0,
                     0,0,0,0,0,0,
                     0,0,0,0,0,0,
                     0,0,0,0,0,0,
                     0,0,0,0,0,0,
                     k = 0.07697 ),
          random = list(ptid = pdDiag(B + D + k ~ 1),
            response = pdDiag(B + D ~ 1)),
           fixed = c(B + D ~ vaccine * response, k ~ 1),
           data = data.long_upd,
           method = 'ML',
           control = nlmeControl(
             maxIter = 500,
             pnlsMaxIter = 500,
             msMaxIter = 500,
             tolerance = 1e-6,
             niterEM = 500,
             pnlsTol = 1e-6,
             msTol = 1e-6,
             msVerbose = T,
             opt = c("nlminb")
           ))
```

Figure 114: R code which implements Model 9, a NLMEM.

```
# Data manipulation required by cplm
data.long_upd_bs$ptid <- as.factor(data.long_upd_bs$ptid)
# Fitting the model
model_11 <- cpglmm(
  value ~
    vaccine * response +
    vaccine * response * bs1 +
    vaccine * response * bs2 +
    vaccine * response * bs3 +
    (1 | ptid / response) +
    (bs1 - 1 | ptid) +
    (bs2 - 1 | ptid),
  data = data.long_upd_bs,
  control = list(max.iter = 5000, trace = 1),
  optimizer = "nlminb",
  doFit = TRUE,
  nAGQ = 1L
)
```

Figure 115: R code which implements Model 11, a univariate Tweedie GLMEM.

```
# Setting up the data for Boral
bspline_timePoint <-as.data.frame( bs(
  data_upd$timePoint,
  df = 3,
  degree = 3))
names(bspline_timePoint) <- c('bs1','bs2','bs3')
y <- data_upd[,responses]
data_upd$vacc4 <- ifelse(data_upd$vaccine == '4',1,0)
data_upd$vacc5 <- ifelse(data_upd$vaccine == '5',1,0)
data_upd$vacc7 <- ifelse(data_upd$vaccine == '7',1,0)
X <- data.frame( data_upd[,c('vacc4','vacc5','vacc7')],
        bspline_timePoint)
X_to_use_1 <-model.matrix(~vacc4+vacc5+vacc7+bs1+bs2+bs3-1,
                  data = X)
traits <- matrix(
  c(0,0,1,
    0,1,0,
    0,1,1,
    1,0,0,
    1,0,1,
    1,1,0,
    1,1,1),
  nrow = 7
)
which_traits_1 <- vector("list",ncol(X_to_use_1)+1)
for(i in 1:length(which_traits_1))
  which_traits_1[[i]] <- 1:ncol(traits)
row.ids <-  matrix(c(as.integer(as.factor(data_upd$ptid))),
   nrow = length(c(as.integer(as.factor(data_upd$ptid)))))
# Fitting the model
model_14 <- boral(y,
                      X = X_to_use_1,
                      traits = traits,
                      which.traits = which_traits_1,
                      family = "tweedie",
                      lv.control = list(num.lv = 2),
                      row.eff = 'random',
                      calc.ics = T,
                      save.model = T,
                      row.ids = row.ids
)
```

Figure 116: R code which implements Model 14, a MGLMEM with occasion specific covariates and response traits.

BIBLIOGRAPHY

Abel, B., Tameris, M., Mansoor, N., Gelderbloem, S., Hughes, J., Abrahams, D., Makhethe, L., Erasmus, M., Kock, M. d., van der Merwe, L. *et al. (2010). The Novel Tuberculosis Vaccine, AERAS-402, Induces Robust and Polyfunctional CD4+ and CD8+ T Cells in Adults*, American Journal of Respiratory and Critical Care Medicine **181**(12): 1407–1417.

Akaike, H. (1974). A New Look at the Statistical Model Identification, Selected Papers of Hirotugu Akaike, *Springer, pp. 215–222.*

Carlin, B. P. and Louis, T. A. (2008). Bayesian Methods for Data Analysis, *Chapman and Hall/CRC.*

Davidian, M. and Giltinan, D. M. (2003). Nonlinear Models for Repeated Measurement Data: An Overview and Update, Journal of Agricultural, Biological, and Environmental Statistics **8**(4): 387.

Day, C. L., Tameris, M., Mansoor, N., van Rooyen, M., de Kock, M., Geldenhuys, H., Erasmus, M., Makhethe, L., Hughes, E. J., Gelderbloem, S. et al. (2013). Induction and Regulation of T-cell Immunity by the Novel Tuberculosis Vaccine M72/AS01 in South African adults, American Journal of Respiratory and Critical Care Medicine **188**(4): 492–502.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals, Journal of Computational and Graphical Statistics **5**(3): 236–244.

Dunn, P. K. and Smyth, G. K. (2005). Series evaluation of Tweedie exponential dispersion model densities, Statistics and Computing **15**(4): 267–280.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). Longitudinal Data Analysis, *CRC press.*

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). Bayesian Data Analysis, *2 edn, Chapman and Hall/CRC.*

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, Staff Report 148, *Federal Reserve Bank of Minneapolis.*
***URL:** https://www.minneapolisfed.org/research/sr/sr148.pdf*

Greenacre, M. J. (2010). Biplots in Practice, *Fundacion BBVA.*

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, *2 edn, Springer.*

Houben, R. M. and Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation using Mathematical Modelling, PLoS Medicine **13**(10).

Hui, F. K. (2016). boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R, Methods in Ecology and Evolution **7**(6): 744–750.

Hui, F. K. C. (2018). boral: Bayesian Ordination and Regression AnaLysis. *R package version 1.7.*

Hui, F. K., Taskinen, S., Pledger, S., Foster, S. D. and Warton, D. I. (2015). Model-Based Approaches to Unconstrained Ordination, Methods in Ecology and Evolution **6**(4): 399–411.

Jamil, T., Ozinga, W. A., Kleyer, M. and ter Braak, C. J. (2013). Selecting Traits that Explain Species–Environment Relationships: A Generalized Linear Mixed Model Approach, Journal of Vegetation Science **24**(6): 988–1000.

Jasenosky, L. D., Scriba, T. J., Hanekom, W. A. and Goldfeld, A. E. (2015). T Cells and Adaptive Immunity to Mycobacterium Tuberculosis in Humans, Immunological Reviews **264**(1): 74–87.

Jørgensen, B. (1987). Exponential Dispersion Models, Journal of the Royal Statistical Society: Series B (Methodological) **49**(2): 127–145.

Legendre, P., Borcard, D. and Peres-Neto, P. R. (2005). Analyzing Beta Diversity: Partitioning the Spatial Variation of Community Composition Data, Ecological Monographs **75**(4): 435–450.

Letten, A. D., Keith, D. A., Tozer, M. G. and Hui, F. K. (2015). Fine-Scale Hydrological Niche Differentiation Through the Lens of Multi-Species Co-occurrence Models, Journal of Ecology **103**(5): 1264–1275.

Lewinsohn, D. A., Lewinsohn, D. M. and Scriba, T. J. (2017). Polyfunctional CD4+ T Cells As Targets for Tuberculosis Vaccination, Frontiers in Immunology **8**: 1262.

Lindstrom, M. J. and Bates, D. M. (1988). Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data, Journal of the American Statistical Association **83**(404): 1014–1022.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data, Biometrics *pp. 673–687.*

Mauff, K. (2011). Multivariate Multi-Level Non-Linear Mixed-Effect Models and their Application to the Modeling of Drug-Concerntration Time Curves, *Master's thesis, University of Cape Town, Cape Town, South Africa.*

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, *2 edn, Chapman and Hall, Boca Raton.*

Mearns, H., Geldenhuys, H. D., Kagina, B. M., Musvosvi, M., Little, F., Ratangee, F., Mahomed, H., Hanekom, W. A., Hoff, S. T., Ruhwald, M. et al. (2017). H1: IC31 Vaccination is Safe and Induces Long-Lived TNF-α+ IL-2+ CD4 T Cell Responses in M. Tuberculosis Infected and Uninfected Adolescents: A Randomized Trial, Vaccine **35***(1): 132–141.*

Penn-Nicholson, A., Tameris, M., Smit, E., Day, T. A., Musvosvi, M., Jayashankar, L., Vergara, J., Mabwe, S., Bilek, N., Geldenhuys, H. et al. (2018). Safety and Immunogenicity of the Novel Tuberculosis Vaccine ID93+ GLA-SE in BCG-Vaccinated Healthy Adults in South Africa: A Randomised, Double-Blind, Placebo-Controlled Phase 1 Trial, The Lancet Respiratory Medicine **6***(4): 287–298.*

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2019). nlme: Linear and Nonlinear Mixed Effects Models. *R package version 3.1-139.*

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model, Journal of Computational and Graphical Statistics **4***(1): 12–35.*

Pinheiro, J. C. and Bates, D. M. (2000). Mixed-effects Models in S and S-PLUS, *Springer, New York.*

Plummer, M. et al. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing, *Vienna, Austria.*

R Core Team (2019). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing, Vienna, Austria.*

Rodo, M. J. (2017). Grouping and Characterising Novel Tuberculosis Vaccine Candidates, *Master's thesis, University of Cape Town, Cape Town, South Africa.*

Rodo, M. J., Rozot, V., Nemes, E., Dintwe, O., Hatherill, M., Little, F. and Scriba, T. J. (2019a). A Comparison of Antigen-Specific T Cell Responses Induced by Six Novel Tuberculosis Vaccine Candidates, PLoS Pathogens **15***(3).*

Rodo, M. J., Rozot, V., Scriba, T., Hatherill, M., Nemes, E., Little, F. and Dintwe, O. B. (2019b). Vaccines.
***URL:** https://zivahub.uct.ac.za/articles/Vaccines/7133513*

Savic, R. M., Jonker, D. M., Kerbusch, T. and Karlsson, M. O. (2007). Implementation of a Transit Compartment Model for Describing Drug Absorption in Pharmacokinetic Studies, Journal of Pharmacokinetics and Pharmacodynamics **34***(5): 711–726.*

Schwarz, G. et al. (1978). Estimating the Dimension of a Model, The Annals of Statistics **6***(2): 461–464.*

Scriba, T. J., Tameris, M., Mansoor, N., Smit, E., van der Merwe, L., Isaacs, F., Keyser, A., Moyo, S., Brittain, N., Lawrie, A. et al. (2010). Modified Vaccinia Ankara-Expressing Ag85A, a Novel Tuberculosis Vaccine, is Safe in Adolescents and Children, and Induces Polyfunctional CD4+ T Cells, European Journal of Immunology **4***0(1): 279–290.*

Scriba, T. J., Tameris, M., Smit, E., van der Merwe, L., Hughes, E. J., Kadira, B., Mauff, K., Moyo, S., Brittain, N., Lawrie, A. et al. (2012). A Phase IIa Trial of the New Tuberculosis Vaccine, MVA85A, in HIV- and / or Mycobacterium Tuberculosis–Infected Adults, American Journal of Respiratory and Critical Care Medicine **1***85(7): 769–778.*

Suliman, S., Geldenhuys, H., Johnson, J. L., Hughes, J. E., Smit, E., Murphy, M., Toefy, A., Lerumo, L., Hopley, C., Pienaar, B. et al. (2016). Bacillus Calmette–Guerin (BCG) Revaccination of Adults with Latent Mycobacterium Tuberculosis Infection Induces Long-Lived BCG-Reactive NK Cell Responses, The Journal of Immunology **1***97(4): 1100–1110.*

Suliman, S., Luabeya, A. K. K., Geldenhuys, H., Tameris, M., Hoff, S. T., Shi, Z., Tait, D., Kromann, I., Ruhwald, M., Rutkowski, K. T. et al. (2019). Dose Optimization of H56: IC31 Vaccine for Tuberculosis-Endemic Populations. A Double-Blind, Placebo-controlled, Dose-Selection Trial, American Journal of Respiratory and Critical Care Medicine **1***99(2): 220–231.*

Tweedie, M. C. (1984). An Index which Distinguishes Between Some Important Exponential Families, Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pp. 579–604.*

Voss, G., Casimiro, D., Neyrolles, O., Williams, A., Kaufmann, S. H., McShane, H., Hatherill, M. and Fletcher, H. A. (2018). Progress and Challenges in TB Vaccine Development, F1000Research **7**.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C. and Hui, F. K. (2015). So Many Variables: Joint Modeling in Community Ecology, Trends in Ecology & Evolution **3***0(12): 766–779.*

Weigend, S., Mielenz, N. and Lamont, S. (1997). Application of a Nonlinear Regression Function to Evaluate the Kinetics of Antibody Response to Vaccines in Chicken Lines Divergently Selected for Multitrait Immune Response, Poultry Science **7***6(9): 1248–1255.*

World Health Organisation (2015). Implementing the End TB Strategy: The Essentials.

World Health Organisation (2019). Global Tuberculosis Report 2019.

Zhang, Y. (2013). Likelihood-Based and Bayesian Methods for Tweedie Compound Poisson Linear Mixed Models, Statistics and Computing **23**(6): 743–757.