UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Longitudinal analysis of viral shedding in astronauts before, during, and after a mission to the International Space Station

Frederico Moreira

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof. Doutor Nuno Sepúlveda
Prof.ª Doutora Marília Antunes

2022

Acknowledgements

I want to thank my supervisors, professor Nuno and professor Marília for the help and support with every question and difficulty that this project presented.

I also want to thank the Immune Stats group for the materials provided.

Quero agradecer à minha família, em particular aos meus pais por todo o apoio ao longo deste último ano.

Aos meus amigos que me ajudaram e motivaram a chegar a esta fase, e aos meus colegas de residência que me acompanharam durante esta tese, obrigado.

Frederico Moreira,
Porto de Mós, Março de 2022.

Resumo

Vírus são encontrados onde quer que haja vida e provavelmente existem desde a evolução das primeiras células vivas. Os vírus são parasitas microscópicos, normalmente muito menores que bactérias, que não têm capacidade de prosperar e de se reproduzir fora de um corpo hospedeiro. Eles infetam todo o tipo de formas de vida, desde animais e plantas até microrganismos, incluindo as bactérias e arquea. Vírus são encontrados em quase todos os ecossistemas na Terra e são a entidade biológica mais numerosa. Os herpesvírus são os vírus membros da família *Herpesviridae* que causam infeções e certas doenças em animais e humanos. Os vírus Epstein-Barr (EBV), varicela-zóster (VZV), citomegalovírus (CMV), herpes simplex vírus 1 (HSV1) e herpes simplex vírus 2 (HSV2) estão extremamente difundidos entre os humanos. Mais de 90% dos adultos já foi infetado por pelo menos um destes vírus, e uma forma latente deles permanece em quase todos os humanos que foram infetados. Outro herpesvírus relevante é herpesvírus humano (HHV6). Estes vírus, quando reativados, já foram associados a patologias como varicela, cancro da próstata, cancro da mama, pneumonia, esclerose múltipla, síndrome de fadiga crónica, entre outros. A reativação pode ser provocada por uma combinação de estímulos celulares internos e/ou externos. A capacidade do sistema imunitário de combater infeções fica reduzida em situações crónicas de stress. Um exemplo dessas situações é o caso das missões à Estação Espacial Internacional onde as condições não naturais sentidas por astronautas podem torná-los particularmente vulneráveis à reativação desses vírus que, em situações normais, estariam num estado de dormência. Assim, uma questão importante consiste em perceber os mecanismos associados à reativação destes vírus visando desenvolver agentes terapêuticos contra infeções virais e consequentes doenças.

Um aumento na reativação de alguns herpesvírus latentes, incluindo EBV, VZV e CMV, foi inicialmente reportado por um estudo em astronautas numa missão de curta duração (10—16 dias) num vaivém espacial. Num estudo subsequente, os vírus cuja reativação estava sob investigação foram EBV, VZV, CMV, HSV1, HSV2 e HHV6, em amostras de saliva e urina. As amostras de saliva foram analisadas para detetar EBV, VZV, HSV1, HSV2 e HHV6, e as amostras de urina foram analisadas para detetar CMV. Nesse estudo participaram 18 homens e 5 mulheres e a missão espacial foi de aproximadamente 180 dias. Não foi detetada reativação de HSV1, HSV2 e HHV6 em qualquer astronauta. O objetivo deste estudo era determinar se os astronautas a participar numa missão de longa duração se acostumariam às condições de stress associadas ao voo espacial e se o seu sistema imunitário conseguiria mitigar a reativação destes vírus nas últimas fases da missão. No final dessa análise foi sugerido que ocorreu o oposto, ou seja, os dados pareciam indicar um aumento na proporção e amplitude da reativação nas últimas fases da missão. Os vírus em estudo reativaram independentemente uns dos outros.

Para EBV e VZV foram executados sete momentos de medição: dois antes (*L−180*, *L−45*), três durante (*Early*, *Mid*, *Late*) e dois depois (*R+0*, *R+30*) do voo espacial. Para CMV, foram realizadas as mesmas medições antes e depois da missão, mas durante o voo apenas uma amostra foi recolhida (*During*). As medições recolhidas consistem no número de contagens detetadas em cada momento de estudo. EBV e VZV têm dados omissos nos três momentos de medição durante a missão. Os dados deste estudo apresentam três dificuldades particulares: pequeno tamanho amostral, inflação do número de zeros e a presença de dados omissos. O estudo original foi apenas observacional e exploratório, e focou-se em analisar a prevalência viral nos diferentes

momentos de medição sem analisar os dados omissos ou modelar os dados dos diferentes vírus. O objetivo deste projeto é estender essa análise usando os mesmos dados que foram partilhados na respetiva publicação. Por simplicidade de análise, os dados utilizados neste projeto são os dados binários representando reativação e não-reativação de um vírus num determinado instante de tempo. Os métodos usados neste projeto foram os seguintes: intervalos de confiança para proporções, teste de McNemar na sua versão exata, análise de modelos lineares para dados categorizados, imputação múltipla de dados e aplicação de modelos de regressão logística mistos (LRMM). O *software* usado nas análises estatísticas foi o R, versão 4.0.1. Os principais *packages* disponíveis no *software* R utilizados neste projeto foram `proportion` 2.0.0, `exact2x2` 1.6.5, `ACD` 1.5.3, `mice` 3.13.0 e `lme4` 1.1-26. O nível de significância considerado ao longo deste projeto foi de 5%.

No momento antes da missão *L−180* não houve qualquer deteção de CMV nas amostras recolhidas. As proporções de reativação deste vírus nos momentos *L−180*, *L−45*, *During*, *R+0* e *R+30* foram, respetivamente, 0, 0.304, 0.522, 0.261 e 0.087. Para CMV foi detetado um aumento significativo na proporção de reativação durante o voo (*During*) quando comparado com as medições antes (*L−180*) e depois (*R+30*) do voo espacial. O teste de McNemar levou à deteção de diferenças significativas entre o momento *During* e os momentos *L−180* (p=<0.001) e *R+30* (p=0.006). Foi estimado um LRMM para os dados binários de CMV onde os efeitos fixos foram os vários momentos de medição e o efeito aleatório foi o astronauta. Nesse LRMM não houve coeficientes significativos, provavelmente porque o nível de referência do modelo era o momento *L−180* onde não foram detetadas reativações do vírus. Então, foi ajustado um novo LRMM onde o efeito de referência estava associado ao momento *L−45*. Nesse segundo modelo o momento *R+30* era significativo (p=0.043), no entanto o momento de referência *L−45* (p=0.074) e o momento *During* (p=0.078) tinham p-values pequenos. O erro-padrão do coeficiente *L−180* continuava a ser muito grande então um terceiro LRMM foi ajustado sem esse momento. Nesse terceiro LRMM os coeficientes significativos mantiveram-se: o único coeficiente significativo estava associado ao momento *R+30* (p=0.043), apesar de o momento de referência *L−45* (p=0.074) e o coeficiente *During* (p=0.078) terem um p-value perto do nível de significância usual de 5%.

Nos momentos pré-missão *L−180* e *L−45* não houve deteção de reativação de VZV. Para EBV, houve deteção de reativação em todos os momentos em estudo. As proporções de reativação de EBV foram 0.130, 0.435, 0.056, 0.227, 0.450, 0.391 e 0.217 para os momentos *L−180*, *L−45*, *Early*, *Mid*, *Late*, *R+0* e *R+30*, respetivamente. Para VZV as proporções de reativação para os mesmos momentos foram, respetivamente, 0, 0, 0.500, 0.318, 0.684, 0.435 e 0.087. Para ambos os vírus, foi detetado um aumento na amplitude das reativações no momento *Late*. As proporções de reativação viral estimadas para EBV são mais estáveis do que para VZV ao longo do tempo. Para VZV houve diferenças significativas nas proporções de reativação entre os momentos pré-missão *L−180* e *L−45*, e os momentos de medição *Early*, *Mid*, *Late*, e *R+0*. O teste de McNemar conduziu à deteção de três pares de momentos para EBV com diferenças significativas entre os momentos de medição e 11 para VZV. Para VZV, a maioria destas diferenças significativas devem-se ao facto de nos momentos de medição antes do voo (*L−180* e *L−45*) não terem sido detetadas reativações desse vírus. A análise do padrão de omissão de dados levou a que não fosse descartável a suposição do mecanismo de omissão ser completamente aleatório (i.e.,

missing completely at random), ou seja, o facto de existirem dados omissos é independente dos dados observados e não observados, para EBV (p=0.490) e VZV (p=0.070). Cinquenta imputações de dados binários foram estimadas usando imputação multivariada por equações encadeadas para ambos os vírus. Para EBV, as probabilidades de reativação estimadas a partir do agrupamento das imputações de dados foram 0.126, 0.239 e 0.454 para *Early*, *Mid* e *Late*, respetivamente. Para VZV, as mesmas probabilidades para os mesmos momentos foram 0.488, 0.330 e 0.617, respetivamente. O LRMM estimado a partir do agrupamento das imputações de dados para EBV, onde os momentos em estudo eram os efeitos fixos e o astronauta o efeito aleatório, era significativo no nível de referência *L−180* (p=0.004), em *L−45* (p=0.029) e em *Late* (p=0.022), apesar de *R+0* (p=0.053) estar perto do nível de significância considerado. Para VZV, o LRMM estimado da mesma maneira que para EBV não apresentou significância em qualquer coeficiente, então um segundo LRMM foi estimado onde o nível de referência era o momento *Early*. Nesse segundo modelo o único momento significativo foi *R+30* (p=0.002) e os erros-padrão dos coeficientes dos momentos onde não foi observada qualquer reativação (*L−180* e *L−45*) continuavam a ser muito grandes. Assim, um terceiro LRMM foi estimado sem os coeficientes *L−180* e *L−45*. Nesse LRMM o único coeficiente significativo continuou a ser o do momento *R+30* (p=0.002).

Em conclusão, a disseminação viral já foi associada a algumas patologias e é necessário entender as condições nas quais estes vírus são reativados para prevenir a propagação de doenças. O conjunto de dados de pequena dimensão, com inflação de zeros e com dados omissos ofereceu dificuldades à sua análise. No geral, existem razões para acreditar que as condições de stress associadas a uma missão espacial influenciam a reativação de vírus latentes nos astronautas.

Palavras-chave: dados omissos, imputação múltipla baseada em equações encadeadas, análise de dados categóricos com respostas omissas, modelos longitudinais

Abstract

Herpesviruses were measured in 23 astronauts with the objective of understanding their reactivation pattern during a long-duration space mission. The measurements consisted of the number of viral copies of cytomegalovirus (CMV), Epstein-Barr (EBV) and varicella-zoster (VZV) viruses collected at different moments: two before (*L−180*, *L−45*), three during (*Early*, *Mid*, *Late*) and two after (*R+0*, *R+30*) a spaceflight. These data present three difficulties: small sample size, zero-inflation and missing responses. The methods used were confidence intervals for proportions, McNemar's exact test, linear models for categorical data, multiple imputation using chained equations (MICE), and logistic regression mixed models (LRMM).

CMV was only measured once during the flight (*During*). There was significant increase in the reactivation proportion during flight compared to before and after flight measures *L−180* and *R+30*. The LRMM fitted for binary CMV that had moments with reactivation as fixed effects and random effect subject was significant at coefficient *R+30* (p=0.043), although *During* (p=0.078) had a p-value close to the statistical significance of 5%.

EBV and VZV were measured in saliva samples and have missing responses for the inflight moments. An increase of the amplitude of reactivation was detected at *Late* for both viruses. The data seemed to follow a missing-completely-at-random mechanism for both viruses (p=0.490 and 0.070 for EBV and VZV, respectively). Fifty imputed data sets were generated for each virus. For EBV the pooled estimates for the reactivation probability were 0.126, 0.239, 0.454 for *Early*, *Mid*, *Late*, respectively, and for VZV 0.488, 0.330, 0.617, respectively. The pooled LRMM of EBV with *L−180* as baseline was significant at *L−45* (p=0.029) and *Late* (p=0.022), and *R+0* (p=0.053) was close to significance. For VZV, *R+30* was the only significant different from baseline *Early*(p=0.002).

In conclusion, the stress conditions of the spaceflight affected the reactivation dynamics of all three viruses.

Keywords: missing data, multiple imputation based on chained equations, categorical data analysis with missing responses, longitudinal models

Other production

This project was presented at the XXV Congress of the Portuguese Statistical Society. This event was organised by the University of Évora and occurred on 13–16 of October of 2021. The abstract for the conference [65] and application for the participation grant are in Appendix C.

Contents

List of Tables

List of Figures

List of Abbreviations

CMV	Cytomegalovirus
EBV	Epstein-Barr virus
GLMM	Generalised linear mixed model
HHV6	Human herpes virus 6
HSV1	Herpes simplex virus 1
HSV2	Herpes simplex virus 2
ISS	International Space Station
LRMM	Logistic regression mixed model
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multivariate imputation by chained equations
MNAR	Missing not at random
VZV	Varicella-zoster virus

Chapter 1

Introduction

Viruses are found wherever there is life and have probably existed since living cells first evolved [1]. A virus is a microscopic parasite, generally much smaller than bacteria, that lacks the capacity to thrive and reproduce outside of a host body [2, 3]. They infect all life forms, from animals and plants to microorganisms, including bacteria and archaea [4, 5]. Viruses are found in almost every ecosystem on Earth and are the most numerous type of biological entity [6, 7].

Herpesviruses are the set of viruses members of the family *Herpesviridae* which is a large family of DNA viruses that cause infections and certain diseases in animals, including humans [8–10]. Epstein-Barr virus (EBV), varicella-zoster virus (VZV), cytomegalovirus (CMV), herpes simplex virus 1 (HSV1) and herpes simplex virus 2 (HSV2) are extremely widespread among humans. More than 90% of adults have been infected with at least one of these, and a latent form of the virus remains in almost all humans who have been infected [11, 12]. Another relevant and common herpesvirus is human herpesvirus 6 (HHV6). Among themselves, these viruses have been associated with conditions like cold sores [11], varicella [13], prostate cancer [14], breast cancer [15], encephalitis and pneumonitis [16]. In particular, EBV has been associated with conditions as lymphohistiocytosis [17], hairy leukoplakia, central nervous system lymphomas [18, 19], rheumatoid arthritis, Sjögren's syndrome [20, 21], multiple sclerosis [22–25], chronic fatigue syndrome [26–29] and diminished cell-mediated immunity [30]. About 200 000 cancer cases globally per year are thought to be attributable to EBV [31].

Viral reactivation may be provoked by a combination of external and/or internal cellular stimuli [32]. The immune system's ability to fight this antigens is reduced under stress conditions, hence the body becomes more susceptible to infections. In particular the stress conditions associated with space missions, such as loss of gravity, lack of exposure to natural light, confinement, among others, can induce reactivation of latent herpesviruses in astronauts. Understanding the mechanism by which viruses reactivate is essential in developing future therapeutic agents against viral infection and subsequent disease.

1.1 The original study

Increased reactivation of some naturally occurring latent herpesviruses including EBV, VZV and CMV was previously demonstrated in astronauts during short-duration (10–16 days) space shuttle flights [33]. In Mehta et al. (2017) [34] the viruses whose reactivation was under investigation were EBV, VZV and CMV, along with HSV1, HSV2 and HHV6. The measurements were collected in saliva and urine samples. The saliva samples were analysed for EBV, VZV, HSV1, HSV2 and HHV6, and the urine samples were analysed for CMV.

This study was conducted with 23 astronauts (18 males and 5 females) with overall mean age ± SD = 53 ± 4.9 years old. This long-duration mission to the International Space Station (ISS) had a length of approximately 180 days. Two crew members participated in shorter missions of approximately 60–90 days. Twenty apparently healthy subjects, matched for age and gender (16 males and 4 females, mean age ± SD of 49.3 ± 4.9 years) participated as ground-based viral reactivation controls. None of the 20 control subjects shed VZV or CMV and only two of them shed EBV. No astronauts or control subjects shed HSV1, HSV2 or HHV6 at any time throughout the study. The viruses reactivated independently from each other [34].

Table 1.1: Number of viral copies per species collected from 23 astronauts before, during and after a long-duration mission to the ISS. The values presented are the highest copy number of the four samples taken at each time point for salivary EBV, salivary VZV and urinary CMV.
Notes: S – Subject, L – Launch of spaceflight, R – Return of spaceflight, NA – Missing.

S	EBV							VZV							CMV				
	Before launch		During flight			After return		Before launch		During flight			After return		Before launch		During flight	After return	
	L−180	L−45	Early	Mid	Late	R+0	R+30	L−180	L−45	Early	Mid	Late	R+0	R+30	L−180	L−45	During	R+0	R+30
1	0	0	0	640	NA	128	0	0	0	368	0	NA	0	0	0	0	0	0	0
2	0	0	0	0	630	88	0	0	0	45	0	816	660	606	0	0	450	0	0
3	0	0	0	0	NA	98	0	0	0	0	0	NA	0	0	0	0	300	0	0
4	0	0	0	450	770	65	0	0	0	816	0	482	220	0	0	0	250	40	0
5	87	0	0	0	321	70	0	0	0	60	0	1300	0	0	0	48	50	90	0
6	0	150	0	0	0	71	0	0	0	61	0	480	180	0	0	136	0	120	89
7	0	34	0	0	0	102	0	0	0	0	0	0	0	0	0	45	120	70	0
8	0	110	100	0	NA	0	126	0	0	130	380	560	0	0	0	345	0	0	0
9	89	0	0	1020	1215	0	120	0	0	0	0	NA	0	0	0	0	400	0	0
10	0	46	0	0	687	0	117	0	0	0	0	0	0	0	0	0	0	0	0
11	0	143	0	0	814	0	390	0	0	0	370	570	0	0	0	0	350	0	0
12	0	98	0	0	900	0	0	0	0	200	290	110	0	0	0	0	0	0	0
13	66	0	0	0	0	0	0	0	0	0	0	NA	60	0	0	0	0	0	0
14	0	68	NA	0	400	496	0	0	0	NA	0	69	125	0	0	0	560	0	0
15	0	0	NA	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
16	0	0	NA	498	0	0	0	0	0	NA	0	120	300	0	0	0	0	0	0
17	0	81	0	0	0	0	0	0	0	160	180	150	230	0	0	57	378	0	0
18	0	0	0	0	400	0	0	0	0	0	156	0	155	0	0	0	0	0	0
19	0	0	0	0	0	350	0	0	0	0	276	150	0	0	0	0	0	0	0
20	0	0	0	490	0	0	0	0	0	280	340	356	200	0	0	56	467	0	340
21	0	55	0	NA	0	0	0	0	0	0	NA	414	120	63	0	0	0	0	0
22	0	50	NA	0	0	0	0	0	0	NA	0	0	0	0	0	60	80	40	0
23	0	0	NA	0	0	0	487	0	0	NA	0	0	0	0	0	0	790	450	0

This study was observational and exploratory, making use of all data from participating astronauts, and as such was not designed to achieve any particular level of power for detecting pre-specified effects [34].

The goal of this study (Mehta et al., 2017) [34] was to determine whether the astronauts participating in a long duration mission to the ISS would get accustomed to the space conditions and hence this adaptation of their immune system could mitigate the reactivation of these viruses in

the later stages of the spaceflight. After the analysis, it was possible to see that the exact opposite occurred and that an increase was detected in the proportion and amplitude of reactivation of the viruses [34].

1.2 The data

The data set consists of the number of viral copies measured for each virus and time point (Table 1.1). Note that the value 0 could mean that reactivation did not occur or that the viral count was below the detection limit, hence, not detected. A visual inspection of the data set shows a high abundance of zeros which suggests a relatively low frequency of viral shedding. The time points by which the viral reactivation were measured at different moments: two before, three during and two after the flight. Except for CMV because there was only one sample of urine collected from each astronaut during the flight so there are only 5 time points in study for this virus. Before the mission, the times of study were 180 and 45 days before the launch, the time points during the mission were stated as *Early* (about 14 days after launch), *Mid* (between mission days 60–120) and *Late* (about 180 days of mission), and the last two time points were 3 hours after the landing and 30 days after the landing. For CMV the during flight measurement (*During*) was collected at 60–120 flight days. For each moment in study, four samples of saliva were collected and the recorded number of viral copies refers to the highest of the counts. For the two crew members with 60–90 days of mission only two samplings were taken during flight, with data aligned with the *Early* and *Mid* for the 180–days crew members.

Table 1.2: Detected reactivation status of the 23 astronauts before, during and after a long-duration mission to the ISS. The value 1 represents detected viral shedding and 0 the opposite.
Notes: S – Subject, L – Launch of spaceflight, R – Return of spaceflight, NA – Missing.

S	EBV							VZV							CMV				
	Before launch		During flight			After return		Before launch		During flight			After return		Before launch		During flight	After return	
	L−180	L−45	Early	Mid	Late	R+0	R+30	L−180	L−45	Early	Mid	Late	R+0	R+30	L−180	L−45	During	R+0	R+30
1	0	0	0	1	NA	1	0	0	0	1	0	NA	0	0	0	0	0	0	0
2	0	0	0	0	1	1	0	0	0	1	0	1	1	1	0	0	1	0	0
3	0	0	0	0	NA	1	0	0	0	0	0	NA	0	0	0	0	1	0	0
4	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0	1	1	0
5	1	0	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	1	0
6	0	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	1	1
7	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0
8	0	1	1	0	NA	0	1	0	0	1	1	1	0	0	0	1	0	0	0
9	1	0	0	1	1	0	1	0	0	0	0	NA	0	0	0	0	1	0	0
10	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
11	0	1	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0
12	0	1	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0
13	1	0	0	0	0	0	0	0	0	0	0	NA	1	0	0	0	0	0	0
14	0	1	NA	0	1	1	0	0	0	NA	0	1	1	0	0	0	1	0	0
15	0	0	NA	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
16	0	0	NA	1	0	0	0	0	0	NA	0	1	1	0	0	0	0	0	0
17	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	0	0
18	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0
19	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
20	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	1	1	0	1
21	0	1	0	NA	0	0	0	0	0	0	NA	1	1	1	0	0	0	0	0
22	0	1	NA	0	0	0	0	0	0	NA	0	0	0	0	0	1	1	1	0
23	0	0	NA	0	0	0	1	0	0	NA	0	0	0	0	0	0	1	1	0

For EBV, reactivation was observed in every moment in study, while for VZV, there was no evidence of reactivation found in the two before the flight time points. The reactivation of CMV was observed in all time points with the exception of the first one. At first glance there appears

to be an increased number of copies detected during the mission when compared to before and after the flight. It is also worth mentioning that, for EBV and VZV, there are 5.59% and 6.21% inflight measurements that are missing, respectively. For the entire data set, the percentage of missingness is 4.35%. There are three difficulties associated with this data set: the presence of missing data, the large number of zeros and the small sample size.

1.3 Objectives and outline

The original study was observational, exploratory and focused on analysing the prevalence of the viruses at different time points without analysing the missing data or modelling the data of the different viruses. The objective of the current project is to extend this analysis, using the data shared within the respective publication (Table 1.1). The count data of the original study offered many difficulties, hence, for this project, the data used were the binary data for simplicity (Table 1.2). Different statistical approaches were applied to the data to accomplish this such as McNemar's test, confidence intervals for binomial proportions, linear models for categorical data, multiple imputation for missing data and logistic regression mixed models (LRMM). The exploratory analysis made to the viruses' data focused on analysing the proportion of viral reactivation and the average viral counts given that reactivation had occurred. McNemar's test was used to assess statistical differences in the reactivation dynamics between two time points as complementary of Friedman's test used in Mehta et al. (2017) [34]. For the viruses with missing responses, the missingness mechanism present in the data was studied and a test of homogeneity of marginal probabilities was made. Missing data can introduce a substantial amount of bias if the data are not missing completely at random, make the handling and analysis of the data more arduous, and create reductions in efficiency [35], so data imputation was performed on missing observations using multivariate imputation by chained equations (MICE). Data imputation is the process of replacing missing data with substituted values. Several methods of confidence intervals for binomial proportions were used to study the probabilities of viral reactivation. Lastly, generalised linear mixed models (GLMM), in particular LRMM, were fitted to the data for the complete case scenario (CMV) and for the imputed data sets (EBV and VZV) with interest of estimating the probability of viral reactivation for each time point.

The following chapter will focus on the statistical theory and methods used (Chapter 2). Next, the exploratory analysis is made, McNemar's test is applied and LRMMs are fitted for CMV (Chapter 3). Afterwards, the exploratory analysis, the complete case analysis, the categorical data analysis with missing responses and data imputation are presented for EBV and VZV (Chapter 4). These viruses are analysed separately from CMV because they exhibit missing responses. Ultimately, the different methodologies and results are discussed and commented (Chapter 5). All statistical analyses and inference tests were done in the software R, version 4.0.1.

Chapter 2

Materials and methods

The statistical approaches used throughout this project are introduced in this chapter. In Section 2.1 the methods explained are performed to the complete categorical data. In this section the statistical methods are the McNemar's exact binomial test, different methods of confidence interval for binomial proportions and the application of LRMMs. In Section 2.2 the statistical methods applied to the incomplete categorical data are explained. This section is divided into analysis without imputation and analysis with imputation. The first one describes the procedures used to assess the missingness mechanism present in the data and significant differences between viral reactivation dynamics for the inflight time points, and the second describes a method of imputation for the categorical data and an evaluation process of the imputation.

2.1 Analysis of complete categorical data

For this section, the methods presented were applied to EBV and VZV data without the missing responses, and to the CMV data.

2.1.1 McNemar's test

To compare the differences in viral shedding between time points, McNemar's test was used. This is a symmetry test used on paired nominal data and it is applied to 2×2 contingency tables, to determine whether the row and column marginal probabilities are equal, that is, whether there is marginal homogeneity.

Contingency tables were built for all the viruses with interest of testing the consistency of every pair of time points. These contingency tables used the binary data where 1 represents detected reactivation and 0 where there is no evidence of reactivation. If an astronaut had a missing value they were not considered for the tables where the missingness occurred, so the frequencies in the tables do not sum up to 23 for the pairs with at least one inflight time point. All the contingency tables built follow the shape represented in Table 2.1 as an example.

Table 2.1: Generic format of the contingency table used between any two time points of a virus to assess their marginal frequencies, excluding the missing responses.

		Time point 2		
		0	1	
Time point 1	0	n_{00}	n_{01}	$n_{00}+n_{01}$
	1	n_{10}	n_{11}	$n_{10}+n_{11}$
		$n_{00}+n_{10}$	$n_{01}+n_{11}$	N_c

In Table 2.1, n_{00}, n_{01}, n_{10} and n_{11} represent the observed frequencies of their respective combination of responses from both time points. The sampling model generating the generic data from Table 2.1 is the multinomial distribution with total sample size N_c and probability vector $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$. The symmetry hypothesis implies the same marginal probabilities. The hypothesis in study are:

$$H_0 : \pi_{00} + \pi_{01} = \pi_{00} + \pi_{10} \ \textit{vs.}\ H_1 : \pi_{00} + \pi_{01} \neq \pi_{00} + \pi_{10} \ ,$$

which is equivalent to

$$H_0 : \pi_{01} = \pi_{10} \ \textit{vs.}\ H_1 : \pi_{01} \neq \pi_{10} \ ,$$

where π_{01} represents the probability of having no reactivation in the first time point and having detected reactivation in the second, and π_{10} represents the probability of having detected reactivation in the first time point and having no reactivation in the second. Note that $\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1$.

By reformulating the hypothesis it becomes:

$$\begin{aligned} H_0 : \frac{\pi_{01}}{\pi_{01}+\pi_{10}} = \frac{\pi_{10}}{\pi_{01}+\pi_{10}} \ \textit{vs.}\ H_1 : \frac{\pi_{01}}{\pi_{01}+\pi_{10}} \neq \frac{\pi_{10}}{\pi_{01}+\pi_{10}} \\ H_0 : \frac{\pi_{01}}{\pi_{01}+\pi_{10}} = 1 - \frac{\pi_{01}}{\pi_{01}+\pi_{10}} \ \textit{vs.}\ H_1 : \frac{\pi_{01}}{\pi_{01}+\pi_{10}} \neq 1 - \frac{\pi_{01}}{\pi_{01}+\pi_{10}} \ , \end{aligned}$$

which can be written as

$$H_0 : \gamma = \frac{1}{2} \ \textit{vs.}\ H_1 : \gamma \neq \frac{1}{2} \ ,$$

where $\gamma = \frac{\pi_{01}}{\pi_{01}+\pi_{10}}$.

In order to test if the detected reactivation is significantly different for the time points in study, only the number of discordant pairs of time points were used, n_{01} and n_{10}, since the other pairs of time points are not necessary to find whether there is detected differences between time points or not [36].

The way the test was applied for these data was through the exact p-value. In this method an exact binomial test can then be used. This method is not usually used because it involves higher computational effort, nevertheless since the data set is small with only 23 subjects, it is better than the alternative asymptotic test.

Let n_{01} conditional on $n_{01} + n_{10}$, represent the test statistic for this test. By conditioning on $n_{01} + n_{10}$, note that, under the null hypothesis, $n_{01} \sim Bin\,(n_{01} + n_{10}, \gamma)$. The probability of observing n_{01} pairs, conditional on $n_{01} + n_{10}$ discordant pairs and the probability parameter

$\gamma = \frac{1}{2}$, under the null hypothesis, is given by

$$f(n_{01}|n_{01}+n_{10},\gamma=\frac{1}{2}) = \binom{n_{01}+n_{10}}{n_{01}}\left(\frac{1}{2}\right)^{n_{01}}\left(1-\frac{1}{2}\right)^{n_{10}} = 2^{-(n_{01}+n_{10})}\binom{n_{01}+n_{10}}{n_{01}} \ . \quad (2.1)$$

The exact McNemar two sided p-value is defined as

$$\text{p-value} = \min\left(1, 2\times\min\left(F(n_{01}),\overline{F}(n_{01})\right)\right) \ , \quad (2.2)$$

where

$$F(n_{01}) = \sum_{i=0}^{n_{01}} f(i\,|n_{01}+n_{10},\gamma=\frac{1}{2}) \quad (2.3)$$

and

$$\overline{F}(n_{01}) = 1 - F(n_{01}-1) \ . \quad (2.4)$$

Given $n_{01}+n_{10}$, the critical region is defined by the values of n_{01} that provide the expression $2\times\min\left(F(n_{01}),\overline{F}(n_{01})\right) \leq 0.05$.

Effectively, the exact binomial test evaluates the imbalance in the discordant n_{01} and n_{10}. The estimations of the tests for all the pairs of time points were made using the `mcnemarExactDP` function available in the package `exact2x2`, version 1.6.5 for the R software [36].

The usual asymptotic McNemar's test statistic is

$$Q_M(n_{01},n_{10}) = \frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}} \ , \quad (2.5)$$

which for large samples is distributed like a chi-squared distribution with 1 degree of freedom, under the null hypothesis [36].

2.1.2 Confidence intervals for binomial proportions

It is often necessary to obtain an interval estimate for an unknown proportion p, based on binomial sampling. In this project it is interesting to assess proportions and estimate probabilities of viral reactivation. To complement this, confidence intervals are useful to assess uncertainty inside these statistics. However, the usual approximation is known to be poor when the true p is close to zero or to one [37].

For the confidence intervals presented, n represents the sample size, X represents the number of positive viral reactivations, where $0 \leqslant X \leqslant n$. Note that $\hat{p} = X/n$. The estimation of these confidence intervals was made using the package `proportion` version 2.0.0 for the R software [38].

Wald method

The Wald confidence interval is the most basic interval for proportions. This confidence interval is infamous for low coverage in practical scenarios when p is on the extreme side (near to 0 or 1) and/or the sample size (n) is not that large. Both these scenarios occur is these data. Also note that, when $X = 0$ or $X = n$, the Wald interval has zero length. For these reasons other confidence interval methods were considered. The limits of this confidence interval are:

$$\left(\max\left(\frac{X}{n} - z\sqrt{\frac{X}{n^2}\left(1 - \frac{X}{n}\right)}, 0\right), \quad \min\left(\frac{X}{n} + z\sqrt{\frac{X}{n^2}\left(1 - \frac{X}{n}\right)}, 1\right)\right) , \tag{2.6}$$

where z represents $z_{1-\alpha/2}$, which is the $1 - \alpha/2$ percentile of the $\mathrm{N}(0, 1)$ distribution.

Wilson's score method

The Wilson score interval is asymmetric, it does not suffer from problems of zero-width intervals that afflict the Wald confidence interval, and it may be safely employed with small samples and skewed observations [39]. This method is good when n is as low as 10 [37]. The limits of this confidence interval are:

$$\left(\frac{2X + z^2 - z\sqrt{z^2 + 4X(1 - X/n)}}{2\left(n + z^2\right)}, \quad \frac{2X + z^2 + z\sqrt{z^2 + 4X(1 - X/n)}}{2\left(n + z^2\right)}\right) , \tag{2.7}$$

where z represents $z_{1-\alpha/2}$, which is the $1 - \alpha/2$ percentile of the $\mathrm{N}(0, 1)$ distribution.

Clopper-Pearson method

The Clopper-Pearson interval, also known as exact interval, is based on the exact binomial distribution and not on the large sample normal approximation like of the Wald interval. This interval is conservative in the sense that its estimates are likely to be wider than other confidence interval methods. The limits of this confidence interval are:

$$\left(\begin{cases} 0 & , \ X = 0 \\ (\alpha/2)^{1/n} & , \ X = n \\ B_{X,n-X+1;\alpha/2} & , \ \text{otherwise} \end{cases}, \quad \begin{cases} 1 - (\alpha/2)^{1/n} & , \ X = 0 \\ 1 & , \ X = n \\ B_{X+1,n-X;1-\alpha/2} & , \ \text{otherwise} \end{cases}\right) , \tag{2.8}$$

where $B_{\theta_1,\theta_2;\gamma}$ is γ percentile of the $\mathrm{Beta}(\theta_1, \theta_2)$ distribution.

Arcsine method

The arcsine transformation maintains its good behaviour for n as low as 10 [37]. This method might be problematic when p is close to 0 or 1. The limits of this confidence interval are:

$$\left(\begin{cases} 0 & , \quad X = 0 \\ \sin^2\left(\arcsin\sqrt{\frac{X}{n}} - \frac{z}{2\sqrt{n}}\right) & , \quad \text{otherwise} \end{cases} , \right.$$
$$\left. \begin{cases} 1 & , \quad X = n \\ \sin^2\left(\arcsin\sqrt{\frac{X}{n}} + \frac{z}{2\sqrt{n}}\right) & , \quad \text{otherwise} \end{cases} \right) , \quad (2.9)$$

where z represents $z_{1-\alpha/2}$, which is the $1-\alpha/2$ percentile of the $\mathrm{N}(0,1)$ distribution.

2.1.3 Logistic regression mixed model

A logistic regression mixed model (LRMM) is a model in which the linear predictor contains random effects in addition to the usual fixed effects [40]. Mixed models are applied in many disciplines where multiple correlated measurements are made on each unit of interest. The fixed effects are constant across individuals while random effects vary [41].

Let the random variable X_t describe the frequency of individuals with detected viral reactivation at time point t, among the sampled. This binomial distribution is included in the exponential family of distributions, with parameters n_t and $p_t = E\left(\frac{X_t}{n_t}\right)$, describing the total number of astronauts at time point t and the probability of an astronaut having viral reactivation at that time point, respectively.

A GLMM allows to build a linear relationship between the response and predictors, even though their underlying relationship is not linear. This is made possible by using a link function, which links the response variable to a linear model. The link function used in this model is the logit function and is defined by $g(p_t) = \ln\left(\frac{p_t}{1-p_t}\right)$. The purpose of the logit link function is to take a linear combination of the covariate values (which may take any value between $\pm\infty$) and convert those values to the scale of a probability, i.e., between 0 and 1 [42].

The response variable of the model is *Reactivation* – whether reactivation was detected or not – this variable is a factor with two levels: 0 and 1. The covariate *Subject* is a factor with 23 levels and the covariate *Time point* is a factor with five levels for CMV and seven levels for EBV and VZV. This variable was modelled by creating dummy variables associated with each time point present in the LRMM that took the value 1 when the probability being estimated referred to their respective time point and 0 otherwise. In the models created, the variable *Time point* represents the fixed effects while the variable *Subject* explains the random effect associated with each astronaut.

A LRMM seems adequate to explain the reactivation status as a function of the time point. The

corresponding LRMM is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1 x_{1a} + ... + \beta_K x_{Ka}\,, \quad t = 0, ..., K,\ a = 1, ..., 23\ , \tag{2.10}$$

where p_{ta} is the probability of viral reactivation at the t-th time point by the a-th astronaut, $\beta_0, ..., \beta_K$ are the unknown coefficients, with $\beta_1, ..., \beta_K$ being the regression coefficients associated with the levels $x_{1a}, ..., x_{Ka}$ of the covariate *Time point*, respectively, and β_0 representing the intercept as the effect of the baseline time point. The parameter K represents the number of coefficients for each model. While for CMV $K = 4$, for EBV and VZV $K = 6$. The A_{0a} parameter represents the random effect associated with subject a. Note that $A_{0a} \sim N\left(0, \tau^2\right)$.

The fitting of these models was made using the package `lme4` version 1.1-26 for the R software [43].

2.2 Analysis of incomplete categorical data

The methods presented in this section were only applied to the EBV and VZV data sets, since CMV has no missing data.

2.2.1 Analysis without data imputation

Missingness mechanism

Missing data can introduce potential bias in parameter estimation and weaken the generalisability of the results [44]. Ignoring cases with missing data leads to the loss of information which in turn decreases statistical power and increases standard errors [45].

When considering the potential impact of the missing data, it is interesting to understand the mechanism by which the data are missing and the underlying reasons for why the data are missing. Missing data mechanisms are typically grouped into three categories [46]:

MCAR: Missing completely at random. When data are MCAR, the fact that the data are missing is independent of the observed and unobserved data [47]. In other words, no systematic differences exist between participants with missing data and those with complete data.

MAR: Missing at random. When data are MAR, the fact that the data are missing is systematically related to the observed yet not the unobserved data [47].

MNAR: Missing not at random. When data are MNAR, the fact that the data are missing is systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured or controlled by the researcher.

Multiple imputation assumes the data are at least MAR, so while MCAR and MAR mechanisms are ignorable for the analysis, MNAR cannot be ignored in performing longitudinal data analysis [48].

Admit that the random vector $\mathrm{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_t)'$ of response variables can assume R values, corresponding to combinations of the levels of its components. In the current case, only the three inflight time points were considered, hence $\mathrm{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, \mathrm{Y}_3)'$ and each time point may assume two different values, so $R = 2 \times 2 \times 2 = 8$. For this study, the viruses were individually studied with only one subpopulation each. The R response categories are indexed by r, and the S subpopulations, by s. For subpopulation s, F_s defines missingness patterns as follows: the set of units with no missing data (i.e., with complete classification) is indexed by $f = 1$ and the sets that have some degree of missingness, by $f = 2, ..., F_s$. Also assume that each unit with the f-th missing pattern is recorded in one of R_{sf} response classes C_{sfc}, $c = 1, ..., R_{sf}$. The total number of response classes for units with some missingness pattern in the s-th subpopulation is represented by $l_s = \sum_{f=2}^{F_s} R_{sf}$ [49]. The vector $\mathrm{N}_s = (\mathrm{N}'_{sf}, f = 1, ..., F_s)'$ encloses all the observed frequencies corresponding to the s-th subpopulation [49].

Let $\mathrm{Z}_s = (\mathrm{Z}_{sf}, f = 1, ..., F_s)$ denote an $R \times (R + l_s)$ matrix corresponding, columnwise, the indicator vectors to all response classes for units with all missingness patterns in the s-th subpopulation [49]. In other words, each element of each column vector takes the value 1 if there is correspondence between its responses in the column and row. The responses in each row correspond to the complete response classes and for the columns there are the complete and incomplete responses.

Let $\pi_{r(s)}$ be the marginal probability that a unit selected at random from the s-th subpopulation is classified in the r-th response category and $\hat{\pi}$ represent the estimates of the marginal probabilities under the MCAR hypothesis. Goodness-of-fit tests for the MCAR mechanism, conditionally on the MAR assumption and considering there is only one subpopulation ($S = 1$), can be obtained either from Wilks' likelihood ratio statistic

$$Q_L(\mathrm{MCAR}|\mathrm{MAR}) = -2 \sum_{f=1}^{F} \sum_{c=1}^{R_f} n_{fc} \left[\ln(\mathrm{z}'_{fc} \widehat{\pi}) - \ln \left(\frac{n_{fc}}{n_{f.}} \right) \right] , \tag{2.11}$$

or from the Pearson (Q_P) statistics

$$Q_P(\mathrm{MCAR}|\mathrm{MAR}) = \sum_{f=1}^{F} \sum_{c=1}^{R_f} \frac{\left(n_{fc} - n_{f.}\, \mathrm{z}'_{fc} \hat{\pi} \right)^2}{n_{f.}\, \mathrm{z}'_{fc} \hat{\pi}} . \tag{2.12}$$

Under the MCAR hypothesis, both statistics follow an asymptotic $\chi^2_{(g)}$ distribution, with $g = S + \sum_{s=1}^{S} (l_s - F_s)$ degrees of freedom [49]. The application of these methods was done using the package `ACD` version 1.5.3 for the R software [50].

Test of homogeneity of marginal probabilities

Consider fitting a model based on functions of the marginal probabilities of categorisation. In this context, the linear model is expressed as

$$M : \mathrm{J}\pi = \mathrm{W}\beta \ , \tag{2.13}$$

where J is a $v \times SR$ matrix defining the v linear functions of interest, W is a $v \times k$ model specification matrix and $\beta = (\beta_1, ..., \beta_k)'$ is a $k \times 1$ vector that contains the unknown coefficients [49].

If the parameter of interest is π, the likelihood ratio statistic for the goodness-of-fit test of model M under the assumption MCAR and considering there is only one subpopulation ($S = 1$) is

$$Q_L(M|\mathrm{MCAR}) = -2 \sum_{f=1}^{F} \sum_{c=1}^{R_f} n_{fc} \left[\ln(\mathrm{z}'_{fc} \hat{\pi}(M)) - \ln(\mathrm{z}'_{fc} \hat{\pi}) \right] \ , \tag{2.14}$$

where $\hat{\pi}(M)$ represents the estimates of the marginal probabilities for model M under the MCAR hypothesis. The Pearson statistic for the same model is

$$Q_P(M|\mathrm{MCAR}) = \sum_{f=1}^{F} \sum_{c=1}^{R_f} \frac{\left(n_{f.} \, \mathrm{z}'_{fc} \hat{\pi} - n_{f.} \, \mathrm{z}'_{fc} \hat{\pi}(M) \right)^2}{n_{f.} \, \mathrm{z}'_{fc} \hat{\pi}(M)} \ . \tag{2.15}$$

The p-values for the goodness-of-fit model tests were estimated using the function `linML` of the package `ACD` for the R software [50].

2.2.2 Analysis with data imputation

Missing data may seriously compromise inferences from small data sets, especially if missing data are ignored. Most statistical and computational methods are not shaped to handle missing data, and data imputation offers a good alternative to this problem. Missing data that occur in multiple variables present a special challenge. The process of specifying the imputation model is a scientific modelling activity on its own, that comes with its own model building principles [51]. The imputation model is conditional on the type of incomplete variable to be imputed. The potential bias due to missing data depends on the mechanism causing the data to be missing. In addition, most statistical procedures are designed for complete data [52], so it is relevant to analyse the missing data.

Data imputation methods are divided into two categories: single imputation and multiple imputation. Common single imputation methods for quantitative data are to replace missing data with the mean or median of each variable. For qualitative variables the missing values could be imputed with the mode of the variable or by values from similar subjects from another data set. Single imputation does not take into account the uncertainty in the imputations. Multiple imputation creates multiple copies of the data set, with the missing values replaced by imputed values. It aims to allow for the uncertainty about the missing data by creating several different

plausible imputed data sets and appropriately combining results obtained from each of them. Multiple imputation is the method of choice for complex incomplete data problems [44, 53].

A general approach for imputing multivariate data has emerged: fully conditional specification, also known as MICE [54]. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, MICE draws imputations by iterating over the conditional densities [54]. In other words, this method starts by creating m imputation chains (m to be defined by the user) which are iteratively updated from a set of initial guesses for the missing data until the convergence of a chosen statistic is attained. At the end of the imputation procedure, the m collections of imputed data sets are analysed using the appropriate methods for complete data and the subsequent estimates of the quantities of interest are pooled together and the respective uncertainty estimated.

Rubin's method

In order to deal with the problem of increased uncertainty due to imputation, Rubin [44] developed a method for averaging the outcomes across multiple imputed data sets to account for this. All multiple imputation methods follow three steps:

Imputation: Similar to single imputation, missing values are imputed. However, the values are imputed m times rather than just once. At the end of this step, there should be m completed data sets.

Analysis: Each of the m data sets is analysed by complete data methods. At the end of this step there should be m analyses and estimates of the parameters of interest.

Pooling: The m estimates are consolidated into one result by estimating the mean, variance and confidence interval of the parameter of interest [55, 56].

For this project, the data imputation was made using the MICE method applied to the binary data set. In this particular case, the MICE algorithm creates a different logistic regression model at each iteration where the missing values of the dependent variable are replaced with new predictions and the independent variables are all the others. This process is repeated for all variables with missing responses. After convergence of the algorithm had been met and the m imputations estimated, the probability of viral reactivation and their respective confidence intervals were estimated for each imputation. To assess a definite estimate for the probability of viral reactivation there is need to pool the individual estimates of all imputations to a single estimate. This step results in statistically valid estimates that translate the uncertainty caused by the missing data into the width of the confidence interval [51]. To this end, let the probability of reactivation at time point t be represented by p_t. These probabilities of reactivation were estimated by the mean of the respective postimputation estimates, that is,

$$\overline{p_t} = \sum_{i=1}^{m} \widehat{p_{ti}}/m, \quad t = 3, 4, 5 \ , \tag{2.16}$$

where $\widehat{p_{ti}}$ is the estimate of p_t using the i-th imputed data set and m represents the number of imputations of each virus. The associated standard errors were given by

$$\text{SE}(\overline{p_t}) = \sqrt{\frac{\sum_{i=1}^{m} \text{var}(\widehat{p_{ti}})}{m} + \frac{m+1}{m} \times \frac{\sum_{i=1}^{m} (\widehat{p_{ti}} - \overline{p_t})^2}{m-1}} \ , \tag{2.17}$$

where $\text{var}(\widehat{p_{ti}}) = \frac{1}{n_{mt}-1} \sum_{d=1}^{n_{mt}} (b_{tid} - \overline{b_{ti}})^2$, where n_{mt} represents the number of missing responses at time point t, b_{tid} represents the viral reactivations index (1 if there is reactivation and 0 if there is not) of the d-th imputed response at time point t and imputation i, and $\overline{b_{ti}}$ represents the proportion of estimated reactivation for the n_{mt} missing responses at time point t and the i-th imputation.

Logistic regression mixed model

A LRMM was fitted to include both fixed and random effects for each of the m imputed data sets. These models follow the same structure of the complete case scenario presented in subsection 2.1.3. The fixed parameters are the time points where the samples were collected. The grouping variable (random effect) is *Subject*, since each subject might have a different response to the stressful conditions and affect the results.

After the fitting of all models, there is need to pool all the results of the imputations into one LRMM with the same characteristics as the models that constitute it and studying the viral reactivation dynamic for each virus [44]. The coefficients of the pooled model were estimated by averaging all the coefficients of the m models, that is,

$$\overline{\beta_k} = \sum_{i=1}^{m} \widehat{\beta_{ki}}/m, \quad k = 0, ..., 6 \ , \tag{2.18}$$

where $\widehat{\beta_{ki}}$ represents the estimate of coefficient β_k using the i-th imputed data set at the t-th time point. The associated standard errors were given by

$$\text{SE}(\overline{\beta_k}) = \sqrt{\frac{\sum_{i=1}^{m} \text{var}(\widehat{\beta_{ki}})}{m} + \frac{m+1}{m} \times \frac{\sum_{i=1}^{m} (\widehat{\beta_{ki}} - \overline{\beta_k})^2}{m-1}} \ . \tag{2.19}$$

The random parameters follow a normal distribution with mean zero and a variance τ_i^2. The mean of these distributions keeps constant at zero after pooling, on the other hand the standard deviation was given by

$$\overline{\tau} = \sqrt{\frac{\sum_{i=1}^{m} \tau_i^2}{m}} \ . \tag{2.20}$$

Chapter 3

Analysis of cytomegalovirus

The data from CMV were collected from samples of urine. During the flight only one sample was collected and it was collected 2–4 months after launch, so the measurements were made in only five time points as opposed to the seven of the other viruses. The analysis of CMV throughout this chapter starts with the exploratory analysis of the number of viral copies and the proportions of viral reactivation in Section 3.1. In Section 3.2, McNemar's exact test is applied to assess significant differences between the time points in study. Finally, in Section 3.3, LRMMs are fitted to the binary data to predict the probabilities of reactivation of this virus. The analysis of CMV differs from the other viruses because the data of this virus have no missing responses. The significance level considered throughout this chapter is 5%.

3.1 Exploratory Analysis

The exploratory analysis presented uses results from Mehta et al. [34] and new results inferred from the data. The objective of the analysis was to detect patterns about the proportions and amplitude of the viral reactivation. The analysis started with a simple plot expressing the numbers of detected viral copies over time. The spacing between the time points in the axis of the next plots represent the time passed between the measures.

In the Figure 3.1, each line represents one astronaut and the values each point takes are the number of detected viral copies of CMV along the time points, for all individuals. It is easy to see that the inflight time point *During* has much higher values of detected viral copies than all the other time points. Both the *During* adjacent and the *R+30* time points have subjects with big values of viral copies observed.

Since this is a zero-inflated data set, a plot like this does not dispose much information outside the amplitude of the highest viral count values. To better interpret the count values for CMV let us look at the plots in Figures 3.2 and 3.3. The first shows the median and average of the positive numbers of viral copies for all individuals with detected viral reactivation. Since there was no detected viral reactivation for any astronaut at the moment *L−180* the median and

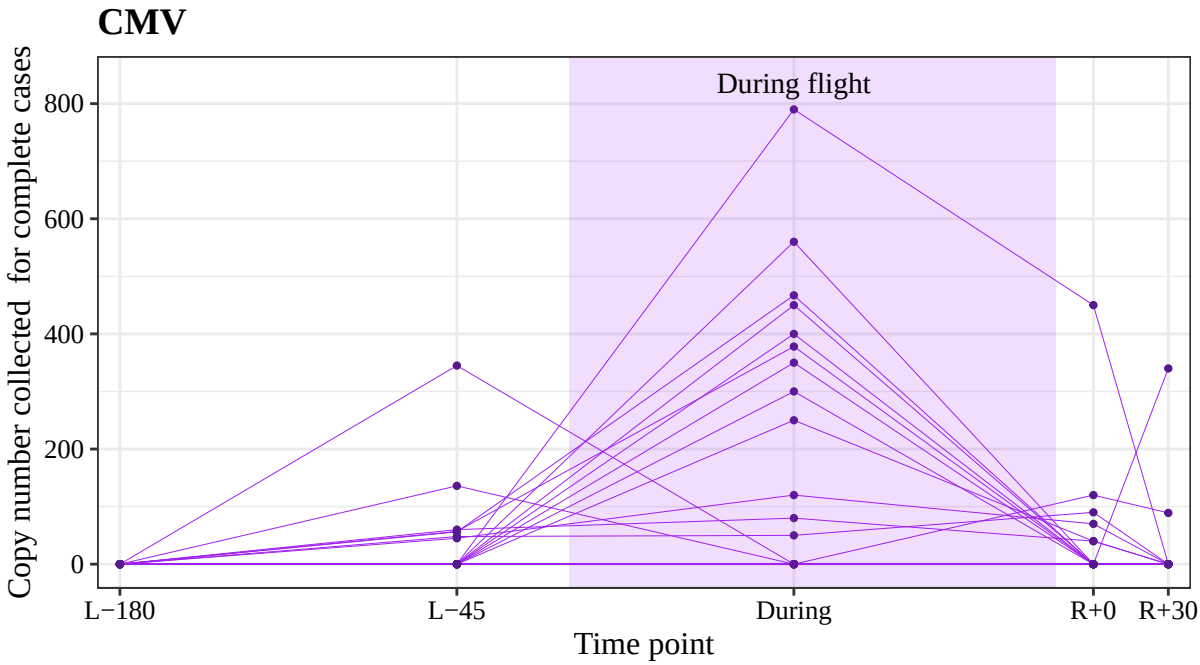

Figure 3.1: Graphical representation of the viral copy number for CMV over time. Each line represents one subject.

average of the positive viral counts here are not represented. The *During* adjacent time points have very similar average and median values of the number of detected copies of this virus. The second largest average of positive numbers of viral copies happens at the last moment measured. The time point with a higher average of positive viral copies was *During*, so no mitigation of shedding appears to occur for the inflight measurements. The median of the numbers of positive viral copies are close to the average in all time points, suggesting that their distributions are symmetric. The quartiles represented show that the *During* time point has a big dispersion for the number of positive viral copies, while all the others have a relatively small one.

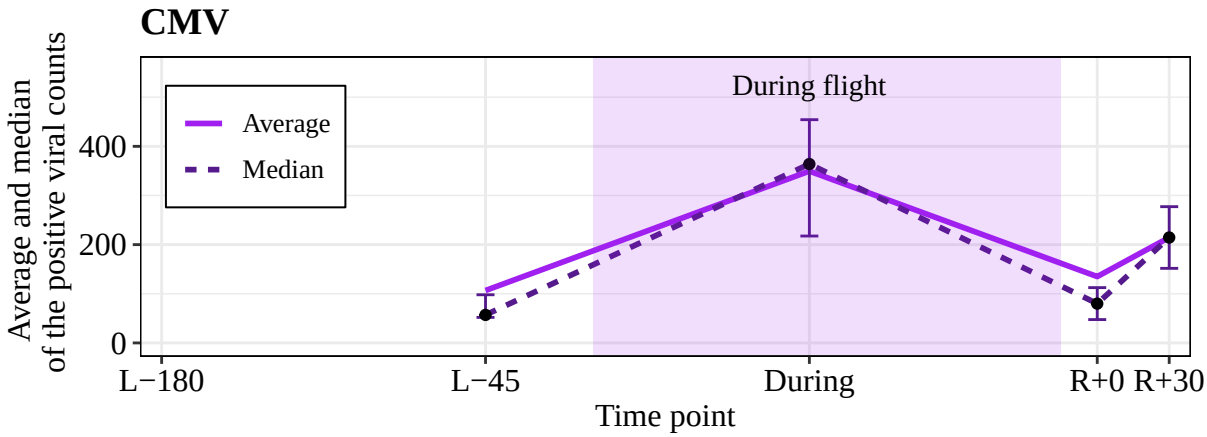

Figure 3.2: Representation of the average and median of the values of positive viral counts for CMV over time. Represented along the median are the dispersion measures of the first and third quartiles.

In Table 3.1 and Figure 3.3 are represented the proportions of reactivation and the binomial 95% confidence intervals using the Wilson's score method of the viral shedding probabilities at all time points for all astronauts. This confidence interval method is preferred over the usual

Wald's method when the observed proportions are close to 0 or 1 [37]. For the time point *L−180* there was no detected reactivation for any astronaut while for the other before the flight time point, *L−45*, 7 subjects had detectable reactivation. During the flight there was only one measurement taken and there were 12 detected reactivations, this is a significant value because more than half of the astronauts had detectable reactivation. For the time points after the return the number of detected reactivations decreased. For *R+0* the number of reactivations was 6 and for *R+30* that number diminished to 2. The 95% confidence intervals of the probabilities of viral reactivation of time points *L−180* and *L−45* do not intercept, which indicates that there are significant differences between the probabilities of viral shedding of these time points. The same happens for time points *L−180* with *During* and *During* with *R+30*. The probability of viral reactivation of time point *During* is statistically different from the probabilities of time points *L−180* and *R+30*, so during the spaceflight the number of reactivations increased compared to these measurements on Earth.

Table 3.1: Observed proportions of viral reactivation and binomial 95% confidence intervals for the theoretical proportions using Wilson's score method for subjects shedding CMV

Time point	Estimates	
	Proportion of reactivation, %	Confidence interval, %
L−180	0.0	(0.0, 14.3)
L−45	30.4	(15.6, 50.9)
During	52.2	(33.0, 70.8)
R+0	26.1	(12.5, 46.5)
R+30	8.7	(2.4, 26.8)

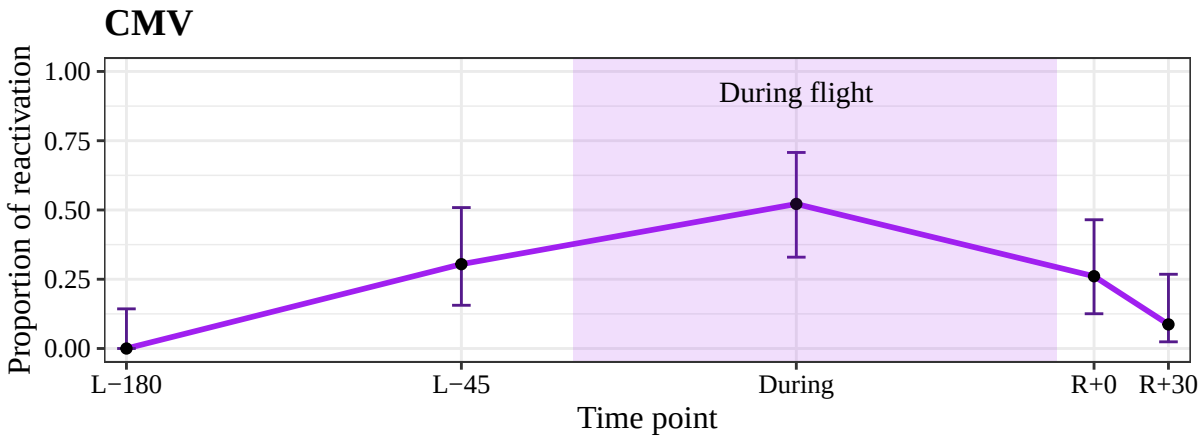

Figure 3.3: Visual representation of the observed proportions of viral reactivation and binomial 95% confidence intervals for the theoretical proportions using Wilson's score method for subjects shedding CMV

For most of the time points the increase or decrease in the viral frequency and amplitude appears to be related, which means that when there is an increase in the average of positive viral counts it can be expected that there was also an increase in the proportion of reactivations. However, this is not true for the time point *R+30* where compared to the other time points it has the second smallest proportion of detected viral reactivation and also has the second highest number for the average the positive viral copies. Viral shedding increased in frequency and amplitude

for the *During* time point when compared with the others.

3.2 Friedman's and McNemar's tests

There was no viral shedding detected at 180 days before flight yet there was viral shedding 45 days before, during and after flight for this virus. In the original study the Friedman test was used to compare the copy numbers between time points. This test is a non-parametric statistical test used for repeated measures analysis of data by ranks [57, 58]. The test was used to assess significant differences in the viral number of copies of all time points at once. The abundance of zeros in the data set causes numerous ties, reducing the power of the test. The Friedman test comparing copy numbers for CMV was significant (p-value<0.0001) [34], so there were evidence that there are significant differences for the reactivations dynamic between time points. Even when the time point with no shedding was excluded from the analysis, there were still significant differences between the remaining time points (p-value=0.0008) [34].

Another useful test to study the reactivation differences between time points is the McNemar's binomial exact test. This test compares the imbalance in the discordant pairs using the binary data where 1 represents detected shedding and 0 the opposite. The results of this statistical test are represented in Table 3.2.

Table 3.2: P-values of the McNemar's test for CMV for all pairs of time points. In parenthesis are the number of discordant pairs, (n_{01},n_{10}), respectively. P-values smaller than the considered significance level are highlighted in bold.

	L−180	L−45	During	R+0	R+30
L−180	–	**0.016** (7,0)	<**0.001** (12,0)	**0.031** (6,0)	0.500 (2,0)
L−45		–	0.180 (7,2)	1.000 (2,3)	0.062 (0,5)
During			–	0.070 (1,7)	**0.006** (1,11)
R+0				–	0.219 (1,5)
R+30					–

The pairs of time points with significant differences in the observed discordant pairs are *L−180* with *L−45*, for *L−180* with *During*, for *L−180* with *R+0* and for *During* with *R+30*. The *During* time point has significant differences with the before flight time point *L−180* and the after flight *R+30*.

Multiple McNemar's tests were made for the same data in this section, this raises the multiple comparisons problem. The Bonferroni correction is a suitable method to address this problem. This correction method works by adjusting the significance level considered in accordance with the number of tests made. The new significance level is calculated by dividing the original significance level by the number of tests, which in this case is $0.05/10 = 0.005$. After this correction the only pair of time points with significant differences in the reactivation is *L−180* and *During*.

Both tests found significant differences in the reactivation dynamics between time points. Friedman's test compared all time points and found them to be significantly different, while McNe-

mar's test individually compared all pairs and found four pairs to be significantly different.

3.3 Logistic regression mixed model

For this virus the estimated probabilities of reactivation are 0, 0.304, 0.522, 0.261 and 0.087 for the time points *L−180*, *L−45*, *During*, *R+0* and *R+30*, respectively. It would be interesting to build a model that could estimate the probability of reactivation given a time point based on the data for CMV. The best option for a model is a LRMM where the fixed effects are the time points and the random variable is the subject as shown in equation (2.10). In this model, the response variable is *Reactivation* – whether reactivation was detected or not – and the explanatory variables are *Subject* and *Time point*. The variable *Time point* is a factor with five levels: *L−180*, *L−45*, *During*, *R+0* and *R+30*. Time point *L−180* is used as the reference level. The equation of the model fitted for CMV is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1 x_{1a} + \beta_2 x_{2a} + \beta_3 x_{3a} + \beta_4 x_{4a} \ , \tag{3.1}$$

where p_{ta} is the probability of viral reactivation at the time point t by astronaut a, $\beta_0, ..., \beta_4$ are the unknown coefficients, with $\beta_1, ..., \beta_4$ being the regression coefficients associated with each time point in the model and β_0 representing the intercept as the effect of the baseline time point *L−180*. The $x_{1a}, ..., x_{4a}$ represent the dummy variables associated with the time points *L−45*, *During*, *R+0* and *R+30*, respectively. The A_{0a} parameter represents the random effect associated with subject a.

The estimated coefficients and their respective significance are represented in Table 3.3. The estimated coefficients of the model are all positive, meaning that they all have a higher probability of reactivation than for the baseline *L−180*. The *During* time point has the biggest parameter estimate and its associated probability estimate of reactivation is also the biggest. The time point *R+30* has the smallest coefficient estimate and also has the smallest probability estimate of viral shedding. The bigger the probability estimate of viral reactivation, the bigger the parameter estimate is going to be. This relation is maintained for all time points.

Table 3.3: Adjusted LRMM fitted to CMV binary data with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *L−180*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*L−180*)	−21.291	213.519	0.921
β_1	(*L−45*)	19.954	213.519	0.926
β_2	(*During*)	21.414	213.520	0.920
β_3	(*R+0*)	19.624	213.519	0.927
β_4	(*R+30*)	17.797	213.520	0.934

All the standard errors are big and because of that all the p-values are close to 1. The reason for this could be the fact that there were no observed viral reactivations at the baseline *L−180*, hence, complete separation could be occurring. By having no detected reactivations at time

point *L−180* the algorithm estimating the model's parameters is not able to converge, resulting in the lack of significance of the coefficients. Since this is the case, it is worth trying to fit a different model where the reference level is the time point *L−45* because viral reactivation was both detected and not detected at this time point. The equation of this new model is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1\, x_{1a} + \beta_2\, x_{2a} + \beta_3\, x_{3a} + \beta_4\, x_{4a}\ , \tag{3.2}$$

where p_{ta} is the probability of viral reactivation at the time point t by astronaut a, $\beta_0, ..., \beta_4$ are the unknown coefficients, with β_0 representing the effect of the intercept *L−45* and $\beta_1, ..., \beta_4$ are the regression coefficients associated with the time points in the model. The $x_{1a}, ..., x_{4a}$ are the dummy variables associated with the time points *L−180*, *During*, *R+0* and *R+30*, respectively. The random effect associated with subject a is represented by the parameter A_{0a}.

Table 3.4: Adjusted LRMM fitted to CMV binary data with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *L−45*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*L−45*)	−1.338	0.749	0.074
β_1	(*L−180*)	−20.147	724.077	0.978
β_2	(*During*)	1.460	0.828	0.078
β_3	(*R+0*)	−0.329	0.815	0.686
β_4	(*R+30*)	−2.157	1.066	0.043

The estimated coefficients of this new model and their respective significance are represented in Table 3.4. Right way, it is possible to see that the standard errors decreased considerably. All the coefficients, except for *During*, are negative, meaning that their respective probabilities of viral reactivation are smaller than for the reference level *L−45*. *During* is the time point with the highest estimated probability of viral reactivation so this is not a surprise. The coefficient *L−180* has the smallest estimated probability of viral reactivation and also has the smallest parameter estimate. This parameter is the only one with a p-value in the order of the previous model because of its large standard error. The only significant parameter in this model is *R+30*, although the intercept *L−45* and the *During* coefficients have p-values in the borderline of the 5% significance level.

It is clear that this model is explaining the data better than the previous one, yet it is still interesting to fit a new model without the coefficient whose standard error of the respective estimate was too large. The equation of this third model is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1\, x_{1a} + \beta_2\, x_{2a} + \beta_3\, x_{3a}\ , \tag{3.3}$$

where p_{ta} is the probability of viral reactivation at the time point t by astronaut a, $\beta_0, ..., \beta_3$ are the unknown coefficients, with $\beta_1, ..., \beta_3$ being the regression coefficients associated with each time point in the model and β_0 representing the effect of the intercept *L−45*. The $x_{1a}, ..., x_{3a}$ represent the dummy variables associated with the time points *During*, *R+0* and *R+30*, respectively. The A_{0a} parameter represents the random effect associated with subject a.

Table 3.5: Adjusted LRMM fitted to CMV binary data excluding time point *L−180* with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *L−45*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*L−45*)	−1.338	0.749	0.074
β_1	(*During*)	1.460	0.828	0.078
β_2	(*R+0*)	−0.329	0.815	0.686
β_3	(*R+30*)	−2.157	1.066	0.043

The estimated coefficients of this model and their respective significance are represented in Table 3.5. Aside from removing the coefficient *L−180*, all parameter estimates, standard errors and p-values stay the same as the model in Table 3.4. In this model, the coefficient *R+30* is the only one that is significant even though the intercept and *During* parameters have p-values close to the significance level of 5%. The parameter *R+0* has a large associated p-value and is not significant, in fact, the probability of viral reactivation of the parameter *R+0* ($p_2 = 0.261$) is close to the probability of viral reactivation of the intercept *L−45* ($p_0 = 0.304$).

Of the three models fitted to the data, the first (Table 3.3) had no significant variables because complete separation might be occurring, hence is not useful for the analysis. The second and third models (Table 3.4 and Table 3.5, respectively) are alike, aside from parameter *L−180*. Being this similar and given that it has less parameters, the model in Table 3.5 appears to be a better choice to fit to these data. For all three models the standard deviation associated to the random effect *Subject* is 1.878.

3.4 Summary

The higher values of the number of viral copies were reported at the *During* time point, although its adjacent and *R+30* time points also had big values observed. Only the first time point, *L−180*, had no astronaut present detectable values of viral copies. The time points *L−45* and *R+0* had similar values for the average of viral copies given that shedding had occurred, and *R+30* had a small increased in that value compared to them. The *During* time point had the biggest number of viral copies detected of all time points. The median and average of the positive viral copies have similar values at all time points, so their distribution appear to be symmetric. For the proportions of viral shedding, the *During* time point has the highest number of astronauts testing positive for reactivation, with a percentage of reactivation of 52.2%. The proportions of viral shedding for *L−45* and *R+0* were estimated to be similar with percentages of 30.4% and 26.1%, respectively, and *R+30* had even fewer reactivations with only 8.7% being observed. For these data there appears to be a direct correlation between the proportion of reactivation and the number of viral copies detected. There are significant differences between the probability of viral reactivation of pairs of time points *L−180* with *L−45*, *L−180* with *During* and *During* with *R+30*.

Friedman's test found significant differences in the number of viral copies for all time points

(p-value<0.0001) [34]. Even after removing the time point with no reactivation, there were still significant differences between the remaining time points (p-value=0.0008) [34]. McNemar's exact binomial test was applied to the binary data and found four pairs with significant differences of reactivation dynamics. The differences were found in pairs of time points *L−180* with *L−45*, *L−180* with *During*, *L−180* with *R+0* and *During* with *R+30*. After the Bonferroni correction, only the pair of time points *L−180* and *During* was significant.

The first LRMM applied to the data, with the time points as fixed effects and the subject as the random effect, had large p-values for all the parameters which suggests that the algorithm estimating the model's coefficients was not converging. The reference level of this model was the parameter *L−180*. The observed reactivation status of this time point were all negative, and because this might be influencing the significance of the model, another model was fitted where the reference level was *L−45*. In this second model the standard errors decreased in general. The only significant parameter was *R+30*, although, the intercept (*L−45*) and *During* parameters were close to the significance level considered. This model was better than the first, yet the coefficient *L−180* still had a very large associated standard error, so a third model was fitted with the same data excluding the time point *L−180*. In this third model all the parameter estimates, standard errors and p-values kept the same values of the previous model. The parameter *R+0* had a large associated standard error and p-value, since the probabilities of viral reactivation for this time point and the reference level *L−45* are close. This model appears to be the better choice to fit to these data, because of the better statistical significance and smaller number of parameters.

Chapter 4

Analysis of Epstein-Barr and varicella-zoster viruses

The data from EBV and VZV were collected from samples of saliva. Both viruses had seven distinct time points of measure and both viruses had missing responses at the three inflight time points. The analysis performed in this chapter uses both the complete observed data and the data with missing responses. The study of these viruses starts with the exploratory analysis of the number of viral copies and the proportions of viral reactivation done individually to EBV and VZV in Section 4.1. In Section 4.2, the analysis is performed ignoring the missing responses in the data. McNemar's exact test is applied to these data to study significant differences between the time points for the data of both viruses. In Section 4.3 the analysis of categorical data is presented, this analysis is made using the binary data with missing responses of both viruses. The missingness mechanism present in the data is studied and a test of homogeneity of marginal probabilities was applied. Ultimately, in Section 4.4, the data imputation of the categorical data using the MICE method is explained, in particular, the procedure for the convergence of the algorithm, the pooling of the data imputations along with the estimates for the probabilities of reactivation and the fitting of LRMMs to the pooled data sets for both viruses. The significance level considered throughout this chapter is 5%.

4.1 Exploratory analysis

The exploratory analysis for these viruses presents results from the original study [34] and new results deduced from the data. The purpose of the exploratory analysis was to study the proportion and amplitude of the viral reactivation. In this section the data that are being considered are only the complete cases, so for a lot the statistics estimated the number of individuals is smaller than the original 23 astronauts.

Epstein-Barr virus

The exploratory analysis of EBV began with the inspection of a plot expressing the numbers of detected viral copies along the time points. In the next plots, the time passed between the measurements is represented with the spacing between the time points in the axis.

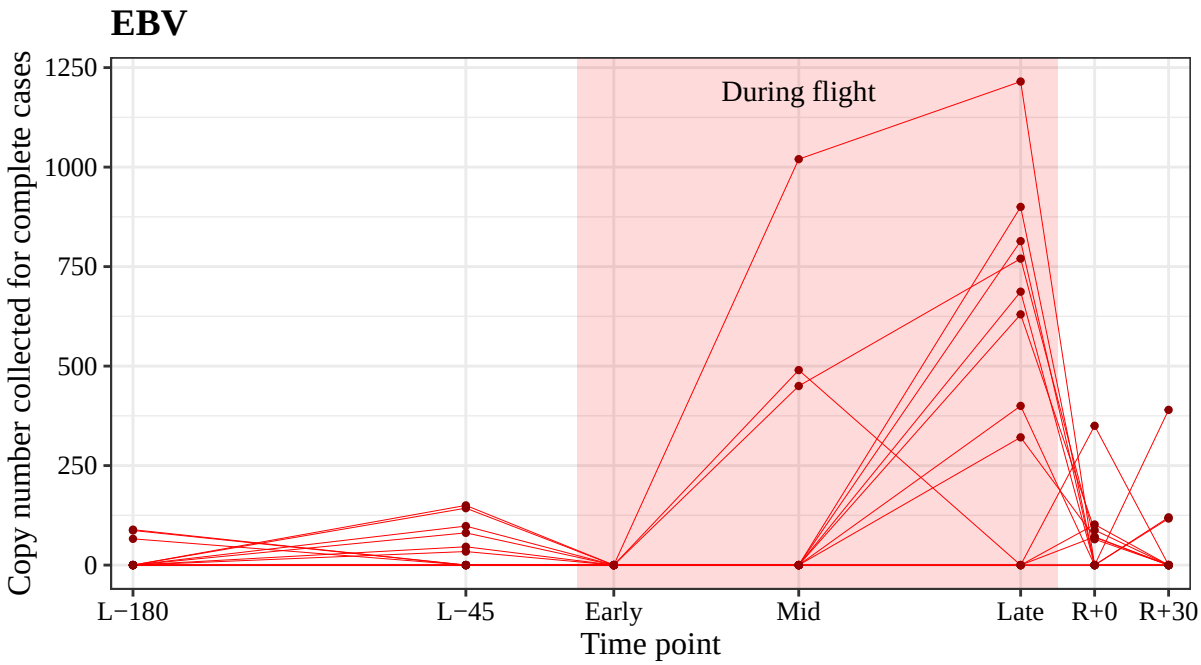

Figure 4.1: Graphical representation of the viral copy number for EBV over time for subjects with no missing data. Each line represents one subject.

In Figure 4.1 each line represents one astronaut and each dot represents the number of detected copies of EBV in each sample along the time points. In this plot only the subjects with complete responses are represented. The before flight time points *L−180* and *L−45* have a very stable number of positive viral copies detected with none of their values exceeding 150 viral counts. For the inflight time points, while the *Early* time point has no detected reactivation the other two, *Mid* and *Late*, have the biggest number of detected viral copies even though there were only three subjects that reactivated in the *Mid* time point. For the *Late* time point the number of occurrences of viral shedding and their amplitude increased compared to the other moments, so right way it appears that the astronauts are not accommodating to the stress conditions and mitigating the stressors effects. The after flight time points *R+0* and *R+30* both have few reactivations and their biggest number of viral copies is around 375.

Because of the difference in amplitude of the observed number of viral copies caused by the large number of zeros, this plot is only useful to assess the behaviour for the higher numbers of viral copies. The plots in Figures 4.2 and 4.3 introduce statistics that help complement the analysis of Figure 4.1. In Figure 4.2, the average of the positive numbers of viral copies for the first three time points is very stable with all their values being around 100 viral copies. The last two inflight time points (*Mid* and *Late*) have the biggest values of the average of positive viral counts with both values between 600 and 700 counts. This high values were expected because of their high amplitude of the number of viral copies. After the return, the average number of detected viral

copies decreases to values around 200 viral copies with the *R+30* time point having an average slightly higher than *R+0*. The median of the positive numbers of viral copies is very close to their average for all time points which indicates that their distributions are symmetric. The dispersion measures keep close to the median except for the *Mid*, *Late* and *R+30* time points. For the time points Mid and R+30 this large dispersion only happens in the third quartile.

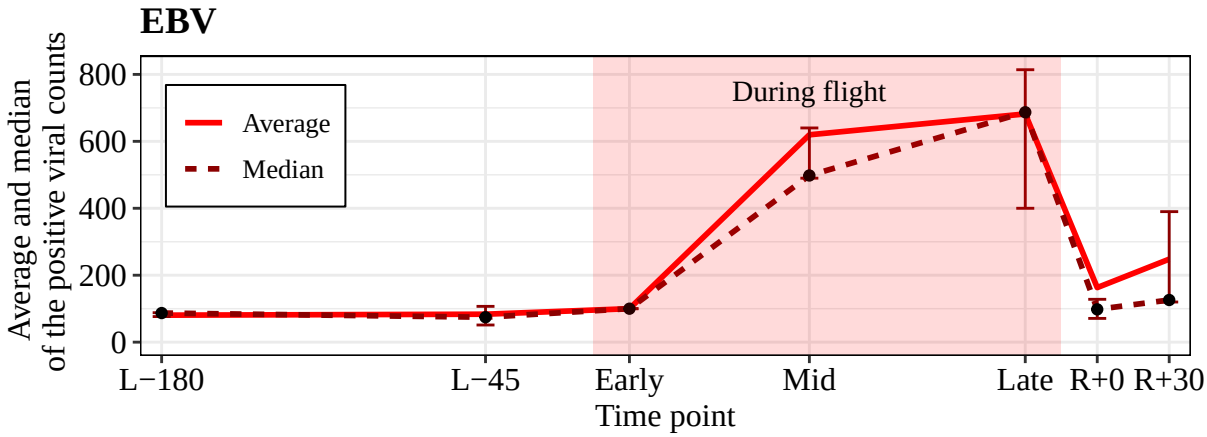

Figure 4.2: Representation of the average and median of the values of positive viral counts for EBV over time ignoring missing responses. Represented along the median are the dispersion measures of the first and third quartiles.

In Table 4.1 and Figure 4.3 are represented the proportions of reactivation along with the binomial 95% confidence intervals using the Wilson's score method at all time points. The proportions there represented ignore the existence of missing values, so the proportions of the inflight time points are measured using the number of detected viral reactivations divided by the number of observed samples at each time point. The proportions for the before and after flight time points all consider the 23 astronauts. At the time point *L−180* there were 3 detected reactivations while the *L−45* time point had 10 detected reactivations, which was the biggest number of detected reactivations. The *Early* time point is the one with fewer detected reactivation with only 1, while the *Mid* time point had 5 reactivations and the *Late* time point with 9 detected reactivations had the highest proportion of reactivation detected. For the after flight time points there is a slight decrease in the proportions. The *R+0* time point has 9 detected reactivations, the same as the previous time point *Late*, while the time point *R+30* has 5 detected reactivations. All the 95% confidence intervals intercept with each other, although, the interval of time point *Early* just barely intercepts the intervals of timepoints *L−45* and *Late*, even so, this suggests that the probabilities of viral reactivation did not change over time.

For this virus it is difficult to assess any relation between the proportion of detected reactions and their amplitude. While the inflight time points *Mid* and *Late* are clearly the ones with the highest average of the positive number of copies, their proportions of reactivation do not stand out very much from the others. The proportion of reactivation of the *Late* time point is very close the proportion of reactivation of the *L−45* time point and their average numbers of viral copies are very different, and the same happens for the *Mid* and *R+30* time points. The interpretation of these plots suggests that for the *Mid* and *Late* time points there is and increase in the amplitude of the viral number of copies.

Table 4.1: Observed proportions of viral reactivation and binomial 95% confidence intervals for the theoretical proportions using Wilson's score method for subjects shedding EBV

Time point	Estimates	
	Proportion of reactivation, %	Confidence interval, %
L−180	13.0	(4.5, 32.1)
L−45	43.5	(25.6, 63.2)
Early	5.6	(1.0, 25.8)
Mid	22.7	(10.1, 43.4)
Late	45.0	(25.8, 65.8)
R+0	39.1	(22.2, 59.2)
R+30	21.7	(9.7, 41.9)

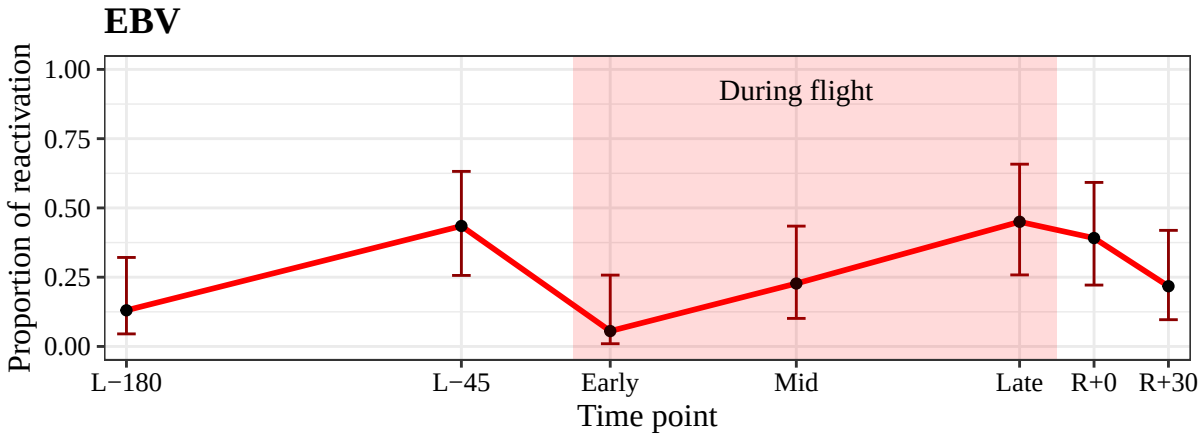

Figure 4.3: Visual representation of the observed proportions of viral reactivation and binomial 95% confidence intervals for the theoretical proportions using Wilson's score method for subjects shedding EBV.

Varicella-zoster virus

The exploratory analysis of VZV started with the examination of a simple plot expressing the numbers of detected viral copies over time. The spacing between the time points in the axis of the next plots represent the time passed between the measures.

In Figure 4.4 the number of detected viral copies is represented by each one of the dots along the time points and each line represents one astronaut. The data at use for this plot are only the complete cases, so all the lines with missing data were removed. The two before the flight time points had no detected reactivation yet all the others had big values of the number of viral copies. The inflight time points had large values of the number of viral copies. The *Early* time point has one astronaut whose number of viral copies stands out for being large and the *Late* time point had the largest values of detected viral copies when compared to the other time points. For the after flight time points, aside from one subject that scored very high values for both time points, most individuals had little numbers of viral copies.

In Figure 4.5 are represented the median and average of the positive number of viral copies. The

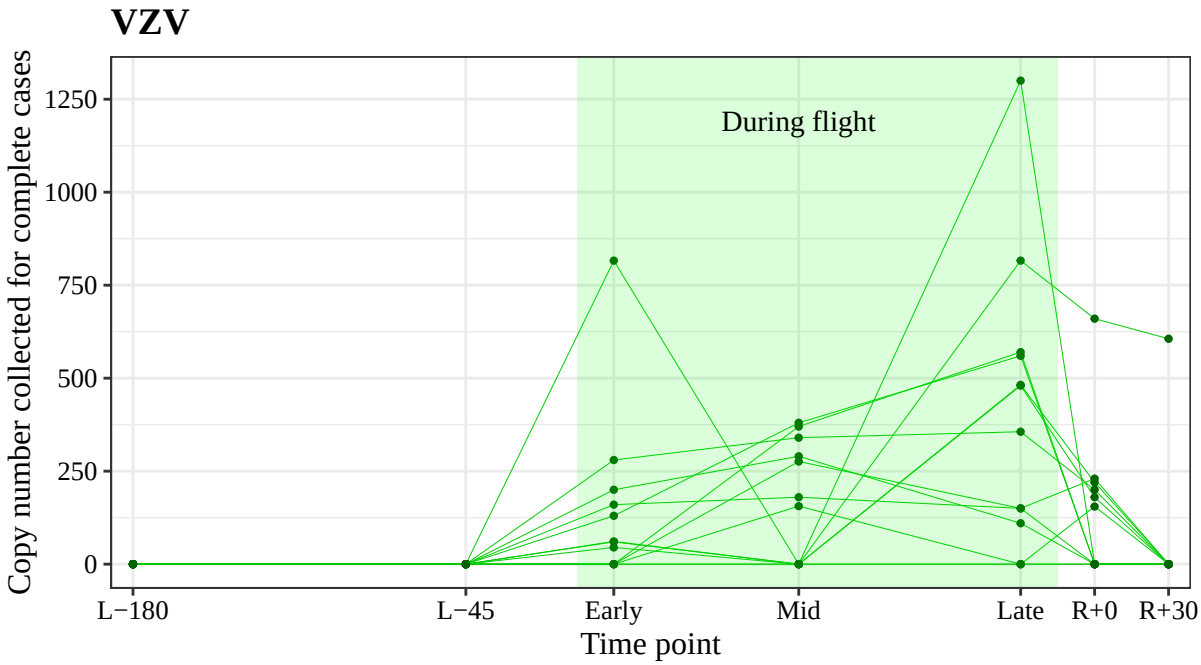

Figure 4.4: Graphical representation of the viral copy number for VZV over time for subjects with no missing data. Each line represents one subject.

first two time points had no positive reactivation so they have no representation in the plot. The inflight time points *Early* and *Mid* have a similar value for the average of positive viral counts, and the *Late* time point has the biggest value for all time points. Immediately after return the value of the average of positive numbers of viral copies decreases yet for *R+30* increases again. Aside from the first two time points that showed no reactivation, the average of the positive number of viral copies for VZV are more or less among 200 and 450. The median and average of the positive number of viral copies appear to be near each other for all time points, which suggests that their distribution is symmetric. The time points *Early*, *Late* and *R+30* present a big dispersion considering the quartiles, when compared to the other time points.

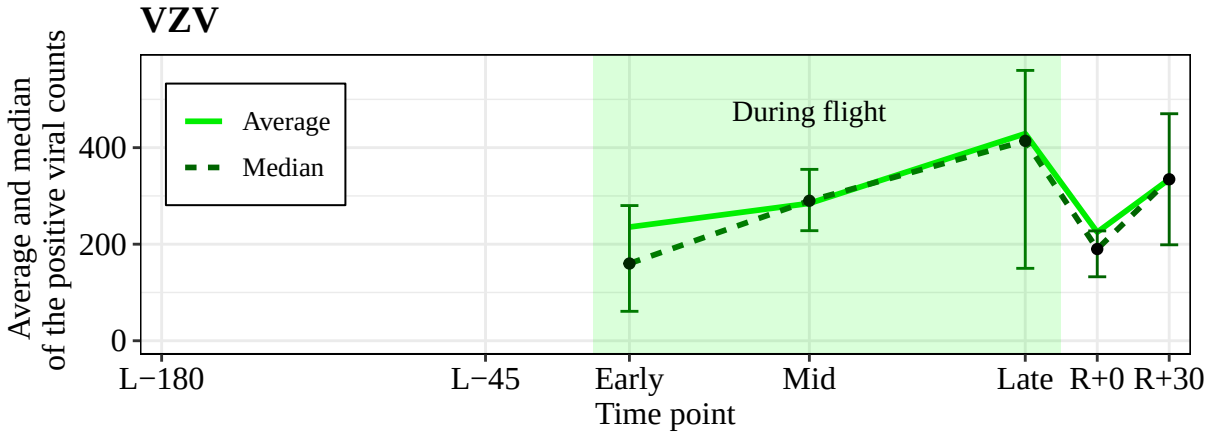

Figure 4.5: Representation of the average and median of the values of positive viral counts for VZV over time ignoring missing responses. Represented along the median are the dispersion measures of the first and third quartiles.

In Table 4.2 and Figure 4.6 are represented the proportions of reactivation and the binomial 95%

confidence intervals using Wilson's score method. The proportion represented were estimated the same way they were for EBV in Figure 4.3. So the statistics in the inflight time points do not use all individual, yet the before and after flight use all 23 astronauts. The first two before flight time points have no detected reactivation so their proportion is 0. For the inflight time points there was detected reactivation. The *Early* time point had 9 detected reactivations, the *Mid* time point had 7 and finally the *Late* time point had 13 detected reactivations. Right away it is possible to see that the inflight time points of VZV had big proportions of reactivation when compared to the other viruses. The after flight time point *R+0* had 10 detected reactivations and the *R+30* time point had 2. The two before flight time points had no detected viral reactivation and their 95% confidence intervals of the proportions of viral reactivation do not intercept with the intervals of any inflight and *R+0* time points, which suggests their proportions of viral reactivation are statistically different. The 95% confidence interval of time point *R+30* also does not intercept with the confidence intervals of time points *Early* and *Late*, suggesting a difference in their probabilities of viral shedding. Because of these significant differences in the confidence intervals of the probabilities of viral reactivation, one might admit that the probabilities of viral reactivation of the inflight time points, in particular *Early* and *Late*, are larger than the time points when the astronauts were on Earth.

Table 4.2: Observed proportions of viral reactivation and binomial 95% confidence intervals for the theoretical proportions using Wilson's score method for subjects shedding VZV

Time point	Estimates	
	Proportion of reactivation, %	Confidence interval, %
L−180	0.0	(0.0, 14.3)
L−45	0.0	(0.0, 14.3)
Early	50.0	(29.0, 71.0)
Mid	31.8	(16.4, 52.7)
Late	68.4	(46.0, 84.6)
R+0	43.5	(25.6, 63.2)
R+30	8.7	(2.4, 26.8)

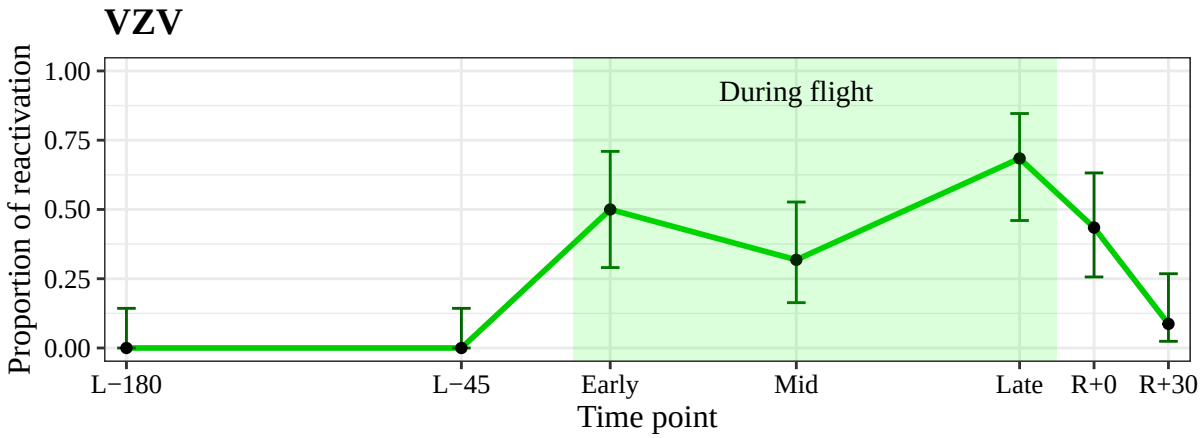

Figure 4.6: Visual representation of the observed proportions of viral reactivation and binomial 95% confidence intervals for the theoretical proportions using Wilson's score method for subjects shedding VZV.

The first thing that is possible to observe is that VZV has bigger proportions of detected reactivations than the other viruses with its two biggest proportions of reactivation being 50.0% and 68.4%. Except for the last time point, *R+30*, there appears to be a correlation between the average of the positive number of viral copies and the proportion of detected reactivation. The *Late* time point has the highest value of both the average number of positive viral copies and the proportion of reactivation, this indicates that the astronauts did not acclimate to the stress conditions for the later days of their flight.

4.2 Complete case analysis

In this section only the subjects that had no missing data were analysed. McNemar's exact binomial test was used to assess differences between reactivation dynamics for any pair of time points. The results of Friedman's test from the original study are also presented.

Friedman's and McNemar's tests

There was considerable variation of the shedding proportions over the time points suggesting an overall mission effect on the reactivation of these viruses [34]. In the original study, the Friedman test was used to study the variation of the reactivations. This test is a non-parametric statistical test that is used to detect differences in treatments across multiple paired samples. Non-parametric tests are applied to quantitative data where it is not expected to have many ties, therefore, the abundance of zeros in the data set could be affecting the result. When comparing copy numbers between time points, the Friedman test did not show a significant difference between time points for EBV (p-value=0.064), although the p-value is very close to the significance level of 0.05. Indeed, EBV was shed at all seven sample collection time points [34]. For VZV, no shedding occurred at both 180 days and 45 days before flight yet shedding was found in *Early*, *Mid* and *Late* time points during flight as well as at landing and 30 days after landing. The Friedman test comparing copy number distributions was significant for VZV (p-value<0.0001). Even after excluding time points with no detected reactivation from the analysis, there were still significant differences between the remaining time points (p-value=0.0027) [34].

Table 4.3: P-values of the McNemar's test for EBV for all pairs of time points. In parenthesis are the number of discordant pairs, (n_{01},n_{10}), respectively. P-values smaller than the considered significance level are highlighted in bold.

	L−180	L−45	Early	Mid	Late	R+0	R+30
L−180	–	0.092 (10,3)	0.625 (1,3)	0.687 (4,2)	0.070 (7,1)	0.109 (8,2)	0.687 (4,2)
L−45		–	**0.016** (0,7)	0.424 (5,9)	1.000 (5,5)	1.000 (6,7)	0.180 (2,7)
Early			–	0.375 (4,1)	**0.008** (8,0)	**0.039** (8,1)	0.250 (3,0)
Mid				–	0.180 (7,2)	0.344 (7,3)	1.000 (4,4)
Late					–	0.727 (3,5)	0.125 (1,6)
R+0						–	0.424 (5,9)
R+30							–

Another applicable test to study the differences in the reactivation dynamics between time

points is the McNemar's binomial exact test. This test uses the binary data where 1 represents detected shedding and 0 the opposite, to compare the disparity in the discordant pairs. As is seen in Table 4.3, for EBV the pairs of time points with significant differences in the observed discordant pairs are *L−45* with *Early*, *Early* with *Late* and *Early* with *R+0*.

For VZV (Table 4.4), because *L−180* and *L−45* have the exact same viral copy numbers, the p-value of their test is considered to be 1 and they are significantly different from every other time point except *R+30* because there are only 2 detected reactivation at that time point. The other time points with significant differences are *R+30* with every other time point except *L−180*, *L−45* and *Mid*.

Table 4.4: P-values of the McNemar's test for VZV for all pairs of time points. In parenthesis are the number of discordant pairs, (n_{01},n_{10}), respectively. P-values smaller than the considered significance level are highlighted in bold.

	L−180	L−45	Early	Mid	Late	R+0	R+30
L−180	–	1.000 (0,0)	**0.004** (9,0)	**0.016** (7,0)	<**0.001** (13,0)	**0.002** (10,0)	0.500 (2,0)
L−45		–	**0.004** (9,0)	**0.016** (7,0)	<**0.001** (13,0)	**0.002** (10,0)	0.500 (2,0)
Early			–	0.727 (3,5)	0.250 (3,0)	1.000 (3,4)	**0.039** (1,8)
Mid				–	0.125 (6,1)	0.754 (6,4)	0.070 (1,7)
Late					–	0.219 (1,5)	**0.001** (0,11)
R+0						–	**0.008** (0,8)
R+30							–

Multiple simultaneous McNemar's tests were made for both viruses and this raises the multiple comparisons problem. To address this problem, the Bonferroni correction can be used to adjust the significance level. The new significance level is calculated by dividing the original significance level by the number of tests for each virus, which in this case is $0.05/21 \approx 0.0024$. After the correction, for EBV there are no pairs of time points with significant differences in the reactivation dynamics between themselves. For VZV, the pairs of time points with significant differences after the Bonferroni correction are *L−180* with *Late*, *L−180* with *R+0*, *L−45* with *Late*, *L−45* with *R+0* and *Late* with *R+30*.

For EBV, the Friedman test did not find significant differences between the viral copy numbers of the time points, although the p-value is very close to the significance level considered. The McNemar's exact test found three pairs of time points to be significantly different for EBV. After the Bonferroni correction there were no significant pairs of time points for EBV. For VZV, the Friedman test found significant differences between measurements. Even after removing the time points with no detected reactivation, the test was still significant. The McNemar's exact test found 11 pairs of time points that were significantly different before the Bonferroni correction and five after for VZV, which is a big difference when compared to EBV and even CMV.

4.3 Categorical data analysis with missing responses

For this section, two parallel analysis were made, one for each virus, because the analysis' results were more interesting considering each individual virus than both together. Since only

one subpopulation is considered the index s is not relevant and will be changed with the index ψ denoting the virus being considered. The index ψ takes values e and v for EBV and VZV, respectively. The time points considered were only the inflight time points because it is computationally easier and to avoid loss of significance in the tests. It is relevant to mention that the methods used are asymptotic tests and they may be imprecise due to the small frequencies observed in Table 4.5.

4.3.1 Missingness mechanism

The analysis began with the construction of a table to assess the vectors of frequencies N_ψ (Table 4.5). These frequencies are the number of occurrences that each reactivation status are observed. The missingness patterns with two or three non-observed responses were not considered because they were not observed.

The sum of the elements of the vectors N_e and N_v have to be the total number of astronauts, which is 23. Because it is a small sample size, all the frequencies are a reasonably small number. Note that $R_{\psi 1} = R = 8$ and $R_{\psi 2} = R_{\psi 3} = R_{\psi 4} = 4$.

The matrix Z_ψ, denoting the correspondence to all response classes for units with all missingness patterns, is the following

$$\mathrm{Z}_\psi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that the columns 1 to 8 correspond to missing pattern $f = 1$, the columns 9 to 12 correspond to the missing pattern $f = 2$, the columns 13 to 16 correspond to missing pattern $f = 3$ and the columns 17 to 20 correspond to missing pattern $f = 4$. Each row of the matrix is associated with each combination of the missing pattern $f = 1$.

For the pattern with no missing responses, when the frequencies were 0, that number was substituted by a small value. This replacement had to be done because null values do not allow information from other missingness patterns to be incorporated [49]. The numbers replacing null frequencies started as 1 and were iteratively set smaller using the formula 10^{-y}, where y represents the number of the iteration. Note that y started as 0 and was increased iteratively. This algorithm run until the difference between two consecutive iterations' p-values of the test statistics Q_L(MCAR|MAR) and Q_P(MCAR|MAR) were both smaller than 10^{-3}. The degrees of freedom for the test statistics are 9 for both viruses. The resulting p-values for EBV and VZV are presented in Table 4.6.

For both viruses the algorithm converged at the seventh iteration. Both methods produce very

Table 4.5: Observed frequencies of astronauts classified by reactivation status and virus. Note: non-reactivation is represented by −, reactivation by + and missing by NA.

Missingness pattern, f	Reactivation status, Y_ψ			Frequency, N_ψ	
	Early	Mid	Late	EBV	VZV
1	−	−	−	5	2
	−	−	+	6	0
	−	+	−	1	1
	−	+	+	2	2
	+	−	−	0	0
	+	−	+	0	4
	+	+	−	0	0
	+	+	+	0	4
2	−	−	NA	1	3
	−	+	NA	1	0
	+	−	NA	1	1
	+	+	NA	0	0
3	−	NA	−	1	0
	−	NA	+	0	1
	+	NA	−	0	0
	+	NA	+	0	0
4	NA	−	−	3	3
	NA	−	+	1	2
	NA	+	−	1	0
	NA	+	+	0	0

Table 4.6: Convergence of the test statistics of MCAR test for likelihood ratio and Pearson methods and respective p-values in parenthesis, with a consequent smaller zero replacement for EBV and VZV.

10^{-y}	EBV		VZV	
	Likelihood R.	Pearson	Likelihood R.	Pearson
1	6.716 (0.667)	5.734 (0.766)	14.303 (0.112)	12.459 (0.189)
0.1	7.720 (0.563)	7.811 (0.553)	15.505 (0.078)	14.561 (0.104)
0.01	8.320 (0.502)	8.382 (0.496)	15.791 (0.071)	15.469 (0.079)
0.001	8.427 (0.492)	8.433 (0.491)	15.838 (0.070)	15.780 (0.072)
0.0001	8.442 (0.490)	8.438 (0.491)	15.845 (0.070)	15.875 (0.070)
0.00001	8.444 (0.490)	8.439 (0.491)	15.845 (0.070)	15.894 (0.069)
0.000001	8.444 (0.490)	8.439 (0.491)	15.845 (0.070)	15.894 (0.069)

similar p-values for each corresponding virus. For EBV the estimated test p-values are far greater than the significance level considered, while for VZV the p-values are greater than yet close to that cut-off. The estimated p-values of both tests are not significant so the null hypothesis is not rejected for either one. The MCAR missing mechanism will be assumed for the remaining analysis for EBV and VZV.

4.3.2 Test of homogeneity of marginal probabilities

The test of homogeneity of marginal probabilities will be performed by fitting a linear model M to the marginal probabilities of categorisation. This model for categorical data is defined by matrices J defining linear functions of interest and W as the model specification matrix, and its expression is in equation (2.13). The way it works is by creating a specification matrix with linear functions of interest and testing their goodness-of-fit to the data. The test hypothesis are

$$H_0 : \pi_{1..} = \pi_{.1.} = \pi_{..1} \textit{ vs. } H_1 : \pi_{1..} \neq \pi_{.1.} \text{ or } \pi_{1..} \neq \pi_{..1} \ ,$$

where $\pi_{1..} = \pi_{111} + \pi_{110} + \pi_{101} + \pi_{100}$, $\pi_{.1.} = \pi_{111} + \pi_{110} + \pi_{011} + \pi_{010}$ and $\pi_{..1} = \pi_{111} + \pi_{101} + \pi_{011} + \pi_{001}$. These parameters represent the probability of viral reactivation at first, second and third time points, respectively.

The matrix J defining the linear functions considers a linear association model and it is

$$\mathrm{J} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} .$$

The W specification matrix was derived, under the null hypothesis, from the following equations: $\pi_{111}+\pi_{110}+\pi_{101}+\pi_{100} = \pi_{111}+\pi_{110}+\pi_{011}+\pi_{010} = \pi_{111}+\pi_{101}+\pi_{011}+\pi_{001}$. By developing these equations in order to isolate each of the parameters, a possible result is $\pi_{110} = -\pi_{100}+\pi_{011}+\pi_{001}$, $\pi_{101} = -\pi_{100} + \pi_{011} + \pi_{010}$, $\pi_{100} = -\pi_{101} + \pi_{011} + \pi_{001}$, $\pi_{011} = \pi_{101} + \pi_{100} - \pi_{010}$ and $\pi_{010} = \pi_{101} + \pi_{100} - \pi_{010}$. Which in matrix form becomes

$$\mathrm{W} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & 0 & 0 \end{pmatrix} .$$

Again, for the pattern with no missing responses the frequencies 0 were iteratively substituted by smaller values using the method explained before. The parameters of the model were estimated taking this into account and they are represented in Tables 4.7 and 4.8 for EBV and VZV, respectively. The algorithm run until the differences between two values of the same parameters were smaller than 10^{-4}. The test statistics applied to both viruses were the likelihood ratio and Pearson tests under the MCAR mechanism for EBV and VZV.

The estimated p-values of the test of homogeneity of marginal probabilities applied were very close to 1 for both methods applied to EBV and VZV, so the null hypothesis is not rejected. There are no evidences indicating significant differences in the probability of reactivation of the three inflight time points studied for EBV and VZV.

Table 4.7: Convergence of the parameters of the model with a consequent smaller zero replacement for EBV. Using these estimated parameters, the test of homogeneity of marginal probabilities was applied to the inflight time points of EBV.

10^{-y}	Parameters						
	π_{111}	π_{110}	π_{101}	π_{100}	π_{011}	π_{010}	π_{001}
1	0.3237	0.0640	0.0727	0.0803	0.0935	0.0726	0.2562
0.1	0.3829	0.0539	0.0645	0.0551	0.1024	0.0680	0.2689
0.01	0.3867	0.0627	0.0647	0.0523	0.1144	0.0576	0.2612
0.001	0.3868	0.0645	0.0648	0.0520	0.1165	0.0557	0.2596
0.0001	0.3868	0.0647	0.0648	0.0520	0.1168	0.0555	0.2594
0.00001	0.3868	0.0648	0.0648	0.0520	0.1168	0.0555	0.2594
0.000001	0.3868	0.0648	0.0648	0.0520	0.1168	0.0555	0.2594

Table 4.8: Convergence of the parameters of the model with a consequent smaller zero replacement for VZV. Using these estimated parameters, the test of homogeneity of marginal probabilities was applied to the inflight time points of VZV.

10^{-y}	Parameters						
	π_{111}	π_{110}	π_{101}	π_{100}	π_{011}	π_{010}	π_{001}
1	0.2398	-0.0967	0.0973	0.1519	0.0167	0.1736	0.2635
0.1	0.3262	-0.1961	0.0886	0.1639	-0.0366	0.2434	0.2390
0.01	0.3429	-0.2130	0.0872	0.1706	-0.0429	0.2569	0.2246
0.001	0.3467	-0.2164	0.0870	0.1730	-0.0435	0.2599	0.2194
0.0001	0.3476	-0.2172	0.0870	0.1738	-0.0435	0.2607	0.2177
0.00001	0.3478	-0.2174	0.0870	0.1739	-0.0435	0.2609	0.2174
0.000001	0.3478	-0.2174	0.0870	0.1739	-0.0435	0.2609	0.2174

The test here applied could be imprecise due to the elevated number of small frequencies in N_ψ, in particular, the large number of zeros observed. A frequency is considered small when it is inferior to five [49]. The null frequencies were iteratively replaced with smaller numbers and the resulting p-values of the homogeneity test did not change over the iterations. This test was imprecise because while for EBV the estimated probabilities of the model are all positive at all iteration, for VZV that is not the case. In Table 4.8, the marginal probability π_{110} was estimated to be negative even when the zeros were replaced by 1, and π_{011} has the only positive estimate at the first iteration.

The estimates for these probabilities converged on negative values given that the data set has a small size that leads to small frequencies of N_ψ. In Poleto et al. (2014) [49], replacing the null frequencies by 10^{-6} bypassed the problem of the negative probabilities. However for the data of this project even when the null frequencies were replaced by 10^{-14} the problem still persisted. The numerical algorithm used in the estimation of the models' parameters has no restrictions imposed on the values of the estimated marginal probabilities and cannot cope with the elevated number of zeros in N_ψ. Clearly there are not guarantees in the usefulness of the algorithm since replacing the null frequencies is ineffective in obtaining realistic estimates for the marginal probabilities. Given that this method to test the homogeneity of the marginal probabilities is unreliable, the inferences made in this subsection will not be considered in the analysis.

4.4 Data imputation

This analysis started by using the binary data sets for detected reactivation and non-detected reactivation. The parameters required by the function `mice`, from the `mice` package version 3.13.0 [59], are the number of imputations, the method, the maximum number of iterations and the seed. The number of imputed data sets chosen was 50. This number was set high because it is recommended in applications where high statistical power is needed [60]. The parameter `method` refers to the form of the imputation model. It was set to logistic regression for each binary variable conditional on the remaining binary variables. The maximum number of iterations was set at 50 because it was considered to be large enough to observe convergence. In order for the results of the data imputations to be able to be replicated the argument `seed` had to be fixed, so the `seed` chosen was 1997.

4.4.1 Convergence

With iterative data imputation, the validity of the inference depends on the state-space of the algorithm at the final iteration. This introduces a potential threat to the validity of the imputations. What is often done is to plot one or more parameters against the iteration number [54]. The plots built this way appear to show no convergence (Figures 4.7 and 4.8), so it is challenging to arrive upon a single point at which one can assume convergence has been reached.

It is clear from both Figures 4.7 and 4.8 that this method is ineffective at assessing convergence in the MICE algorithm. The plots of the standard deviations associated with the data imputation are in Appendix A.

Although, iterative imputation algorithms can yield correct outcomes, even when a converged state has not yet formally been reached [61]. To assess the convergence it was necessary to find a statistic that could study the behaviour of the algorithm along the iterations. This statistic needed to indicate that the chained imputations estimated by the algorithm were conducting similarly and keeping steady across iterations. So the statistic considered was

$$w_j = \sum_{i=1}^{j} \frac{h_i}{j} \ , \tag{4.1}$$

where j = 1,...,total number of iterations, w_j represents the average of the proportion of reactivation at the iteration j and h_i is the proportion of the reactivation at iteration i. What this statistic does is, for a certain iteration, averages the proportions given at that iteration and all the previous iterations for a certain imputation chain. This statistic was used for all data imputations.

The plots representing the values originated from this statistic are the coloured lines in Figure 4.9 and Figure 4.10. Each line represents a chain of the algorithm and since all imputation tend to the same values as the iterations increase we can conclude that all the imputations behave similarly.

In Figure 4.9 and Figure 4.10 there is also represented a black line for each plot. These line represents the estimated probabilities of reactivation for each iteration. The lines were estimated by averaging the probabilities of reactivation of all the imputations, at each iteration. In general these lines keep stable along the iterations.

Also represented in these plots, are lines indexed with *max* and *min* at the top and the bottom of the graphic. These lines represent the maximum and the minimum possible values of proportions of reactivation: the *max* considers all non-observed values have reactivation and the *min* considers that all the non-observed values have no reactivation.

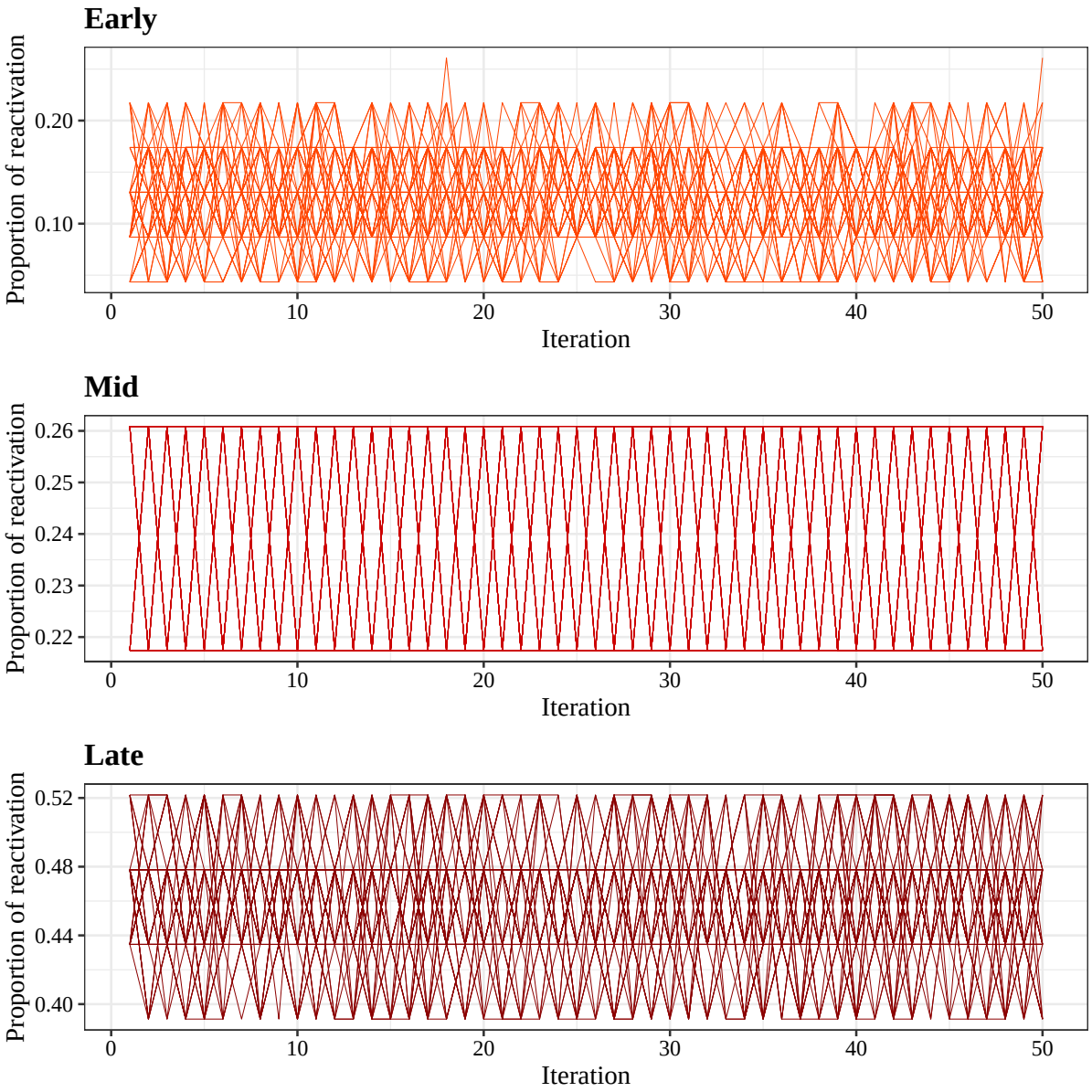

Figure 4.7: Graphic representation of the MICE algorithm without convergence statistics for EBV. Each line represents one chain of imputation and it takes the value of the proportion of viral reactivation (detected and imputed) along the iterations. In these plots there is no indication that convergence is occurring.

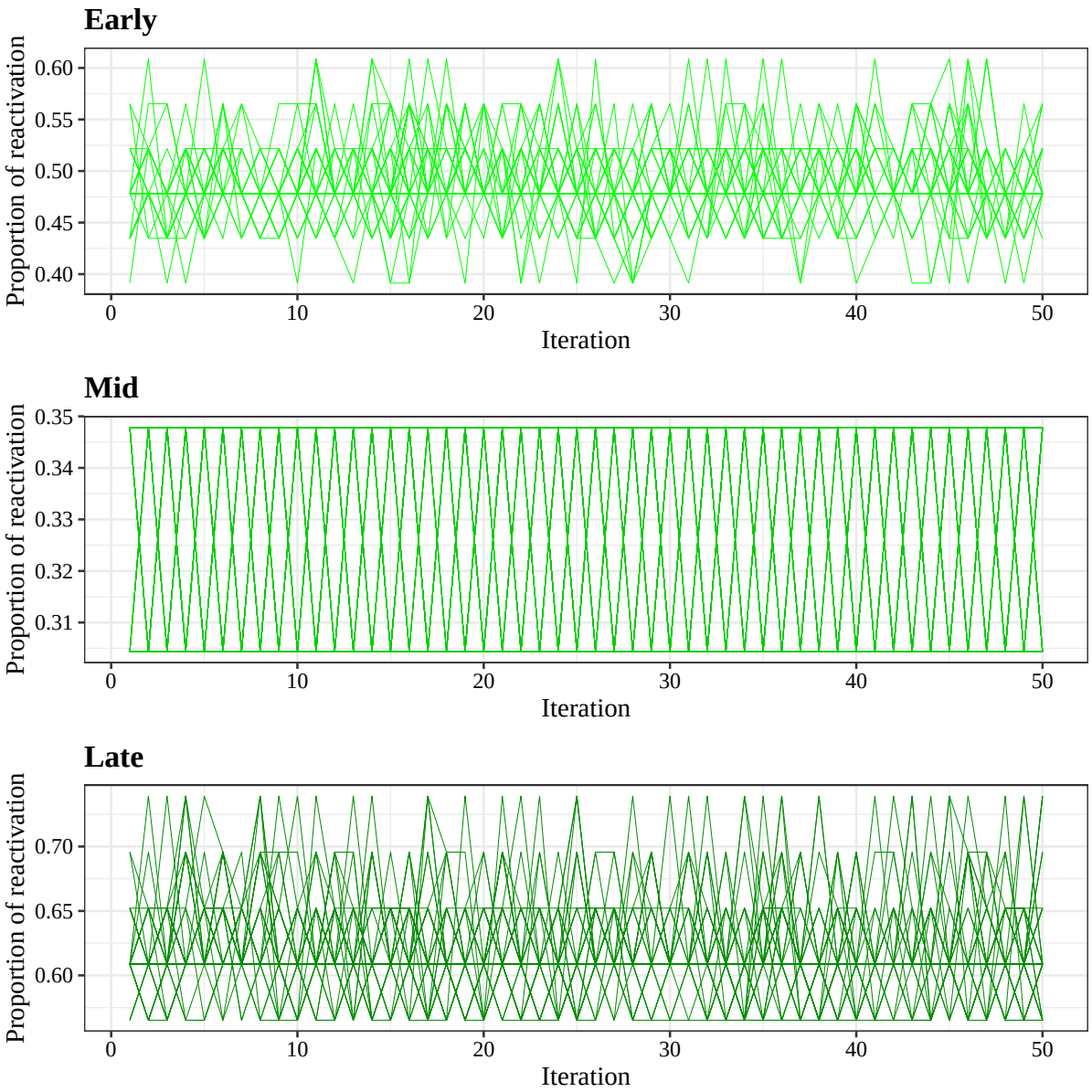

Figure 4.8: Graphic representation of the MICE algorithm without convergence statistics for VZV. Each line represents one chain of imputation and it takes the value of the proportion of viral reactivation (detected and imputed) along the iterations. In these plots there is no indication that convergence is occurring.

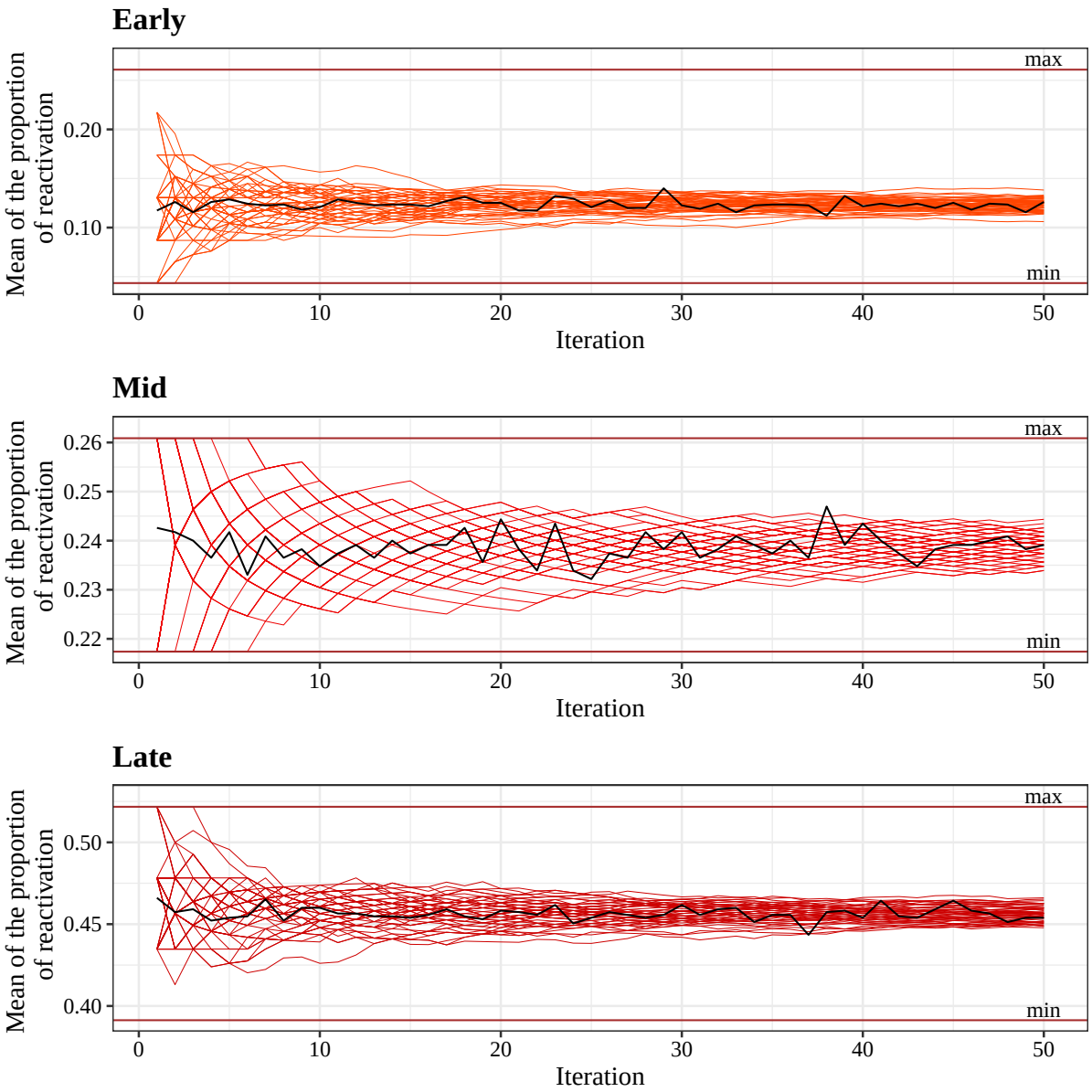

Figure 4.9: Representation of the statistics used to assess the convergence of MICE algorithm for EBV. Each of the coloured lines represents one chain of imputation submitted to the statistic in (4.1) and the black line represents the average of the proportion of reactivation for all chain of imputations at each iteration. The *min* and *max* lines represent the proportions when all the imputed data were replaced by 0 and 1, respectively.

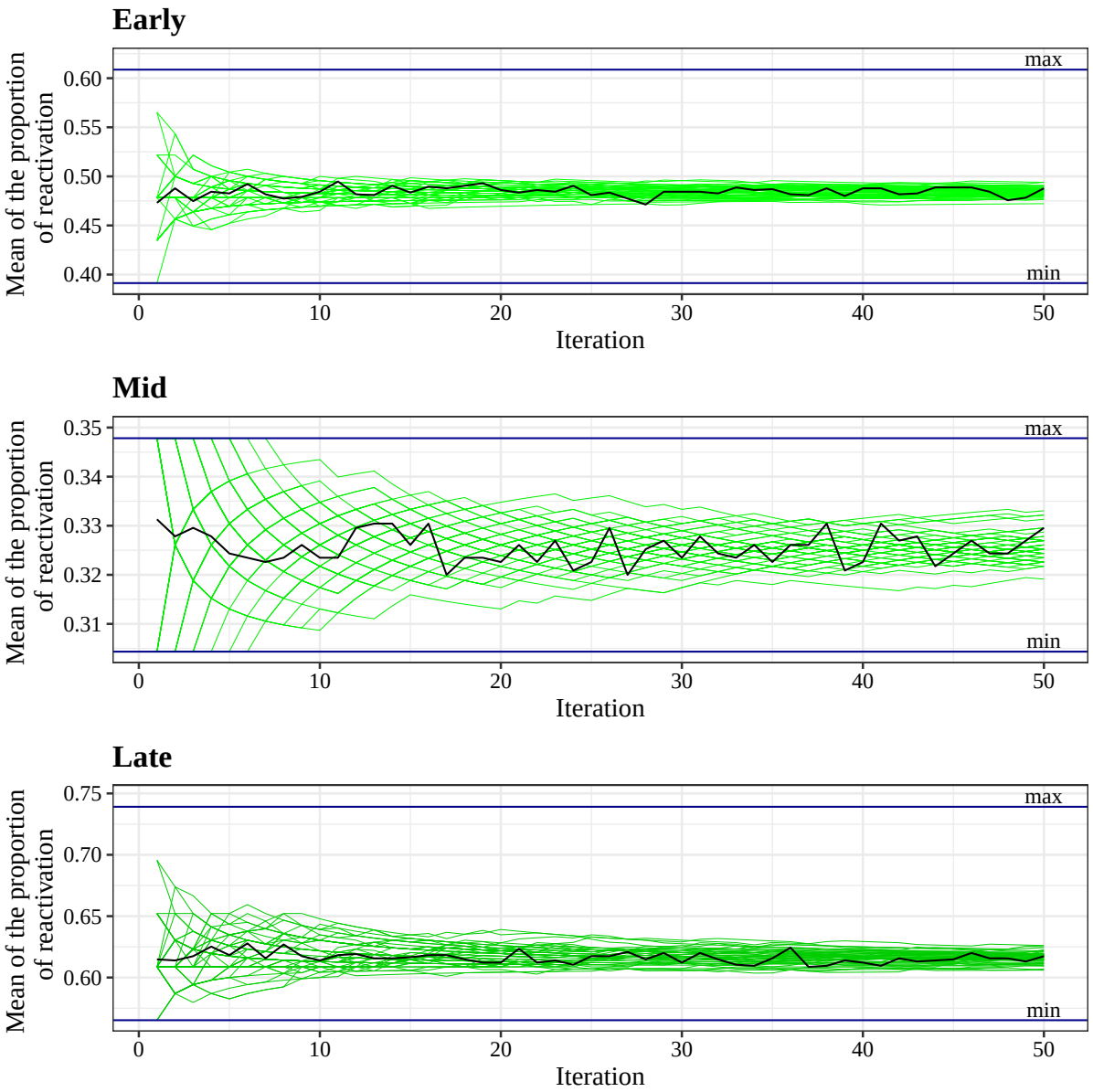

Figure 4.10: Representation of the statistics used to assess the convergence of MICE algorithm for VZV. Each of the coloured lines represents one chain of imputation submitted to the statistic in (4.1) and the black line represents the average of the proportion of reactivation for all chain of imputations at each iteration. The *min* and *max* lines represent the proportions when all the imputed data were replaced by 0 and 1, respectively.

Confidence intervals

For each of the 50 data imputations created, the probabilities of reactivations and their associated standard deviations were measured. In Figure 4.11 the 95% Wilson's score method confidence intervals for the probabilities of reactivation of the EBV are represented. This method is preferred over others because it behaves well for small samples with proportions close to zero [37, 39]. For this virus, the estimated probabilities of reactivation are, in the *Early* time point, between 0.043 and 0.261, in the *Mid* time point the proportions are between 0.217 and 0.261 and for the *Late* time point the proportions of reactivation are between 0.391 and 0.522. There is an apparent increase in the number of reactivation over time for the inflight time points of EBV.

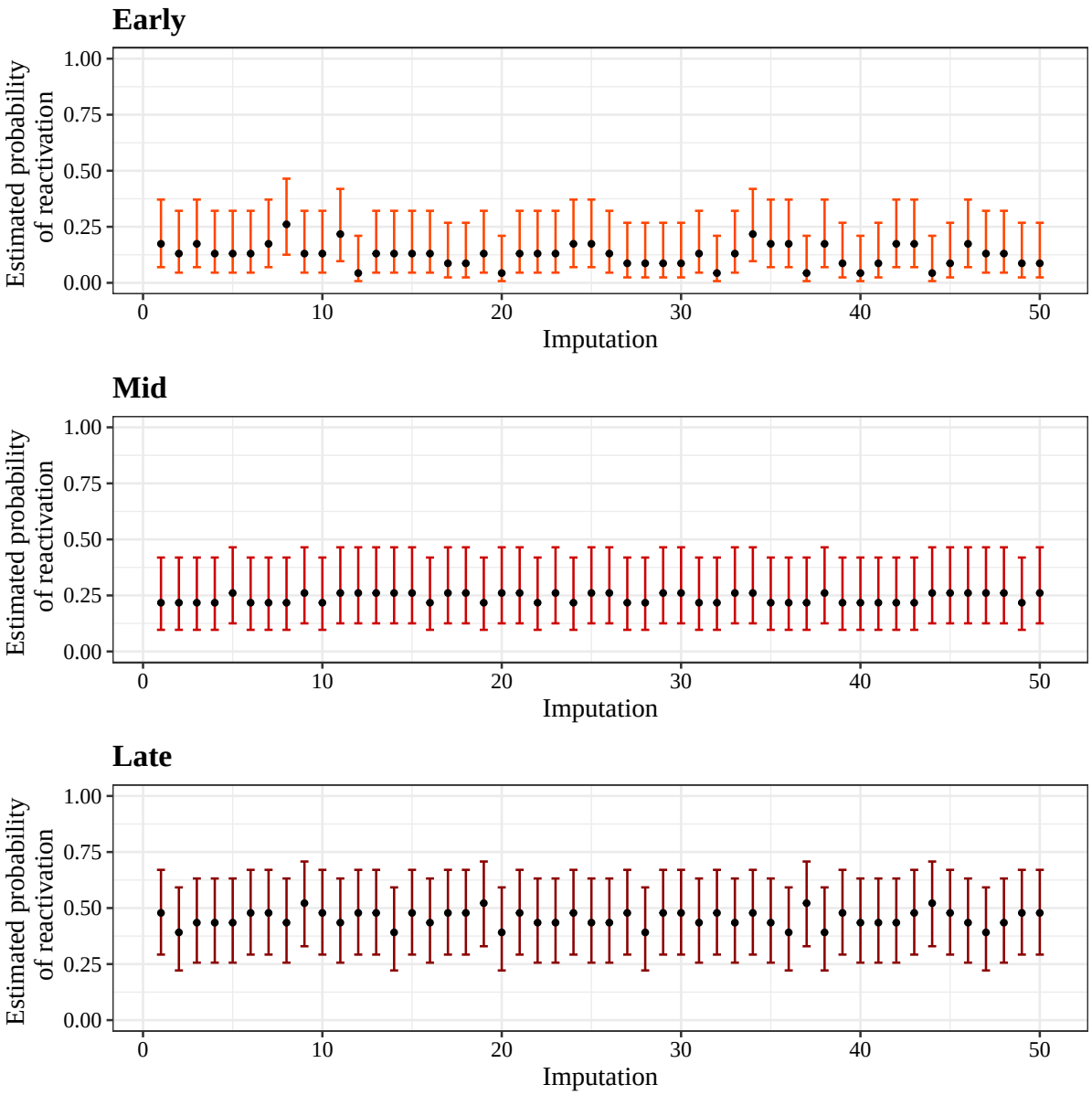

Figure 4.11: Estimated 95% confidence intervals of the probability of viral reactivation at the last iteration of MICE algorithm for all imputations for EBV. The confidence intervals were estimated using Wilson's score method.

In Figure 4.12 the 95% Wilson's score method confidence intervals for the probabilities of reactivation of the VZV are represented. The estimated probabilities of reactivation, for this virus, in the *Early* time point are between 0.435 and 0.565, in the *Mid* time point the probabilities are between 0.304 and 0.348 and for the *Late* time point the probabilities are between 0.565 and 0.739. For VZV, there is a large escalation of the number of detected reactivations with more than half of the astronauts for all imputations for the *Late* time point.

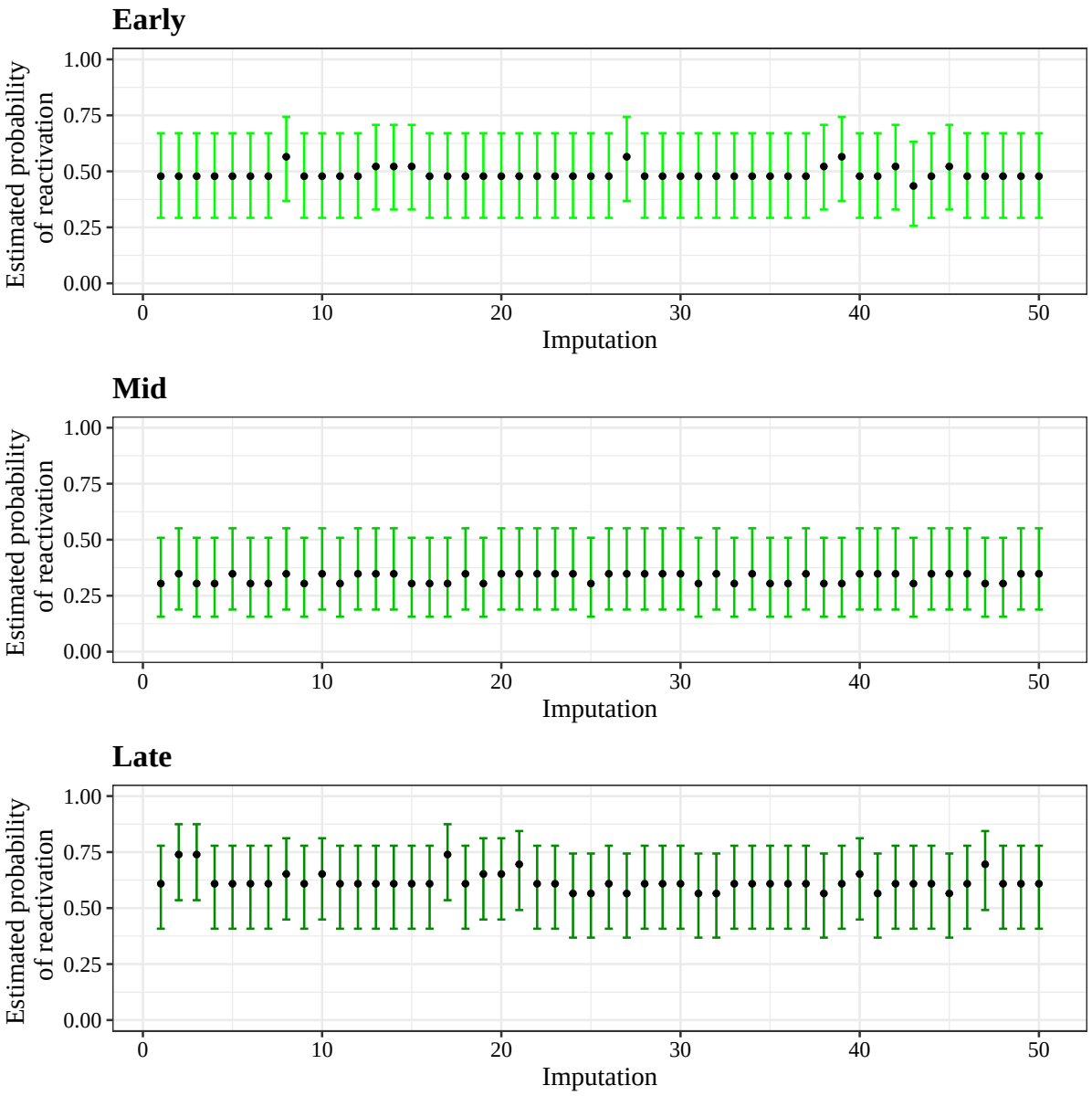

Figure 4.12: Estimated 95% confidence intervals of the probability of viral reactivation at the last iteration of MICE algorithm for all imputations for VZV. The confidence intervals were estimated using Wilson's score method.

For EBV the estimated probabilities of reactivation vary between themselves, while for VZV they are more stable. In the inflight time points and for all the imputations, the estimated probabilities of viral reactivations for VZV are greater than for EBV. The next step is to pool all the estimated probabilities of viral shedding and their associated standard errors into single estimates.

4.4.2 Pooling the data

Rubin [44] developed a set of rules for combining the separate estimates and standard errors from each of the 50 imputed data sets into an overall estimate with standard error and confidence intervals. These rules are based on asymptotic theory on the normal distribution.

The applications of equations (2.16) and (2.17) to the estimates of the probabilities of viral reactivation and their standard errors for all imputations attains the values presented in Table 4.9. For a sufficiently large number of imputed data sets, the combined estimates $\overline{p_t}$ approximately follow a Gaussian distribution [62].

Table 4.9: Pooled estimates for probability of reactivation and standard error for the inflight time points of EBV and VZV.

Time point	Estimates			
	Probability of reactivation		Standard error	
	EBV	VZV	EBV	VZV
Early	0.126	0.488	0.484	0.541
Mid	0.239	0.330	0.022	0.022
Late	0.454	0.617	0.511	0.453

For EBV the pooled probabilities of reactivation are 0.126, 0.239 and 0.454 for the *Early*, *Mid* and *Late* time points, respectively, while for VZV the pooled probabilities of reactivation for the same time points are 0.488, 0.330 and 0.617, respectively.

For the complete case scenario the proportions of reactivation for EBV were 0.056, 0.227 and 0.450, respectively for the same time points. The *Mid* and *Late* time points have similar estimated probability of reactivation and a proportion of detected reactivation, while for the *Early* time point the relative difference is bigger. An explanation for this is the fact that the *Early* time point has 5 non-observed responses, and this allied with the fact that there was only one detected reactivation amongst the studied astronauts makes any imputed reactivation have a big impact on the estimated probability of reactivation. For VZV the proportions of reactivation for the *Early*, *Mid* and *Late* time points were 0.500, 0.318 and 0.684, respectively. The differences between the estimated probabilities of reactivation and the proportions of detected reactivation for this virus are relatively small. The imputation process should preserve the relations in the data and the uncertainty about those relation [54]. Since this appears to happen for the imputation performed it is safe to say that the imputation worked successfully.

Confidence intervals

Using the estimations of the probability of reactivation and its associated standard error, the next step is to estimate its confidence intervals. However, the usual approximations of the confidence interval are known to be poor when the true $\overline{p_t}$ is close to zero or to one [37]. To avoid this problem, alternatives with better properties were used to estimate the confidence intervals for both viruses. In Figure 4.13 the confidence intervals that are represented were

estimated using the Wilson's score method.

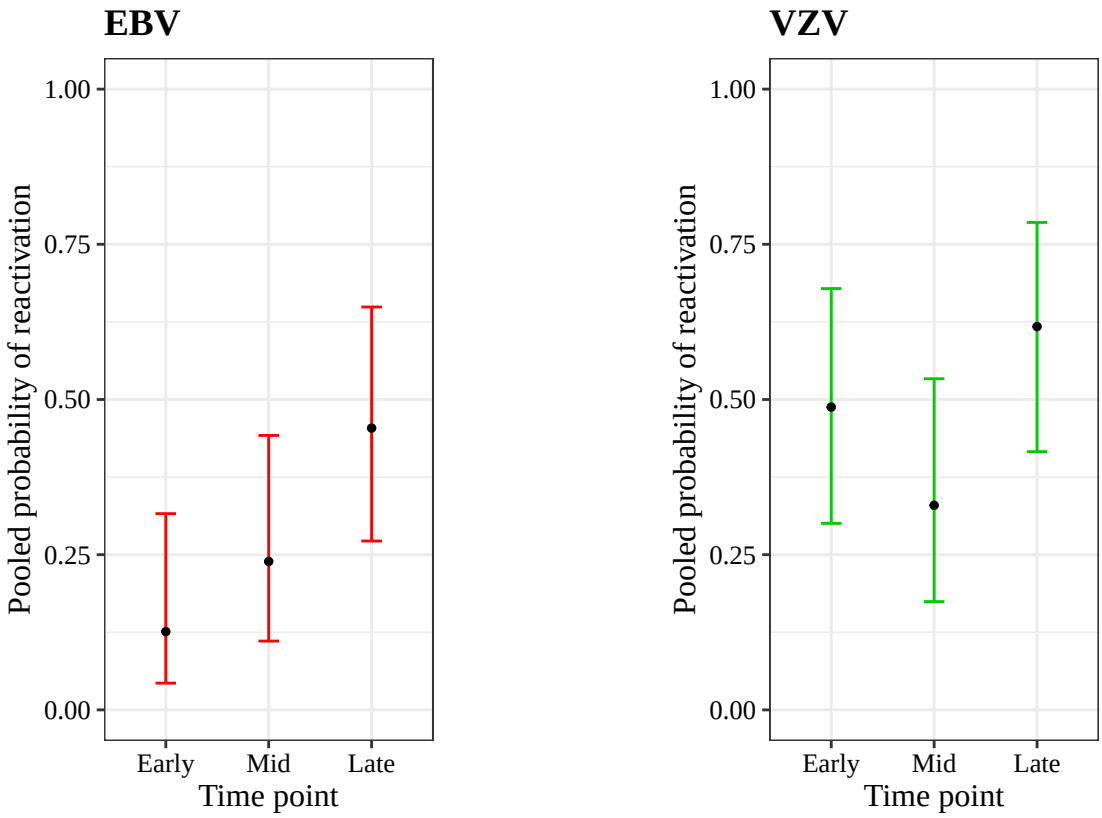

Figure 4.13: Pooled 95% confidence intervals of the probabilities of viral reactivation for the inflight time points of EBV and VZV using Wilson's score method.

For EBV the estimated probabilities of reactivation increase along the inflight time points. The *Early* time point has an estimated probability of viral shedding of 0.126 and the 95% Wilson's score confidence interval of this statistic was estimated to be (0.043, 0.316). For the *Mid* time point the 95% Wilson's score confidence interval of the probability of reactivation was estimated to be (0.111, 0.442) and the probability of reactivation is 0.239. The *Late* time point has the biggest estimated probability of viral shedding of this virus, estimated to be 0.454, which is almost 50% of reactivation and its estimated 95% Wilson's score confidence interval is (0.272, 0.649).

VZV has higher estimated probabilities of reactivations than EBV for all time points. For the *Early* time point, the estimated probability of reactivation is 0.488 and its 95% Wilson's score confidence interval is (0.300, 0.679). The 95% Wilson's score confidence interval for the probability of viral shedding for the *Mid* time point is (0.174, 0.533) and its estimated probability of reactivation is 0.330. The probability of viral shedding for the *Late* time point was estimated to be 0.617, which is the highest value of reactivation for all time points of all viruses, and its associated 95% Wilson's score confidence interval was estimated to be (0.416, 0.785).

The Tables 4.10 and 4.11 have examples of other confidence intervals that can be used for the probability of viral shedding of these viruses [37]. The first method in the tables, the Wilson's score method, is represented in Figure 4.13 since this is a method with good properties [63, 64].

The Wald test is the most commonly used though it is not much reliable when the proportion is close to 0 or 1 [37].

Table 4.10: Alternative 95% confidence intervals of the proportions of viral reactivation built from different methods for EBV and respective interval length. Represented are the Wilson's score, Wald, Clopper-Pearson and arcsine methods.

Method	Time point					
	Early		Mid		Late	
	Conf. interval	Length	Conf. interval	Length	Conf. interval	Length
Wilson's score	(0.043, 0.316)	0.273	(0.111, 0.442)	0.331	(0.272, 0.649)	0.377
Wald	(0.000, 0.262)	0.262	(0.065, 0.413)	0.348	(0.250, 0.657)	0.407
Clopper-Pearson	(0.026, 0.331)	0.305	(0.088, 0.461)	0.373	(0.248, 0.672)	0.424
Arcsine	(0.025, 0.289)	0.264	(0.091, 0.430)	0.339	(0.260, 0.656)	0.396

For EBV the lower limit of the Wald interval for *Early* is a lower aberration, which means that the real value was negative and was substituted by 0. This method differs from the others at time points *Early* and *Mid* because its estimates are slightly lower then the other methods, indeed, these two particular time points have the two lowest probabilities of viral reactivation and the Wald method is not reliable when the proportion is close to 0 or 1 [37]. As expected, the Clopper-Pearson method produces the biggest intervals. The arcsine method estimates intervals not too different from the Wilson's score method, although slightly lower for *Early* and *Mid*. The arcsine method produces good results when n is small [37], yet it is not very reliable for p close to 0 or 1.

Table 4.11: Alternative 95% confidence intervals of the proportions of viral reactivation built from different methods for VZV and respective interval length. Represented are the Wilson's score, Wald, Clopper-Pearson and arcsine methods.

Method	Time point					
	Early		Mid		Late	
	Conf. interval	Length	Conf. interval	Length	Conf. interval	Length
Wilson's score	(0.300, 0.679)	0.379	(0.174, 0.533)	0.359	(0.416, 0.785)	0.369
Wald	(0.284, 0.692)	0.408	(0.137, 0.522)	0.385	(0.419, 0.816)	0.397
Clopper-Pearson	(0.276, 0.703)	0.427	(0.150, 0.555)	0.405	(0.394, 0.810)	0.416
Arcsine	(0.290, 0.687)	0.397	(0.157, 0.530)	0.373	(0.415, 0.801)	0.386

For VZV the confidence intervals produced by the methods presented appear to be much more similar than they were for EBV, this might be because the probabilities of reactivation for VZV are further from zero. For the four methods, both the intervals and the interval lengths are similar. In general all the confidence intervals appear to be reliable since none of them differ much from one another.

Modelling the data

To model the data, a LRMM was fitted to each of the 50 data sets created after the data imputations by MICE were generated for both viruses, where the fixed parameters are the

study's time points and the random parameter is the subject. These models were fitted as is described in subsection 2.1.3. The explanatory variables in the models are *Subject* and *Time point* and the response variable is *Reactivation* – whether reactivation was detected or not. The variable *Time point* is a factor with seven levels: *L−180*, *L−45*, *Early*, *Mid*, *Late*, *R+0* and *R+30*. The equation of the models fitted for each data imputation of both viruses is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1\, x_{1a} + \beta_2\, x_{2a} + \beta_3\, x_{3a} + \beta_4\, x_{4a} + \beta_5\, x_{5a} + \beta_6\, x_{6a}\ , \quad (4.2)$$

where p_{ta} is the probability of viral reactivation at the time point t by astronaut a, $\beta_0, ..., \beta_6$ are the unknown coefficients, with $\beta_1, ..., \beta_6$ being the regression coefficients associated with each time point in the model and β_0 representing the intercept as the effect of the baseline time point *L−180*. The $x_{1a}, ..., x_{6a}$ represent the dummy variables associated with the time points *L−45*, *Early*, *Mid*, *Late*, *R+0* and *R+30*, respectively. The A_{0a} parameter represents the random effect associated with subject a.

After the fitting of all the models, the next step is to pool them all into a single LRMM with the characteristics for the models that constitute it. The parameters of all the models are represented in Appendix B. The pooling methods are the same that were used for the probabilities and their standard deviations. For the coefficients, the equation (2.18) was applied and for the standard errors, equation (2.19) was used for both viruses.

The pooled estimates for the probability of reactivation and standard errors and estimated p-values of the coefficients of the LRMM for EBV are represented in Table 4.12. For EBV the estimated probabilities of reactivation after pooling the data for all time points were 0.130, 0.435, 0.126, 0.239, 0.454, 0.391 and 0.217, for the *L−180*, *L−45*, *Early*, *Mid*, *Late*, *R+0* and *R+30* time points, respectively.

Table 4.12: Pooled LRMM fitted to EBV data with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *L−180*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*L−180*)	−1.897	0.619	0.004
β_1	(*L−45*)	1.635	0.749	0.029
β_2	(*Early*)	−0.127	0.925	0.891
β_3	(*Mid*)	0.736	0.790	0.351
β_4	(*Late*)	1.711	0.748	0.022
β_5	(*R+0*)	1.455	0.752	0.053
β_6	(*R+30*)	0.616	0.799	0.441

In the model of Table 4.12 the intercept parameter is *L−180*. For the remaining parameters, the larger the estimated probability of viral reactivation, the larger the value of its respective coefficient. The only time point with a negative value of the parameter is *Early*, so it is the only time point with an estimated probability of viral reactivation smaller than the intercept *L−180*, although it is a small difference and hence its large p-value. The *Late* time point has the largest parameter value and also the highest estimated probability of viral shedding. Time points *L−45* and *Late* have a p-value smaller than the significance level of 0.05, and so, their

estimated probabilities of viral reactivation are significantly larger than the baseline *L−180*, although, the parameter *R+0* is also close to the significance level considered. The parameters *Mid* and *R+30* have large p-values so their associated probabilities of viral reactivation are close to the probability of the reference level.

The random effect *Subject* follows a normal distribution with mean 0 and standard error 3.676×10^{-9}, this is a small value for the standard error which means that the effect each individual astronaut has on the model is null.

In Table 4.13 the pooled coefficients, standard deviations and estimated p-values of the coefficients of the LRMM for VZV are represented. For VZV the estimated probabilities of reactivation after pooling the data for all time points are 0, 0, 0.488, 0.330, 0.617, 0.435 and 0.087, for the *L−180*, *L−45*, *Early*, *Mid*, *Late*, *R+0* and *R+30*, respectively.

Table 4.13: Pooled LRMM fitted to VZV data with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *L−180*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*L−180*)	−21.149	5922.770	0.997
β_1	(*L−45*)	0.006	8323.595	1.000
β_2	(*Early*)	21.017	5922.770	0.997
β_3	(*Mid*)	20.058	5922.770	0.997
β_4	(*Late*)	21.813	5922.770	0.997
β_5	(*R+0*)	20.704	5922.770	0.997
β_6	(*R+30*)	17.928	5922.770	0.998

For this model of VZV data, the intercept parameter is *L−180*. The parameters of the model are all positive, so they all have a higher probability of reactivation than for the baseline *L−180*, although, the time point *L−45* has an estimated coefficient value close to zero. The reason for the coefficient value of *L−45* to be so small is because both before flight time points (*L−180* and *L−45*) had no detected viral reactivation for any astronaut. The *Late* time point has the highest parameter value and also the highest estimated probability of reactivation, the same as it were for EBV.

The standard errors of the parameters are large, which causes the p-values to be close 1. The reason for this could be the fact that the baseline *L−180* had no detected reactivations, thus, complete separation could be occurring at that time point. This time point and *L−45* have no detected reactivation of VZV, which might be causing the algorithm estimating the model's parameters not to be able to converge.

It is worth trying to fit a different model where the reference level is the time point *Early* because viral reactivation was both detected and not detected at this time point. For this mixed model the fixed effects remain the time points and the subject as the random effect. The new model was built in the same way as the model in Table 4.13 where first, a LRMM was fitted to each of the 50 data imputations with *Early* as the reference level. Then this 50 models were pooled into one model with the same characteristics as the ones that constitute it, using equations (2.18)

and (2.19) to achieve this. The equation of this new model is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1 x_{1a} + \beta_2 x_{2a} + \beta_3 x_{3a} + \beta_4 x_{4a} + \beta_5 x_{5a} + \beta_6 x_{6a} , \quad (4.3)$$

where p_{ta} is the probability of viral reactivation at the time point t by astronaut a, $\beta_0, ..., \beta_6$ are the unknown coefficients, with β_0 representing the effect of the intercept *Early* and $\beta_1, ..., \beta_6$ are the regression coefficients associated with the time points in the model. The $x_{1a}, ..., x_{6a}$ represent the dummy variables associated with the time points *L−180*, *L−45*, *Mid*, *Late*, *R+0* and *R+30*, respectively. The random effect associated with subject a is represented by the parameter A_{0a}.

Table 4.14: Pooled LRMM fitted to VZV data with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *Early*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*Early*)	−0.132	0.618	0.831
β_1	(*L−180*)	−21.035	1598.924	0.990
β_2	(*L−45*)	−20.802	1677.534	0.990
β_3	(*Mid*)	−0.959	0.738	0.194
β_4	(*Late*)	0.796	0.740	0.282
β_5	(*R+0*)	−0.313	0.716	0.662
β_6	(*R+30*)	−3.089	1.014	0.002

The pooled coefficients, standard deviations and estimated p-values of the coefficients of this LRMM are represented in Table 4.14. Straight way, it is clear that all the standard errors of the coefficients decreased considerably. All the parameters, except for *Late*, are negative, this is not a surprise because their respective estimated probabilities of viral reactivation are smaller than for the reference level *Early*. The only significant parameter in this model is *R+30*, indeed, the estimated probability of viral reactivation at this time point ($p_6 = 0.087$) and the reference level *Early* ($p_0 = 0.488$) are considerably different. All other parameters, except for *L−180* and *L−45*, have estimated probabilities of viral reactivation close to the probability at the reference level *Early*, hence they are not significant. The particularity of time points *L−180* and *L−45* not having detected viral reactivation makes their associated standard errors large, hence not significant in the model, even though their probabilities of viral reactivation are a great deal different from the reference level *Early*.

This model (Table 4.14) is clearly explaining the data better than the model in Table 4.13, nevertheless it is still interesting to fit a new model without the parameters *L−180* and *L−45*, whose standard errors of the respective estimate were too large. This third model was build the same way as two previous VZV models and its equation is:

$$\ln\left(\frac{p_{ta}}{1-p_{ta}}\right) = \beta_0 + A_{0a} + \beta_1 x_{1a} + \beta_2 x_{2a} + \beta_3 x_{3a} + \beta_4 x_{4a} , \quad (4.4)$$

where p_{ta} is the probability of viral reactivation at the time point t by astronaut a, $\beta_0, ..., \beta_4$ are the unknown coefficients, with $\beta_1, ..., \beta_4$ being the regression coefficients associated with each time point in the model and β_0 representing the effect of the intercept *Early*. The $x_{1a}, ..., x_{4a}$ represent the dummy variables associated with the time points *Mid*, *Late*, *R+0* and *R+30*,

respectively. The A_{0a} parameter represents the random effect associated with subject a.

Table 4.15: Pooled LRMM fitted to VZV data, excluding time points *L−180* and *L−45*, with respective parameter estimates, standard error and p-value. The reference level of the model is the time point *Early*.

Coefficient		Parameter estimate, $\widehat{\beta}$	Standard error	p-value
β_0	(*Early*)	−0.132	0.623	0.832
β_1	(*Mid*)	−0.959	0.745	0.198
β_2	(*Late*)	0.796	0.747	0.287
β_3	(*R+0*)	−0.313	0.722	0.665
β_4	(*R+30*)	−3.089	1.017	0.002

The pooled coefficients, standard deviations and estimated p-values of the coefficients of the third LRMM fitted to VZV data are represented in Table 4.15. This model was created using the same techniques as the two previous models of VZV. Aside from removing the coefficients *L−180* and *L−45*, all parameter estimates stay the same as the model in Table 4.14. The standard errors of the parameters increased slightly and consequently, so did their p-values. Again, all parameters, except for *R+30*, are not significant because their associated probabilities of viral reactivation are close to the probability of the reference level *Early*. For time point *R+30* the probability of viral reactivation is much smaller than the reference level, hence, this parameter is significant. All except coefficient *Late* have a negative parameter estimate, meaning that their respective probabilities of viral reactivation are smaller than for the reference level *Early*.

Of the three models fitted to the data, the first (Table 4.13) had large standard errors and no significant variables, likely because the algorithm estimating the model's parameters was not able to converge. The second and third models (Table 4.14 and Table 4.15, respectively) are very similar other than parameters *L−180* and *L−45* being excluded. Being this similar and given that parameters *L−180* and *L−45* are not significant to the model, the model in Table 4.15 appears to be a better choice to estimate the probability of viral reactivation. For all three models of VZV the standard error associated to the random effect *Subject* is 1.645.

4.5 Summary

For EBV there was detected viral reactivation at all time points and higher number of viral copies were detected at the *Mid* and *Late* time points. The average of the number of viral copies, given that shedding had occurred, was the highest at the *Mid* and *Late* time points, for EBV. The time points with the highest proportion of detected viral reactivation are *L−45*, *Late* and *R+0* with values around 40%. For EBV there appears not to be a relation between the average of the number of positive viral copies and the proportion of detected reactivations. The 95% confidence intervals fitted to the time points all intercept with each other, which suggests that the probabilities of viral reactivation are constant across the study. For VZV all the inflight time points and the *R+0* time point had big numbers of detected viral copies, while the before flight time points had no detected viral reactivation. The average of the number of viral copies, given that shedding had occurred, had big values in general, with the biggest value happening

at the time point *Late*. The *Early* and *Late* time points have the biggest proportions of viral reactivation with at least half of the subjects testing positive for viral shedding. For VZV there appears to be a correlation between the average of the positive number of viral copies and the proportion of detected reactivation. The 95% confidence intervals of *L−180* and *L−45* do not intercept with the intervals of *Early*, *Mid*, *Late* and *R+0*, suggesting there are differences in their probabilities of viral reactivation. The same happened for the intervals of time points *R+0* with *Early* and *Late*. It became visible that the astronauts did not acclimate to the stress conditions for the later days of their flight.

The Friedman test did not find significant differences between the viral copy numbers of time points for EBV [34]. McNemar's exact binomial test for complete categorical data found three time points to be significantly different between themselves for the same virus. For VZV, Friedman's test found significant differences between the reactivation dynamics of the time points [34]. Even after removing the time points with no detected reactivation the result of the test was still significant [34]. McNemar's exact binomial test for the complete categorical data of EBV found no significant pairs of time points after the Bonferroni correction. For VZV, the test found 11 pairs of time points that were significantly different from each other before the Bonferroni correction and five after, which is a big difference when compared to EBV and even CMV.

It is interesting to understand the mechanism by which the data are missing and the underlying reasons for why the data are missing. The test applied to assess the missingness mechanism present for the inflight time points suggests a MCAR mechanism for both viruses. This was not a surprise given that the test applied was asymptotic and the sample was small, so the test can lose power. The test of homogeneity of marginal probabilities using categorical data with missing responses to compare significant differences between the reactivation dynamics of the inflight time points was found to be impracticable because of the amount of null frequencies caused by the small data set.

For the data imputation, the algorithm's parameters were 50 data imputations, a maximum of 50 iterations and the imputation model was set to logistic regression for EBV and VZV. The convergence of the MICE algorithm was not evident, so statistics were used to confirm it. The imputations were pooled using Rubin's rules [44]. The estimates for the probability of viral reactivation for EBV are 0.126, 0.239 and 0.454 for the time points *Early*, *Mid* and *Late*, respectively. For VZV the estimates for the probability of viral reactivation for the same time points were, respectively, 0.488, 0.330 and 0.617. It was estimated that it was more probable to have positive viral reactivation for VZV than for EBV for the inflight time points.

A LRMM was fitted for each of the 50 imputations for both viruses that considered the time points as fixed effects and the subject as random effect. The LRMMs were then pooled into a single LRMM using Rubin's rules [44]. For the model of EBV the coefficients that were statistically significantly were *L−45* and *Late*, although coefficient of *R+0* was close to being significant. These parameters verified an increase in their probability of viral reactivation, when compared with the before flight reference level *L−180*. The random effect of this model followed a distribution $A_{0a} \sim \mathrm{N}(0, 3.676 \times 10^{-9})$. A standard error this small suggested a null effect of each astronaut on the model. For VZV, the model had large p-values for all the parameters

which suggested that the algorithm estimating the model's parameters was not converging. So a different model with *Early* as the reference level was estimated the same way as the previous model. For this new model the standard errors of all parameters decreased and the only significant coefficient was *R+30*. All other coefficients had an estimated probability of viral reactivation close to the reference level. Coefficients *L−180* and *L−45* were not significant even though their probabilities of viral reactivation were significantly different from the reference level *Early*, so a new model was created without this coefficients. The coefficients of this third model were all equal to the ones of the previous model, only their standard errors increased slightly and consequently their p-values, yet the significance of the coefficients did not change. Being so similar and given that parameters *L−180* and *L−45* are not significant to the model, this third model appears to be a better choice to estimate the probability of viral reactivation. The random effect of all these three models followed a distribution $A_{0a} \sim \mathrm{N}(0, 1.645)$.

Chapter 5

Discussion

5.1 Summary

The main objective of this project was to extend the analysis of Mehta et al. (2017) [34] that had the goal of determining whether the astronauts participating in a long duration mission would get accustomed to the stress conditions and hence mitigate the viral reactivation for the later moments of the spaceflight. Various statistical methods were applied and tested, based on the data set with the viral reactivation status collected from 23 astronauts, that was shared in the original study. The particularities of these data that made their analysis challenging is the fact that only 23 individuals participated in the study, the presence of missing responses and the large number of zeros observed.

First, the exploratory analysis of CMV was made using the data of the viral number of copies along with the binary data, where 1 represents detected viral reactivation. This analysis suggested a positive correlation between the proportion of reactivation and the number of viral copies detected for CMV. The 95% Wilson's score method confidence intervals of the probabilities of viral reactivation of time points *L−180* and *L−45* do not intercept, suggesting that there are statistical differences between their proportions of reactivation. The same happens for the pairs of time points *L−180* with *During* and *During* with *R+30*. The McNemar's test applied to the binary data found four pairs of time points with reactivation dynamics significantly different from each other. After the Bonferroni correction, only one pair of time points was significant.

A LRMM was built to infer about the probabilities of having a positive viral reactivation for CMV. For this model the time points were represented as the fixed effects and the random effect delineating different levels of the model was the subject. The first LRMM had very high standard errors and p-values for all coefficients, so a new model with a different reference level was fitted. In this second model, the coefficient of *R+30* was significant, although the intercept (*L−45*) and *During* parameters were close to being significant. The standard error of *L−180* at this second model was still very high, so a third model without this parameter was fitted. In this third model all parameter estimates and standard errors were the same as the second model, so given that it has less parameters this third model seems like a better choice to fit to

the CMV data.

The analysis of EBV and VZV was divided into analysis without missing responses, categorical data analysis with missing responses and data imputation. The exploratory analysis of EBV and VZV was made using the data of the viral number of copies along with the binary data. For EBV this analysis suggested that for the *Mid* and *Late* time points there is an increase in the amplitude of the viral number of copies. All the 95% Wilson's score method confidence intervals of the probabilities of viral reactivation for this virus intercepted with each other, suggesting that the proportion of viral reactivation did not change over time. For VZV the analysis suggested the existence of a correlation between the average of the positive number of viral copies and the proportion of detected reactivation. The 95% Wilson's score method confidence intervals of the probabilities of viral reactivation for this virus at time points *L−180* and *L−45* did not intercept with the intervals of the time points *Early*, *Mid*, *Late* and *R+0*, which suggests that the spaceflight had an impact on the probabilities of reactivation of VZV. The confidence interval of *R+30* also did not intercept with the intervals of time points *Early* and *Late*. For both viruses it appears that that the astronauts did not acclimate to the stress conditions for the later stages of their flight.

The McNemar's test applied to the binary data found three pairs of time points with reactivation dynamics significantly different from each other for EBV and 11 pairs for VZV. After the Bonferroni correction, no pairs of time points were significant for EBV and only five were for VZV. To assess the mechanism by which the data were missing, a test was applied to the inflight time points of the binomial data. This test suggested that the missingness mechanism present was MCAR for EBV and VZV. Using categorical data with missing responses, the test of homogeneity of marginal probabilities was found to be impracticable because of the large amount of null frequencies caused by the small data set.

After the convergence of the MICE algorithm had been met using supplementary statistics, the parameter estimates from different imputed data sets were pooled using Rubin's rules [44]. The estimates for the probability of viral reactivation for EBV are 0.126, 0.239 and 0.454 for the time points *Early*, *Mid* and *Late*, respectively, while for VZV the estimates for the same time points were, respectively, 0.488, 0.330 and 0.617. A LRMM was fitted for each of the 50 imputations for both EBV and VZV that considered the time points as fixed effects and the subject as random effect. The pooled model of EBV was statistically significantly at the intercept (*L−180*), *L−45* and *Late* coefficients. The only time point with an estimated probability of viral reactivation is *Early*. There are statistical differences between the before flight time point *L−180* and time points *L−45* and *Late* for EBV. The random effect followed a distribution $\mathrm{N}(0, 3.676 \times 10^{-9})$. A variance this small suggests a null effect of each astronaut on the model. The pooled model of VZV had large p-values for all the parameters which suggested that the algorithm estimating the model's coefficients was not converging. A new model with *Early* as the reference level was created for this virus. In this second model all the standard errors decreased considerably yet the only coefficient that was significant was *R+30*. For the time points with no detected viral reactivation (*L−180* and *L−45*) the standard error was still very large, so a third model without these parameters was created. In this third model all the parameter estimates stayed the same and the standard errors increased slightly, so the coefficient *R+30* was the only one that was

significant. The *Late* coefficient was the only one with a positive parameter estimate, which suggests that its estimated probability of viral reactivation is greater than for the reference level *Early*. In this third model all parameter estimates were the same as the second model and standard errors only increased slightly, so given that it has less parameters this third model seems like a better choice to estimate the probability of viral reactivation. The random effect subject had a distribution $N(0, 1.645)$ for the three models of VZV. These models that had time points with no observed reactivations and were unable to converge must be used carefully, especially because of the small sample used to fit them, since this lack of convergence could cause the algorithm to estimate a solution outside the parameter space.

5.2 Conclusion

Viral shedding of common herpesviruses has been associated with many diseases. It is necessary to understand the conditions by which these viruses shed in order to prevent them from spreading pathologies. Various approaches were applied with the objective of furthering the knowledge there is of these viruses. The small, zero-inflated data set with missing responses used offered difficulties in the analysis made.

Data imputation, such as MICE, was an important tool to efficiently deal with missing data. For these data it was specially important to use data imputation to avoid reducing the small data set even more. Missing data can introduce bias and reduce efficiency in the analysis [35], so data imputation allowed for the data set to be analysed using standard techniques for complete data like LRMMs. Multiple imputation was important to capture the intrinsic uncertainty in the missing values by estimating several different imputations. Multiple imputation was preferable over single imputation, because it reduced the bias associated with the small data set.

Many data sets have missing responses, particularly biological data sets, for diverse reasons, such as people not responding to surveys, some things not being easy to measure or even data entry errors. The `ACD` package of the R software was a very useful tool to assess the MCAR mechanism present in the data. The MICE algorithm created several imputations whose analysis resulted in reliable estimates for the probability of viral reactivation for the inflight time points of EBV and VZV. The LRMMs applied to the imputed data allowed for the probability of viral shedding to be estimated based on the time point in question.

The data used were not a random sample from the population, but from specific individuals with above average physical capabilities. Therefore, results and estimates made throughout this project should be used carefully and considered only in similar contexts. Ideally, the sample collected would have been bigger in order to increase the underlying statistical power. The development and use of imputation techniques could help to further explore the possibilities to deal with data with missing responses without losing statistical significance.

5.3 Final remarks

The main goals of this project were, originally, to extend the statistical analysis in Mehta et al. (2017) [34] by using longitudinal zero-inflated Poisson and negative-binomial models, together with different missing data imputation techniques, such as MICE. Although the imputation using MICE was made, the longitudinal zero-inflated Poisson and negative-binomial models were not created. The categorical data analysis performed was not in the original plan, nevertheless, it was considered best to assess the missingness mechanism and to further understand the marginal probabilities of categorisation.

Rubin's rules were useful to estimate the probabilities of viral reactivation at each time point, their associated variance and to build the LRMMs of the imputed data. There was another method considered during the realisation of this project. The method consisted in creating a new data set with the introduction of the variable *Imputation* as to have all observed and imputed values in one set. After that, fit a LRMM with the time points as fixed effects and the subject and imputation as random effects. The equation of this model was $\ln\left(\frac{\pi_{tai}}{1-\pi_{tai}}\right) = \beta_0 + b_{ai} + b_{0i} + (\beta_1 + b_{1i})x_1 + ... + (\beta_t + b_{ti})x_t,\ t = 1, ..., 6,\ a = 1, ..., 23,\ i = 1, ..., 50$. Where π_{tai} represents the probability of reactivation at time point t by subject a at imputation i, b_{ai} represents the variability effect associated with the variable subject a from imputation i, b_{0i} represents the variability effect associated with the variable imputation and b_{ti} represents the effect on the slope of the time point t and imputation i. This method was later abandoned because it over-fitted the model to the data.

The next step for this project is to continue the statistical analysis proposed, in particular, build the models suggested.

References

[1] Iyer, L., Balaji, S., Koonin, E., and Aravind, L. (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research,* **117**, 156–184.

[2] Vidyasagar, A. (2020) *What Are Viruses?*. Available at: *https://www.livescience.com/53272-what-is-a-virus.html*.

[3] Wu, K. (2020) *There are more viruses than stars in the universe. Why do only some infect us?*. Available at: *https://www.nationalgeographic.com/science/article/factors-allow-viruses-infect-humans-coronavirus*.

[4] Koonin, E., Senkevich, T., and Dolja, V. (2006) The ancient Virus World and evolution of cells. *Biology Direct,* **1**(1), 29.

[5] Zimmer, C. (2020) *The Secret Life of a Coronavirus - An oily, 100-nanometer-wide bubble of genes has killed more than two million people and reshaped the world. Scientists don't quite know what to make of it.* Available at: *https://www.nytimes.com/2021/02/26/opinion/sunday/coronavirus-alive-dead.html*.

[6] Lawrence, C., Menon, S., Eilers, B., Bothner, B., Khayat, R., Douglas, T., and Young, M. (2009) Structural and functional studies of archaeal viruses. *The Journal of Biological Chemistry,* **284**(19), 12599–12603.

[7] Edwards, R. and Rohwer, F. (2005) Viral metagenomics. *Nature Reviews Microbiology,* **3**(6), 504–510.

[8] Ryan, K. and Ray, C. (2004) *Sherris Medical Microbiology*, McGraw Hill, 4th edition.

[9] Mettenleiter, T. and Sobrino, F. (2008) *Animal Viruses: Molecular Biology*, Vol. 5, Caister Academic Press, 1st edition.

[10] Sandri-Goldin, R. (2006) *Alpha Herpesviruses: Molecular and Cellular Biology*, Caister Academic Press, 1st edition.

[11] Chayavichitsilp, P., Buckwalter, J., Krakowski, A., and Friedlander, S. (2009) Herpes simplex. *Pediatrics in Review,* **30**(4), 119–129.

[12] Staras, S., Dollard, S., Radford, K., Flanders, W., Pass, R., and Cannon, M. (2006) Seroprevalence of cytomegalovirus infection in the United States, 1988–1994. *Clinical Infectious Diseases,* **43**(9), 1143–1151.

[13] Khan, Z. (2020) *Varicella-Zoster Virus (VZV).* Available at: *https://emedicine.medscape.com/article/231927-overview.*

[14] Geder, L., Sanford, E., Rohner, T., and Rapp, F. (1977) Cytomegalovirus and cancer of the prostate: in vitro transformation of human cells. *Cancer Treat Rep.,* **61**(2), 139–146.

[15] Kumar, A., Tripathy, M., Pasquereau, S., Al Moussawi, F., Abbas, W., Coquard, L., Khan, K., Russo, L., Algros, M.-P., Valmary-Degano, S., and Adotevi, O. (2018) The Human Cytomegalovirus Strain DB Activates Oncogenic Pathways in Mammary Epithelial Cells. *EBioMedicine,* **30**, 167–183.

[16] Yoshikawa, T. (2004) Human herpesvirus 6 infection in hematopoietic stem cell transplant patients. *British Journal of Haematology,* **124**(4), 421–432.

[17] Rezk, S., Zhao, X., and Weiss, L. (2018) Epstein-Barr virus (EBV)-associated lymphoid proliferations, a 2018 update. *Human Pathology,* **79**, 18–41.

[18] Maeda, E., Akahane, M., Kiryu, S., Kato, N., Yoshikawa, T., Hayashi, N., Aoki, S., Minami, M., Uozaki, H., Fukayama, M., and Ohtomo, K. (2009) Spectrum of Epstein-Barr virus-related diseases: a pictorial review. *Japanese Journal of Radiology,* **27**, 4–19.

[19] Cherry-Peppers, G., Daniels, C., Meeks, V., Sanders, C., and Reznik, D. (2003) Oral manifestations in the era of HAART. *Journal of the National Medical Association,* **95**, 21S–32S.

[20] Dreyfus, D. (2011) Autoimmune disease: A role for new anti-viral therapies?. *Autoimmunity Reviews,* **11**(2), 88–97.

[21] Pender, M. (2012) *CD8+ T-Cell Deficiency, Epstein-Barr Virus Infection, Vitamin D Deficiency, and Steps to Autoimmunity: A Unifying Hypothesis.* Available at: *https://www.hindawi.com/journals/ad/2012/189096/.*

[22] Ascherio, A. and Munger, K. (2010) Epstein-barr virus infection and multiple sclerosis: a review. *Journal of Neuroimmune Pharmacology,* **5**(3), 271–277.

[23] Moreno, M., Or-Geva, N., Aftab, B., Khanna, R., Croze, E., Steinman, L., and Han, M. (2018) Molecular signature of Epstein-Barr virus infection in MS brain lesions. *Neurology,* **5**(4).

[24] Hassani, A., Corboy, J., and Al-Salam, S ad Khan, G. (2018) Epstein-Barr virus is present in the brain of most cases of multiple sclerosis and may engage more than just B cells. *PLoS One,* **13**(2).

[25] Bjornevik, K., Cortese, M., Healy, B. C., Kuhle, J., Mina, M. J., Leng, Y., Elledge, S. J., Niebuhr, D. W., Scher, A. I., Munger, K. L., and Ascherio, A. (2022) Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science,* **375**(6578).

[26] Sepúlveda, N., Carneiro, J., Lacerda, E., and Nacul, L. (2019) Myalgic Encephalomyelitis/Chronic Fatigue Syndrome as a Hyper-Regulated Immune System Driven by an Interplay Between Regulatory T Cells and Chronic Human Herpesvirus Infections. *Frontiers in Immunology,* **10**.

[27] Domingues, T. D., Grabowska, A. D., Lee, J.-S., Ameijeiras-Alonso, J., Westermeier, F., Scheibenbogen, C., Cliff, J. M., Nacul, L., Lacerda, E. M., Mouriño, H., and Sepúlveda, N. (2021) Herpesviruses Serology Distinguishes Different Subgroups of Patients From the United Kingdom Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Biobank. *Frontiers in Medicine,* **8**.

[28] Ruiz-Pablos, M., Paiva, B., Montero-Mateo, R., Garcia, N., and Zabaleta, A. (2021) Epstein-Barr Virus and the Origin of Myalgic Encephalomyelitis or Chronic Fatigue Syndrome. *Frontiers in Immunology,* **12**.

[29] Loebel, M., Strohschein, K., Giannini, C., Koelsch, U., Bauer, S., Doebis, C., Thomas, S., Unterwalder, N., von Baehr, V., Reinke, P., Knops, M., Hanitsch, L., Meisel, C., Volk, H.-D., and Scheibenbogen, C. (2014) Deficient EBV-Specific B- and T-Cell Response in Patients with Chronic Fatigue Syndrome. *PLOS ONE,* **9**(1), e85387.

[30] Glaser, R., Pearson, G., Bonneau, R., Esterling, B., Atkinson, C., and Kiecolt-Glaser, J. (1993) Stress and the memory T-cell response to the Epstein-Barr virus in healthy medical students. *Health Psychology,* **12**(6), 435–442.

[31] Khan, G., Fitzmaurice, C., Naghavi, M., and Ahmed, L. (2020) Global and regional incidence, mortality and disability-adjusted life-years for Epstein-Barr virus-attributable malignancies, 1990-2017. *BMJ Open,* **10**(8).

[32] Traylen, C., Patel, H., Fondaw, W., Mahatme, S., Williams, J., Walker, L., Dyson, O., Arce, S., and Akula, S. (2011) Virus reactivation: a panoramic view in human infections. *Future Virology,* **6**(4), 451–463.

[33] Mehta, S., Laudenslager, M., Stowe, R., Crucian, B., Sams, C., and Pierson, D. (2014) Multiple latent viruses reactivate in astronauts during Space Shuttle missions. *Brain Behavior and Immunity,* **41**, 210–217.

[34] Mehta, S., Laudenslager, M., Stowe, R., Crucian, B., Feiveson, A., Sams, C., and Pierson, D. (2017) Latent virus reactivation in astronauts on the international space station. *npj Microgravity,* **3**(11), 1–8.

[35] Barnard, J. and Meng, X. (1999) Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research,* **8**(1), 17–36.

[36] Fay, M. (2020) *Exact McNemar's Test and Matching Confidence Intervals.* Available at: *https://cran.r-project.org/web/packages/exact2x2/vignettes/exactMcNemar.pdf.*

[37] Pires, A. and Amado, C. (2008) Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods. *REVSTAT - Statistical Journal,* **6**(2), 165–197.

[38] Subbiah, M. and Rajeswaran, V. (2017) proportion: A comprehensive R package for inference on single Binomial proportion and Bayesian computations. *SoftwareX,* **6**, 36–41.

[39] Wallis, S. (2013) mice: Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics,* **20**(3), 178–208.

[40] Breslow, N. and Clayton, D. (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association,* **88**(421), 9–25.

[41] Gelman, A. (2005) Analysis of variance - Why it is more important than ever. *The Annals of Statistics,* **33**(1), 1–53.

[42] MacKenzie, D., Nichols, J., Royle, J., Pollock, K., Bailey, L., and Hines, J. (2017) *Occupancy Estimation and Modeling*, Academic Press, 2nd edition.

[43] Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,* **67**(1), 1–48.

[44] Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc., 1st edition.

[45] Peng, C., Harwell, M., Liou, S., and Ehman, L. (2006) *Advances in missing data methods and implications for educational research. In: Real data analysis*, Sawilowsky S.S., Charlotte, North Carolina.

[46] Rubin, D. (1976) Inference and Missing Data. *Oxford University Press,* **63**(3), 581–592.

[47] Little, R. and Rubin, D. (1987) Statistical Analysis with Missing Data. *Journal of Educational Statistics,* **16**(2), 150–155.

[48] Liu, X. (2015) *Methods and Applications of Longitudinal Data Analysis*, Academic Press, 1st edition.

[49] Poleto, F., Singer, J., and Paulino, C. (2014) A product-multinomial framework for categorical data analysis with missing responses. *Brazilian Journal of Probability and Statistics,* **28**(1), 109–139.

[50] Poleto, F., Singer, J., Paulino, C., Correa, F., and Jelihovschi, E. (2015) *Package 'ACD'.* Available at: *https://cran.r-project.org/web/packages/ACD/ACD.pdf.*

[51] van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research,* **16**, 219–242.

[52] Schafer, J. and Graham, J. (2002) Missing data: Our view of the state of the art. *Psychol Methods,* **7**(2), 147–177.

[53] Rubin, D. (1996) Multiple Imputation after 18+ Years. *Journal of the American Statistical Association,* **91**(434), 473–489.

[54] van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* **45**(3), 1–67.

[55] Yuan, Y. (2010) Multiple Imputation for Missing Data: Concepts and New Development. *AS Institute Inc., Rockville, MD.,* **49**, 1–11.

[56] van Buuren, S. (2012) *Flexible Imputation of Missing Data*, Chapman and Hall/CRC, 1st edition.

[57] Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association,* **32**(200), 675–701.

[58] Friedman, M. (1939) A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association,* **34**(205), 109.

[59] van Buuren, S. (2021) *Package 'mice'.* Available at: *https://cran.r-project.org/web/packages/mice/mice.pdf.*

[60] Graham, J., Olchowski, A., and Gilreath, T. (2007) Structural and functional studies of archaeal viruses. *How many imputations are really needed? Some practical clarifications of multiple imputation theory,* **8**(3), 206–213.

[61] Oberman, H., van Buuren, S., and Vink, G. (2021) *Missing the Point: Non-Convergence in Iterative Imputation Algorithms.* Available at: *https://www.researchgate.net/publication/355493989_Missing_the_Point_Non-Convergence_in_Iterative_Imputation_Algorithms.*

[62] Sepúlveda, N., Manjurano, A., Drakeley, C., and Clark, T. (2014) On the Performance of Multiple Imputation Based on Chained Equations in Tackling Missing Data of the African $\alpha^{3.7}$-Globin Deletion in a Malaria Association Study. *Annals of Human Genetics,* **78**, 277–289.

[63] Agresti, A. and Coull, B. (1998) Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician,* **52**, 119–126.

[64] Newcombe, R. (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine,* **17**, 857–872.

[65] Moreira, F., Sepúlveda, N., and Antunes, M. (2021) Longitudinal analysis of viral shedding in astronauts before, during and after a mission to the International Space Station. In *Book of Abstracts of SPE 2021, Évora, Portugal, 13–16 October 2021* (p. 103). Sociedade Portuguesa de Estatística.

Appendices

A Plots of convergence of MICE algorithm

Epstein-Barr virus

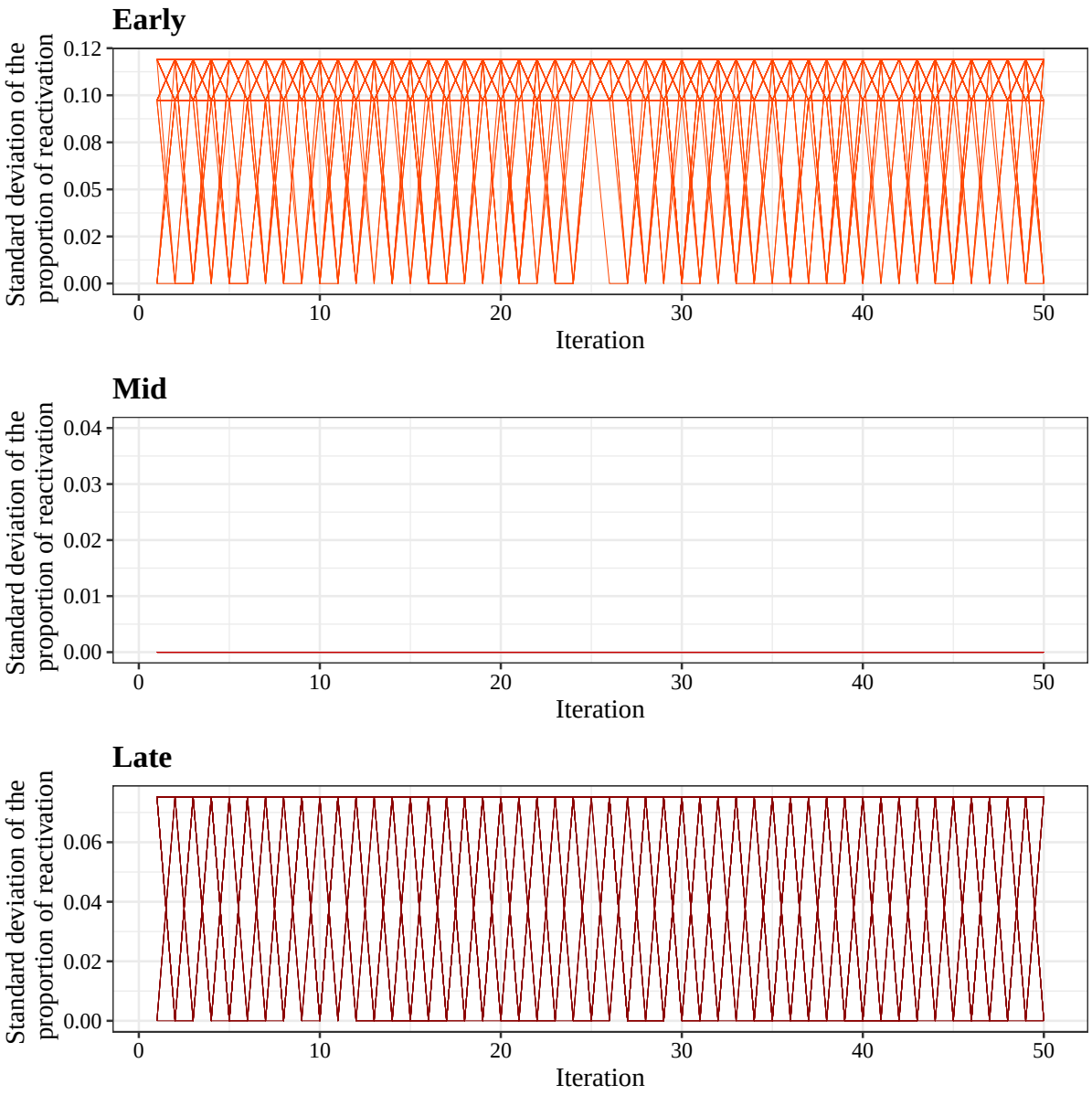

Figure A.1: Representation of the comportment of the standard deviation along the iterations for the inflight time points of EBV. The *Mid* time point only has one missing value so its associated standard deviation is constantly zero. In this plot there is no indication that convergence is occurring.

Varicella-zoster virus

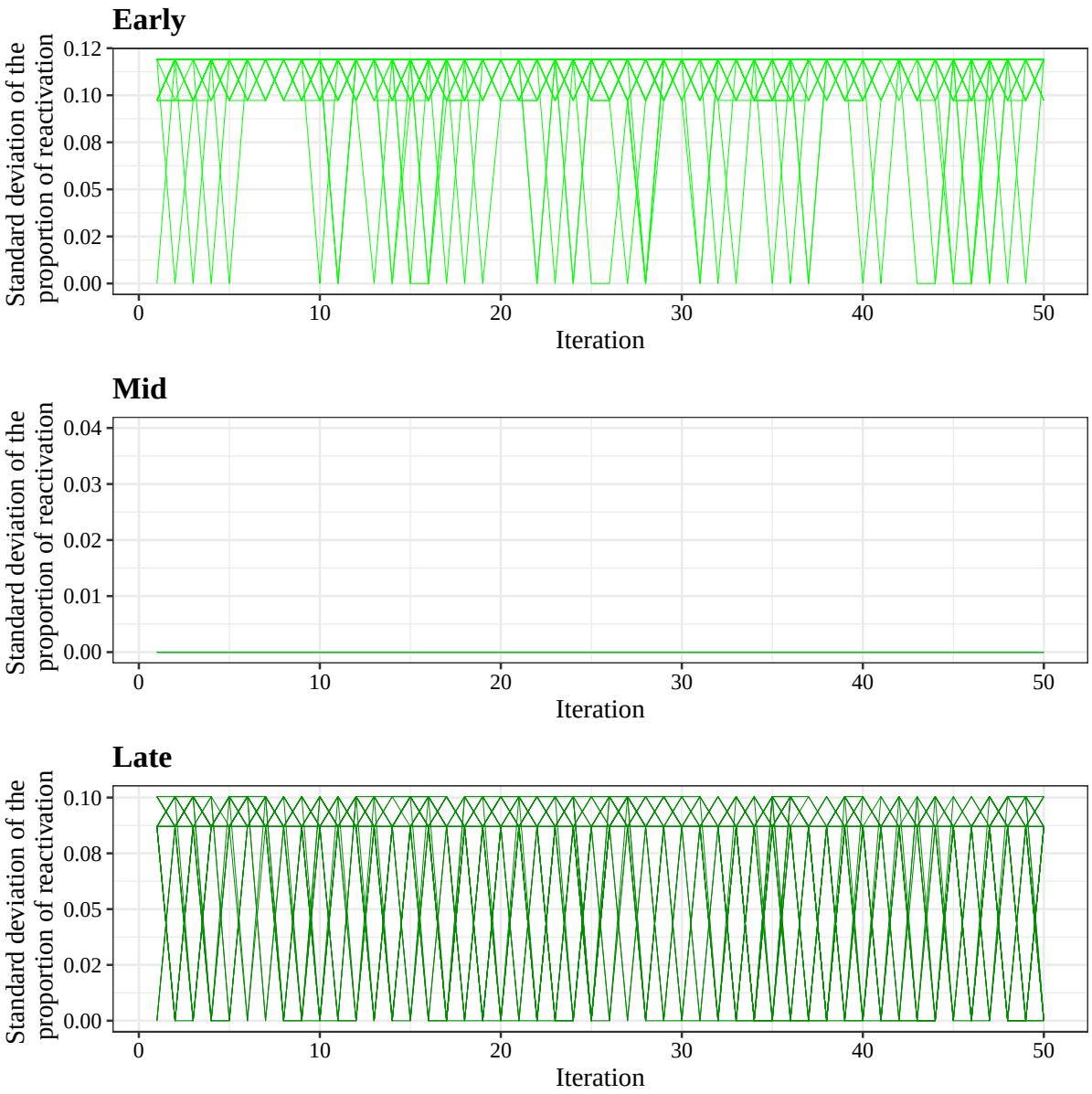

Figure A.2: Representation of the comportment of the standard deviation along the iterations for the inflight time points of VZV. The *Mid* time point only has one missing value so its associated standard deviation is constantly zero. In this plot there is no indication that convergence is occurring.

B Coefficients of LRMMs fitted to all data imputations

Epstein-Barr virus

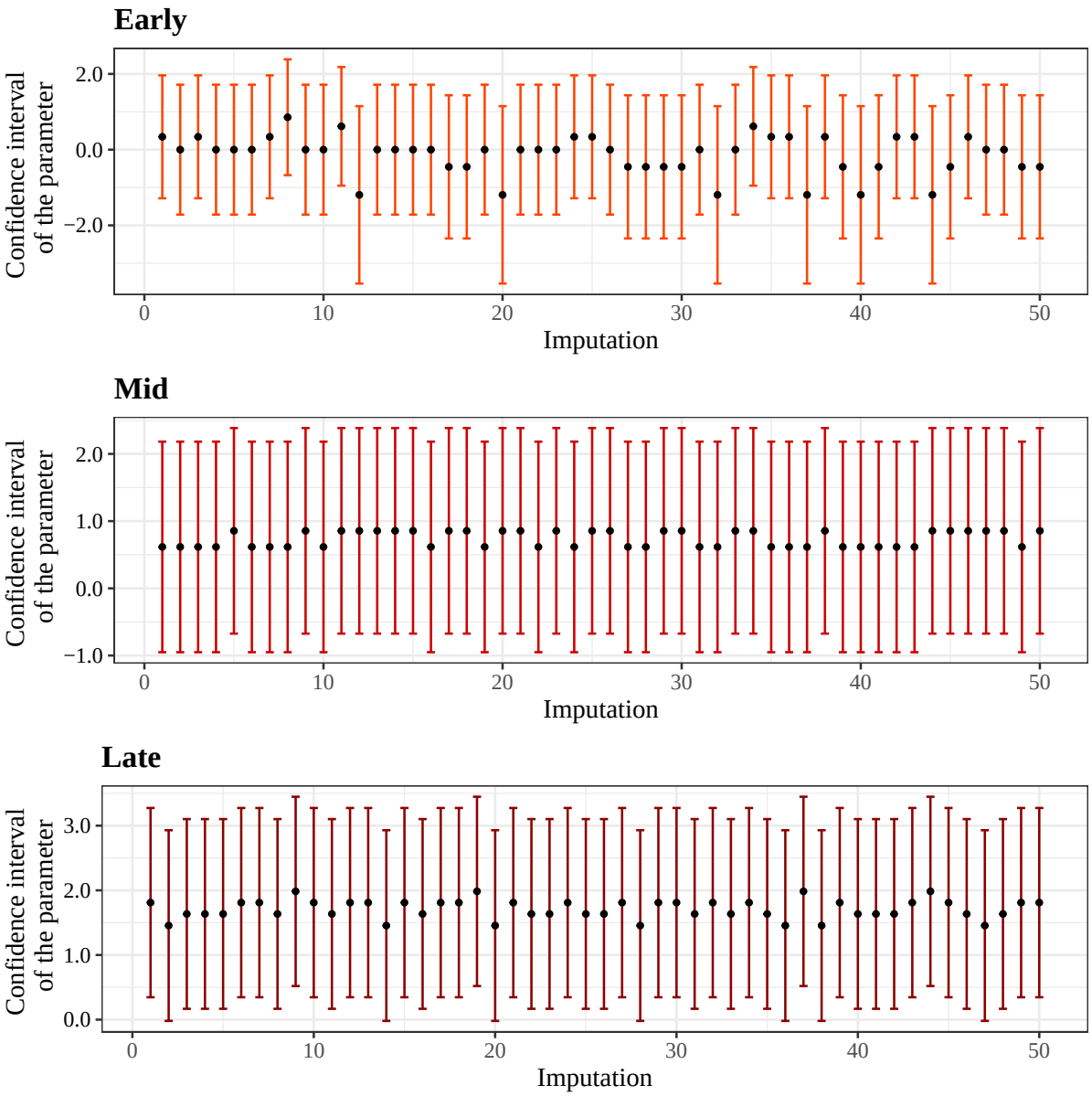

Figure B.1: Confidence intervals of the inflight coefficients of LRMM fitted to all data imputations for EBV. The method used is the 95% confidence interval for the mean. The reference level of the model is *L−180*. The parameters have a slight variation in the confidence intervals.

Varicella-zoster virus

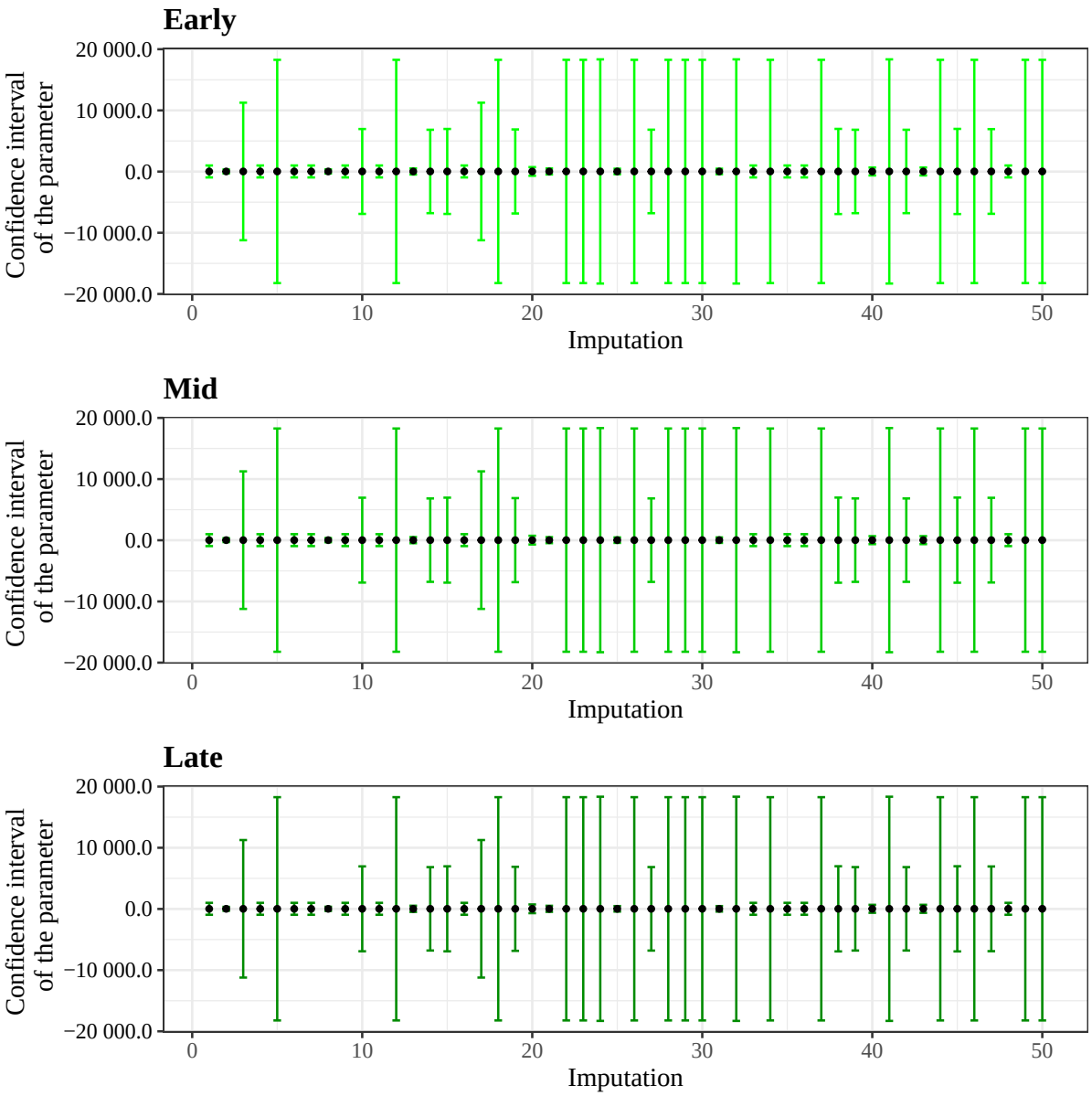

Figure B.2: Confidence intervals of the inflight coefficients of LRMM fitted to all data imputations for VZV. The method used is the 95% confidence interval for the mean. The reference level of the model is *L−180*. All the parameters have variation in the confidence intervals between themselves and there are big variation in their associated standard errors.

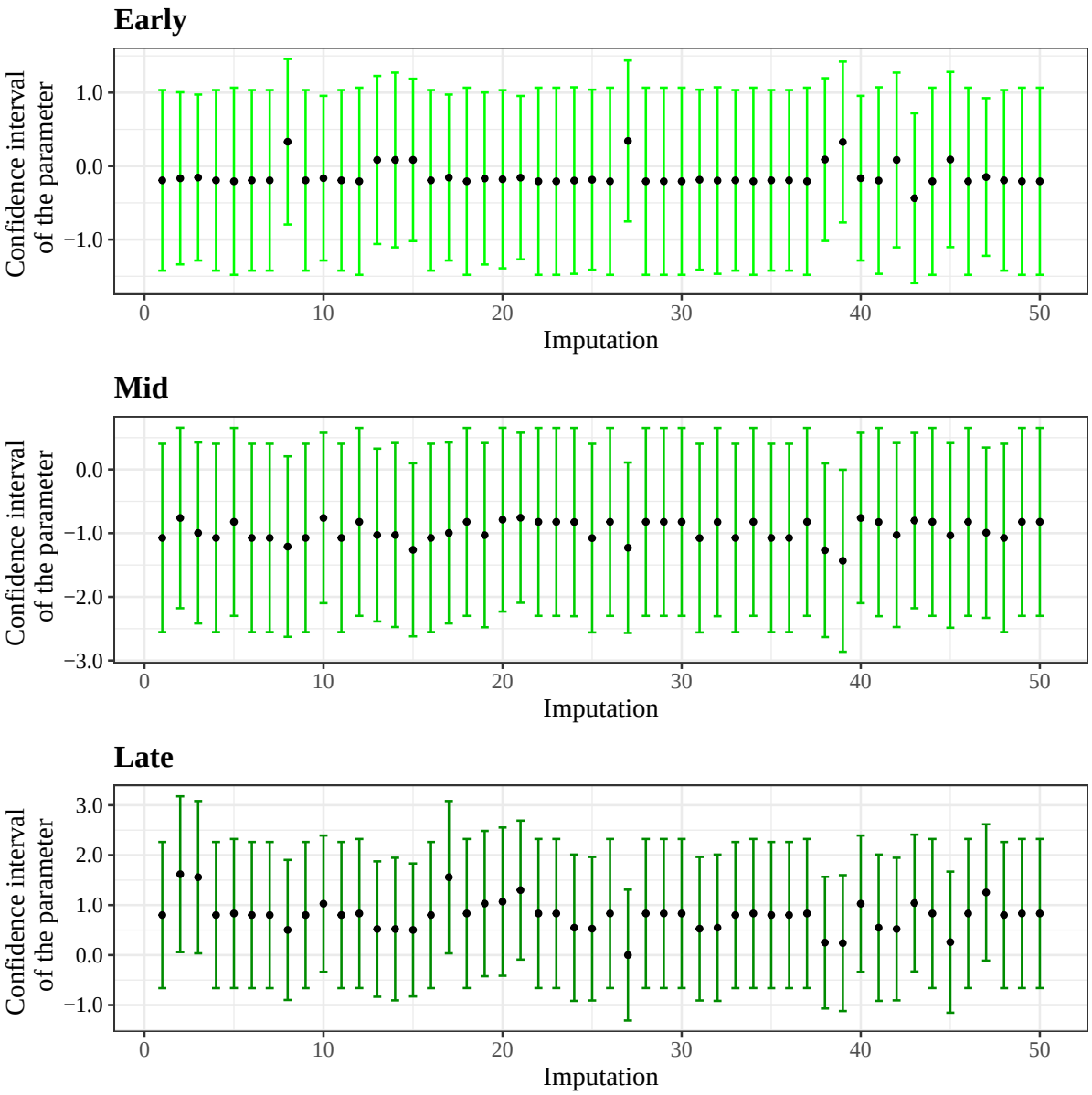

Figure B.3: Confidence intervals of the inflight coefficients of LRMM fitted to all data imputations for VZV. The method used is the 95% confidence interval for the mean. The reference level of the model is *Early*. The parameters have a slight variation in the confidence intervals. Compared to the original model fitted to VZV data, this model has much stabler parameter values and associated standard errors.

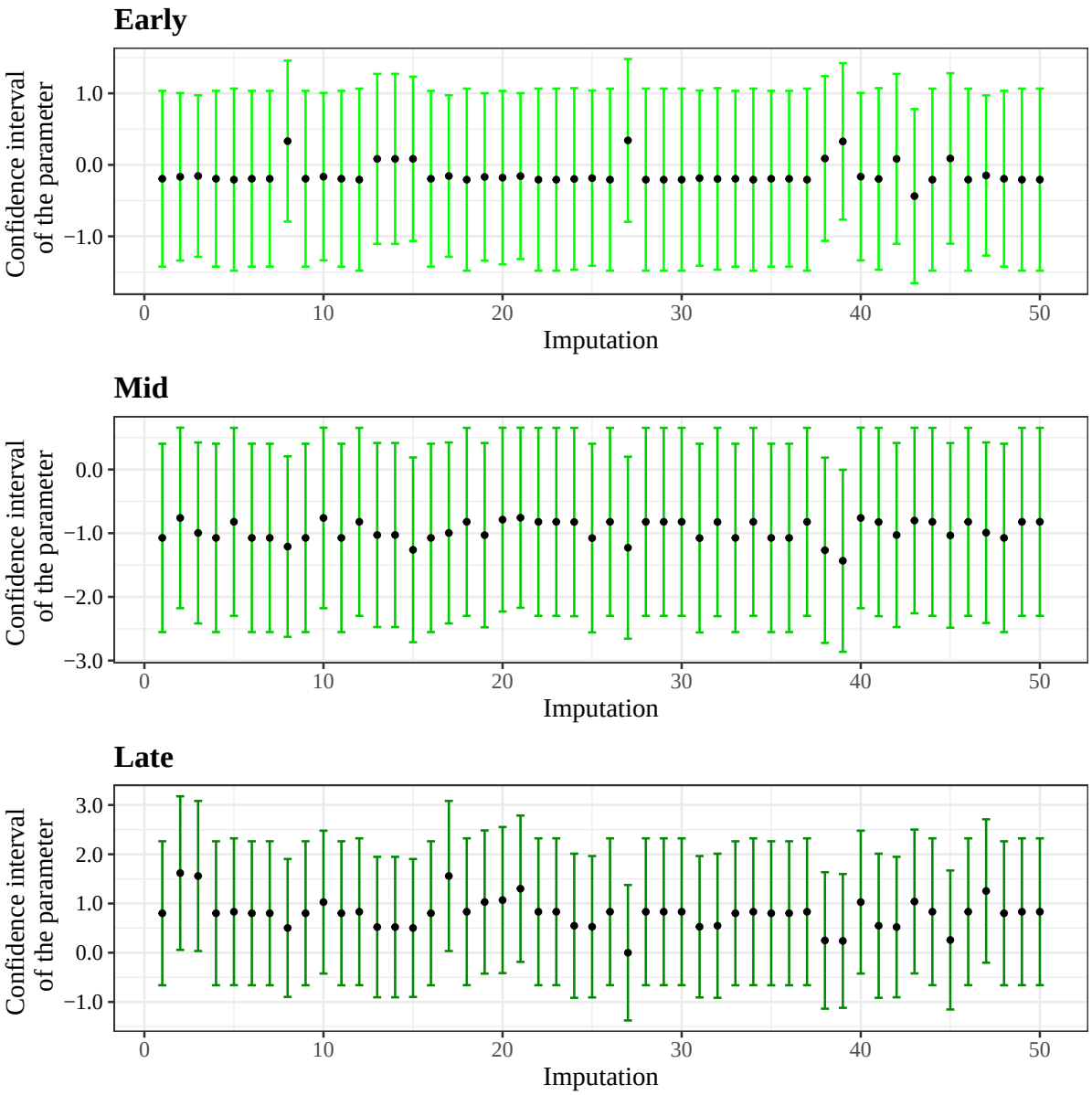

Figure B.4: Confidence intervals of the inflight coefficients of LRMM fitted to all data imputations for VZV excluding time points $L-180$ and $L-45$. The method used is the 95% confidence interval for the mean. The reference level of the model is *Early*. The parameters have a slight variation in the confidence intervals. Compared to the original model fitted to VZV data, this model has much stabler parameter values and associated standard errors.

Abstract for the conference [65]

Longitudinal analysis of viral shedding in astronauts before, during, and after a mission to the International Space Station

Frederico Moreira [a], Marília Antunes [a,b], Nuno Sepúlveda [b]
fc48046@alunos.fc.ul.pt, mcreis@fc.ul.pt, nunosep@gmail.com

[a] *Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*

Keywords: missing data, multiple imputation based on chained equations, longitudinal models, space flight

Abstract: In space, astronauts experience loss of gravity, lack of exposure to natural light and confinement. These stressful conditions could lead to the reactivation of dormant infections by common herpesviruses. These viral reactivations have been linked to different pathologies, such as chronic fatigue syndrome, cancer and arteriosclerosis. To understand these viral reactivations, a recent study analysed data on viral shedding in 23 astronauts participating in a long-duration mission to the International Space Station [1]. Viral shedding occurs when a virus is able to replicate productively within the host cell. The data of this study refers to viral counts measured before, during and after the flight and there were some missing data during the flight. The analysis of this data started with the imputation of the missing values using the method of multiple imputation by chained equations (MICE) [2]. Initially, MICE was used with the transformed binary data where 1 represented a detected reactivation in that given moment and 0 otherwise. In general MICE creates m imputation chains which are iteratively updated until a convergence is achieved. For each imputation a logistic mixed model was fitted describing the probability of reactivation in each timepoint. The pooling method proposed by Rubin estimated that the probabilities of reactivation for the Epstein-Barr virus were 0.126, 0.239 and 0.453 for the inflight timepoints "early", "mid" and "late", respectively. After testing the hypothesis, the only inflight timepoint with a model parameter significantly different from the baseline (referring to measurements taken 180 days before launch) was the "late" timepoint.

References

[1] Mehta S. K., Ladenslager M.L., Stowe R.P., Crucian B.E., Feiveson A.H., Sams C.F., Pierson D.L., Latent virus reactivation in astronauts on the International space station, *Npj Microgravity* 3:11, 2017.

[2] Van Buuren, S., Groothuis-Oudshoorn, K., mice: Multivariate Imputation by Chained Equations, *R. Journal of Statistical Software*, 45(3), 1-67, 2011.

[3] Rubin, D. B., Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91:434, 473-489, 1996.

Application for the participation grant

Longitudinal analysis of viral shedding in astronauts before, during, and after a mission to the International Space Station

Frederico Moreira [a], Marília Antunes [a,b], Nuno Sepúlveda [b,c]
fc48046@alunos.fc.ul.pt, mcreis@fc.ul.pt, nunosep@gmail.com

[a] *Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal*
[b] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*
[c] *Institute for Medical Immunology, Charité-Universitatsmedizin Berlin, Berlin, Germany*

Astronauts live under considerably challenging conditions in space, such as loss of gravity, lack of exposure to natural light and confinement, among others. These stressful conditions, like others occurring on earth, provide the perfect environment for the reactivation of latent infections by common herpesviruses. In turn, the reactivation of these viruses has been linked to many severe pathologies, including cancer, arteriosclerosis and the neglected Chronic Fatigue Syndrome. Increased reactivation of some naturally occurring latent herpesviruses including Epstein-Barr virus (EBV), varicella-zoster virus (VZV) and cytomegalovirus (CMV) was previously demonstrated in astronauts during short-duration (10-16 days) space shuttle flights (Mehta et al, 2014). It is interesting to know whether long-duration (up to 180 days) spaceflight aboard the international space station (ISS) would allow astronauts to acclimate to spaceflight and mitigate the impact of spaceflight-associated stressors on crewmembers. To understand the conditions by which these viruses could be reactivated, a recent study analysed data on viral shedding in 23 astronauts who had experienced a long-duration mission to the International Space Station (Mehta et al, 2017). Viral shedding occurs when a virus is able to replicate productively with the host cell.

In this study (Mehta et al, 2017), the data set was analysed by estimating the prevalence of viral shedding at different timepoints. These timepoints were organized by two, three and two moments of measurement before, during, and after the flight, respectively. Before the mission, the times of study were 180 and 45 days before the launch, the timepoints during the mission were stated as early (about 14 days after launch), mid (between mission days 60-120) and late (about 180 days of mission), and the last two time points were 3 hours and 30 days after the landing.

The viruses whose reactivation is under investigation were Epstein-Barr virus (EBV), varicella-zoster virus (VZV) and cytomegalovirus (CMV) and three other herpes viruses (HSV1, HSV2, HHV6). The measurements were collected in saliva and urine samples. The saliva samples were analysed for EBV, VZV, HSV1, HSV2 and HHV6 and the urine samples were analysed for CMV. For each moment in study, four samples of saliva were collected and the recorded number of viral copies refers to the highest of the counts. There was only sample of urine collected from each astronaut during flight so there are only 5 timepoints in study for that virus.

This study was conducted with 18 men and 5 women with overall mean age ± SE = 53 ± 4.9 years old. The nominal mission duration was approximately 180 days. Two crew members participated in shorter missions of approximately 2–3 months. For those crewmembers only two samplings were taken during flight, with data aligned with the Early and Mid for the 6-month crewmembers. Twenty apparently healthy subjects, matched for age and gender (16 males and 4 females, mean age ± SE of 49.3 ± 4.9 years) participated as ground-based viral reactivation controls. None of the 20 control subjects shed VZV or CMV and only two of them shed EBV. No astronauts or control subjects shed HSV1, HSV2, or HHV6 at any time throughout the study.

In this study the data set was basically analysed by estimating the prevalence of viral shedding at different timepoints. The goal of the current project is to extend this analysis, using the same data, which was shared within the respective publication (Table 1). The data sets consist of the number of viral copies measured for each virus and timepoint where the value 0 means no viral reactivation.

	EBV_data							VZV_data							CMV_data				
	Before Flight		During Flight			After Flight		Before Flight		During Flight			After Flight		Before Flight		During Flight	After Flight	
Subject	L-180	L-45	Early	Mid	Late	R+0	R+30	L-180	L-45	Early	Mid	Late	R+0	R+30	L-180	L-45	During	R+0	R+30
1	0	0	0	640	NA	128	0	0	0	368	0	NA	0	0	0	0	0	0	0
2	0	0	0	0	630	88	0	0	0	45	0	816	660	606	0	0	450	0	0
3	0	0	0	0	NA	98	0	0	0	0	0	NA	0	0	0	0	300	0	0
4	0	0	0	450	770	65	0	0	0	816	0	482	220	0	0	0	250	40	0
5	87	0	0	0	321	70	0	0	0	60	0	1300	0	0	0	48	50	90	0
6	0	150	0	0	0	71	0	0	0	61	0	480	180	0	0	136	0	120	89
7	0	34	0	0	0	102	0	0	0	0	0	0	0	0	0	45	120	70	0
8	0	110	100	0	NA	0	126	0	0	130	380	560	0	0	0	345	0	0	0
9	89	0	0	1020	1215	0	120	0	0	0	0	NA	0	0	0	0	400	0	0
10	0	46	0	0	687	0	117	0	0	0	0	0	0	0	0	0	0	0	0
11	0	143	0	0	814	0	390	0	0	0	370	570	0	0	0	0	350	0	0
12	0	98	0	0	900	0	0	0	0	200	290	110	0	0	0	0	0	0	0
13	66	0	0	0	0	0	0	0	0	0	0	NA	60	0	0	0	0	0	0
14	0	68	NA	0	400	496	0	0	0	NA	0	69	125	0	0	0	560	0	0
15	0	0	NA	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
16	0	0	NA	498	0	0	0	0	0	NA	0	120	300	0	0	0	0	0	0
17	0	81	0	0	0	0	0	0	0	160	180	150	230	0	0	57	378	0	0
18	0	0	0	0	400	0	0	0	0	0	156	0	155	0	0	0	0	0	0
19	0	0	0	0	0	350	0	0	0	0	276	150	0	0	0	0	0	0	0
20	0	0	0	490	0	0	0	0	0	280	340	356	200	0	0	56	467	0	340
21	0	55	0	NA	0	0	0	0	0	0	NA	414	120	63	0	0	0	0	0
22	0	50	NA	0	0	0	0	0	0	NA	0	0	0	0	0	60	80	40	0
23	0	0	NA	0	0	0	487	0	0	NA	0	0	0	0	0	0	790	450	0

Table 1 - Salivary VZV, Salivary EBV and urinary CMV copies in the 23 international space station crew-members before, during and after the flight.

L – Launch of spaceflight; R – Return of spaceflight.
Notes: Highest copy number of the four samples taken at each time point was given in this Table.
There was only one urine sample inflight, hence only one CMV measurement is shown.

A visual inspection of Table 1 shows a high abundance of zeros which suggested a relatively low frequency of viral shedding. For EBV, reactivation was observed in every

moment in study, while for VZV, there was no evidence of reactivation found in the two before the flight timepoints. The reactivation of CMV was observed in all timepoints with the exception of the first one. At the first glance there appears to be an increased number of copies detected during the mission when compared to before and after the flight. It is also worth mentioning that, for EBV and VZV, there are a few inflight measurements that are missing.

The goal of the original study (Mehta et al, 2017) was to determine whether the astronauts participating in a long duration mission would get accustomed to the stress conditions and hence mitigate the reactivation of these viruses in the late timepoint. After the analysis, it was possible to see that the exact opposite occurred and that an increase was detected in the reactivation of the viruses.

The analysis of this data set started with the imputation of the missing values. For this purpose, the method of multiple imputation by chained equations was used (Van Buuren & Groothuis-Oudshoorn, 2001). Multiple imputation (Rubin, 1996) is a good method for complex incomplete data problems. Missing data that occur in more than one variable presents a special challenge and so the philosophy behind the MICE methodology is that multiple imputation is best done in a sequence of small steps. This method should account for the process that generated the missing data - in the current case, the missing data was assumed to be missing completely at random; preserve the relations in the data and preserve the uncertainty about these relations. Initially, MICE was used with the transformed binary data where 1 represented a detected reactivation in that moment and 0 when there was no reactivation detected. MICE starts by creating m imputation chains (m to be defined by the user) which are iteratively updated from a set of initial guesses for the missing data until the convergence of a chosen statistic is attained. In the current case, the statistic for assessing the convergence of the imputation process was the estimated probability of reactivation using both observed and imputed data at a given iteration of the imputation process. At the end of the imputation procedure, the m collections of imputed data sets are analysed and the subsequent estimates of the quantities of interest are pooled and the respective uncertainty calculated. The analysis of each imputed data set was made using generalized linear mixed models to describe the probability of viral reactivation in the same individual over time. These models were based on Binomial distribution as the random component. The structural component of the model was a linear combination of fixed effects referring to the different timepoints plus an additional random effect at the level of the individual. The logit link function connected the probability of reactivation with the structural component of the model. After fitting the model to different imputed data set, the final step of the analysis consisted in pooling the estimates for the probability of reactivation at different time points. With this purpose, two pooling methods were used: (i) the classical pooling method proposed by Rubin; (ii) the pooling made by fitting a joint generalized mixed model to all imputed data sets using a random effect to describe the effect of each imputed data set on the estimation of the probability of reactivation.

For the first pooling method, the estimated probabilities of reactivation for EBV were 0.1522, 0.2304 and 0.4609 for the timepoints Early, Mid and Late, respectively. Although, when testing the hypothesis that each fixed effect parameter is equal to zero using the Wald test statistic, there was evidence for a significant effect of the timepoint Late with respect to the reference timepoint L-180. For the second pooling method, the estimated probabilities of reactivation for the same virus were 0.1336, 0.2103 and 0.4566 for the timepoints Early, Mid and Late, respectively. When the Wald test was applied to test the significance of fixed effects, all of these effects were considered different from zero at the significance level of 5%.

After analysing the data with binary variables related to reactivation and non-reactivation events, the idea is to perform the data imputation in terms of the number of viral copies. To this end, the imputation procedure is supposed to be based on zero-inflated models, such as the zero-inflated Poisson and the zero-inflated Negative Binomial models (Molenberghs & Verbeke, 2005). However, after a simple analysis of the mean and the variance of the data, it was concluded that a zero-inflated Negative Binomial model would suit better for this scenario, because it has an extra parameter that account for overdispersion in relation to what is expected from a zero-inflated Poisson model. The last step of the analysis is to make a longitudinal analysis of each virus over time and a joint longitudinal analysis of the multiple viruses.

Bibliography:

Mehta, S.K. et al. (2014), *Multiple latent viruses reactivate in astronauts during space shuttle missions*. Brain Behav. Immun. 41, 210–217.

Mehta S. K., Ladenslager M.L., Stowe R.P., Crucian B.E., Feiveson A.H., Sams C.F., Pierson D.L. (2017), *Latent virus reactivation in astronauts on the International space station*. Npj Microgravity 3:11.

Van Buuren, S., Groothuis-Oudshoorn, K. (2011), *mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software*, 45(3), 1-67. https://www.jstatsoft.org/v45/i03.

Rubin, Donald B. (1996), *Multiple Imputation After 18+ Years*. Journal of the American Statistical Association, Vol. 91, No.434, 473-489.

Molenberghs, G., Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer, USA.