

Data Science in HIV

Statistical approaches for therapeutic HIV vaccine data

Yovaninna Alarcón Soto

PhD Thesis
Department of Statistics and Operations Research



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Barcelona, 2021

Data Science in HIV

Statistical approaches for therapeutic HIV vaccine data

Yovaninna del Carmen Alarcón Soto

PhD Thesis directed by

Klaus Langohr

Guadalupe Gómez Melis

Department of Statistics and Operations Research

Universitat Politècnica de Catalunya

BARCELONATECH



Thesis presented in partial fulfillment of the requirements for the
Degree of Doctor by the Universitat Politècnica de Catalunya
Statistics and Operations Research

2020–2021

This research was supported by: Agencia Nacional de Investigación y Desarrollo (Chile) under Becas de Doctorado en el extranjero Becas Chile (2016-2020), Ministerio de Economía y Competitividad (Spain) under Grant MTM2015-64465-C2-1-R; Ministerio de Ciencia e Innovación (Spain) under Grant PID2019-104830RB-I00; Departament d'Empresa i Coneixement de la Generalitat de Catalunya (Spain) under Grant 2017 SGR 622 (GRBIO).

DEDICATION

Para mis padres, Edith y Sergio.

Para mi hermana Marioly.

AGRADECIMIENTOS

Desde pequeña jugaba con mi hermana y nuestras muñecas a que yo era la profesora, siempre era además la profesora de matemáticas. Creo que no fue coincidencia que al terminar la Secundaria decidiera estudiar Pedagogía en esta área. En la malla curricular de mi carrera había solo dos cursos de Estadística, y en el último de éstos conocí a Álvaro Cortínez quien gracias a su pasión por la Estadística y su tradicional dicho: “La Estadística es como la vida misma” hicieron encariñarme con la asignatura y por la amistad que aún nos une. Al decidir el tema del trabajo de fin de grado, le pedí a él que sea mi profesor tutor y aceptó. Nuestro trabajo era sobre muestreo en poblaciones finitas.

Después de la defensa de este trabajo, me aventuré a hacer el Magíster en Estadística Aplicada de la Universidad de Concepción donde conocí a la profesora Katia Sáez, a quien debo mi pasión por la Bioestadística. Agradezco su entusiasmo y la facilidad con la que explica en clases. En los cursos de consultoría que teníamos me orientó en el trabajo con doctorandos de medicina, enfermería, biología marina, entre otros. La sensación de estar aprendiendo algo que parecía no tener nada que ver con la estadística era fascinante, saber un poco más de algas, de estudios en hospitales y de conceptos médicos en general, realmente me cautivó. La profesora Katia me habló de una ex-alumna suya que estaba en ese momento estudiando en Barcelona y aquel comentario hizo que me imaginara estudiando aquí la línea de bioestadística.

Fue así que al terminar mis estudios de Magíster, decidí postular al Doctorado de la Universitat Politècnica de Catalunya donde me dijeron que no y me recomendaron que hiciese primero el Máster, que da acceso inmediato al Doctorado, y así lo hice, aunque costó. Postulé 3 veces a Becas Chile, con la misma ilusión cada vez, de poder salir de mi país y conocer otra realidad. En la tercera oportunidad me gané la beca. Estaba en casa con mis padres cuando lo supe, todos lloramos. Sé que nuestros motivos eran distintos: Yo de felicidad, al ver que mi sueño iba camino de realizarse. Ellos porque su hija se iba lejos, pero también de orgullo. Les agradezco a mis padres y a mi hermana por todo el apoyo y la comunicación continua que mantenemos a distancia, por todas nuestras videollamadas y las risas, a pesar de que un océano nos separa. Indiscutiblemente no estaría donde estoy si ustedes no me hubiesen enseñado el valor del esfuerzo y la humildad, a reír y luchar por mis sueños.

También doy gracias porque me dijeron que no al doctorado en primera instancia, pues había muchas cosas que debía aprender primero. Seguí el área de la bioestadística, y tuve cursos de Supervivencia, de Ómicas, de Epidemiología, de R, entre otros, que hoy se plasman también en esta tesis.

Durante este trayecto dos de mis profesores llamaron profundamente mi atención. Klaus, porque

preparaba tan bien las clases, que cuando me surgía una duda, él ya estaba explicando la respuesta. Siempre fue además muy recto en su palabra, si decía que algo iba a estar corregido el viernes después de las 14h, a las 14:01 podías asegurar que allí estaría corregido. Además siempre da un toque de humor en las clases, que resultan muy entretenidas. Por otro lado Lupe, no sé si lograré algún día hacer lo que ella hace. Te puede explicar el teorema más difícil, de la manera más fácil.

Klaus y Lupe son mis directores de tesis. Aprovecho para darles aquí las gracias porque aceptaron dirigirme. No podría haber escogido mejor. Klaus con su infinita paciencia nunca me recibió mal aunque tocara su puerta con dudas cinco veces en el mismo día. Lupe porque gracias a ella hoy formo parte de distintas sociedades, del grupo GRBIO, por sacar adelante los proyectos que tenemos y que nos han permitido ser visibles en el área. A ambos por toda la ayuda en este proceso, por todas las reuniones que tuvimos para clarificar ideas, y por alentarme aquellas veces que me sentí insegura, e incluso por otorgar tiempo fuera del horario de la universidad, para hablar por Skype conmigo cuando estuve en Boston, y por mantenerse en contacto durante la pandemia del 2020. ¡Os admiro mucho!

In 2018 I did a secondment at the T.H Chan School of Public Health, in Harvard University, thanks to Rebecca Betensky who said yes to my stay. I would also say thanks to Sebastien Haneuse, who supported me and have many talks with me at the Department of Biostatistics. He helped me to connect with different people in Harvard and encourage me to participate in the Working Group of HIV. Those three months in Boston I will treasure forever. Three cold months, plenty of activities, meetings, lunch seminars with pizza, a lot of work and new friendships.

Durante esta tesis pude aprender más del VIH y de datos ómicos, teniendo contacto directo con médicos y doctores. Gracias a Felipe García y a su doctorando Csaba Fehér por las reuniones, los emails, por facilitarme datos de interés y poder explicarme las variables clínicas. Para entender mejor los datos ómicos he de agradecer a dos profesores que también admiro mucho, Malu Calle y Alex Sánchez por sentarse conmigo más de una vez o hacer una videollamada para tratar de resolver mis dudas.

Agradezco también a IrsiCaixa y la Fundació Lluita contra la Sida (mi actual trabajo) por los datos facilitados, las reuniones de discusión de variables y metodología, por las idas a laboratorio y por la oferta laboral. Agradezco especialmente a Christian Brander, a Bea Mothe y Pepe Moltó, a quienes admiro por su vocación y por ayudar a que la ciencia siga avanzando.

Agradezco a mis compañeros del GRBIO por todas las reuniones, los ensayos, los proyectos, la divulgación. He podido ampliar mis conocimientos con vuestras presentaciones y hemos tenido bonitas discusiones. Hemos compartido en los retreat y también por videollamada últimamente. Me han regalado también la ilusión de poder participar en mi primer libro *L'Alfabet de l'Estadística*.

Agradezco a Sonia, por su infinita paciencia y disposición. A Toni por su amabilidad y ayuda. A Fran, Lore, Marta y Ceci por esos café y conversaciones. A Ceci, Klaus, Lesly y David por ser mis compis de gimnasio. A Lesly por las conversaciones y el apoyo en todo este proceso. David, te agradezco especialmente por todo el apoyo, la compañía, las revisiones del inglés, el té, las salidas de cultura y gastronomía (principalmente patatas bravas) por la ciudad.

A mi familia aquí, Carmencita, Jormy, Anil, Jessica Cabezas, Aldo, Vivi, Jessica Ibarra. Por todas las veces que se han preocupado por mí y por cómo va mi tesis. Por las conversaciones y los viajes compartidos. A Tere, Claudio y Pame por hacer la vida en Plaza Zurbarán mucho más alegre y fácil, sobre todo en cuarentena.

Finalmente, a toda mi gente en Chile, mis amigos que me mantienen actualizada de sus vidas, del contexto nacional y por su permanente interés en mi desarrollo personal y cuidado emocional. Mención especial para mi amigo Rodrigo, por toda la ayuda en la última etapa.

Tengo el corazón lleno, la felicidad de mi sueño por cumplir, ¡voy por ti Doctorado!

ABSTRACT

The present dissertation contributes to Data Science in the Human Immunodeficiency Virus (HIV) field, addressing specific issues related to the modelling of data coming from three different clinical trials based on the development of HIV therapeutic vaccines. The biological questions that these studies raise are identify biomarkers that predict HIV viral rebound; explain the time to viral rebound as a consequence of antiretroviral therapy (cART) stop considering the variability of data sources; and find the relationship between spot size and spot count from Enzyme-Linked Immunosorbent spot (ELISpot) assays data. To handle these problems from a statistical perspective, in this thesis we: adapt the elastic-net penalization to the accelerated failure time model with interval-censored data, fit a mixed effects Cox model with interval-censored data, and improve statistical methodologies to deal with ELISpot assays data and a binary response, respectively.

In order to address the variable selection among a vast number of predictors to explain the time to viral rebound, we consider an elastic-net penalization approach within the accelerated failure time model. Elastic-net regularization considers a possible correlation structure among covariates, which is the case of messenger RNA (mRNA) data. For this purpose, we derive the expression of the penalized log-likelihood function for the special case of the interval-censored response (time to viral rebound). Following, we maximize this function using distinct approaches and optimization methods. Finally, we apply these approaches to the Dendritic Cell-Based Vaccine clinical trial, and we discuss different numerical methods for the maximization of the log-likelihood.

To explain the time to viral rebound in the context of another study with data from several clinical trials, we use a mixed effects Cox model to account for the data heterogeneity. This model allows us to handle the heterogeneity between the Analytical Treatment Interruption (ATI) studies and the fact that the patients had different number of ATI episodes. Our method proposes the use of a multiple imputation approach based on a truncated Weibull distribution to replace the interval-censored by imputed survival times. Our simulation studies show that our method has desirable properties in terms of accuracy and precision of the estimators of the fixed effects parameters. Concerning the clinical results, the higher the pre-cART VL, the larger the instantaneous risk of a viral rebound. Our method could be applied to any data set that presents both interval-censored survival times and a grouped data structure that could be treated as a random effect.

We finally address two different issues that have arisen when analyzing the BCN02 clinical trial. On one hand, we fit univariate log-binomial models as an alternative to the usual logistic regression.

On the other hand, we use one/two- way unbalanced ANOVA to analyze the variability of the main outcomes from the ELISpot assays across time. Although these assays are widely used in the context of the HIV study, the relationship between spot size or spot count and other variables has not been studied until now.

In this thesis, we propose, develop, and apply different statistical approaches that contributes to answer diverse clinical questions that are relevant in several clinical trials. We have tried to highlight that to be able to choose the appropriate methodology, make correct clinical interpretations and contribute to a meaningful scientific progress, a narrow collaboration with scientists is necessary. We expect that the original results from this thesis will contribute to the path of development and evaluation of a therapeutic HIV vaccine, helping to improve the way of living of HIV-infected people.

RESUMEN

La presente tesis contribuye a la ciencia de datos abordando problemas biológicos relevantes en el desarrollo de vacunas terapéuticas para el Virus de Inmunodeficiencia Humana (VIH) mediante la modelización de datos procedentes de tres ensayos clínicos diferentes. Algunas de las cuestiones suscitadas en estos estudios y que esta tesis aborda son: identificar biomarcadores para estudiar los factores de riesgo del rebote viral del VIH, explicar el tiempo transcurrido hasta el rebote viral como consecuencia del cese de la terapia antirretroviral (cART) considerando la variabilidad de las fuentes de datos y estudiar la relación entre las variables spot size y spot count en ensayos inmunoabsorbentes (ELISpot). Para abordar cada uno de estos interrogantes desde una perspectiva estadística, en esta tesis hemos adaptado una penalización de red elástica para el modelo de vida acelerada (AFT) con datos censurados en un intervalo, ajustado un modelo de Cox de efectos mixtos con datos censurados en un intervalo y mejorado las metodologías estadísticas existentes para tratar los datos de los ensayos ELISpot y de respuesta binaria, respectivamente.

En primer lugar, hemos abordado el problema de tener más de cinco mil ARN mensajeros (ARNm) para explicar el tiempo hasta el rebote viral. Para ello, hemos considerado un enfoque de penalización de red elástica para el modelo de vida acelerada. Esta regularización considera una posible estructura de correlación entre las covariables, como sucede con los ARNm. Para este objetivo, primero derivamos la expresión de la función de verosimilitud penalizada considerando una respuesta censurada en un intervalo (tiempo hasta el rebote viral). A continuación, maximizamos esta función utilizando distintos enfoques y métodos de optimización. Finalmente, aplicamos estos métodos al ensayo clínico DCV2 y discutimos sobre diferentes enfoques numéricos para la maximización de la verosimilitud.

En segundo lugar, para explicar el tiempo hasta el rebote viral proponemos ajustar un modelo de Cox de efectos mixtos. Dado que el tiempo hasta el rebote viral está censurado en un intervalo utilizamos imputación múltiple basada en una distribución de Weibull truncada. Este modelo nos permite controlar la heterogeneidad entre los estudios de interrupción analítica del tratamiento (ATI) y el hecho de que los pacientes tengan diferente número de episodios ATI. Según el estudio de simulación que realizamos, nuestro método tiene propiedades deseables en términos de exactitud y precisión de los estimadores de los parámetros de efectos fijos.

Finalmente abordamos dos problemas diferentes dentro del ensayo clínico BCN02. Por un lado, ajustamos modelos log-binomiales univariados como alternativa a la clásica regresión logística. Por otro lado, utilizamos un modelo ANOVA no balanceado para analizar la variabilidad de los resultados

principales de los ensayos ELISpot a lo largo del tiempo. Aunque los ensayos ELISpot se usan a menudo en el estudio del VIH, la relación entre variables como el spot size, spot count y otras no se había estudiado hasta ahora.

En esta tesis hemos propuesto y desarrollado diferentes enfoques estadísticos que han dado respuesta a preguntas biológicas planteadas en tres ensayos clínicos. En este trabajo se destaca la importancia de que los distintos miembros de un equipo científico multidisciplinar colaboren estrechamente, para así poder determinar la metodología apropiada, hacer correctas interpretaciones clínicas de los resultados de éste y, de esta forma, contribuir a un progreso científico significativo. Esperamos que los resultados originales de esta tesis contribuyan al desarrollo y la evaluación de una vacuna terapéutica del VIH, lo cual ayudaría notablemente a mejorar la calidad de vida de las personas infectadas por VIH.

RESUM

La present tesi contribueix a la ciència de dades abordant problemes biològics rellevants en el desenvolupament de vacunes terapèutiques per al virus d'immunodeficiència humana (VIH) mitjançant la modelització de dades procedents de tres assaigs clínics diferents. Algunes de les qüestions suscitées en aquests estudis i que aquesta tesi aborda són: identificar biomarcadors per estudiar els factors de risc del rebot viral de VIH, explicar el temps transcorregut fins al rebot viral com a conseqüència de la cessació de la teràpia antiretroviral (cART) considerant la variabilitat de les fonts de dades i estudiar la relació entre les variables spot size i spot count en assajos immunoabsorbents (ELISPOT). Per abordar cadascun d'aquests interrogants des d'una perspectiva estadística, en aquesta tesi hem adaptat una penalització de xarxa elàstica per al model de vida accelerada amb dades censurades en un interval, ajustat un model de Cox d'efectes mixtos amb dades censurades en un interval i millorat les metodologies estadístiques existents per tractar les dades dels assajos ELISPOT i de resposta binària, respectivament.

En primer lloc, hem abordat el problema d'haver-hi més de cinc mil ARN missatgers (ARNm) per explicar el temps fins al rebot viral. Per a això, hem considerat un enfocament de penalització de xarxa elàstica per al model de vida accelerada. Aquesta regularització considera una possible estructura de correlació entre les covariables, com succeeix amb els ARNm. Per a aquest objectiu, primer derivem l'expressió de la funció de versemblança penalitzada considerant una resposta censurada en un interval (temps fins al rebot viral). A continuació, maximitzem aquesta funció utilitzant distints enfocaments i mètodes d'optimització. Finalment, apliquem aquests mètodes a l'assaig clínic DCV2 i discutim sobre diferents enfocaments numèrics per a la maximització de la versemblança.

En segon lloc, per explicar el temps fins al rebot viral proposem ajustar un model de Cox d'efectes mixtos. Atès que el temps fins al rebot viral està censurat en un interval utilitzem imputació múltiple basada en una distribució de Weibull truncada. Aquest model ens permet controlar l'heterogeneïtat entre els estudis d'interrupció analítica del tractament (ATI) i el fet que els pacients tinguin diferent nombre d'episodis ATI. Segons l'estudi de simulació que vam realitzar el nostre mètode té propietats desitjables en termes d'exactitud i precisió dels estimadors dels paràmetres d'efectes fixos.

Finalment abordem dos problemes diferents dins de l'assaig clínic BCN02. D'una banda, ajustem models log-binomials univariats com a alternativa a la clàssica regressió logística. D'altra banda, utilitzem un model ANOVA no balancejat per analitzar la variabilitat dels resultats principals dels assajos ELISPOT al llarg del temps. Tot i que els assajos ELISPOT s'usen sovint en l'estudi de VIH, la relació

entre variables com el spot size, spot count i altres no s'havia estudiat fins ara.

En aquesta tesi hem proposat i desenvolupat diferents enfocaments estadístics que han donat resposta a preguntes biològiques plantejades en tres assaigs clínics. En aquest treball es destaca la importància que els diferents membres d'un equip científic multidisciplinari col·laborin estretament, per així poder determinar la metodologia apropiada, fer correctes interpretacions clíniques dels resultats d'aquest i, d'aquesta manera, contribuir a un progrés científic significatiu. Esperem que els resultats originals d'aquesta tesi contribueixin al desenvolupament i l'avaluació d'una vacuna terapèutica de VIH, la qual cosa ajudaria notablement a millorar la qualitat de vida de les persones infectades per VIH.

“If you understand AIDS, you understand public health.
There’s almost no aspect of behavior, policy,
basic science, statistics, epidemiology, nutritional
interventions –everything– that does not touch HIV/AIDS.”

Max Essex

Harvard Public Health Review Spring/Summer 2011

CONTENTS

Contents	xvi
List of Tables	xx
List of Figures	xxii
1 Introduction	1
2 Data science in biomedicine	5
2.1 Introduction	7
2.2 Data science: global impact and dissemination	9
2.3 Data science in the biomedical field	15
2.3.1 Biomedical data science in the Web of Science	17
2.3.2 Multidisciplinary environment for biomedical data science	18
2.3.3 Standardization of information	20
2.4 Conclusions	22
3 Overview of HIV	23
3.1 History of HIV	23
3.2 HIV transmission	24
3.3 HIV RNA viral load	24
3.4 CD4 T cells and their role	25
3.5 Latent HIV reservoir	26
3.6 The HIV life cycle	26
3.7 Stages of HIV infection	27
3.8 Combination antiretroviral treatment	28
3.9 HIV RNA viral rebound	29
3.10 Therapeutic HIV vaccine	29
3.11 HIV cure and viral eradication	30
4 Concepts of survival analysis and omics data analysis	31
4.1 Survival analysis	31

4.1.1	Basic concepts	31
4.1.2	Interval-censored data	32
4.1.3	Nonparametric estimation of the survival function	33
4.1.4	Proportional hazards model	34
4.1.5	Accelerated failure time model	35
4.2	Omics data analysis	36
4.2.1	Transcriptome	36
4.2.2	Introduction to microarrays	38
4.2.3	Pipeline for mRNA analysis	38
5	Elastic-net approach for the accelerated failure time model	45
5.1	Introduction	45
5.2	From the ordinary least squares to the elastic net penalized regression model	46
5.2.1	Ordinary least squares	47
5.2.2	Ridge regression	48
5.2.3	Least absolute shrinkage and selection operator (LASSO)	49
5.2.4	Elastic net	50
5.2.5	Selection of the optimal tuning parameter λ_{OPT}	51
5.2.6	Elastic-net extension: the adaptive elastic net	52
5.3	State of the art	52
5.4	Elastic net approach with the proportional hazards model	55
5.4.1	Estimation of the model parameters with right-censored data	55
5.4.2	Estimation of the model parameters with interval-censored data	57
5.5	Elastic net approach for the accelerated failure time model	58
5.5.1	Relation between Weibull distribution and the log linear model	58
5.5.2	Log-likelihood function for Weibull model	59
5.5.3	The optimization problem	60
5.5.4	Different approaches to maximize the elastic-net penalized log-likelihood function	62
5.6	DCV2 dataset	64
5.6.1	Description of the design	64
5.6.2	Baseline clinical parameters	65
5.6.3	Viral rebound of HIV-infected patients in DCV2 trial	65
5.6.4	Time to viral rebound analysis using midpoint imputation	67
5.6.5	Fit of the AFT model by means of ad-hoc methods	71
5.7	Discussion	75
6	Mixed effects Cox model	77
6.1	Introduction	79

6.2	Notation and preliminaries	82
6.3	Parameter estimation in the mixed effects Cox model	83
6.3.1	Imputation of interval-censored survival times	84
6.3.2	Fit of the mixed effects Cox model	85
6.3.3	Pooling the results	86
6.3.4	Software issues	86
6.4	Effect of gender and pre-cART VL on the time to HIV RNA viral rebound considering multiple random effects	87
6.4.1	Descriptive analysis of the ATI dataset	87
6.4.2	Fit of the mixed effects Cox model	89
6.5	Simulation study	91
6.5.1	Simulation settings and data generation	91
6.5.2	Evaluation criteria	93
6.5.3	Simulation results	94
6.6	Discussion	94
7	Statistical methodologies applied to BCN02 clinical trial	101
7.1	BCN02 clinical trial	101
7.2	Clinical and survival data of BCN02	102
7.2.1	Descriptive analysis of clinical covariates	103
7.2.2	Survival models for time until viral rebound	105
7.3	Log-binomial regression model to study the patient profile	106
7.3.1	Log-binomial regression model	106
7.3.2	Fitted univariate log-binomial regression models	107
7.4	ELISpot assays for BCN02	107
7.4.1	What is an ELISpot assay?	108
7.4.2	Plate organization and settings for BCN02	109
7.5	Statistical methodologies to work with ELISpot assay data	111
7.5.1	Data management from BCN02 ELISpot assay	111
7.5.2	Distribution of spot size and spot count over time and its variability	112
7.5.3	Results	115
7.6	Discussion	118
8	Conclusions and further research	121
	Bibliography	125
	Appendices	138
A	Supplementary information for Chapter 2	139

B	Supplementary information for Chapter 5	141
C	R code used to fit the mixed effects Cox model	145
D	Additional information on the ATI data set	149
E	More information of BCN02 clinical trial	153
	E.1 Additional tables of BCN02 clinical trial	153
	E.2 Results for OUT and HTI region	159
F	Glossary	169

LIST OF TABLES

2.1	Current journals in the Data Science field up to March 2020.	15
3.1	Main types of drugs for HIV infection.	28
5.1	State-of-the-art of maximization methods.	54
5.2	Description of clinical covariates of the DCV2 study.	66
5.3	Coefficients of the selected mRNAs for each group of treatment and overall.	68
5.4	Official full name of the selected mRNAs in each group of treatment.	71
5.5	Coefficients of the selected mRNAs for each method (Approach A).	72
5.6	Official full name of the selected mRNAs with each method (Approach A).	72
5.7	Coefficients of the selected mRNAs for each method (Approach B).	74
5.8	Official full name of the selected mRNAs with each method (Approach B).	74
6.1	Description of the eight studies in ATI data set.	89
6.2	Estimation of the fixed effects parameters and the standard deviation of the random effects of Model (6.1) using the three-step imputation method.	91
6.3	Settings of the simulation study.	92
6.4	Fixed parameters estimators without right-censored observations and 15 imputations ($\beta_1 = 0.5, \beta_2 = 0.6$).	95
6.5	Fixed parameters estimators considering 10% right-censored observations and 15 imputations ($\beta_1 = 0.5, \beta_2 = 0.6$).	96
7.1	Demographic, clinical, and treatment characteristics of study patients at study entry. . .	103
7.2	Summary of continuous covariates for BCN02.	104
7.3	Type of outputs from the ImmunoSpot reader.	111
7.4	Analysis of variance for the model in (7.5).	115
A.1	Number of publications associated with the topics “Data Science”, “Big Data” and “Cloud Computing” in different countries from 2004 to 2019.	139
B.1	Set of predictors for the 5 subsets defined using iregnet and midpoint imputation. . . .	141
D.1	Inclusion criteria and treatment for the different studies of the ATI data set.	151

E.1	Summary of continuous covariates of BCN02.	153
E.2	Univariate fitted survival models of BCN02.	154
E.3	Univariate log-binomial regression models for patient profile.	155
E.4	Pool of peptides for HIVconsv in ELISpot assay of BCN02.	158
E.5	Pool of peptides for OUT in ELISpot assay of BCN02.	158
E.6	Pool of peptides for HTI in ELISpot assay of BCN02.	158
E.7	Cells per well in each timepoint for every subject in BCN02.	159

LIST OF FIGURES

2.1	Data science scheme based on the Conway's Venn diagram.	8
2.2	Google trends for the terms "Data Science", "Big Data", and "Cloud Computing" for global queries.	11
2.3	Google trends for the terms "Data Science", "Big Data", and "Cloud Computing" for some countries of Europe.	12
2.4	Google trends for the terms "Data Science", "Big Data", and "Cloud Computing" for United States and some of its states.	13
2.5	Google trends for the terms "Data Science", "Big Data", and "Cloud Computing" in some countries of Asia and in Australia.	14
2.6	Healthcare field process in which a data scientist is involved.	16
3.1	HIV life cycle	27
3.2	HIV RNA viral rebound	29
3.3	"Kick and kill" therapeutic approach	30
4.1	mRNA's role in protein synthesis.	37
4.2	A simplified view of a gene expression matrix.	38
4.3	Visualization of ten microarrays.	39
4.4	Values of standard deviations along all samples for all genes ordered from the smallest to the biggest.	42
5.1	Estimation picture for the LASSO and ridge regression.	50
5.2	Flowchart of patients in the DCV2 trial.	64
5.3	DCV2 clinical trial design.	65
5.4	Lengths of the ordered interval-censored times (weeks) until viral rebound of DCV2.	66
5.5	Lengths of the time (weeks) until viral rebound per intervention group.	67
5.6	Turnbull's estimations of survival functions of times to viral rebound.	67
5.7	Cross validated error plots for the three groups of DCV2.	69
5.8	Cross validated error plot for the overall DCV2 set.	70
6.1	Viral load dynamics during cART and the first ATI episode.	80
6.2	Lengths of the weeks until viral rebound of the 229 ATI episodes.	88

6.3	Non-parametric estimation of the distribution function of the time until viral rebound. . . .	90
7.1	BCN02 diagram	102
7.2	Estimation of the survival function of the time to viral rebound.	105
7.3	Estimate relative risks for in the univariate log-binomial regression model	107
7.4	The ELISpot assay workflow.	108
7.5	The ELISpot assay tools.	109
7.6	Pool of peptides organization in half a plate for BCN02.	110
7.7	Schematic representation of the selected conserved regions in the HIV proteome.	110
7.8	Scenario 1: spot count with or without replicate.	113
7.9	Scenario 2: spot size without replicate.	113
7.10	Scenario 3: spot size with replicate.	114
7.11	Mean and 95% confidence interval for spot counts in HIVconsv region.	116
7.12	Mean and 95% confidence interval for spot size in HIVconsv region.	117
7.13	Mean and 95% confidence interval for spot size and spot count in HIVconsv region.	119
7.14	Spearman correlation between spot size and spot count in HIVconsv region.	120
E.1	Mean and 95% confidence interval for spot counts in OUT region (1).	160
E.2	Mean and 95% confidence interval for spot counts in OUT region (2).	161
E.3	Mean and 95% confidence interval for spot counts in HTI region.	162
E.4	Mean and 95% confidence interval for spot size in OUT region (1).	163
E.5	Mean and 95% confidence interval for spot size in OUT region (2).	164
E.6	Mean and 95% confidence interval for spot size in HTI region.	165
E.7	Mean and 95% confidence interval for spot size and spot count in OUT region (1).	166
E.8	Mean and 95% confidence interval for spot size and spot count in OUT region (2).	167
E.9	Mean and 95% confidence interval for spot size and spot count in HTI region.	168

LIST OF ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
AFTM	Accelerated Failure Time Model
ANOVA	Analysis of Variance
ATI	Analytical Treatment Interruption
BDS	Biomedical Data Science
cART	Combination Antiretroviral Therapy
CTL	Cytotoxic T lymphocytes
ELISPOT	Enzyme-Linked Immunospot
HIV	Human Immunodeficiency Virus
IPL	Integrated Partial Likelihood
LASSO	least Absolute Shrinkage and Selection Operator
LRA	Latency Reversing Agent
MAP	Monitored Antiretroviral Pause
miRNA	Micro ribonucleic acid
MLE	Maximum Likelihood Estimator
mRNA	Messenger ribonucleic acid
MSE	Mean Squared Error
NPMLE	Non-Parametric Maximum Likelihood Estimator
OLS	Ordinary Least Squares
PBMC	Peripheral Blood Mononuclear Cell

RMD	Romidepsin
RNA	Ribonucleic Acid
SIV	Simian Immunodeficiency Virus
SFC	Spot Forming Cell
VL	Viral Load

INTRODUCTION

The use of different technological devices in our lives produces tons of data every second. Data Science is a discipline that arises from the need to draw conclusions and knowledge from these data. It can be seen as the confluence of distinct fields, including Statistics, Informatics, and a specific field of research. In Data Science, the combination of multidisciplinary team work and the individual skills of the team members are equally essential to obtain a complete vision of a certain problem, the methodology to approach that problem, the correct interpretation of the research results, and the participation in the decision-making process based on these results. It is common to find the concept of Data Science linked with areas such as business or finance, however, it is not really clear if it can be related to other knowledge fields. In this thesis, we aim to explore some applications of Data Science in a specific biomedical discipline: the research on Human Immunodeficiency Virus (HIV-1), and more specifically, on the development of a therapeutic vaccine for HIV-1-infected patients.

To deepen my knowledge on HIV-1 and its dynamics, I have collaborated for 4 years with different institutions and hospitals in Barcelona, such as the IrsiCaixa AIDS Research Institute, the Fight AIDS and Infectious Diseases Foundation, and the Hospital Clínic. During these collaborations I have worked with distinct groups of people working in various clinical trials that explore the development of a therapeutic vaccine for HIV-1. These multidisciplinary groups are composed by physicians, senior researchers, postdoctoral researchers, laboratory technicians, and nurses, among others. Moreover, during the last years, I have attended several meetings and seminars on HIV-related topics. I also visited a biological safety laboratory to better understand the practicalities of a type of assays, the Enzyme-Linked ImmunoSpot (ELISpot) assays, that were particularly relevant to the goals of my thesis. These activities and interactions helped me to become acquainted with the steps, factors, and people involved in a clinical trial. They also helped me to understand the dynamics of the virus in HIV-1-infected patients. Moreover, I started to understand the language that clinicians use, which was

key to have a fluid communication with them.

In 2018, I spent 3 months at the Harvard T.H. Chan School of Public Health at Harvard University, Boston, USA. During these months, I had meetings with prominent scientists in research areas such as Biostatistics and Bioinformatics and attended some plenary talks they gave. Also, I interacted with different postdoctoral researchers working in these areas. Interestingly, I attended a seminar with the participation of Timothy Ray Brown, the “Berlin patient”, who was until that year the only patient cured of HIV-1. Thanks to this seminar, I found out how an HIV-1-infected person copes with the infection and how his experience may help the scientific community to better understand the virus. During that same stay, I participated in the HIV Working Group’s seminars organized by the Department of Biostatistics of the Harvard T.H. Chan School of Public Health. These seminars made me approach various ongoing clinical trials, where diverse techniques were applied depending on the hypothesis to be tested.

All these experiences and collaborations have greatly contributed to the development of this work. The general goal of this thesis is to provide adequate and rigorous statistical methodology for different HIV-1 studies. In particular, we have made various types of contributions to three different clinical trials that address particular questions related to HIV-1-research: the DCV2 clinical trial ([García et al., 2013](#)), the ATI study ([Leal et al., 2017](#)), and the BCN02 clinical trial (*Study to Evaluate the Safety and Effect of HIVconsu Vaccines in Combination With Histone Deacetylase Inhibitor Romidepsin on the Viral Rebound Kinetic After Treatment Interruption in Early Treated HIV-1 Infected Individuals. ClinicalTrials.gov Identifier: NCT02616874, 2018*). In each of these trials, we aimed to identify biomarkers in order to detect potential risk factors of HIV viral rebound, explain the time to viral rebound as a consequence of stopping the combination antiretroviral treatment (cART) considering the variability of data sources, and improve statistical techniques to deal with Enzyme-Linked Immunosorbent Spot (ELISpot) assays data and a binary response, respectively.

The DCV2 clinical trial was led by Felipe García at the Hospital Clinic, Barcelona. In this study, HIV-1-infected patients were induced to specific immune responses with a therapeutic immunization treatment based on dendritic cells. This treatment aimed to control viral replication after the discontinuation of complementary antiretroviral therapy. The ATI study, also coordinated by Felipe García at the Hospital Clínic (Barcelona), is not is not a clinical trial but a recompilation of retrospective data from eight Analytical Treatment Interruption (ATI) clinical trials where HIV-1-infected patients were treated with cART. The BCN02 clinical trial, conducted in the Hospital Universitari Germans Trias i Pujol (Badalona) and in the Hospital Clínic (Barcelona), was led by Beatriz Mothe Pujadas and sponsored by the IrsiCaixa AIDS Research Institute. Here, we had two main goals: to study the patient profile (controller or rebounder) and to analyze some variables coming from a specific immunoassay called ELISpot.

Although these three clinical trials tested specific hypotheses that led us to use different statistical methodologies, they presented some common characteristics. They aimed to develop a therapeutic vaccine for HIV-1-infected patients. In addition, the outcomes of interest were the times to a particular

event, such as the interval-censored times to viral rebound. To approach these hypotheses, we used different methods that we detail below.

Thesis structure and contents

Given that this thesis has been carried out in a multidisciplinary context, we have added three chapters, Chapters 2, 3, and 4, that will provide the background and main concepts to better understand the core chapters of this thesis. These core chapters are Chapters 5, 6, and 7, which correspond to our original research contributions.

In Chapter 2, we show the importance of Data Science in the biomedical field. Using Google Trends, we carry out a search to explore the evolution of the terms “Data Science” along with “Big Data” and “Cloud Computing” until December 2019 in different parts of the world. We also show an overview of the journals dedicated to this discipline and introduce some examples of the applications of Data Science within the biomedical field. Chapter 3 introduces the main variables of interest in this thesis and the vocabulary that we will employ throughout this document, regarding the HIV field. Here, we describe HIV-related vocabulary that we will employ throughout this document. Moreover, we give an overview of the history of HIV, its life cycle, the forms of transmission, the stages of HIV infection, and the current HIV-cured patients. In Chapter 4, we provide the background of survival analysis and omics data analysis. We have grouped these two fields of study because they are the theoretical foundations underpinning the following chapters.

Chapter 5 is the first of our three original research contributions. Here, we develop and apply an elastic-net penalization for the accelerated failure time model in the context of the DCV2 clinical trial. Using this approach, we aim to study the interval-censored times to viral rebound considering more than five thousands mRNAs as possible predictors. More specifically, we start the chapter presenting the main concepts related to elastic-net penalization. Then, we review various approaches (parametric, semiparametric, and piecewise exponential) in the case of complete and right-censored data for the proportional hazards model and the accelerated failure time model (AFTM). To the best of our knowledge, elastic-net penalization has not been used in the AFTM with interval-censored data. We derive the expression of the penalized log-likelihood function considering an interval-censored response. Following, we maximize this function by means of an Expectation- Maximization (EM)-based algorithm. Next, we show a simulation study that we performed to examine the properties of our approach. To carry out our procedure we wrote various functions in R since, up to our knowledge, there is no package that implements elastic-net penalization for accelerated failure time model using interval-censored data. Finally, we apply this methodology to the data coming from the DCV2 clinical trial.

In Chapter 6, we present a mixed effects Cox model considering a multiple imputation approach for interval-censored times to viral rebound. The use of the mixed effects Cox model is motivated by the ATI study. As data come from distinct studies and each patient can have more than one assessment, we handle this variability using the mixed effect Cox model. This model considers a random intercept

per subject and a correlated intercept and slope for pre-cART viral load per study. We adopt a multiple imputation approach based on a truncated Weibull distribution to replace the interval-censored data by imputing right-censored or exact survival times. Following, we perform a simulation study to investigate its properties in terms of bias and mean squared error. In the end, we apply this methodology to the ATI study mentioned above. This research gave rise to a publication in the Biometrical Journal ([Alarcón-Soto et al., 2019](#)).

In Chapter 7, we provide different statistical techniques applied to the BCN02 clinical trial. We divide these methodologies into two parts. In the first part, we use univariate log-binomial regression models to identify covariates that can explain the profile of the patient, classified as controller or rebounder, according to whether they keep the viral load controlled at week 12 after monitored antiretroviral pause (MAP), and discuss the advantages of using these models instead of the classical logistic regression. Log-binomial regression uses the Risk Ratio (RR) as an association measurement. This analysis is presented in a manuscript that has recently been published in *Frontiers in Immunology* ([Mothe et al., 2020](#)). The second part of the chapter shows the statistical approach applied to the immunology variables involved in this trial that come from the ELISpot assays. Until now, the replicates presented in these type of assays were averaged and then, the variability was obtained without considering the replicates. To take into account the replicates in the assay, we use an unbalanced one/two-way ANOVA, so that we can compute the variability of the main variables considering the effect of the main factors. Although ANOVA is widely known, it is generally not used in these type of trials.

Finally, Chapter 8 reviews the main results of this thesis and provides ideas for further research based on the methodologies explored.

DATA SCIENCE IN BIOMEDICINE

We highlight the role of Data Science in Biomedicine. This chapter goes from the general to the particular, presenting a global definition of Data Science and showing the trend for this discipline together with the terms of cloud computing and big data. In addition, since Data Science is mostly related to areas like economy or business, we describe its importance in biomedicine. Biomedical Data Science (BDS) presents the challenge of dealing with data coming from a range of biological and medical research, focusing on methodologies to advance the biomedical science discoveries, in an interdisciplinary context.

The contents of this chapter have been published in arXiv:1909.04486v1:

Alarcón-Soto, Y., Espasandín-Domínguez, J., Guler, I., Conde-Amboage, M., Gude-Sampedro, F., Langohr, K., Cadarso-Suárez, C., & Gómez-Melis, G. (2019). Data Science in Biomedicine. *arXiv preprint arXiv:1909.04486v1*.

This chapter is based on the above manuscript. We updated the dates for the search (from 2018 to 2019) and the references, thus it differs slightly from the one used in the original document.

arXiv.org > stat > arXiv:1909.04486v1

Statistics > Other Statistics

Data Science in Biomedicine

[Yovaninna Alarcón-Soto](#), [Jenifer Espasandín-Domínguez](#), [Ipek Guler](#), [Mercedes Conde-Amboage](#), [Francisco Gude-Sampedro](#), [Klaus Langohr](#), [Carmen Cadarso-Suárez](#), [Guadalupe Gómez-Melis](#)

(Submitted on 9 Sep 2019)

We highlight the role of Data Science in Biomedicine. Our manuscript goes from the general to the particular, presenting a global definition of Data Science and showing the trend for this discipline together with the terms of cloud computing and big data. In addition, since Data Science is mostly related to areas like economy or business, we describe its importance in biomedicine. Biomedical Data Science (BDS) presents the challenge of dealing with data coming from a range of biological and medical research, focusing on methodologies to advance the biomedical science discoveries, in an interdisciplinary context.

2.1 Introduction

In the last 10 years, we have observed an important increase in the number of job offers requesting data scientists. Data science was already recognized as a science more than 5 decades ago by John Tukey. In the article *The Future of Data Analysis* he points out that more emphasis should be placed on using data to suggest hypotheses to test and reflects on the existence of an as-yet unrecognised science, whose subject of interest was learning from data (Donoho, 2017) and that lays the foundation of today *data science* area. “Data analysis”, includes

“(...) among other things: procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analysing data.” (Tukey, 1962)

Due to the technological explosion of the last few years, massive amounts of data are generated every day in different areas. This new era requires the development of new techniques to analyse and draw reliable conclusions from these data. In this context, the figure of the data scientist emerges, proclaimed by Davenport & Patil (2012) as “the Sexiest Job of the 21st Century”. But, what exactly is a data scientist?

This question has been already addressed by many other researchers, such as Schutt & O’Neil (2013) or Donoho (2017), and it has been the topic of many columns and discussions in important media such as The Guardian or The New York Times.

To provide a definition of data science in our own terms, we start by referring to the definition of data scientist found in the Oxford Dictionary (Oxford University Press, 2008):

“A person employed to analyse and interpret complex digital data, such as the usage statistics of a website, especially in order to assist a business in its decision-making.”

We will follow the very helpful data science scheme created by Conway (2010) to explore the different attributes a data scientist should convey (Figure 2.1). First, knowledge in Mathematics and Statistics is necessary. Mathematics gives a universal language and is essential for solving real-world problems. From Statistics comes the understanding and experience to work with data, selecting the appropriate techniques to deal with it, to pre-process, summarize, analyse and draw conclusions. Second, computer science knowledge is also fundamental. Not only getting computers to do what you want them to do requires intensive hands-on experience, but also computer scientists must be adept at modelling and analysing problems. They must also be able to design solutions and verify that they are correct. Problem solving requires precision, creativity, and careful reasoning. Computer science has a wide range of sub-areas. These include computer architecture, software systems, graphics, artificial intelligence, computational science, and software engineering. Drawing from a common core of computer science knowledge, each of these areas focuses on particular challenges.

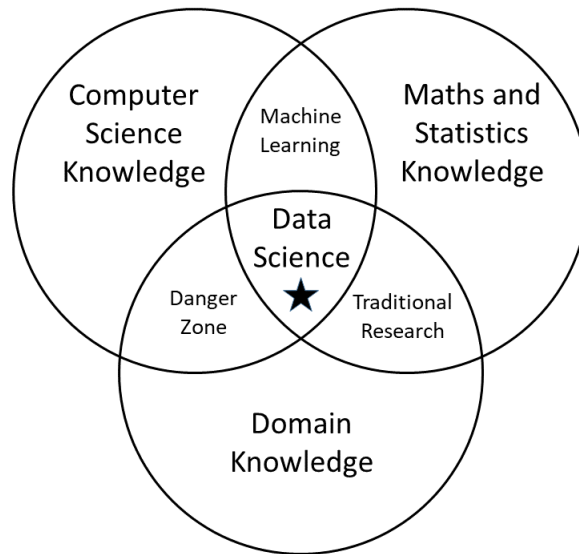


Figure 2.1: Data science scheme based on the Conway’s Venn diagram (Conway, 2010).

The third and not least important characteristic of a data scientist is domain knowledge, a thorough understanding of the field in which the research is being developed is needed to understand the research context and more important to be able to provide realistic and responsible answers to the questions at hand. Examining what these three areas have in common, the intersection between mathematical and statistical knowledge and domain knowledge is the most common, from which traditional research emerges, whereas Machine Learning arises from the intersection of mathematical and statistical knowledge and computer science knowledge. The name Machine Learning, coined by Samuel (1959), is a field of computer science that uses statistical techniques to give computer systems the ability to learn with data. Nevertheless, if there is not enough statistical knowledge to choose the appropriate methods and analyses for the pertinent research objectives, mixing expertise in the field of research with computer science knowledge might lead us to a danger zone.

This overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created (Conway, 2010). As Wilson (1927) stated

“(...) it is largely because of lack of knowledge of what statistics is that the person untrained in it trusts himself with a tool quite as dangerous as any he may pick out from the whole armamentarium of scientific methodology.”

We believe, however, that further soft skills are required by a data scientist. For this reason, we have added a star in the intersection of the three areas, in the core of the Data Science concept. A data scientist needs not only to be an expert in his or her area, but also a good communicator, collaborator, leader, advocate, and scholar. As a communicator, the key competencies are active listening

and asking questions, explaining advantages or shortcomings of statistical and computer methods, and interpreting results in a meaningful way in the context of the application. He or she has to be a fine collaborator, because he or she will have to work in interdisciplinary teams. In addition, being a leader is the key to successfully influence multidisciplinary research, the data scientist will have to advocate to use his or her expertise, and given that science is continuously developing, a data scientist has to be a scholar. In a recent paper by Zapf et al. (2018) these soft skills are already identified for being a successful biostatistician, and they can be generalized to any data scientist.

Therefore a data scientist needs to master a set of skills—mathematical, statistical, computational, communication skills—that are not easy to develop for a single person. Given the scarcity of people with such a complete profile, there is a need to create multidisciplinary working groups formed by different specialists who add their qualities to make room for data science itself.

The chapter is organized as follows: in Section 2.2, we analyze the global impact of Data Science by updating the research of Kane (2014) in which the author analyzes the search-term usage of “Data Science” over time until 2014 adding “Cloud Computing” and “Big Data” to the search, until 2019 and using Google Trends. This section includes an overview of the Data Science journals. Following, in Section 2.3, we describe Data Science in Biomedicine, or Biomedical Data Science (BDS), present a web search restricted to the biomedical area, and include some examples of BDS studies. Finally, the main findings are summarized in Section 6.6.

2.2 Data science: global impact and dissemination

Cleveland (2014) proposes an action plan for statistics, in which he elevates the role of the statistician to the level of a researcher who should not limit him or herself to providing only statistical calculations and p-values, but should, also, be involved in the interpretation of these.

Data science has become very popular in recent years as a tool in many fields such as Economics (business analytics, fraud and risk detection), internet search, digital advertisements, image and speech recognition, delivery logistics, gaming, price comparison websites, airline route planning, robotics, among others. To contextualize the impact of this new discipline all over the world, we have used Google Trends to update the research of Kane (2014). Kane analyses the search-term usage of “Data Science”, “Cloud Computing” and “Big Data” until 2014 (see Figure 2.1). “Cloud Computing” and “Big Data” were added because of their close relation with Data Science, their intrinsic relation with the computational techniques and to frame the evolution of the impact of the Data Science. It must be taken into account that Google Trends is an online search tool that allows the user to see how often specific keywords, subjects, and phrases have been queried over a specific period of time and provides information about Google searches all over the world. Search trends show how the interest for a given term has evolved over time by assigning a score between 0 and 100 to search terms on a year-by-year basis.

To visualize the progress of the terms “Data Science”, “Cloud Computing”, and “Big Data”, we present the results obtained both worldwide and in some countries in Europe, the United States (and some of its states), in Asia, and in Australia over time. The results are summarized in Figures 2.2 - 2.5. All the searches were performed using the R package `gtrendsR` (Massicotte & Eddelbuettel, 2019), which is an interface for retrieving and displaying the information returned online by Google Trends. The R script to perform the analysis can be accessed in <http://doi.org/10.5281/zenodo.3735059>.

An up-tick in Data Science is not produced until approximately the year 2012. It is precisely in this year that the interest for the term “Big Data” starts to grow at high rate. On the other hand, by the end of 2014 and the beginning of 2015, the trend for searches on “Big Data” begins to stagnate, and we can observe an almost exponentially increasing interest for the term “Data Science”. On the other hand, the term “Cloud Computing”, had its main boom around 2011, and since then, its influence has been decreasing.

However, in some countries such as Spain, no real peak for the term “Data Science” is observed until the year 2015. Even though there is also an increase in searches about this concept, the growth is much less pronounced than in other European countries such as Germany, where the interest for “Data Science” is equal to that of “Big Data”, or the United Kingdom, where the trend for “Data Science” begins to unseat that of “Big Data” (Figure 2.3). The trend is even more pronounced in United States, in particular in some of its states such as Massachusetts or California, where the main universities and research centers are. In these US states, the trend for “Big Data” is decreasing sharply coinciding with a growing interest in “Data Science” (see Figure 2.4). In other countries such as China, India, or Japan, the pattern of interest on these terms is similar but with a certain slowness with respect to other countries. It seems that the interest in “Data Science” in these countries as well as in Spain, has not yet reached the same level as in other parts of the world (Figure 2.5).

With this search, we reassert the findings presented in Kane (2014): i) The trend for the term “Data Science” is eclipsing the popularity of the infrastructure on which it is based (cloud computing, big data, computational skills, etc.) specially in the more technological countries; ii) The interest for Data Science is increasing worldwide and it appears that the trend is that this growth will continue in the coming years.

Journals of data science

In this new field, there are only nine scientific journals directly related with the data science (up to March 2020); see Table 2.1. Notice that we do not consider journals that are only related to Big Data Analysis or Machine Learning for the reasons exposed in Section 6.1.

The goal of the *Journal of Data Science* is to enable scientists to do their research on applied science and through the effective use of data. Regarding the *International Journal of Data Science and Analytics*, the main topics addressed are data mining and knowledge discovery, database management, artificial intelligence (including robotics), computational biology/bioinformatics, and business information systems. The related industry sectors are: electronics, telecommunications and IT & Software. The

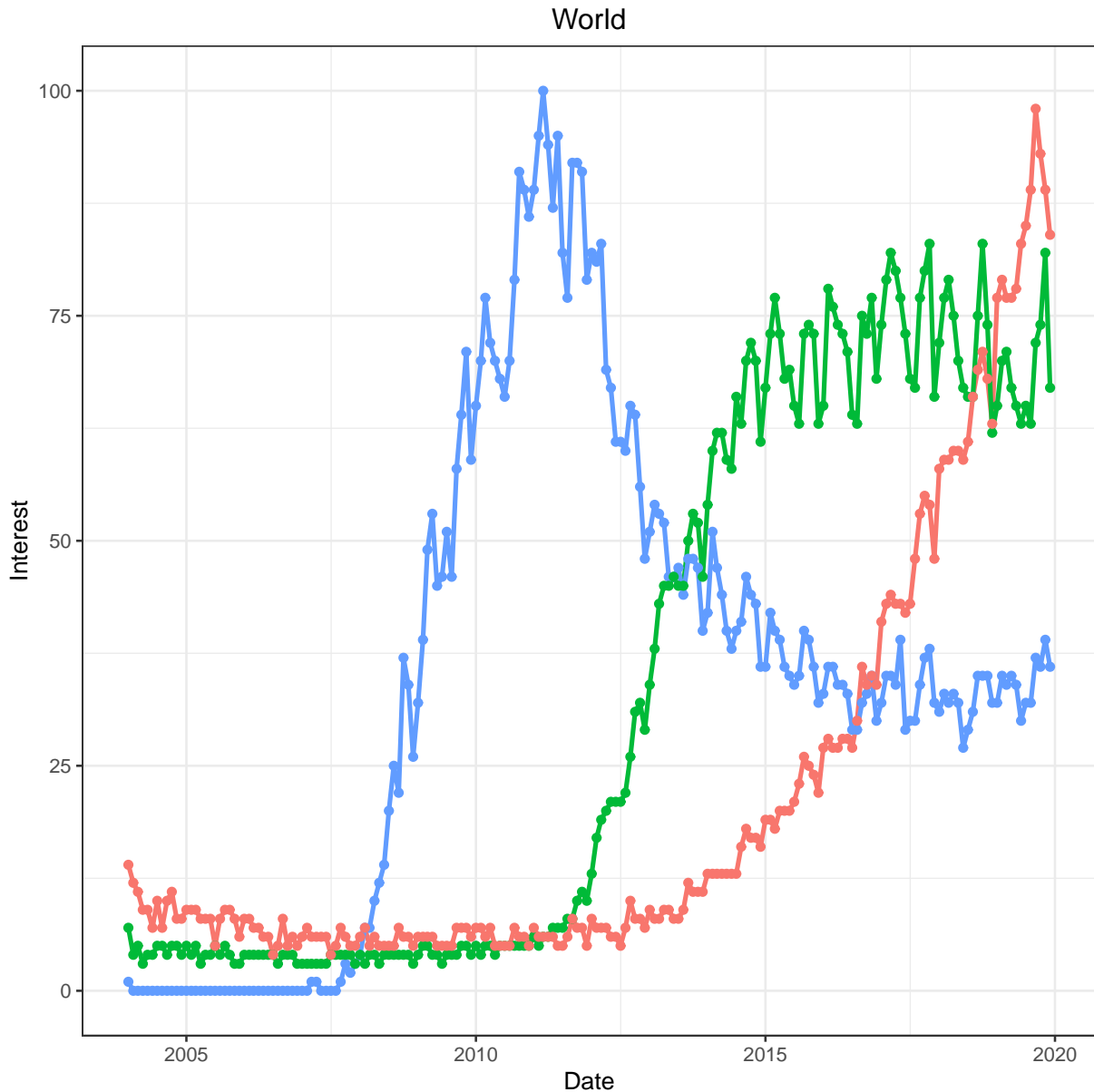


Figure 2.2: Google trends for the terms “Data Science” (red), “Big Data” (green), and “Cloud Computing” (blue) for global queries. The scores assigned by Google Trends on the “interest” ordinate express the popularity of that term over a specified time range, based on the absolute search volume for a term, relative to the number of searches received by Google. The scores have no direct quantitative meaning. For example, two different terms that have been searched 1000 and 20000 times, respectively, could achieve a score of 100. This is because the scores have been scaled between 0 and 100, and a score of 100 always represents the highest relative search volume. Yearly scores are calculated on the basis of the average relative daily search volume within the year.

International Journal of Data Science and Analytics brings together researchers, industry practitioners, and potential users of big data, to promote collaborations, exchange ideas and practices, discuss new

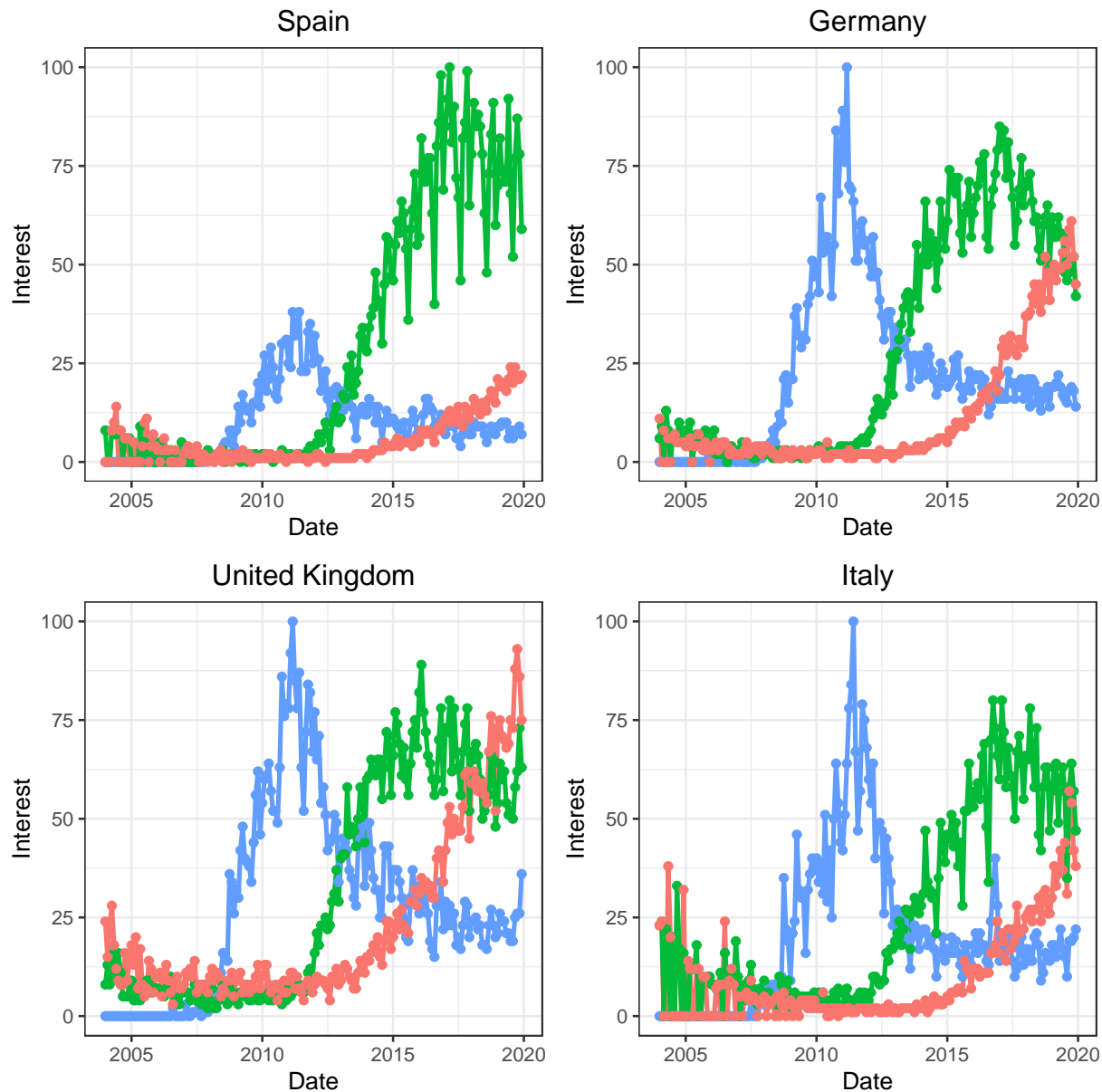


Figure 2.3: Google trends for the terms “Data Science” (red), “Big Data” (green), and “Cloud Computing” (blue) for some countries of Europe.

opportunities, and investigate analytics frameworks. The journal welcomes experimental and theoretical findings on data science and advanced analytics along with their applications to real-life situations. The scope of the *Data Science Journal* includes descriptions of data systems, their publication on the internet, applications and legal issues. All the sciences are covered, including the Physical Sciences, Engineering, the Geosciences, and the Biosciences, along with Agriculture and the Medical Science. The ultimate goal of *Data Science - Methods, Infrastructure and Applications* is to unleash the power of scientific data to deepen our understanding of physical, biological, and digital systems, gain insight

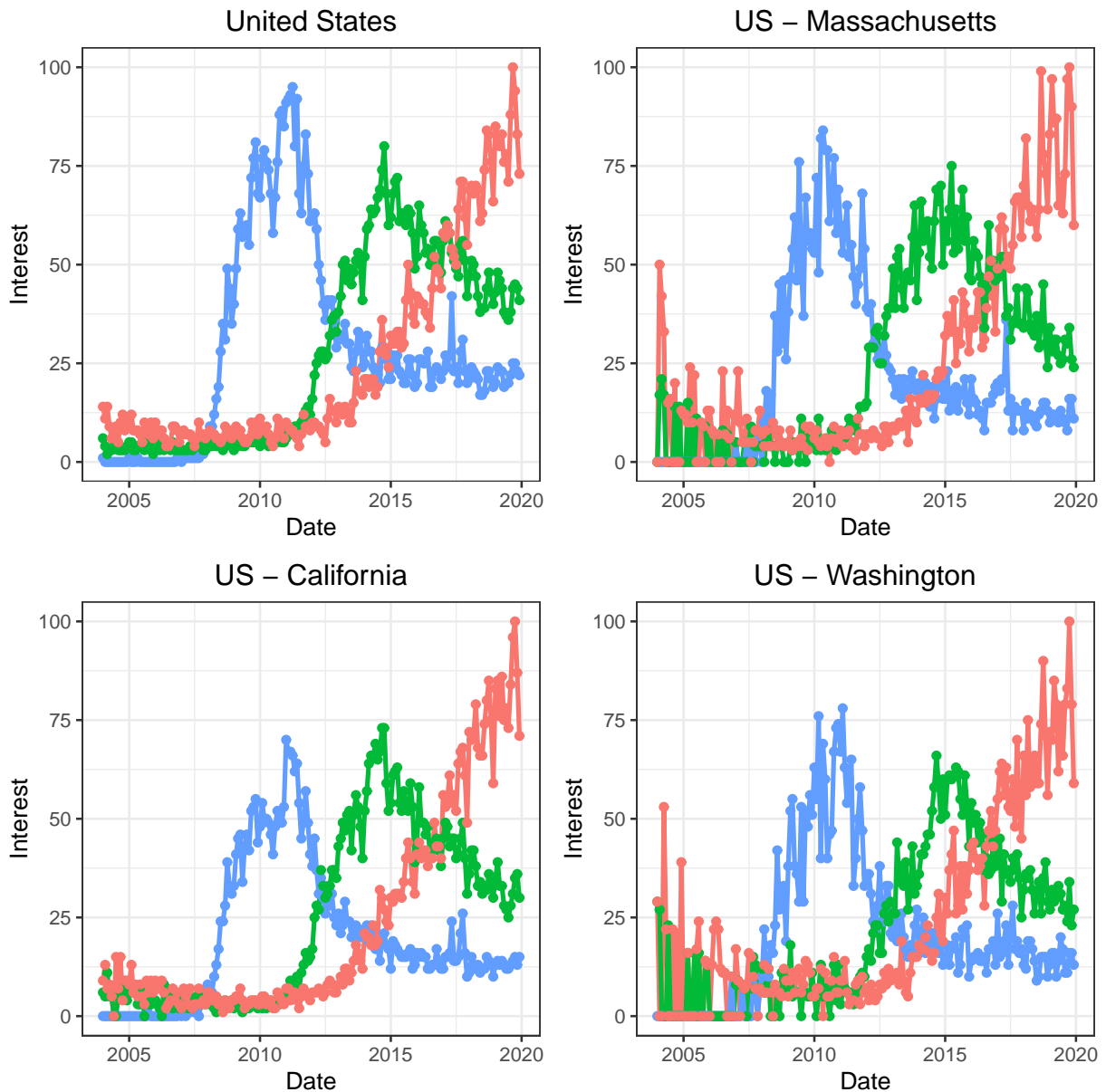


Figure 2.4: Google trends for the terms “Data Science” (red), “Big Data” (green), and “Cloud Computing” (blue) for United States and some of its states.

into human social and economic behaviour, and design new solutions for the future. Additionally, the *EPJ Data Science* covers a broad range of research areas and applications and particularly encourages contributions from techno-socio-economic systems. Topics include, but are not limited to, human behaviour, social interaction (including animal societies), economic and financial systems, management and business networks, socio-technical infrastructure, health and environmental systems, the science of science, as well as general risk and crisis scenario forecasting up to and including policy advice. The *International Journal of Data Science* aims to provide a professional forum for examining the pro-

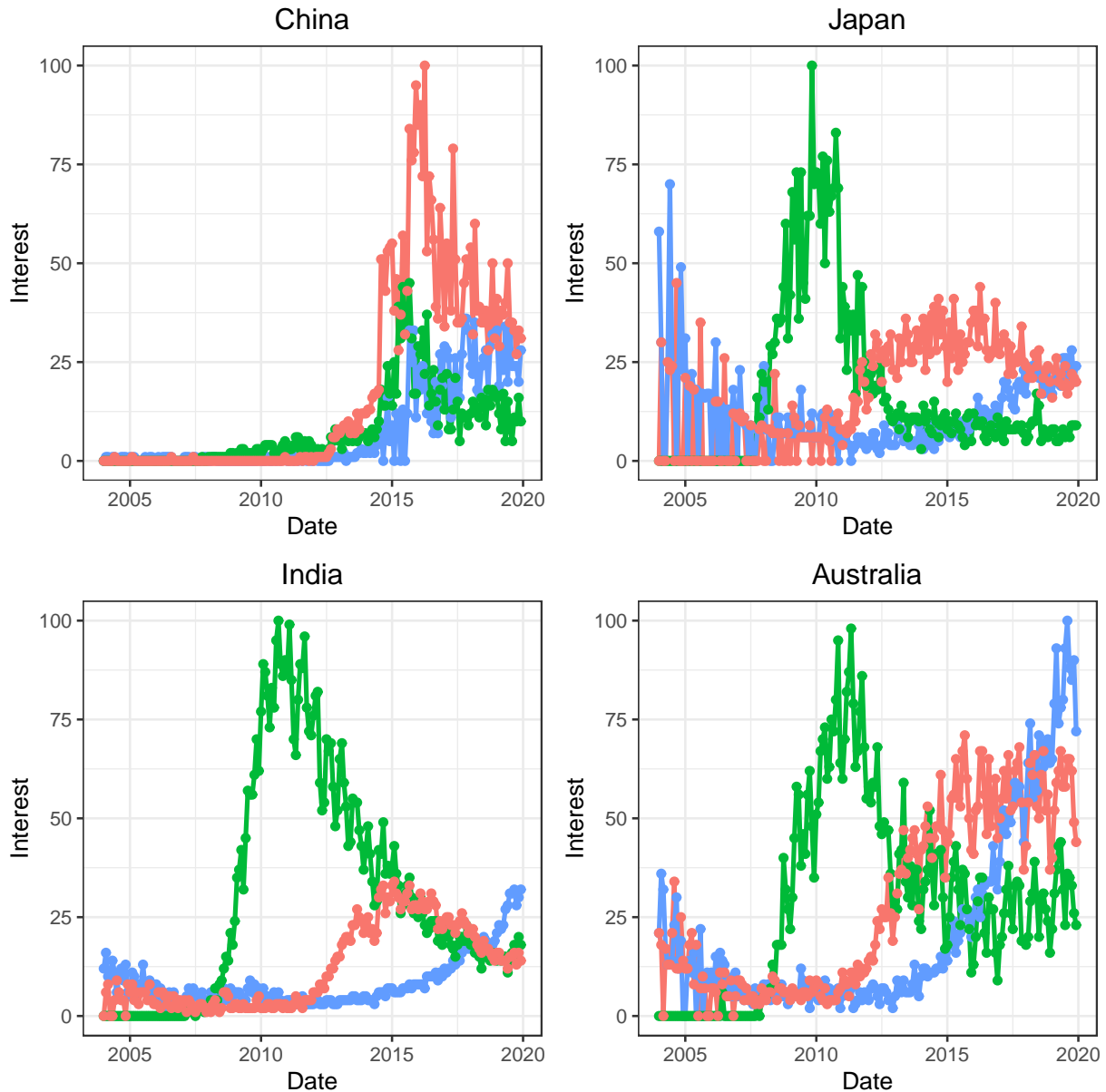


Figure 2.5: Google trends for the terms “Data Science” (red), “Big Data” (green), and “Cloud Computing” (blue) in some countries of Asia and in Australia.

cesses and results associated with obtaining data, as well as munging, scrubbing, exploring, modelling, interpreting, communicating and visualizing data. Data science takes data in cyberspace as a research object. The goal is an integrated and interconnected process designed to form a common ground from which a knowledge-based system can be built, shared, and supported by professionals from different disciplines. The journal *Advances in Data Science and Adaptive Analysis* is an interdisciplinary journal dedicated to report original research results on data analysis methodology developments and their applications, with a special emphasis on the adaptive approaches. The mission of the journal is to elevate

Table 2.1: Current journals in the Data Science field up to March 2020.

Journal and website	Publisher	Scopus	Open access	Bio/health research (explicitly)
Journal of Data Science http://jds-online.com	-	No	No	No
International Journal of Data Science and Analytics https://springer.com/journal/41060	Springer	No	Hybrid	Yes
Data Science Journal https://datascience.codata.org	Uniquity Press	Yes	Yes	Yes
Data Science- Methods, Infrastructure, and Applications https://datasciencehub.net	IOS Press	No	Yes	Yes
EPJ Data Science https://epjdatascience.springeropen.com	Springer Open	Yes	Yes	Yes
International Journal of Data Science http://www.inderscience.com/jhome.php?jcode=ijds	Inderscience	No	Hybrid	No
Advances in Data Science and Adaptive Analysis https://worldscientific.com/worldscinet/adsaa	World Scientific	No	Hybrid	No
Statistical Analysis and Data Mining: The ASA Data Science Journal https://onlinelibrary.wiley.com/journal/19321872	Wiley Online Library	Yes	Hybrid	No
Journal of Data and Information Science https://content.sciendo.com/jdis/	Sciendo	Yes	Yes	No

data analysis from the routine data processing by traditional tools to a new scientific level, which encourages innovative methods development for data science and its scientific research and engineering applications. The journal *Statistical Analysis and Data Mining: The ASA Data Science Journal* addresses the broad area of data analysis, including data mining algorithms, statistical approaches, and practical applications. Finally, the *Journal of Data and Information Science* devotes itself to the study and application of the theories, methods, techniques, services, and infrastructural facilities using big data to support knowledge discovery for decision and policy making.

As we can see in Table 2.1, not all the journals listed above explicitly include health data science and none of them is exclusively dedicated to this area. Following, we provide a proper description of what we consider health or biomedical data science.

2.3 Data science in the biomedical field

A Biomedical Data Scientist should be quantitatively trained including a comprehensive and rigorous proficiency of statistical principles and those computing skills to handle massive and complex data. He/she has to be able to manage and analyse health data to solve emerging problems in public health

and biomedical sciences and to learn how to interpret their findings.

Health data refers to data that come from the biomedical sciences, public health, and any other area related to the “bio” sciences. Examples are data sets from clinical trials, observational studies, genomics and other omics studies, medical records, health care programs, or environmental programs.

Health-related data are also a good example of the legal and ethical concerns that should be taken into consideration regarding sensitive personal data (medical records, genomic profiles, etc.) or digital epidemiology in the context of public health. Thus, ensuring compliance with ethical policies, adequate informed consents, and data use agreements are essential when sharing information and collaboratively using data (Gómez-Mateu et al., 2016).

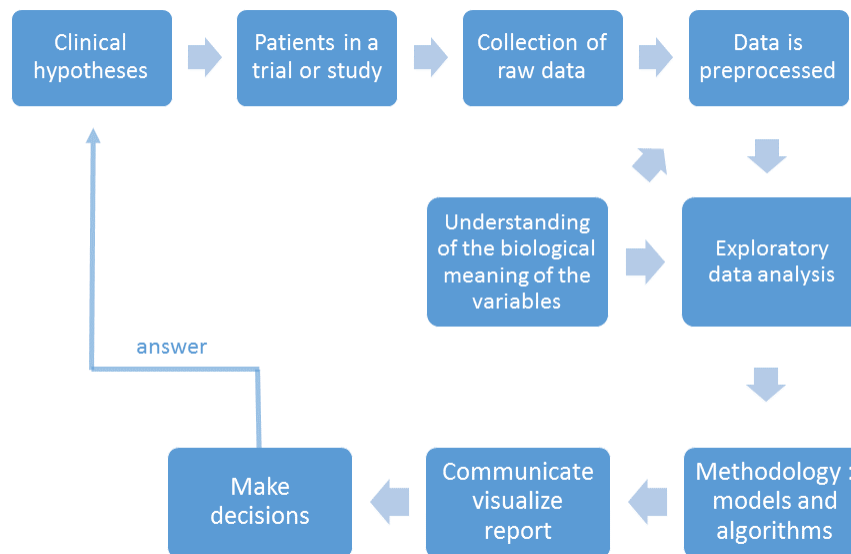


Figure 2.6: Healthcare field process in which a data scientist is involved.

According to the field of study and previous clinical hypotheses, patients who meet the inclusion criteria are recruited into a trial or study, and the raw data such as clinical parameters, demographics, or omics data is collected. Following, preprocessing of the data is done with the objective of cleaning and preparing the data set for exploratory analysis extracting important descriptive statistics. The next step is to find the right methodology to provide an answer to the specific questions for this problem. First, existing models have to be explored for their adequacy to the data and the relevant question. Then, the chosen method has to be implemented and developed. In many occasions new methods have to be developed or old methods have to be adequately adapted. Moreover, the data scientist needs to be able to communicate properly and clearly the results obtained by means of reports and graphical tools. For these reasons, statistics is a fundamental part of the decision making process, it helps draw conclusions and answer the clinical hypotheses. As an added value, the data scientist understands the biological problem, know the biological meaning of the main variables and has to manage a common vocabulary with the rest of the team.

Due to the very likely possibility of potential statistical pitfalls when adapting or developing the chosen methodology, the data scientist should be reliable, coherent and a guide to follow. A small list of these pitfalls are biased samples, overgeneralization, spurious correlation, prediction performance, incorrect analysis choices, and violation of the assumptions for an analysis. A good example of biased samples is cited by Crawford (2013) and shows the data collected in the city of Boston through the StreetBump smartphone app, created with the objective of solving the problem with potholes in this city. This app passively detects bumps by recording the accelerometers of the phone and GPS data while driving, instantly reporting them to the traffic department of the city. Thus, the city could plan their repair and the management of resources in the most efficient possible way. However, one of the problems observed was that some segments of the population, such as people in lower income groups, have a low rate of smartphone use, a rate that is even lower in the older residents, where smartphone penetration is as low as 16%. Therefore, these data provide a big but very biased sample of the population of potholes in the city, with the consequent impact on the underestimation of the number of potholes in certain neighbourhoods and the deficient management of resources. Thus provide a clear instance when having large amounts of data is not synonymous of quality and using the data to solve a problem might result in unfair and not cost effective policies.

Statistical thinking is the central element to avoid the above-mentioned pitfalls. It requires a non-trivial understanding of the real-world problem and the population for whom the research question is relevant. It involves judgements such as those about the relevance and representativeness of the data, about whether the underlying model assumptions are valid for the data at hand, and about causality and the role of confounding variables as possible alternative explanations for observed results. In fact, an essential component of good statistical thinking is the ability to interpret and communicate the results of a statistical analysis so non statisticians can understand the findings (Greenhouse, 2013). In *The Seven Pillars of Statistical Wisdom* Stigler (2016) summarizes Statistical reasoning as an integral part of modern scientific practice and sets forth the foundation of statistics around seven principles. Stigler's second pillar, Information, challenges the importance of "big data" by noting that observations are not all equally important: the amount of information in a data set is often proportional to only the square root of the number of observations, not the absolute number.

2.3.1 Biomedical data science in the Web of Science

Similar to the search presented in Section 2.2, we have analyzed the number of publications associated with "Data Science", "Big Data", and "Cloud Computing" in several countries and along the last fifteen years, using Web of Science (<https://clarivate.com/products/web-of-science/>). The countries considered were Australia, China, Germany, India, Italy, Japan, Spain, the United Kingdom, and the United States. Notice that "publication" refers to articles, reviews, clinical trials, case reports, and books. Moreover, only topics related with the biomedical area, such as Oncology, Respiratory System, or Pediatrics, were considered. The search strategy is available at <http://doi.org/10.5281/zenodo.3735077>, the datasets from the Web of Science can be accessed in <http://doi.org/>

[10.5281/zenodo.3735063](https://doi.org/10.5281/zenodo.3735063) and the R script to perform the analysis can be found at <http://doi.org/10.5281/zenodo.3735059>. The publication counts were obtained at the beginning of 2020 and are presented in Table A.1 (Appendix A).

From the publication counts presented in Table A.1 (see Appendix A), we can conclude that the number of biomedical publications has increased during the last years in the countries considered. Moreover, as might be expected, the number of publications associated with “Data Science” is much larger than the number of publications associated with the topics “Big Data” and “Cloud Computing”. In fact, the publications associated with Data Science represent more than 95% of all the publications analyzed, regardless of the country considered. Most noteworthy is the tremendous increase of record counts in China: 917 publications were registered in 2004, and this number has increased to 12013 in 2017; that is, an increase of more than 10000 publications in only 13 years. Furthermore, the case of Spain is also remarkable because the presence of publications associated with “Data Science” is much lower than in other European countries like Germany, Italy, or the United Kingdom. For instance, in 2017 the number of publications in the United Kingdom and in Germany is approximately three times and twice as high as in Spain, respectively. Although the comparison is not immediate because the population of United Kingdom and Germany is more two times as high as in Spain.

On the other hand, the presence of publications about “Cloud Computing” in the biomedical area is really low: until after 2010, very low number of publications were registered in any of the countries considered. Even in 2017 the number of publications was low compared with the other topics. We can, hence, state that the use of Cloud Computing techniques is not widespread among researchers in the field of Biomedicine. Finally, the explosion of “Big Data” in the last years, seems to have an effect in the Biomedical research because the number of publications in this topic has increased each year in the countries considered. For example, in Australia the number of publications about “Big Data” in 2007 was 17 as compared to 131 publications in 2017, that is, an increase greater than 670%. It is clear that Big Data techniques have been very useful in order to solve biomedical problems.

2.3.2 Multidisciplinary environment for biomedical data science

The confluence of science, technology, and medicine in our dynamic digital era has spawned new data applications to develop prescriptive analytics, to improve healthcare personalization and precision medicine, and to automate the reporting of health data for clinical decisions (Bhavnani et al., 2016). As we mentioned before, several biomedical research institutes are involved in the data science process working on complex data bases in the areas of genomic and proteomic data analysis, infectious and immunological diseases, new therapies in cancer, hormones and cancer, genetics, cellular biology, among others. Most of the research studies need data science techniques to deal with these data sets. Those data science studies that are usually characterized by complex structures or large numbers of variables, require a multidisciplinary environment with biomedical informatics, bioinformatics, biostatisticians, and clinicians. This environment brings together statistics, computer sciences, and computational engineering, and aims to provide a methodologically correct analysis.

Biomedical Data Science can be applied in many different areas such as personalized medicine, genomic research, gene expression analysis, or in cancer drug studies, among others. Following, we present some examples of applications.

Personalized medicine is a medical approach in which patients are stratified in subgroups according to their individual characteristics (genomic alterations, lifestyles, diagnostic markers, clinical profile, response to treatments). With abundant and detailed patient data, medical decisions, such as diagnostic tests or treatments, may be personalized and addressed to these subgroups of patients and not to the whole population. The advantages of personalized medicine are evident: more effective use of therapies and reduction of adverse effects, early disease diagnosis and prevention by using biomarkers, among others. A well-known example is the treatment with trastuzumab (Herceptin, a breast cancer drug) that can only be administered if the HER2/neu receptor is overexpressed in tumor tissue because the drug interferes with this receptor. Another example of those personalized predictions can be the survival probabilities predicted for a future level of a longitudinal biomarker recorded. The joint model approaches to study the association between a longitudinal biomarker and survival data provides dynamic predictions for survival probability coming from the effect of the longitudinal biomarker taken until time t , which can be updated when the patient has new information ([Rizopoulos, 2011, 2012](#)).

Data science helps to examine health disparities because as [Chase & Vega \(2016\)](#) pointed out: “Research examining racial and ethnic disparities in care among older adults is essential for providing better quality care and improving patient outcomes. Yet, in the current climate of limited research funding, data science provides the opportunity for gerontological nurse researchers to address these important health care issues among racially and ethnically diverse groups, groups typically under-represented and difficult to access in research.”

Other example is to use data science for clinical decision making. Clinical laboratories contribute towards the screening, diagnosis and monitoring of many types of health conditions. While it is believed that diagnostic testing may account for just 2% - 4% of all healthcare spending, it may influence 60% - 80% of medical decision-making. The work of [Espasandín-Domínguez et al. \(2018\)](#) is an example of BDS where a very recent extension of the distribution regression model introduced by [N. Klein et al. \(2015\)](#) is applied to a data set of blood potassium concentrations from patients across a Spanish region.

The development of automated workflows that can capture and memorialise extensive experimental protocols, aiding in reproducibility as well as taking data analysis to a new level ([Ludäscher et al., 2006](#)) is a central data science technique. Workflows help support and accelerate scientific discoveries in biomedical research by eliminating the burden of dealing with time-consuming data and software integration. This approach fundamentally frees researchers to concentrate on the scientific questions at hand instead of addressing technical issues involved in setting up, executing, and validating the computational pipeline ([Amaro, 2016](#)).

We can find applications in many other fields. For instance, while studying the consequences of the analytical treatment interruption in HIV-infected patients, [Alarcón-Soto et al. \(2019\)](#) present a method to fit a mixed effects Cox model with interval-censored data to study the viral rebound of HIV. The proposal is based on a multiple imputation approach that uses the truncated Weibull. The authors addressed the fact of having data from eight different studies based on different grounds (see Chapter 6 for further information).

Another application is to quantify spatio-temporal effects to graft failures in organ transplantation. The transplantation of solid organs is one of the most important accomplishments of modern medicine. Yet, organ shortage is a major public health issue. Using data science, the research can investigate early graft failure time. When an organ becomes available from a deceased donor, the allocation policies such as medical urgency, expected benefit and geographical constraints (distance between donor and recipient) are applied to people in the waiting list to select a match. Allocation policies regard the survivability of the organ outside the human body, namely, the cold ischemic time, as an important factor since it is associated with the quality degradation of the organ. Besides, the distance is an important factor on these decisions given that the farther the distance from the donor hospital to the transplant center, the worse might be the quality of the organ ([Pinheiro et al., 2016](#)).

We can even relate data science with mental health. Mental disorders are arguably the greatest “hidden” burden of ill health, with substantial long-term impacts on individuals, carers and society. People with these conditions are often socially excluded and less likely to participate in research studies or remain in follow-up. Complexities around defining diagnoses present particular challenges for mental health research. Richly annotated, longitudinal data sets matched to data science analytics offer an unprecedented opportunity for more robust diagnostics, and also the prediction of outcome, treatment response, and patient preferences to inform interventions ([McIntosh et al., 2016](#)).

Many more examples of BDS are expected to arise in any other field related to health or bio sciences in the near future.

2.3.3 Standardization of information

From the above, we could say that one of the main objectives of Data Science in Biomedicine is to generate valid knowledge through better structuring in the procedures for extracting, analysing and processing data obtained in health and environmental research, supporting the transfer of their results to society. All these disciplines share common goals in terms of improving the quality of life of the people through actions in the promotion of health and in the prevention of disease.

A major challenge that exists in the healthcare domain is the “data privacy gap” between medical researchers and computer scientists. Medical researchers have natural access to healthcare data because their research is paired with a medical practice. Acquiring data is not quite as simple for computer scientists without a proper collaboration with a medical practitioner. There are barriers in the acquisition of data. Many of these challenges can be avoided if accepted protocols, technologies, and safeguards are in place.

On the other hand, people to whom the research efforts are addressed and those responsible for funding agencies need to ensure that research output are used to maximize knowledge and potential benefits. Sharing the data ensures that these are available to the research community, which accelerates the pace of discovery and enhances the efficiency of the research. Believing on these benefits, many initiatives actively encourage investigators to make their data available.

Widely available crowd-sourcing programs such as PatientsLikeMe (www.patientslikeme.com) have amassed participation from more than 400 thousand patients across 2,500 disease conditions who actively share health related data on an open and online platform that tracks and collects important patient-reported outcomes. The United Kingdoms BioBank is a large-scale biomedical data set containing detailed phenotypic, genotypic, and multimodal imaging findings to determine the genetic and nongenetic determinants of health and disease in a contemporary cohort of more than 500,000 participants. Available through open access, research collaborations have advanced our knowledge in the risk prediction of cardiovascular, psychiatric, and cerebrovascular diseases and have identified important anthropometric and genetic traits of metabolic health including diabetes mellitus and obesity.

The objectives for these kind of initiatives are similar to the established data sources such as census and public health data sets, or standardized patient registries such as the National Cardiovascular Data Registry, where data are structured and aggregated. The objective is to monitor population trends, develop guideline-based care, and infer changes to healthcare policy, new citizen science and crowd-sourcing initiatives aim to leverage public and patient participation to collect health data and vital statistics through new massive open, and online data repositories (Bhavnani et al., 2016).

Since 2003, the National Institutes of Health (NIH) has required a data sharing plan for all large funding grants. Similarly, some journals are also requiring the deposit of data and other research documentation associated with published articles (Borgman, 2012; Piwowar et al., 2007).

In May, 2010, the Wellcome Trust and the Hewlett Foundation convened a workshop in Washington, DC, to explore how funders could increase the availability of data generated by their funded research, and to promote the efficient use of those data to accelerate improvements in public health (Walport & Brest, 2011). In this meeting, funders agree to promote greater access to and use of data in ways that are: equitable, ethical and efficient. Equitable refers to recognizing those researchers who generate the data, other analysts reusing these data, meanwhile population and communities expect health benefits arising from research. It should protect the privacy of individuals. Healthcare data is obviously very sensitive because it can reveal compromising information about individuals. Several laws in various countries explicitly forbid the release of medical information about individuals for any purpose, unless safeguards are used to preserve privacy. Finally, it should improve the quality and value of research, and increase its contribution to improving public health.

In June 2018, the NIH releases its first Strategic Plan for Data Science (<https://www.nih.gov/news-events/news-releases/nih-releases-strategic-plan-data-science>). In this plan, “NIH addresses storing data efficiently and securely; making data usable to as many people as possible; developing a research workforce poised to capitalize on advances in data science and information tech-

nology; and setting policies for productive, efficient, secure, and ethical data use. This plan commits to ensuring that all data-science activities and products supported by the agency adhere to the FAIR principles, meaning that data be Findable, Accessible, Interoperable, and Reusable" (Wilkinson et al., 2016).

2.4 Conclusions

Motivated by the remarkable increase of the number of publications on Data Science in the past few years, the purpose of this chapter has been to study the impact of Data Science in the area of biomedicine.

With this objective in mind, we have carried out a search of the terms "Data Science" along with "Big Data" and "Cloud Computing" using Google Trends until December 2019. While Big Data represents the information assets characterized by a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value (De Mauro et al., 2015), Cloud Computing enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell & Grance, 2009). Big Data and Cloud Computing were chosen since they are somewhat related to computing movements and they help to put the "Data Science" search-traffic into perspective (Kane, 2014). According to our search results, in the last years more and more publications in the area of biomedicine make use of the term "Data Science", however, there are large differences among the countries considered.

We have also listed the main journals only related to Data Science to point out the increasing importance of Data Science. However, not all of the journals presented explicitly include Biomedical Data Science (BDS) as their main areas of research. In addition, we have stepped ahead of the contemporary definition of Data Science, directly related to the economics or business world, describing the Data Science in the Biomedical field. We understand BDS as the interdisciplinary field that encompasses the study and pursuit of the effective use of biomedical data, information, and knowledge for scientific inquiry, problem-solving, and decision-making, driven by efforts to improve human health. It investigates and supports reasoning, modelling, simulation, experimentation, and translation across the spectrum, from molecules to individuals to populations.

We strongly believe that the importance of Biomedical Data Science will continue increasing in the near future due to nowadays' possibilities to record enormous quantities of data and the technical facilities to process them. Statistical thinking and knowledge will play a key role in the correct analysis of such data.

OVERVIEW OF HIV

Human immunodeficiency virus (HIV) is the virus that causes HIV infection. HIV attacks and destroys the infection-fighting CD4 cells of our natural defense against pathogens, infections and illnesses ([NAM Publications, 2017](#)), the immune system. The loss of CD4 cells makes it difficult for the body to fight infections and certain cancers. Without treatment, HIV can gradually destroy the immune system and advance to Acquired Immunodeficiency Syndrome (AIDS).

There are two variations or “serotypes” of HIV: HIV-1 and HIV-2, which correspond to two genetic differentiations of the HIV. However, their genomes have only 45% of similarity. It is thought that HIV-2 “jumped” in Africa from simians to men. Today, HIV-2 is present only in countries like Senegal, Gambia, Liberia, Ghana or Nigeria. In this work, from now on, the term HIV refers to HIV-1. For more information about biological concepts presented in this thesis, please see the [Appendix F](#).

3.1 History of HIV

The earliest known case of infection with HIV-1 in a human was detected in a blood sample collected in 1959 from a man in Kinshasa, Democratic Republic of the Congo ([Faria et al., 2014](#)). Genetic analysis of this blood sample suggested that HIV-1 may have stemmed from a single virus in the late 1940s or early 1950s.

In 1981, the United States Centers for Disease Control and Prevention (CDC) reported five cases of Pneumocystis pneumonia in homosexual men living in Los Angeles ([Gottlieb et al., 1981](#)). Although the CDC first believed that the new disease was confined to homosexual men ([Altman, 1981](#)), by the end of the year, several cases had been reported in non-homosexual injecting drug users and outside the United States, such as Haiti and some African countries ([Pitchenik et al., 1983](#); [Clumeck et al., 1983](#)).

In 1983, a retrovirus (which was later termed HIV) was identified from a patient with AIDS in France ([Barré-Sinoussi et al., 1983](#)).

For many years, scientists theorized as to the origins of HIV and how it appeared in the human population, most believing that HIV originated in other primates. In 1999, an international team of researchers reported that they had discovered the origins of HIV. These researchers identified a type of chimpanzee in Central Africa as the source of HIV infection in humans. They believed that the Simian Immunodeficiency Virus (SIV) most likely was transmitted to humans and mutated into HIV when humans hunted these chimpanzees for meat and came into contact with their infected blood. More recent studies show that HIV may have jumped from apes to humans as far back as in the late 1800s ([Sharp & Hahn, 2011](#)).

3.2 HIV transmission

HIV may be transmitted through certain body fluids that may contain high concentrations of HIV. These body fluids include blood, semen and pre-seminal fluid, vaginal and rectal fluids, and breast milk.

There are four main routes of HIV transmission: 1) unprotected vaginal, oral or anal sex (being oral sex the one with small risk), 2) sharing unsterilized injecting drug equipment, 3) from mother-to-baby in pregnancy, childbirth or breastfeeding, and 4) infected blood transfusions, transplants or medical procedures.

Since HIV infection often presents no physical symptoms, the only way to know if a person has HIV is through an HIV test. We highlight three types of tests that check the blood or body fluids to confirm the presence of the virus in the body: antibody screening, antibody/antigen combination, and RNA test. Antibody screening tests, also called immunoassay or ELISA tests, check for a protein that the body produces in response to the HIV from 2 to 8 weeks after the infection happens. These tests are considered very accurate except in the case of early infections, which can be detected with antibody/antigen tests. Antibody/antigen tests check for HIV antigen, a protein called p24 that is part of the virus and shows up 2-4 weeks after infection. In addition, antibody/antigen tests also check for HIV antibodies. Finally, RNA tests look for the virus itself and can diagnose HIV about 10 days after having been exposed to contagion.

3.3 HIV RNA viral load

Viral load is the term used to describe the amount of HIV in the blood. It is essential to measure the viral load, as well as the CD4 count, as these quantities are prognostic indicators of the evolution of patients treated with antiretroviral drugs. These indicators allow to precise specific treatment to follow, when it has to be started, or if any change in medication is needed.

Viral load tests measure the amount of HIV's genetic material in a blood sample. The results of these tests are described as the number of copies of HIV RNA in a millilitre of blood. Viral load tests have a cut-off point below which they cannot reliably detect HIV, called "limit of detection". Most commonly the lower limit of detection is around 40 or 50 copies/ml, but there are some very sensitive tests that can measure below 20 copies/ml. If the viral load is below the established threshold, it is usually said to be undetectable. However, the fact that the level of HIV is too low to be detected does not mean that HIV has disappeared completely from the body (Vanable et al., 2000).

In the first few weeks after contracting HIV, the amount of viral load in the HIV-infected person is very high. A person with high viral load in the blood are likely to also have a high viral load in other body fluids, such as semen or vaginal fluids (Van Dyk, 2010). This is the reason why people with high viral load are potentially more infectious and can pass on HIV more easily. On the other hand, if HIV in the blood is undetectable, it is likely to also be undetectable in semen, vagina fluid or rectum as well. Having an undetectable viral load implies that the risk of HIV being passed on during sex is extremely low. In 2011, a large scientific trial found that HIV treatment reduces the risk of passing on HIV to a regular heterosexual partner by 96% (1 from 28 infections) during sex (M. Cohen et al., 2011). In this trial, the only person that acquired HIV did it only a few days before or after their partner started treatment. In 2014, a study found that no HIV transmission took place in 16,400 and 28,000 sex encounters between gay and heterosexual men, respectively, where the HIV-positive partner had a viral load below 200 copies/ml (M. S. Cohen et al., 2011). Recently, the PARTNER 2 study confirmed that people with HIV on effective antiretroviral treatment cannot pass on the virus (Rodger et al., 2019). The results of this study provided a similar level of evidence on viral suppression and HIV transmission risk for gay men to the level previously observed for heterosexual couples. Additionally, the study suggested that there is no risk of HIV transmission in gay couples through condomless sex when HIV viral load is suppressed. These recent findings support the message of the U=U (Undetectable equals Untransmittable) global campaign, and the benefits of early testing and treatment for HIV.

3.4 CD4 T cells and their role

CD4 cells (also known as CD4+ T cells, T-lymphocytes or helper cells) are white blood cells that play a major role in the immune system.

The CD4 cell count is the number of blood cells in a cubic millimetre of blood. High CD4 cell counts (500 - 1,500) indicate a strong immune system. This number declines when a person is HIV infected. The lower the CD4 cell count, the greater the damage to the immune system and the greater the risk of illness. In addition to HIV, infected people who have a CD4 cell count below 200 are at high risk of developing several illnesses. Current antiretroviral therapy helps HIV-infected people gradually increase the CD4 cells as well as decrease the viral load.

Regarding the relationship between CD4 cells counts and viral load, the study by [Opportunistic Infections Project Team of the Collaboration of Observational HIV Epidemiological Research in Europe](#)

(COHERE) in EuroCoord (2012) have shown that among people with the same CD4 cell counts, those with higher viral load tend to develop HIV symptoms faster than those with lower viral load. In addition, among people with the same viral load, those with lower CD4 cell counts tend to become ill faster.

3.5 Latent HIV reservoir

Latent HIV reservoir is the term used for the group of immune cells in the body that are HIV-infected but are not actively producing new HIV material. Latent HIV reservoirs can be found throughout the body, including the brain, lymph nodes, blood, and the digestive tract.

As mentioned above, HIV attacks immune system cells in the body and uses the cell's machinery to make copies of itself. This process starts when HIV inserts its generic blueprint into the DNA of an immune system cell, such as a CD4 cell. The infected cells start producing HIV proteins, which act as the building blocks for new HIV genetic material. However, some HIV-infected cells shut down and go into a latent (or resting) state. How HIV-1 establishes latent infection in CD4 cells has been unclear (Sengupta & Siliciano, 2018). While in this latent state, the infected cells do not produce new HIV. HIV can “hide” inside these cells for years, forming a latent HIV reservoir. At any time, cells in the latent reservoir can become active again and start producing more HIV.

Finding ways to target and destroy latent reservoirs is one of the major challenges that HIV researchers face. New studies are exploring different strategies for clearing out reservoirs, including the use of gene manipulation to cut out or inactivate the virus in HIV-infected immune cells, the development of drugs or other methods that reactivate latent HIV reservoirs so that the immune system itself or new therapies can effectively eliminate them, and the elaboration of approaches that enhance the immune system's ability to recognize and clear reactivated latent HIV reservoirs, among others (Deeks et al., 2012).

3.6 The HIV life cycle

The HIV life cycle is the process of how the virus attaches itself to a CD4 cells and use them to takes control of the cell's genetic material, replicates itself inside the cell and finally releases more HIV into the blood to continue the multiplication process. It is usually divided into seven stages (as shown in Figure 3.1):

1. **Binding (also called Attachment):** HIV attaches itself to the surface of a CD4 cell.
2. **Fusion:** The HIV envelope and the CD4 cell membrane fuse (join together), which allows HIV to enter the CD4 cell.

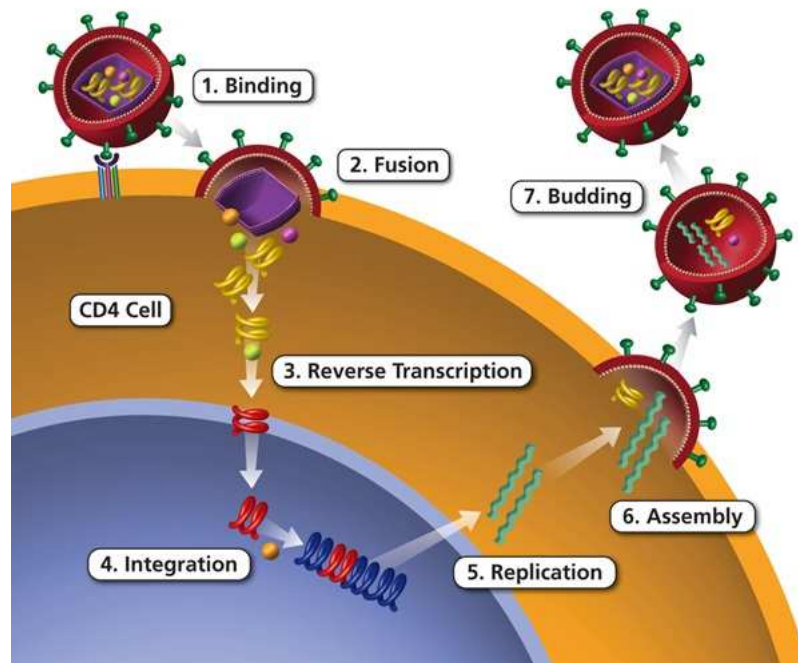


Figure 3.1: HIV life cycle. Source: <https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle>.

3. **Reverse Transcription:** Once inside the CD4 cell, HIV releases and uses an HIV enzyme called reverse transcriptase to convert its genetic material known as HIV RNA into HIV DNA. This conversion allows HIV to enter the CD4 cell nucleus and to combine with the cell's genetic material.
4. **Integration:** Inside the CD4 cell nucleus, HIV releases an enzyme called integrase. HIV uses integrase to insert its viral DNA into the DNA of the CD4 cell.
5. **Replication:** The infected cell produces more HIV proteins that are used to produce more HIV particles inside the cell.
6. **Assembly:** New HIV proteins and HIV RNA move to the surface of the cell and assemble into immature or noninfectious HIV.
7. **Budding:** Newly formed immature HIV pushes itself out of the host CD4 cell. The new HIV releases an enzyme called protease which acts to break up the long protein chains that form the immature virus. The smaller HIV proteins combine to form mature or infectious HIV.

3.7 Stages of HIV infection

There are three stages of HIV infection: acute, chronic, and AIDS (*The stages of HIV infection, n.d.*).

Acute HIV infection is the earliest stage of infection, and it generally develops within 2 to 4 weeks after a person is infected with HIV. During this time, some people have flu-like symptoms, such as fever, headache, and rash. In the acute stage of infection, HIV multiplies rapidly and spreads throughout the body. HIV can be transmitted during any stage of infection, but the risk is greatest during this stage.

Chronic HIV infection, also called asymptomatic HIV infection or clinical latency, is the second stage of infection. During this stage HIV continues to multiply in the body but at very low speed. People with chronic HIV infection may not have any HIV-related symptoms, but they can still spread HIV to others. If not treated, chronic HIV infection usually advances to AIDS in 10 years or longer, though it may take less time in some cases.

AIDS is the final stage of HIV infection. The immune system has been severely damaged, the body cannot fight off opportunistic infections. According to the Centers for Disease Control and Prevention (CDC) definition, a patient has AIDS if he or she are infected with HIV and have either a CD4 cell count below 200 cells/mm³, a CD4 cell percentage of total lymphocytes of less than 14%, or one opportunistic infection. Without treatment, people with AIDS typically survive about 3 years ([U.S. Department of Health and Human Services, 2019](#)).

3.8 Combination antiretroviral treatment

The use of drugs to treat HIV infection is known as combination antiretroviral treatment (cART). cART prevents HIV from multiplying and reduces the amount of HIV viral load in the body.

There are seven main types (or classes) of drugs that work against different parts of the HIV life cycle, as presented in Table 3.1. There are more than 30 HIV drugs and formulations, cART usually combines HIV drugs from at least two different classes, making it very effective at preventing HIV from multiplying. Only a few combinations are now commonly used ([HIV i-Base, 2017](#)).

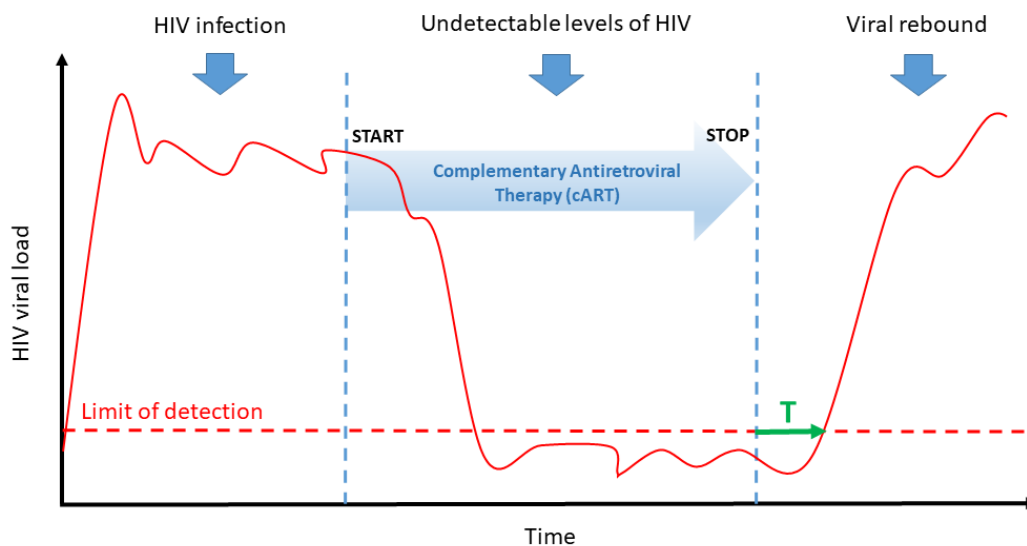
cART helps HIV-infected people live longer, healthier lives. HIV drugs also reduce the risk of HIV transmission. In fact, it has been shown that people who take cART daily as prescribed, and therefore maintain an undetectable viral load, cannot sexually transmit the virus to others ([Eisinger et al., 2019](#); [M. S. Cohen et al., 2011](#)).

3.9 HIV RNA viral rebound

The HIV RNA viral rebound occurs if a person following cART has persistent and detectable levels of HIV in the blood after a period of undetectable levels.

Table 3.1: Main types of drugs for HIV infection.

Abbreviation	Full names
NRTIs/NtRTIs (“nukes”)	Nucleoside/tide reverse transcriptase inhibitors or nucleoside/tide analogues
NNRTIs (“non-nukes”)	Non-nucleoside reverse transcriptase inhibitors
PIs	Protease inhibitors
INIs (or INSTIs)	Integrase (strand transfer) inhibitors
CCR5 inhibitors	CCR5 inhibitors are a type of entry inhibitor
Fusion inhibitors	Fusion inhibitors are a type of entry inhibitor
mAbs	Monoclonal antibodies block HIV entering the T-cell

**Figure 3.2:** HIV RNA viral rebound. Limit of detection is shown in red-dashed line.

The dynamic of HIV RNA viral rebound can be seen in Figure 3.2. As we explained before, viral load tests have a cut-off point below which they cannot reliably detect HIV. We can see in Figure 3.2 that after stopping the HIV treatment there is an increase of circulating virus. When viral load becomes detectable, it can be measured and probably continues increasing to a certain setpoint, unless the patient resumes cART. This is called viral rebound. The time to viral rebound depends on the characteristics of each patient. cART aims the viral load to fall again and become undetectable.

To observe if a specific therapeutic vaccine (see section 3.10) and/or the combination of it with other therapies can influence viral rebound dynamics, through preservation and enhancement of immune response, the cART should be stopped. The main idea is to observe the time to viral rebound as an indicator of the efficiency of the treatment. Once the viral rebound occurs, the patient must resume the treatment.

3.10 Therapeutic HIV vaccine

A therapeutic HIV vaccine is designed to improve the body's immune response to HIV in a person who is already infected with HIV. Researchers are developing and testing these vaccines to slow down the progression of HIV infection (including the progression to AIDS). The ultimate goal of therapeutic vaccines are that they can achieve undetectable levels of HIV without the need for regular cART.

Researchers are also evaluating therapeutic HIV vaccines as part of a larger strategy to eliminate all HIV from the body and fully. This strategy may involve using other drugs and therapies in addition to the therapeutic HIV vaccine.

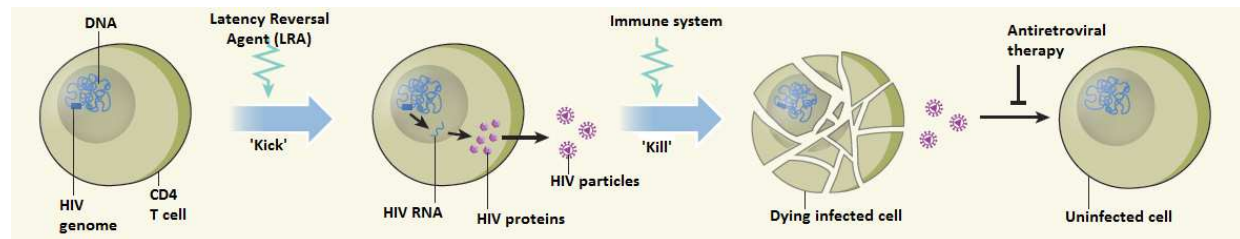


Figure 3.3: “Kick and kill” therapeutic approach. Modified from [Deeks \(2012\)](#).

In this thesis we present some clinical trials based on “kick and kill” strategies ([Ruiz-Riol & Brander, 2019](#); [Lewin & Rasmussen, 2020](#)). “Kick and kill” strategy is based on the fact that an HIV cure might be possible if all infected cells of the latent virus reservoir are forced out of their hidden place (“kick”), leading ultimately to the death of these cells (“kill”). As can be seen in [Figure 3.3](#), current treatments for HIV infection do not eradicate virus because HIV genome remains integrated into the DNA of the CD4 cells. For example, using a Latency Reversal Agent, as Romidepsin, leads to activation of HIV genes (“kick”). This causes the infected cells to be killed by the virus itself or by the patient’s immune system.

3.11 HIV cure and viral eradication

Over the last 15 years, there have been different initiatives aiming at achieving the so-called “HIV cure and viral eradication”. Although these terms are often used freely and interchangeably, it is generally accepted that “cure”, or “functional cure” refers to a state of persistent viral suppression without the need to take antiretroviral drugs (cART), while eradication refers to the complete elimination of any HIV from the body of an infected individual. The discrimination is certainly more than semantics, given that HIV establishes a life-long, largely immunologically silent, latent reservoir shortly after acute infection and that, to date, no effective strategies exist to reactivate and eliminate parts or the entirety of this latent reservoir ([Ruiz-Riol & Brander, 2019](#)).

More than a decade ago, Timothy Ray Brown (the “Berlin patient”) made history as the first person to be “cured” of HIV after a bone marrow stem cell transplant to treat cancer. In 2019, Adam Castillejo, known as the “London patient”, has become the second person to be cured after the same transplant.

In particular, Brown and Castillejo received stem cells from a donor with a CCR5 gene mutation, making them HIV-resistant. Brown came off the cART and had been HIV-free the rest of his life (he died of leukemia on September 2020). Scientists considered him cured ([Hütter et al., 2009](#); [Allers et al., 2011](#)). The London patient stopped cART at the end of 2017 and now he is considered cured ([Gupta et al., 2020](#)). Unfortunately, the procedure can almost never be offered as a cure for HIV infection because stem cell transplants carry several risks ([Gupta et al., 2020](#)).

CONCEPTS OF SURVIVAL ANALYSIS AND OMICS DATA ANALYSIS

Survival and omics data analysis are key components in this thesis. This Chapter aims to explain the main concepts of these specific areas, as they served as basis for different approaches developed during the next chapters of this thesis. Concerning the survival analysis, the definitions of the survival function, distribution function, hazard function, and cumulative hazard function are presented. Moreover, we emphasize the definition of interval-censored variables. Proportional hazards and accelerated failure time model are also introduced. Regarding the omics data analysis, we present the main concepts related to transcriptome. We also address the classical pipeline for analyzing microarray data, in this specific case mRNAs, using R and Bioconductor (<https://www.bioconductor.org/>).

4.1 Survival analysis

4.1.1 Basic concepts

Let T be the time until the event of interest, \mathcal{E} , which, in the present work, corresponds to the time to viral rebound. Formally, T is a non-negative random variable, whose distribution can be characterized by the survival function, $S(t)$, the cumulative distribution function, $F(t)$, the hazard function, $\lambda(t)$, or the cumulative hazard function, $\Lambda(t)$. Each of them serves to illustrate different aspects of the distribution of T . All the concepts described below can be found in [Gómez et al. \(2015\)](#).

The survival function is denoted by $S(\cdot)$. It corresponds to the probability of an individual surviving beyond time t (experiencing the event after time t). It is defined as

$$S(t) = P(T > t) \text{ for } t \geq 0.$$

The survival function can take different forms, but all start from $S(t = 0) = 1$, decrease monotonically and converge to zero when t tends to infinity.

Similarly, the nondecreasing right-continuous distribution function $F(t)$ is defined as the cumulative probability of reaching viral rebound before time t , i.e.,

$$F(T) = P(T \leq t) = 1 - S(t), \quad t \geq 0.$$

Another possibility is to specify the hazard function $\lambda(t)$ which represents the instantaneous risk of viral rebounding

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t), \quad t \geq 0.$$

Intuitively, $\lambda(t)\Delta t$ can be interpreted as the probability that \mathcal{E} occurs in $(t; t + \Delta t]$ given that the event has not occurred before.

The hazard function is estimated by the proportion of people who rebound at time t among those who had not previously rebounded. The hazard function expresses how the instantaneous risk changes over time containing the same information as the survival but in terms of its speed (or rate) of change. When the risk is high, survival declines quickly, whereas if the risk is zero the survival curve is flat.

Finally, the cumulative hazard function, $\Lambda(t)$, when T is absolutely continuous, is defined by

$$\Lambda(t) = \int_0^t \lambda(u) du, \quad t \geq 0$$

and it is very useful graphically, also technically, for example to check a model's goodness-of-fit, but does not have an intuitive interpretation.

4.1.2 Interval-censored data

One difficulty of survival analysis is the incomplete information on the survival of some individuals. For example, when the exact time until the viral rebound is not observed, either because the event of interest \mathcal{E} occurs before the person enters the study, or because when the study ends \mathcal{E} has not happened yet, and in general because all the knowledge is that \mathcal{E} has occurred within a certain time interval. These peculiar characteristics of survival studies are known under the name of censoring. In this thesis we address different clinical trials that consider interval-censored times to viral rebound.

Different censoring mechanisms give rise to interval-censored data of varying nature and the methods and the theoretical developments behind these are also different and not necessarily interchangeable. Basically, we refer to interval-censored data to those situations where instead of observing the

actual value of a random variable T , we only observe a window $(L, R]$ where T has occurred (Gómez et al., 2009). Interval-censored data is frequent in longitudinal studies in many areas of medical research, where the occurrence of the event can often be recorded only at periodic follow-ups. Interval censoring may also arise when an individual misses one or more scheduled visits, and when the disease status has changed when s/he returns. Examples of interval-censored data are found, in particular, in HIV-1 studies when analyzing time to viral rebound, where the event occurs between consecutive visits of a patient. Interval-censored data include right-censored times as a particular case with $R = \infty$, which in this work corresponds to patients whose viral load has not rebounded by the end of the study. More information on this type of censoring can be found in Gómez et al. (2015).

Noninformativity conditions

Most of the methods to work with interval-censored data assume noninformativeness of the censoring mechanism. The non-informative condition establishes that the mechanism that generates the censoring is noninformative for T , the variable of interest. This means the observed interval $(L, R]$ carries no further information on the survival time T other than the fact that T belongs to the interval $(L, R]$. As a consequence of this assumption, censored data can be evaluated without modelling the censoring process.

In this work, we adopt the noninformativity conditions in Oller et al. (2004) where three equivalent characterizations of noninformativeness are given, conditions ensuring that the censoring mechanism cannot affect the distribution of T . These properties describing non-informativeness guarantee that the contribution to the likelihood function of an individual with observed interval (ℓ, r) ,

$$\int_L^R F_{T,L,R}(t, \ell, r) dt = P(T \in (L, R], L \in d\ell, R \in dr)$$

is proportional to $P(T \in (\ell, r])$, that is, the probability that T belongs to $(\ell, r]$ ignoring the censoring mechanism. This probability is denoted as simplified likelihood.

4.1.3 Nonparametric estimation of the survival function

To obtain a non-parametric estimation of the survival function, $S(t) = 1 - F(t)$, under interval censoring, one of the most popular methods is Turnbull's estimator (Turnbull, 1976). For this purpose, we define the so-called Turnbull intervals, denoted by $\mathcal{I} = \{(q_1, p_1], (q_2, p_2], \dots, (q_m, p_m]\}$, where \mathcal{I} are those intervals where all the mass of any non-parametric maximum likelihood estimator (NPMLE) will be concentrated. To obtain these intervals, let $\mathcal{L} = \{L_i, 1 \leq i \leq n\}$ and $\mathcal{R} = \{R_i, 1 \leq i \leq n\}$ be the set of left and right endpoints respectively. We need to derive all the distinct intervals $(q_j, p_j]$ such that $q_j \in \mathcal{L}$, $p_j \in \mathcal{R}$, and that there is no other left or right endpoint between q_j and p_j . The NPMLE for the survival function decreases inside the set \mathcal{I} and is constant outside of them. Specifically, denoting by $w_j = P(q_j < T \leq p_j) = S(q_j) - S(p_j)$ the weight of the j th Turnbull's interval, the NPMLE for $S(t)$ is given by

$$\hat{S}_n(t) = \begin{cases} 1, & \text{if } t \leq q_1 \\ 1 - (\hat{w}_1 + \dots + \hat{w}_j), & \text{if } p_j \leq t \leq q_{j+1}, \quad 1 \leq j \leq m-1 \\ 0, & \text{if } t \geq p_m \end{cases} \quad (4.1)$$

and is not specified within $(q_j, p_j]$, for $1 \leq j \leq m$. Observe that Turnbull's estimate of the survival function has a special shape with horizontal stretches in non-Turnbull's intervals and rectangular boxes indicating areas of equal likelihood in Turnbull's intervals.

4.1.4 Proportional hazards model

Let \mathbf{x} be a vector of covariates including, for example, some clinical covariates or a specific type of omics data. In order to specify how the covariates may affect the time to viral rebound, a regression model is needed. Let $\lambda(t|\mathbf{x})$ be the hazard function at time t for an individual with covariate vector \mathbf{x} . The basic so-called Cox proportional hazards model (Cox, 1972) is given by

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}) = \lambda_0(t) \exp\left(\sum_{k=1}^p \beta_k x_k\right), \quad (4.2)$$

where $\lambda_0(t)$ is an arbitrary unspecified baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the vector of regression parameters.

Under the proportional hazards model, the conditional density and survival functions of T given \mathbf{x} have the forms:

$$f(t; \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}) \exp[\Lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta})]$$

and

$$S(t; \mathbf{x}) = \exp[-\Lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta})] = [S_0(t)]^{\exp(\mathbf{x}'\boldsymbol{\beta})},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ and $S_0(t) = \exp[-\int_0^t \lambda_0(s) ds]$ are the baseline cumulative hazard function and the baseline survival function. The conditional cumulative hazard function of T given \mathbf{x} has the form $\Lambda(t; \mathbf{x}) = \Lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta})$.

Once the model (4.2) has been established and assuming that the primary interest lies in the role played by the fixed effects, the estimation of $\boldsymbol{\beta}$ must be addressed allowing that $\lambda_0(t)$ is arbitrary. Given that the baseline function $\lambda_0(t)$ is not specified, to study the influence of the covariates in the survival times, a modification of the classical theory of maximum likelihood is needed. With this goal and assuming there are no ties among the uncensored survival times, the partial likelihood function has the following expression:

$$L_P(\boldsymbol{\beta}) = \prod_{j=1}^r \frac{\exp\{\boldsymbol{\beta}'\mathbf{x}_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{\beta}'\mathbf{x}_{(j)}\}}, \quad (4.3)$$

where $R(t_{(j)})$ is the risk set at time $t_{(j)}$ and $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ are the r distinct failure times. $L_P(\boldsymbol{\beta})$ is based on that part of the data that does not carry information about $\lambda_0(t)$. The partial likelihood function is especially useful since it is much more simple than the complete likelihood function and it is a good remedy when the general method of maximum likelihood is not adequate due to the presence of many nuisance parameters (Gómez & Cadarso-Suárez, 2017).

4.1.5 Accelerated failure time model

The accelerated failure time model (AFTM) can be expressed as

$$\log(T) = \mathbf{x}'\boldsymbol{\beta} + W, \quad (4.4)$$

where $\boldsymbol{\beta}$ is the unknown parameter vector, and W is an error variable with an unknown distribution function.

Define $W^* = \exp(W)$ and let $\lambda_w(t)$ denote the hazard function of W^* , which is independent of $\boldsymbol{\beta}$. Then $T = \exp(\mathbf{x}'\boldsymbol{\beta})W^*$, and the hazard and survival functions of T given \mathbf{x} have the forms

$$\lambda(t; \mathbf{x}) = \lambda_w(t \exp\{-\mathbf{x}'\boldsymbol{\beta}\}) \exp(-\mathbf{x}'\boldsymbol{\beta})$$

and

$$S(t; \mathbf{x}) = \exp[-\Lambda_w(t \exp\{-\mathbf{x}'\boldsymbol{\beta}\})],$$

respectively, where $\Lambda_w(t) = \int_0^t \lambda_w(s) ds$.

Notice that under the model (4.4), the effects of the covariates is multiplicative as under the proportional hazards model, but on t instead of the hazard function. This means the effect is to change the timescale and therefore to accelerate or decelerate the time to viral rebound. Although the proportional hazard model specifies that the effect of covariates on the hazard is multiplicative, it does not give a direct relationship between \mathbf{x} and T because $\lambda_0(t)$ is arbitrary. In contrast, the model (4.4) specifies a linear relationship between $\log T$ and \mathbf{x} .

Let $\delta_i = 1$, if there is an event at time t_i , and $\delta_i = 0$, if the data is censored at t_i . The usual likelihood function for right-censored data is given by

$$L_i = \prod_{i=1}^n \{f_i(t_i)^{\delta_i} (1 - F_i(t_i))^{1-\delta_i}\}, \quad (4.5)$$

where $f_i(t_i)$ is the density at time t_i and $1 - F_i(t_i)$ corresponds to the probability of survival beyond time t_i (censoring point).

In the case of interval-censored data, T_L and T_R are observed and T_i is not, that is $0 < T_{Li} < T_i < T_{Ri} < \infty$. The likelihood function for interval-censored data is written as

$$L_i = \prod_{i=1}^n \{F_i(t_{Ri}) - F_i(t_{Li})\}. \quad (4.6)$$

4.2 Omics data analysis

In biological context, the suffix “omics” is used to refer to the study of large sets of biological molecules (Smith et al., 2005). In other words, the study of different components participating or regulating complex biological processes, triggering the development of several fields that, together, are described with the term omics. Among them, we can mention genomics, proteomics, or metabolomics. Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. In this thesis we work with transcriptome data.

4.2.1 Transcriptome

Transcriptome is the study of the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell, using high-throughput methods, such as microarray analysis.

Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cells populations, or in response to different treatments. Following, as in this thesis we work with transcriptome data belonging to mRNA and miRNA, we will explain these particular concepts with further details.

Messenger ribonucleic acid (mRNA)

Messenger ribonucleic acid (mRNA) is a subtype of RNA created during transcription, that carries a portion of the DNA code to other parts of the cell for processing. During the transcription process, a single strand of DNA is encoded by RNA polymerase, and mRNA is synthesized. The mRNA's role in protein synthesis can be explained following the Figure 4.1.

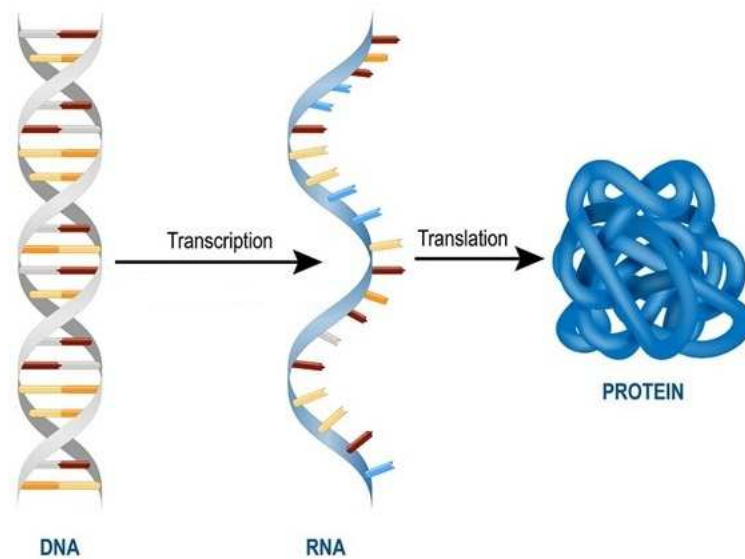


Figure 4.1: mRNA's role in protein synthesis. Source: <https://www.news-medical.net/life-sciences/-Types-of-RNA-mRNA-rRNA-and-tRNA.aspx>.

In a first step, through a process known as transcription, an RNA copy of a DNA sequence for creating a given protein is made. Then, this mRNA copy travels from the nucleus of the cell to the part of the cell known as cytoplasm, which houses ribosomes. Ribosomes are complex machinery in the cells that are responsible for making proteins. Following, through another process known as translation, ribosomes read the mRNA, and follow the instructions, creating the protein step by step. Finally, the cell expresses the protein and it, in turn, carries out its designated function in the cell or the body.

Micro ribonucleic acid (miRNA)

Micro ribonucleic acid (miRNA) represents a class of small, 18- to 28-nucleotide-long, noncoding RNA molecules. Their major role is in the post transcriptional regulation of protein expression, and their involvement was demonstrated in normal and in pathological cellular processes. miRNAs can be described as “multivalent”, with one miRNA able to target multiple genes, thus regulating the expression of several proteins. They were demonstrated to act on several key cellular processes, such as cell differentiation, cell cycle progression, and apoptosis. In tumors, some miRNAs function as oncogenes, others as tumor suppressors, upregulation of oncogenic miRNAs, among others (Tanase et al., 2011).

As the development in high throughput technologies has become more common and accessible, it is becoming usual to take several distinct simultaneous approaches to study the same problem. In practice, this means that data of different types may be available for the same study, highlighting the need for methods and tools to analyse them in a combined way (Sánchez et al., 2012). In fact, the idea that efficient integration of data from different omics can greatly facilitate the discovery of true causes and states of disease is rapidly pervading the biomedical community (Joyce & Palsson, 2006).

4.2.2 Introduction to microarrays

Measuring relative changes in levels of specific mRNAs provide information about what is going on in the cells from which they come. In this thesis, we analyze mRNA data in order to identify biomarkers to predict the viral rebound of HIV-infected patients.

In this work, we can assume that a microarray dataset is a matrix of continuous values that represents the expressions of a set of genes (one gene per row), in a variety of samples (one sample per column), see Figure 4.2

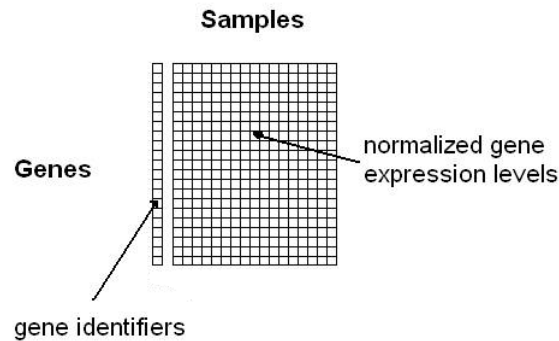


Figure 4.2: A simplified view of a gene expression matrix. Adapted from [Sanz & Sánchez-Pla \(2019\)](#).

To analyze this data, we will use the R statistical software ([R Core Team, 2020](#)). Most packages used for the analysis of high throughput genomic data are part of the Bioconductor, which is based on the R programming language. Bioconductor has become the state-of-the-art way to analyze microarray and other omics data.

The first step of the analysis is to read the .CEL files using the package `oligo` ([Carvalho & Irizarry, 2010](#)). CEL files are the files with the “raw data” originated after microarray scanning and preprocessing with Affymetrix software. Affymetrix, Inc. was an American company that was acquired by Thermo Fisher Scientific in March 2016. Affymetrix makes quartz chips for analysis of DNA Microarrays called GeneChip arrays. Affymetrix is focused on oligonucleotide microarrays. These microarrays are used to determine which genes exist in a sample by detecting specific pieces of mRNA. A single chip can be used to analyze thousands of genes in one assay. Chips can be used only once.

4.2.3 Pipeline for mRNA analysis

Quality control of raw mRNA data

The explanation of the pipeline to analyze mRNA data is explained below, following the steps detailed in [Sanz & Sánchez-Pla \(2019\)](#). The first step when analyzing microarray data is to check its quality, since bad quality data could introduce a lot of noise in the analysis.

The Figure 4.3 shows ten samples of microarrays. Microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence

or gene. The DNA molecules attached to each slide act as probes to detect mRNA transcripts expressed by a group of genes.

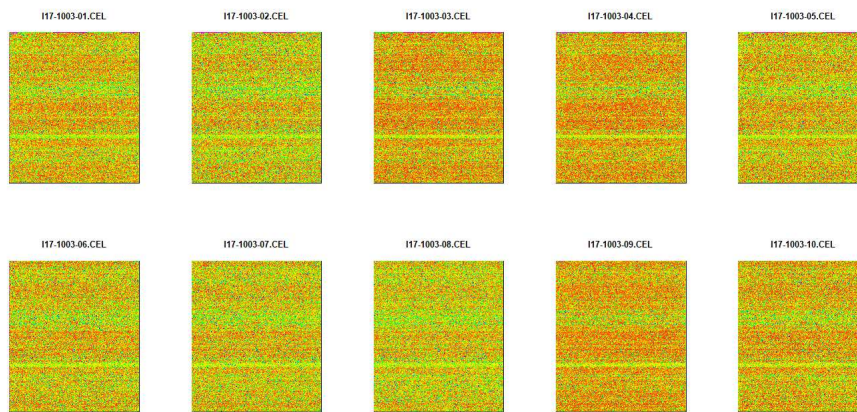


Figure 4.3: Visualization of ten microarrays.

To perform a microarray analysis, mRNA molecules are typically collected from both an experimental sample and a reference sample. The two mRNA samples are then converted into complementary DNA (cDNA), and each sample is labelled with a fluorescent probe of a different color. Both samples are then mixed together and allowed to bind to the microarray slide. The process in which the cDNA molecules bind to the DNA probes on the slide is called hybridization. Following hybridization, the microarray is scanned to measure the expression of each gene printed on the slide. If the expression of a particular gene is higher in the experimental sample than in the reference sample, then the corresponding spot on the microarray appears red. In contrast, if the expression in the experimental sample is lower than in the reference sample, then the spot appears green. Finally, if there is equal expression in the two samples, then the spot appears yellow.

The Bioconductor package `ArrayQualityMetrics` (Kauffmann et al., 2008) generates microarray quality metrics reports for microarray data. This report is a useful tool to observe the existence of outliers in our data. Usually if there is less than three marks in this report the potential problems are small and solved by following the normalization process.

Data normalization

The process of data normalization is necessary in order to make the arrays comparable among them and try to reduce, and if it is possible to eliminate, all the variability in the samples not owing to biological reasons. Normalization process tries to assure that intensity differences present in the array, reflects the differential expression of genes, rather than artificial biases due to technical issues. Normalization process is performed using the function `rma` from the Bioconductor `affy` package (Gautier et al., 2004). RMA is the acronym for Robust Multiarray Average, and consists of three steps:

Step 1: Background correction

A probe pair consists of a Perfect Match probe (PM), which is designed to match exactly the sequence of interest, and a Mismatch probe (MM), which is designed to contain a single base mismatch at the center base position of the 25-mer oligonucleotide probe.

Let's assume PM data is a combination of background and signal

PM = Signal + Background, where

Signal : $S \sim \exp(\lambda)$

Background : $B \sim N(\mu, \sigma^2)$

The background correction is performed on each array separately. The idea is to estimate μ , σ and λ separately in each chip using the observed distribution of PMs. In this way it is possible to obtain an estimate of $E(S|PM)$

$$E(S|PM) = PM - \mu - \lambda\sigma^2 + \sigma \frac{\phi((PM - \mu - \lambda\sigma^2)/\sigma) - \phi((\mu + \lambda\sigma^2)/\sigma)}{\Phi((PM - \mu - \lambda\sigma^2)/\sigma) - \Phi((\mu + \lambda\sigma^2)/\sigma) - 1}$$

for each PM value. These estimates are the background adjusted values.

Step 2: Normalization

The next step corresponds to normalize across all the arrays. To perform this step and correct for array biases rma uses "Quantile normalization". It consist of a repetitive process of replacing the ordered highest values on each chip with the average of the same order value for all the chips.

Step 3: Summarization

Once we have the background corrected, normalized and \log_2 -transformed intensities (Y_{ijn}), being i , j , and n the subscripts for chips, probes (genes) and individuals, respectively.

The next step is to consider the equation

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \epsilon_{ijn}$$

Where Y_{ijn} correspond to the \log_2 -transformed intensities after the background correction and normalization steps. The subscripts for chips, probes (genes) and individuals are i , j , and n , respectively. Moreover, μ_{in} is the log scale expression level (RMA measure), α_{jn} is the probe affinity effect and ϵ_{ijn} corresponds to the independent identically distributed error term (with mean 0).

The main idea of this step is to combine these intensity values Y_{ijn} to get a single intensity value for each gene (probeset). This is done using “median polishing” following three steps: 1) Each chip is normalised to its median. 2) Each gene is normalised to its median. And 3) the previous steps are repeated until medians converge, with a maximum of 5 iterations to prevent infinite loops.

Quality control of normalized data

After the normalization process it is interesting to perform a second quality control and have a visual idea on how the data looks. We can use the `ArrayQualityMetrics` package again and visualize the data using, for example, the boxplots for the distributions of intensities for the normalized data.

Detecting the most variable genes

Selection of differential expressed genes is affected by the number of genes. The higher the number, the greater the necessary adjustment of p-values, which will lead us to end up miscarrying more genes.

If a gene is differentially expressed, it is expected that there is a certain difference between groups, and therefore the overall variance of the genes will be greater than that of those that do not have differential expression. Plotting the overall variability of all genes is useful to decide which percentage of genes shows a variability that can be attributed to other causes than random variation. Figure 4.4 depicts an example of the standard deviations of a set of genes sorted from the smallest to the biggest values. The plot shows that the most variable genes are those with a standard deviation above 90-95% of all standard deviations.

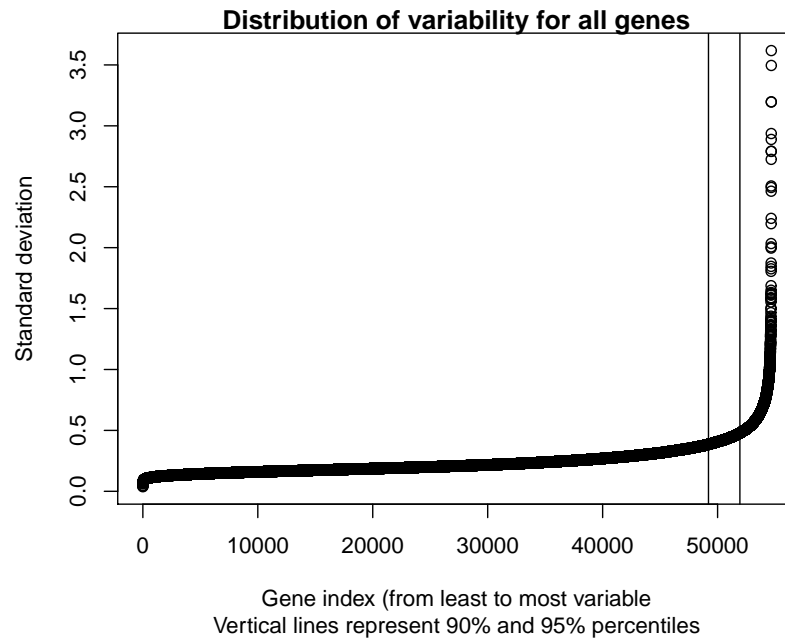


Figure 4.4: Values of standard deviations along all samples for all genes ordered from the smallest to the biggest.

Filtering the least variable genes

Filtering out those genes whose variability can be attributed to random variation, that is the genes that are, reasonably, not expected to be differential expressed, has proven to be useful to reduce the number of tests to be performed with the corresponding increase in power (Hackstadt & Hess, 2009).

We use Entrez Gene ID in our filtering process. Entrez Gene (<http://www.ncbi.nlm.nih.gov/gene>) is National Center for Biotechnology Information (NCBI)'s database for gene-specific information. Entrez Gene maintains records from genomes which have been completely sequenced, which have an active research community to submit gene-specific information, or which are scheduled for intense sequence analysis (Maglott et al., 2005).

We filtered out:

- features without an Entrez Gene ID annotation
- in the case of features mapping to the same Entrez Gene ID, then the feature with the largest value of IQR will be retained and the other(s) removed
- the features values less than the 75th percentile
- Affymetrix quality control probe sets

Model estimation and gene selection

To decide if the genes are differentially expressed, estimating the model, defining the contrasts and performing the significance test are needed. For this purpose, the Bioconductor `limma` package ([Ritchie et al., 2015](#)) is used. This package is used for the analysis of gene expression microarray data, especially the use of linear models for analyzing designed experiments and the assessment of differential expression. `limma` provides the ability to analyze comparisons between many RNA targets simultaneously in different designed experiments.

The analysis provides the usual test statistics such as fold-change t-moderated or adjusted p-values that are used to order the genes from more to less differential expressed.

In order to control the percentage of false positives that may result from the high number of contrasts made simultaneously, the p-values are adjusted so that we have control over the false positive rate using the Benjamini and Hochberg method ([Benjamini & Hochberg, 1995](#)).

Gene Annotation

An additional and useful step is to provide additional information on the features that have been previously selected. This process is called “annotation” and essentially looks for information to associate identifiers, usually corresponding to transcripts, with more familiar names such as the Gene Symbol, the Entrez Gene identifier or the Gene description.

The annotation is an essential step, because until now we only have the gene names in a particular format, but in order to use this information with other databases or only to have more information about these genes, we must annotate the results. In our case we use the annotation package `hgu133plus2cdf` ([Project, 2015](#)).

ELASTIC-NET APPROACH FOR THE ACCELERATED FAILURE TIME MODEL

5.1 Introduction

To identify biomarkers as potential risk factors of HIV viral rebound, it is essential to study low and high-dimensional data. Low-dimensional data includes information about the clinical, viral, analytical parameters per patient as well as their survival. With high-dimensional data we refer to the information on the omics data. Although different approaches to combine low and high dimensional data have been developed, to the best of our knowledge they have not been applied to HIV studies. Because of this, one aim of this thesis is to develop a model that considers omics and survival data.

The development of immunologic interventions to control viral rebound in HIV infection is a major goal of the HIV-1 cure field. In this chapter, we present the DCV2 clinical trial ([García et al., 2013](#)) based on a therapeutic vaccine in HIV-infected patients. In this clinical trial, a therapeutic vaccination with “kick and kill” strategy has been proposed to control viral replication after discontinuation of antiretroviral therapy. For more information on “kick and kill” vaccines, see [Chapter 3](#).

A major challenge to analyze the time to viral rebound in the context of this trial consists of identifying convenient biomarkers. These have to be chosen among more than five thousand messenger RNAs (mRNAs). We address this problem by means of an elastic-net approach for the accelerated failure time (AFT) model. Elastic-net regularization combines the penalizations from ridge regression and LASSO and allows automatic variable selection and continuous shrinkage, as well as the selection of groups of correlated variables. The AFT model has an intuitive physical interpretation and is a useful alternative to the Cox model in survival analysis ([Wei, 1992](#)). The AFT model is a parametric model that is based upon the survival curve rather than the hazard function. In this study, the AFT model

was used to analyze the interval-censored time (weeks) to viral rebound from 35 patients, considering distinct mRNAs as potential predictors.

The chapter is organized as follows. First, we explain the need for the elastic-net penalty, starting with the ordinary least squares and its pitfalls to continue with two penalization techniques: ridge regression and LASSO which help to understand better the elastic-net penalization. Second, we present the state of the art for the elastic-net penalization applied to the PH and the AFT model, considering different approaches. Third, we introduce the elastic-net penalization for the PH model. Fourth, we describe with greater detail the elastic-net approach for the AFT model. Finally, we apply different methods previously described to our DCV2 dataset.

5.2 From the ordinary least squares to the elastic net penalized regression model

The continued development of high-throughput genomic technologies has fundamentally changed the genetic analyses of complex traits and diseases. Nowadays, multi-omics datasets are available for many different diseases (mainly cancers). The information extracted from these type of data can be an avenue for improving the understanding about some specific diseases, for example, discovering new biomarkers to better predict disease risks and prognosis, as well as the development of new therapeutic treatments. In the area of cancer research, many clinical data and meta-dimensional omics data have been generated from large-scale initiatives such as The Cancer Genome Atlas (TCGA), available at <http://gdc.cancer.gov>. Nevertheless, as far as we know, these initiatives have not been taken place in HIV studies. Using a model that combines, in this case, a specific omic layer (transcriptome) and associates it with the time to viral rebound of HIV-infected individuals, could help to accurately identify biomarkers and gain a deeper understanding of the HIV viral dynamics.

The information that omics data provide can be used, as we mentioned before, to develop models for understanding and predicting disease risk and disease prognosis. However, integrating high-dimensional omics data into risk-assessment models is statistically and computationally challenging. High-dimensionality is typically handled via either variable selection, dimension reduction or regularization techniques. A second problem is the relative importance given to the clinical predictors and omics predictors respectively, an issue that we can denote as the combination of low and high-dimensional data (De Bin et al., 2014). In this chapter we introduce the elastic-net penalization to deal with high-dimensional mRNA data. Following, we start presenting the ordinary least squares and how this methodology raises the need for regularization techniques.

5.2.1 Ordinary least squares

The usual linear regression model can be expressed as follows:

$$Y_i = \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i, \quad i = 1, \dots, n$$

or in its matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y} = (y_1, \dots, y_n)'$ is a continuous response variable, \mathbf{X} is the design matrix containing the values of the p independent variables of the n observations, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the parameter vector, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ corresponds to the error term.

The ordinary least squares (OLS) estimates of $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ are obtained by minimizing the residual sum of squares:

$$\boldsymbol{\beta}_{\text{OLS}} = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j X_{ij} \right)^2.$$

In the case that the columns of \mathbf{X} form a linearly independent set, the solution is unique and is given by $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. There are two critical characteristics of estimators to be considered: the bias and the variance. The bias is the difference between the true population parameter and the expected estimator

$$\text{Bias}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = E(\hat{\boldsymbol{\beta}}_{\text{OLS}}) - \boldsymbol{\beta} \quad (5.1)$$

and it measures the accuracy of the estimates. Variance, on the other hand, measures the spread, or uncertainty, in these estimates. It is given by

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}, \quad (5.2)$$

where the unknown error variance σ^2 can be estimated from the residuals as

$$\hat{\sigma}^2 = \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{n-p}, \quad (5.3)$$

where $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. The model's error can be decomposed into three parts: error resulting from a large variance, error resulting from significant bias, and the remainder which is the unexplainable part.

$$MSE = (E(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}))^2 + E(\mathbf{X}\hat{\boldsymbol{\beta}} - E(\mathbf{X}\hat{\boldsymbol{\beta}}))^2 + \sigma^2 = \text{Bias}^2 + \text{Variance} + \sigma^2.$$

The OLS estimator has the desired property of being unbiased. However, it can have a huge variance. This could happen when the predictor variables are strongly correlated with each other or when there are many predictors. The latter is reflected in the Formula (5.3), if p approaches n , the variance

approaches infinity. The general solution to this is to reduce variance at the cost of introducing some bias. This approach is called regularization and is almost always beneficial for the predictive performance of the model. Penalization techniques have been proposed to improve OLS, such as ridge and LASSO regression, as we explain next.

5.2.2 Ridge regression

Ridge regression (Hoerl & Kennard, 1970) is a method that deals with the problem of collinearity in a linear model estimated by OLS. It is well known that forward or backward selection are used with the purpose of obtaining a parsimonious model, decreasing the number of parameters, but these methods are not able to tell anything about the removed variables' effect on the response. Removing predictors from the model can be seen as setting their coefficients to zero. Instead of forcing them to be exactly zero, a penalization is introduced if they are too far from zero, thus enforcing them to be small in a continuous way. Decreasing model complexity while keeping all variables in the model is what ridge regression does.

Ridge regression minimizes the residual sum of squares subject to a bound on the L_2 norm of the coefficients

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j X_{ij})^2 + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}, \quad (5.4)$$

where λ_2 is the parameter of penalization.

The Equation (5.4) can also be expressed in matrix form: $\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I})^{-1}(\mathbf{X}'\mathbf{Y})$, where \mathbf{I} denotes the identity matrix. Notice that as λ_2 tends to zero, the parameter $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ tends to $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

Ridge regression is a continuous shrinkage method, that is, shrinkage reduces the size of the coefficient estimates (shrinking them towards zero). Note that if a coefficient gets shrunk to exactly zero, the corresponding variable drops out of the model. Ridge regression achieves a better prediction performance through a bias-variance trade-off. Incorporating the regularization coefficient in the Formulas (5.1) and (5.2), for bias and variance, respectively, gives us

$$\begin{aligned} \operatorname{Bias}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) &= -\lambda_2 (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I})^{-1} \boldsymbol{\beta} \\ \operatorname{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) &= \sigma^2 (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I})^{-1}. \end{aligned}$$

From the last expressions we can see that as λ_2 becomes larger, the variance decreases, and the bias increases. Since $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ cannot be zero no matter how big the λ_2 value is set, ridge regression cannot produce a parsimonious model, because it always keeps all the predictors in the model.

5.2.3 Least absolute shrinkage and selection operator (LASSO)

The least absolute shrinkage and selection operator (**LASSO**) was proposed by Tibshirani (1996) and it is, conceptually, quite similar to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression, which penalizes sum of squared coefficients (using a L_2 penalty), LASSO penalizes the sum of their absolute values based on a L_1 penalty. As a result, for high values of λ , many coefficients are exactly zeroed under LASSO, which is never the case in ridge regression.

The LASSO solves the problem of OLS by imposing an L_1 penalty on the regression coefficients:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^p |\beta_j| \right\}.$$

Owing to the nature of the L_1 penalty, the LASSO does both continuous shrinkage and automatic variable selection simultaneously. But this method is less efficient for a big number of covariates. Moreover, if there are correlated variables, it tends to select just one of them and to ignore the others, that is, LASSO does not take into account the “grouping effect” (Zou & Hastie, 2005). The property of the grouping effect states that highly correlated features will have similar estimated coefficients. For example, in gene expression studies, genes that have similar functions, or that work together in a pathway to accomplish a certain function, are often correlated.

Comparing ridge regression and LASSO, we can mention that neither of them outperforms the other. LASSO can set some coefficients to zero, thus performing variable selection, while ridge cannot. Both methods allow us to use correlated predictors, but they solve multicollinearity issue differently: in ridge regression the coefficients of correlated predictors are similar, in LASSO one of the correlated predictors has a larger coefficient while the rest are zeroed. LASSO tends to do well if there are a small number of significant parameters and the others are close to zero, this means that only a few predictors actually influence the response. Ridge performs well if there are many large parameters of about the same value, this means that most predictors have an impact on the response.

Why does the LASSO provide variable selection?

Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, be the usual linear regression model with an L_1 penalty on $\hat{\boldsymbol{\beta}}$ and a least squares loss function on $\hat{\boldsymbol{\epsilon}}$. Expanding the expression to be minimized we obtain

$$\operatorname{argmin}\{\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + 2\lambda_1|\hat{\boldsymbol{\beta}}|\}.$$

If $\hat{\boldsymbol{\beta}} > 0$, the penalty term is equal to $2\lambda_1\hat{\boldsymbol{\beta}}$. The derivative of the objective function with respect to $\hat{\boldsymbol{\beta}}$ is $-2\mathbf{Y}'\mathbf{X} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + 2\lambda_1$ which has the solution $\hat{\boldsymbol{\beta}} = (\mathbf{Y}'\mathbf{X} - \lambda_1)/(\mathbf{X}'\mathbf{X})$. By increasing λ_1 we can drive $\hat{\boldsymbol{\beta}}$ to zero. In the case that $\hat{\boldsymbol{\beta}}$ becomes negative, the derivative of the objective function changes to $-2\mathbf{Y}'\mathbf{X} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\lambda_1$ where the flip in the sign of λ_1 is due to the absolute value nature of the penalty term. This leads to the solution $\hat{\boldsymbol{\beta}} = (\mathbf{Y}'\mathbf{X} + \lambda_1)/(\mathbf{X}'\mathbf{X})$, which is inconsistent with $\hat{\boldsymbol{\beta}} < 0$ (given that the least squares

solution is greater than zero), which implies $\mathbf{Y}'\mathbf{X} > 0$ and $\lambda_1 > 0$. With the least squares penalty $\lambda_1 \hat{\boldsymbol{\beta}}^2$, however, the derivative becomes $-2\mathbf{Y}'\mathbf{X} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + 2\lambda_1 \hat{\boldsymbol{\beta}}$ which has solution $\hat{\boldsymbol{\beta}} = (\mathbf{Y}'\mathbf{X})/(\mathbf{X}'\mathbf{X} + \lambda_1)$.

Following the Figure 5.1 from Elements of Statistical Learning by J. Friedman et al. (2001) is very illustrative:

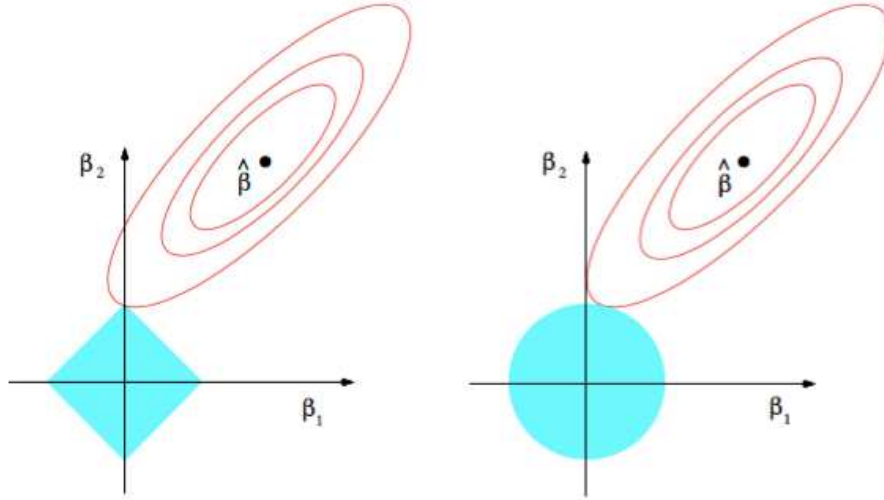


Figure 5.1: Estimation picture for the LASSO (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function (J. Friedman et al., 2001).

The $\hat{\boldsymbol{\beta}}$ is the unconstrained least squares estimate. The red ellipses are the contours of the least squares error function, in terms of parameters β_1 and β_2 . Without constraints, the error function is minimized at the MLE $\hat{\boldsymbol{\beta}}$, and its value increases as the red ellipses out expand. The diamond and disk regions are feasible regions for LASSO and ridge regression, based on L_1 and L_2 penalization, respectively. Heuristically, for each method, we are looking for the intersection of the red ellipses and the blue region as the objective is to minimize the error function while maintaining the feasibility. Considering the diamond feasible region, based on the L_1 constraint, it is more likely to produce an intersection that has one component of the solution equal to zero, this is the sparse model, due to the geometric properties of ellipses, disks, and diamonds.

5.2.4 Elastic net

Similar to the LASSO, the **elastic net** (Zou & Hastie, 2005) simultaneously does automatic variable selection and continuous shrinkage, and similar to the ridge regression, it can select groups of correlated variables. To estimate $\boldsymbol{\beta}$, the naive elastic net uses a mixture of the L_1 (LASSO) and L_2 (ridge regression) penalties:

$$\hat{\boldsymbol{\beta}}_{\text{naive}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j X_{ij})^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}.$$

L_1 regularization tends to produce sparse solutions, but also tends to select the feature most strongly correlated with the outcome and zero out the rest. Moreover, in a dataset with n observations, it can select at most n features. While L_2 regularization is suited to deal with p features. As the naive elastic net uses a double penalization that can introduce bias in the estimation, it is necessary to correct the previous estimation, in this way

$$\hat{\boldsymbol{\beta}}_{\text{enet}} = (1 + \lambda_2) \hat{\boldsymbol{\beta}}_{\text{naive}}.$$

A common reparameterization of the elastic net is to express the regularization parameters in terms of λ , which controls the overall degree of regularization, and α , which controls the balance between the LASSO and ridge penalties

$$\begin{aligned} \lambda_1 &= \alpha \lambda \\ \lambda_2 &= (1 - \alpha) \lambda. \end{aligned}$$

This reparameterization is useful in practice, as it allows one to fix α and then select a single tuning parameter λ , which is more straightforward than attempting to select λ_1 and λ_2 separately. Summing up, the elastic net creates a useful compromise between the ridge regression penalty ($\alpha = 0$) and the LASSO penalty ($\alpha = 1$). In addition, elastic net is not limited by the fact that $p \geq n$ and works efficiently if there is a group of variables among which the pairwise correlations are very high, this allows for those genes (in the microarray context) sharing the same biological pathway, to include whole groups into the model automatically once one gene among them is selected.

5.2.5 Selection of the optimal tuning parameter λ_{OPT}

For the selection of the optimal λ_{OPT} we use leave-one-out cross validation. Cross validation schemes can be implemented in most statistical frameworks and for most estimation procedures. The general idea of cross-validation is to separate m observation from the n subjects (sample size), fit a model based on the remaining $n - m$ observations and test it on the m observations outside the dataset. The left-out group of size m is called the test set, while the remaining group of size $n - m$ is called the training set. The test set is used to validate or assess the performance of the estimators using the mean squared error criterion. The training and test splitting is repeated several times.

In the K -fold cross validation the dataset is divided into k subsets, and the method explained above is repeated k times. Each time, one of the k subsets is used as the test set and the other $k - 1$ subsets are put together to form a training set. Then the average error across all k trials is computed.

Leave-one-out cross validation (LOOCV) is K -fold cross validation taken to its logical extreme, with K equal to N , the number of data points in the set. That means that N separate times, the function is trained on all the data except for one point and a prediction is made for that point. As before the average error is computed and used to evaluate the model. We adopt the LOOCV approach to select the tuning parameter λ that minimizes the CV error in the following methodology applied to the DCV2 trial.

5.2.6 Elastic-net extension: the adaptive elastic net

While the elastic net combines the best of ridge and LASSO regression, it may lack from the desirable oracle property.

Let's suppose that we have p true predictors, $\beta_1^*, \beta_2^*, \dots, \beta_p^*$. We define \mathcal{A} as the subset of indicators for which β_j^* is not null, $\mathcal{A} = \{j : \beta_j^* \neq 0\}$, and assume that $|\mathcal{A}| = p_0 < p$.

Following the definition of [Fan & Li \(2001\)](#), δ is an oracle procedure if $\hat{\beta}(\delta)$ identifies the right subset model \mathcal{A} and δ has the optimal estimation rate, i.e,

$$\sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \Sigma^*),$$

where Σ^* is the covariance matrix of the true subset model.

The adaptive elastic net is an extension of the elastic net that complies with the oracle property. According to [Zou & Zhang \(2009\)](#), the adaptive elastic net can be viewed as a combination of the elastic-net and the adaptive LASSO (for more information, see [Zou \(2006\)](#)). Suppose we first compute the elastic-net estimator $\hat{\beta}(\text{enet})$, and then we construct the adaptive weights by

$$\hat{w}_j = (|\hat{\beta}_j(\text{enet})| + 1/n)^{-\gamma}, \quad j = 1, 2, \dots, p,$$

where γ is a positive constant (any positive γ can be used, according to [Zou \(2006\)](#)). In the next step, we solve the following optimization problem to get the adaptive elastic-net estimates:

$$\hat{\beta}(\text{AdaEnet}) = \left(1 + \frac{\lambda_2}{n}\right) \left[\operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_j X_{ij})^2 + \lambda_1^* \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\} \right]. \quad (5.5)$$

If we force λ_2 to be zero in (5.5), then the adaptive elastic net reduces to the adaptive LASSO. In this thesis, we used the elastic net; the possible use of the adaptive elastic net is discussed in Chapter 8.

5.3 State of the art

In Table 5.1, we present a summary of the current methodology and R packages to address different maximization methods for PH and AFT models considering complete, right-censored, and interval-

censored data. We have identified three main approaches: semiparametric (SP), piecewise exponential (PE), and parametric (PM). In Table 5.1 we also describe if these approaches consider any type of penalization technique and high-dimensional data.

We start reviewing the different methodologies that account for complete and right-censored data. First, for PH models there are different methods of maximization regarding the type of approach. In the case of the semiparametric (SP) approach, J. Friedman et al. (2010) uses the cyclical coordinate descent algorithm to maximize the log-likelihood function. This algorithm has been implemented in the `glmnet` R package (Hastie & Qian, 2014) and it considers the elastic-net penalization and high-dimensional data. In section 5.6.4 we describe an application of this method by using a midpoint imputation approach for interval-censored times to viral rebound in the DCV2 clinical trial. In the case of the piecewise exponential (PE) approach, Wu & Cook (2015) present the Expectation-Maximization (EM) algorithm to maximize the log-likelihood. In this case the authors considers three different penalization techniques: LASSO, adaptive LASSO (ALASSO) and smoothly clipped absolute deviation (SCAD) that deal with high-dimensional data. For the parametric (PM) approach, the R package `eha` (Broström, 2019) can be used, the maximization is done using the Newton-Raphson algorithm, from the `coxph` function, available at the `survival` package (T. Therneau, 2015). The `eha` package does not consider any penalization technique nor high-dimensional data.

If the relationship between the response variable and the explanatory variable is an AFT model then, considering the SP approach, Chen et al. (2016) uses the Stute's weighted least squares (Stute & Wang, 1994) and the group bridge penalty. This method is able to simultaneously carry out feature selection at both the group and within-group individual variable levels. For the PM approach, the `AdapEnetClass` package has been implemented (Khan & Shaw, 2015) using the Stute's weighted least squares, explained in the work of Khan & Shaw (2016). This package uses adaptive elastic net and weighted elastic net and it considers high-dimensional data. Moreover, the package `iregnet` (accessible on Github at: <https://rdr.io/github/anujkhare/iregnet/man/iregnet.html>) uses the cyclical coordinate descent algorithm. However this package is still under development and has not yet been released on CRAN. The main pitfalls of this package is that its function `cv.iregnet` is still not implemented for the Weibull distribution (the one we use in this thesis). The `cv.iregnet` function works for normal, logistic and exponential distributions and allow us to find the optimal tuning parameter λ .

Methods for interval-censored data have not been thoroughly developed and to the best of our knowledge there is no developed methodology or R packages that address this scenario, considering the PH model with elastic-elastic net penalization and high-dimensional data. As we describe above, in the case of the PE, Wu & Cook (2015) present the EM algorithm to maximize the log-likelihood function. They programed a function using R, which is available as supporting information at <https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12302>. We adapted this function to consider the elastic-net penalization and to apply this to our dataset, as we presented in the Section 5.6.4. More complete information regarding the elastic net approach for the proportional hazards model is pre-

Table 5.1: State-of-the-art of maximization methods for the Semiparametric (SP), Parametric (PM) and Piecewise Exponential (PE) approaches considering the Proportional Hazards (PH) and the Accelerated Failure Time (AFT) models .

Data	Approach	Model	Maximization Method	R package	Penalization	High-dim data
Complete and right-censored	SP	PH	Cyclical coordinate descent (CCD) (J. Friedman et al., 2010)	glmnet (Hastie & Qian, 2014)	Elastic net	Yes
		AFT	Stute's weighted least squares (Chen et al., 2016)	No	Group bridge	Yes
	PE	PH	Expectation-Maximization (EM) algorithm (Wu & Cook, 2015)	R function available at (Wu & Cook, 2015)	LASSO, ALASSO, SCAD	Yes
	PM	PH	Newton Raphson from coxph function, survival package	eha (Broström, 2019)	No	No
		AFT	Stute's weighted least squares (Khan & Shaw, 2016)	AdapEnetClass (Khan & Shaw, 2015)	Adaptive elastic net, weighted elastic net	Yes
			CCD (J. Friedman et al., 2010)	iregnet*	Elastic net	Yes
Interval-censored	SP	PH	No	No	No	No
		AFT	No	No	No	No
	PE	PH	EM algorithm (Wu & Cook, 2015)	R function available at (Wu & Cook, 2015)	LASSO, ALASSO, SCAD	Yes
	PM	PH	Newton Raphson	eha (Broström, 2019)	No	No
		AFT	Newton Raphson	survreg function from survival (T. Therneau, 2015)	No	No
			CCD (J. Friedman et al., 2010)	iregnet*	Elastic net	Yes

* iregnet is accessible on <https://rdrr.io/github/anujkhare/iregnet/man/iregnet.html>.

sented in Section 5.4.

Maximization methods for the AFT model using interval-censored data and considering the SP approach is open to further research. Regarding the PM approach, the function `survreg` from the R package `survival` does not consider high-dimensional data nor penalization techniques. In this scenario, we derive and maximize the log-likelihood function for the AFT model considering a Weibull distribution and considering the interval-censoring. We present the model for the log-likelihood in the Section 5.5 and its application to the DCV2 dataset is presented at Section 5.6.5. In this scenario appears also the `iregnet` package as the first package to fit AFT model with PM approach and considering interval-censored data. However, as we described previously is still under development. The issues on this package are clearly stated on Github: <https://github.com/anujkhare/iregnet/issues>.

Finally, it is also important to mention a work which is also related to the use of elastic net in AFT models. Khan & Shaw (2019) proposed four variable selection algorithms based on the Buckley-James and the Dantzig selector methods.

5.4 Elastic net approach with the proportional hazards model

The main goal of survival analysis is to characterize the dependence of the survival time T on a covariate vector $\mathbf{X} = (X_1, \dots, X_p)'$. Let's define the Cox's proportional hazards model as in the Equation (4.2). As survival data with many predictors prevail in clinical trial studies, risk factor identification becomes more important than ever for analyzing high-dimensional survival data. The problem is to select a submodel of (4.2) by providing a sparse estimate of β .

5.4.1 Estimation of the model parameters with right-censored data

Let T denote the time to some event. The data, based on a sample of size n , consists of the triple $(T_i, \delta_i, \mathbf{x}_i)$, $i = 1, \dots, n$ where T_i is the time on study for the i th patient, δ_i is the event indicator for the i th patient ($\delta_i = 1$ if the event has occurred and $\delta_i = 0$ if the lifetime is right-censored) and $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})'$ is the vector of covariates or risk factors for the i th individual at time $t = 0$ which may affect the survival distribution of T . The standard way to estimate the β_j coefficients is by maximizing the partial likelihood function called $L(\beta)$. The estimators obtained comply with generally good properties of the maximum likelihood method.

Suppose there are r times to the event E (e.g. death), $n - r$ times of censoring and no ties. Denote by $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ the r ordered death times, by $R_j = R(t_{(j)}) = \{i : Y_i \geq t_{(j)}\}$ the set of all individuals at risk of dying at time $t_{(j)}$, that is to say, the set of all those individuals who are alive and not censored at time $t_{(j)}$ - and by $n_j = \text{card}(R_j)$ the number of individuals at risk in $t_{(j)}$. Denote also the set containing all the information in the sample $\Gamma = \{(Y_i, \delta_i, \mathbf{x}_i), i = 1, \dots, n\}$.

The basic principle of the deduction of the partial likelihood function resides in the fact that knowledge of r death times $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ with the labels e_1, e_2, \dots, e_r indicating which individual corre-

sponds the death, is equivalent to the original data (this is certainly true if there is no censoring). For more details, see [Gómez et al. \(2015\)](#). The partial likelihood is defined as

$$L(\beta_1, \dots, \beta_p) = \prod_{j=1}^r P\{e_j = i | \Gamma_j\} = \prod_{j=1}^r P\{x_{(j)} = x_{(j)} | \Gamma_j\}$$

and is interpreted as the product, for each time of death, of the conditional probabilities that the individual whose vector of covariates is $x_{(j)}$ dies at time $t_{(j)}$ knowing death has occurred among n_j individuals at risk at time $t_{(j)}$. Therefore, the partial likelihood is equal to

$$L(\beta_1, \dots, \beta_p) = \prod_{j=1}^r \frac{\exp\{\beta' x_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\beta' x_{lj}\}},$$

or equivalently

$$L(\beta_1, \dots, \beta_p) = \prod_{i=1}^n \left(\frac{\exp\{\beta' x_i\}}{\sum_{l \in R(Y_i)} \exp\{\beta' x_l\}} \right)^{\delta_i} = \prod_{j=1}^r \frac{\exp\{\sum_{k=1}^p \beta_k x_{(j)k}\}}{\sum_{l \in R(t_{(j)})} \exp\{\sum_{k=1}^p \beta_k x_{jk}\}}, \quad (5.6)$$

and its logarithm can be expressed as

$$\log L(\beta_1, \dots, \beta_p) = \sum_{i=1}^n \delta_i \left(\beta' x_i - \log \sum_{l \in R(Y_i)} \exp\{\beta' x_l\} \right). \quad (5.7)$$

The estimator $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ is obtained maximizing the partial likelihood (5.6), or equivalently maximizing the logarithm of the partial likelihood (5.7). The maximization of this function using numerical methods provides the corresponding estimators. It can be shown that the maximum likelihood estimator $\hat{\beta}$ obtained from maximizing the partial likelihood is asymptotically unbiased, efficient and normal. For more details, see [Gómez et al. \(2015\)](#).

For classical problems, with many more observations than predictors, the Cox model performs well. However, problems with $p > n$, lead to degenerate behavior; to maximize the partial likelihood, all of the β_i are sent to $\pm\infty$. [Zou & Hastie \(2005\)](#) propose to maximize the expression (5.6) subject to the constraint $\alpha \sum |\beta_i| + (1 - \alpha) \sum \beta_i^2 \leq c$. Notice if $\alpha = 1$ we are in the LASSO case and if $\alpha = 0$ corresponds to the ridge regression

$$\ell(\beta) = \left[\sum_{i=1}^r x'_{j(i)} \beta - \log \left(\sum_{j \in R_i} \exp\{x'_j \beta\} \right) \right].$$

Hence, if we consider the Lagrangian formulation, our problem becomes

$$\hat{\beta} = \operatorname{argmax} \left[\left(\sum_{i=1}^r x'_{j(i)} \beta - \log \left(\sum_{j \in R_i} \exp\{x'_j \beta\} \right) \right) - \lambda P_\alpha(\beta) \right], \quad (5.8)$$

where

$$\lambda P_\alpha(\beta) = \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1}{2} (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right). \quad (5.9)$$

In R, the `glmnet` package solves the maximization problem in (5.8) for right-censored data over a grid of values of λ covering the entire range. The elastic-net penalty is controlled by α , and bridges the gap between LASSO ($\alpha = 1$, the default) and ridge ($\alpha = 0$). The tuning parameter λ controls the overall strength of the penalty. The `glmnet` algorithms use cyclical coordinate descent (J. Friedman et al., 2010; Simon et al., 2011), which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence.

5.4.2 Estimation of the model parameters with interval-censored data

Let T denote the time to an event of interest and \mathbf{x}_i the vector of covariates, as in the previous section. Let $C_i = [L_i, R_i)$ denote the interval known to contain the event for subject i , $i = 1, \dots, n$. For left-censored data $L_i = 0$, for right-censored data $R_i = \infty$, and for interval censored data $0 < L_i < R_i < \infty$.

The likelihood function is

$$L(\beta) = \prod_{i=1}^n [S(L_i | \mathbf{x}_i) - S(R_i | \mathbf{x}_i)]$$

and the corresponding log-likelihood is

$$\begin{aligned} \log L(\beta) &= \sum_{i=1}^n \log [S(L_i | \mathbf{x}_i) - S(R_i | \mathbf{x}_i)] \\ &= \sum_{i=1}^n \log [S_0(L_i)^{\exp\{\beta' \mathbf{x}_i\}} - S_0(R_i)^{\exp\{\beta' \mathbf{x}_i\}}], \end{aligned}$$

where $S_0(t)$ corresponds to the baseline survival function.

When viewing this as a variable selection problem, we are specifically interested in identifying the covariates for which the regression coefficients are non-zero. The main idea is to maximize the penalized likelihood of the form

$$\hat{\beta} = \operatorname{argmax} \left[\frac{1}{m} \log L(\beta) - \lambda P_\alpha(\beta) \right], \quad (5.10)$$

where $\lambda P_\alpha(\beta)$ is defined as in (5.9). The value of the scalar λ is typically found by cross-validation (Shao, 1993) or generalized cross-validation (Golub et al., 1979).

To solve this problem, Wu & Cook (2015) proposed to adopt a flexible piecewise exponential model (M. Friedman, 1982) for the event of interest and penalize the complete data likelihood constructed by treating the interval-censored failure times as known. The authors used an expectation-maximization (EM) algorithm (Dempster et al., 1977) using different penalization techniques including LASSO.

The R function written by (Wu & Cook, 2015) was implemented by authors to be used with LASSO, adaptive LASSO and SCAD penalizations. Our contribution was to extend this function and adapt it for using the elastic-net penalization, which was not implemented, to our dataset considering the midpoint of the censoring intervals. We use the midpoint since the function does not allow for an interval-censored response. The application to our DCV2 trial can be seen in Section 5.4. The R code corresponding to our adaptation can be found in <http://doi.org/10.5281/zenodo.4678278>. The data is accessible upon request due to privacy issues.

5.5 Elastic net approach for the accelerated failure time model

In this section, we derive the expression of the likelihood function assuming an accelerated failure time (AFT) model for the interval-censored times to viral rebound data. We consider an elastic-net penalization and use mRNAs as predictors of this rebound. Besides, we maximize the corresponding penalized likelihood function to obtain an estimation of the model parameters, and apply this method to the specific case of the DCV2 trial.

We consider the accelerated failure time model as in Equation (4.4) and present the likelihood function using the elastic net approach for the AFTM in which T follows a Weibull distribution. We assume T follows a Weibull since this distribution presents a flexible shape and can be used to model a wide range of failure rates. More characteristics of this distribution can be found in Kızılersü et al. (2018).

First, we present the usual expression for the likelihood function for a) right-censored data and for b) interval-censored data. Second, we present the relation between Weibull distribution and the log linear model. Third, we present the log-likelihood according to the Weibull model. Finally, we consider the elastic net penalization approach.

5.5.1 Relation between Weibull distribution and the log linear model

Under the Weibull model, the survival and density functions of T have the forms

$$S(t) = \exp\left[-\left(\frac{t}{\rho}\right)^k\right]$$

and

$$f(t) = \frac{k}{\rho} \left(\frac{t}{\rho}\right)^{k-1} \exp\left[-\left(\frac{t}{\rho}\right)^k\right]$$

respectively.

A Weibull distribution $T \sim \mathcal{W}(\rho, k)$ can also be described by means of a log linear model with parameters μ and σ , where the error distribution W is the standard Gumbel distribution:

$$\log(T) = \mu + \sigma W.$$

The relation between the parameters is $k = 1/\sigma$ and $\rho = \exp(-\mu/\sigma)$. The density and survival functions of the standard Gumbel or extreme value distribution are given by

$$\begin{aligned} f(w) &= \exp(w - e^w) \\ S(w) &= \exp(-e^w). \end{aligned}$$

The expression as a log linear model has the advantage that covariates can be incorporated. The parameters of the Weibull distribution depend then on the value of the covariate Z . Whereas the shape parameter $k = 1/\sigma$ is the same for all conditional survival times T given Z , the location parameter changes with Z : $\lambda(z) = \exp(-(\mu + \beta z)/\sigma)$. Due to the invariance property, given the maximum likelihood estimates $(\hat{\mu}, \hat{\beta}, \hat{\sigma})$, the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\lambda}(z)$ are easily obtained applying the corresponding transformation.

Using the Weibull distribution regression model, the term $\exp(-\beta/\sigma)$ corresponds to the relative risk and $\exp(\beta)$ to the acceleration factor when comparing two individuals, whose covariate values differ by one unit. The interpretation of these terms implies that augmenting the covariate by one, the instantaneous risk of dying increases/decreases ($\beta < 0/\beta > 0$) by the factor $\exp(-\beta/\sigma)$, whereas the median time until the event of interest is decreased/increased ($\beta < 0/\beta > 0$) by the factor $\exp(\beta)$, as explained in [Langohr \(2004\)](#).

5.5.2 Log-likelihood function for Weibull model

Using the subindex T for the Weibull distribution survival time T and W for the Gumbel distributed error of the model (4.4), we have the following relations ([J. P. Klein & Moeschberger, 2006](#)):

$$\begin{aligned} f_T(t) &= \frac{1}{\sigma} f_W\left(\frac{\log(t) - \mu - \beta z}{\sigma}\right) = \frac{1}{\sigma} \exp\left(\frac{\log(t) - \mu - \beta z}{\sigma} - e^{\frac{1}{\sigma}(\log(t) - \mu - \beta z)}\right), \\ S_T(t) &= S_W\left(\frac{\log(t) - \mu - \beta z}{\sigma}\right) = \exp\left(-e^{\frac{1}{\sigma}(\log(t) - \mu - \beta z)}\right). \end{aligned} \tag{5.11}$$

Using the density and survival functions of the Gumbel distribution in (5.11), the contributions of each individual to the log-likelihood functions for right-censored and interval-censored data, respectively can be written as follows:

Right-censored data

$$\begin{aligned}
\ell_i = \log L_i &= \log \prod_{i=1}^n \{f_i(t_i)^{\delta_i} (1 - F(t_i))^{1-\delta_i}\} \\
&= \sum_{i=1}^n [\delta_i \log f_i(t_i) + (1 - \delta_i) \log S_i(t_i)] \\
&= \sum_{i=1}^n \left[\frac{\delta_i}{\sigma} \left(\frac{\log(t) - \mu - \beta z}{\sigma} - e^{\frac{1}{\sigma}(\log(t) - \mu - \beta z)} \right) + (1 - \delta_i) \left(-e^{\frac{1}{\sigma}(\log(t) - \mu - \beta z)} \right) \right].
\end{aligned} \tag{5.12}$$

Interval-censored data

$$\begin{aligned}
\ell_i = \log L_i &= \log \prod_{i=1}^n \{F_i(t_R) - F_i(t_L)\} \\
&= \sum_{i=1}^n [\log(S_i(t_L) - S_i(t_R))] \\
&= \sum_{i=1}^n [\log(\exp(-e^{\frac{1}{\sigma}(\log(t_L) - \mu - \beta z)}) - \exp(-e^{\frac{1}{\sigma}(\log(t_R) - \mu - \beta z)}))].
\end{aligned} \tag{5.13}$$

The contribution of each individual to the penalized log-likelihood function, considering interval-censored data and elastic-net penalization ($\alpha = 0.5$) is given by

$$\ell_i = \log(\exp(-e^{\frac{1}{\sigma}(\log(t_L) - \mu - \beta z)}) - \exp(-e^{\frac{1}{\sigma}(\log(t_R) - \mu - \beta z)})) - \lambda P_\alpha(\beta), \tag{5.14}$$

where

$$P_\alpha(\beta) = \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1}{2} (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right).$$

In this way, the log-likelihood function written as:

$$\ell = \frac{1}{n} \sum_{i=1}^n \log(\exp(-e^{\frac{1}{\sigma}(\log(t_L) - \mu - \beta z)}) - \exp(-e^{\frac{1}{\sigma}(\log(t_R) - \mu - \beta z)})) - \lambda P_\alpha(\beta), \tag{5.15}$$

where

$$P_\alpha(\beta) = \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1}{2} (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right).$$

5.5.3 The optimization problem

The maximization of the log-likelihood function ℓ (5.18) is a non-linear programming optimization problem that can be addressed using different methods. To solve this type of problem considering multidimensional data, we have used gradient based and non-gradient based algorithms.

Gradient based refers to an algorithm to solve minimization problems with search directions defined by the gradient of the function at the current point (Polak, 2012). The algorithms we have used in this thesis are the Conjugate Gradient (CG) (Shewchuk, 1994) and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) algorithms. The CG is an iterative algorithm that uses the first derivative to obtain the gradient for the search direction. Specifically, the search direction s_n of the next point results from the negative gradient of the last point. The basic steps are: first, calculate search direction $s_n = -\nabla\ell$. Second, pick next point x_{n+1} by moving with step size a_n in the search direction (step size a can be fixed or variable). Repeat steps 1 and 2 until $\nabla\ell = 0$ or another stopping criterion. The movement towards the minimum looks like a “zig-zagging” movement. On the other hand, BGFS is a quasi-Newton method (also known as a variable metric algorithm), which uses function values and gradients to build up a picture of the surface to be optimized.

In our case and for the ease of notation, let's rewrite our penalized log-likelihood (5.18) in terms of A_1 and A_2 as follows:

$$\begin{aligned} A_1 &= -\exp\left[\frac{1}{\sigma}(\log(t_{Li}) - \mu - \beta z)\right] \\ A_2 &= -\exp\left[\frac{1}{\sigma}(\log(t_{Ri}) - \mu - \beta z)\right] \end{aligned} \quad (5.16)$$

Replacing (5.16) in the sum of all the n individual contributions described by (5.14) we obtain:

$$\ell = \frac{1}{n} \sum_{i=1}^n \log[\exp(A_1) - \exp(A_2)] - \lambda P_\alpha(\beta)$$

Now, let's define $\nabla\ell = [\partial_\mu(\ell), \partial_{\beta_j}(\ell), \partial_\sigma(\ell)]$ with $j = 1, \dots, p$ as the gradient of ℓ , where

$$\begin{aligned} \partial_\mu(\ell) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma} \cdot \frac{1}{\exp(A_1) - \exp(A_2)} \cdot [A_1 \cdot \exp(A_1) - A_2 \cdot \exp(A_2)] \\ \partial_{\beta_j}(\ell) &= \frac{1}{n} \sum_{i=1}^n \frac{z_j}{\sigma} \cdot \frac{1}{\exp(A_1) - \exp(A_2)} \cdot [A_1 \exp(A_1) - A_2 \cdot \exp(A_2)] + \alpha \lambda \text{sign}(\beta_j) + (1 - \alpha) \lambda \beta_j \\ \partial_\sigma(\ell) &= \frac{1}{n} \sum_{i=1}^n -\frac{1}{\sigma^2} \cdot \frac{1}{\exp(A_1) - \exp(A_2)} \cdot [A_1 \cdot \exp(A_1) \cdot (\log(t_{Li}) - \mu - \beta z) - A_2 \cdot \exp(A_2) \cdot (\log(t_{Ri}) - \mu - \beta z)] \end{aligned} \quad (5.17)$$

The performance of a gradient based method strongly depends on the initial values supplied. Several optimization runs with different initial values might be necessary if no a priori knowledge about the function to optimize can be applied.

Non-gradient methods, on the other hand, do not require gradient information to converge to a solution. Rather, these methods solely use function evaluations of the objective function to converge to a solution (Hare et al., 2013). The non-gradient based algorithm we have used in this thesis is called Nelder-Mead. The Nelder-Mead (Nelder & Mead, 1965) method uses a geometrical shape called a simplex as its “vehicle” of sorts to search the domain. This is why the technique is also called the simplex search method. The Nelder-Mead method attempts to minimize a scalar-valued non-linear function of n real variables using only function values and does not depend on the implicit or explicit expressions of the derivatives of the non-linear function. Each iteration of a simplex-based direct search method begins with a simplex, specified by its $n + 1$ vertices and the associated function values. One or more test points are computed, along with their function values, and the iteration terminates with a new (different) simplex such that the function values at its vertices satisfy some form of descent condition compared to the previous simplex (Lagarias et al., 1998). In an ideal case, the last few iterations of this algorithm would involve the simplex shrinking inwards towards the best point inside it. At the end, the vertex of the simplex that yields that most optimal objective value, is returned.

In the following section we will explain the different approaches we have applied to maximize the elastic-net penalized log-likelihood function.

5.5.4 Different approaches to maximize the elastic-net penalized log-likelihood function

In this section we present two approaches, one based in the implementation of the maximization using the Nelder-Mead method and the other one applying the package `iregnet`.

Approach A

We have programmed the maximization of the log-likelihood function ℓ considering the elastic-net penalization and the interval-censored data, according to Zou & Hastie (2005) and also according to Equation (5.18). Notice that function in (5.18) has the penalization term slightly different from the one presented in Zou & Hastie (2005), the factor in the L_2 norm is 0.5 and not 1. This factor is discussed in section 5.7.

The maximization of the elastic-net penalized log-likelihood function requires a value of the parameter λ . To obtain this value we have used the `cv.glmnet` function of the `glmnet` package using leave-one-out cross-validation to find the optimal λ value in each case. Moreover, to obtain this parameter estimation, we have used midpoint imputation for the interval-censored times to viral rebound, because this package does not allow for interval-censored data. The main advantage of using this approach is that it is well known and described in the literature, such as Law & Brookmeyer (1992) and Kim (2003). When using midpoint imputation approach we reduce the problem of dealing with interval-censored data to exact and right-censored data. We have used the `mle2` function of the R package `bbmle` (Bolker & R Development Core Team, 2020) to maximize ℓ using the method of Nelder-Mead and the Conjugate Gradient considering the previous value of λ . This function returns a result in which none of the parameters is equal to zero. For this reason, we have used a threshold (converge

tolerance = 0.0001) to force the values below this threshold to be equal to zero and with the remaining variables, whose parameter is not set to 0, the algorithm is started again using an updated value of λ . The algorithm ends when the parameter estimates of all the remaining variables lie above the threshold.

Approach B

This proposal use the `iregnet` package to compare with the previous results. To use this package we have different ways to proceed. In each case, we continue using `cv.glmnet` to select an approximation to the hyperparameter λ , since `iregnet` does not have implemented the function `cv.iregnet` for the Weibull distribution, which is our case. First, we fitted the AFT model using interval-censored times to viral rebound. Second, we splitted the full set of mRNAs in five randomly assigned disjoint subsets, and we selected the predictors in each case. Finally, with all these predictors we adjust a final AFT model. Third, we have also tried the previous two analyses but considering exact data by using the midpoint imputation point.

The application of each approach to the DCV2 clinical trial can be seen in the next section. The R code with our algorithm can be found in <http://doi.org/10.5281/zenodo.4678278>.

5.6 DCV2 dataset

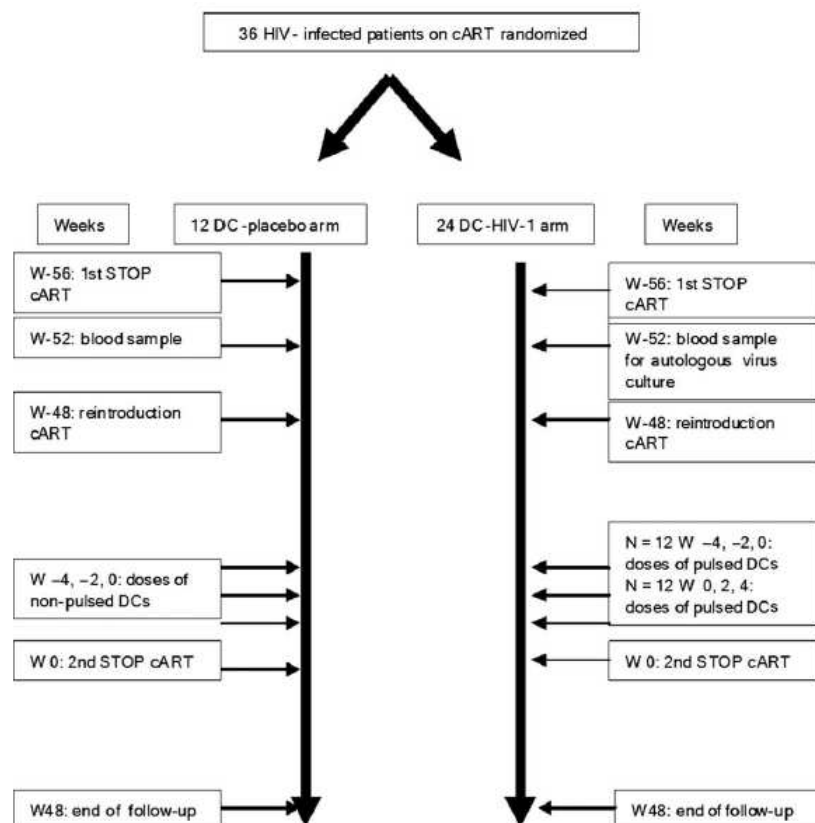


Figure 5.2: Flowchart of patients in the DCV2 trial. Source: [García et al. \(2013\)](#).

The DCV2 study was described by [García et al. \(2013\)](#). Combination antiretroviral therapy (cART) greatly improves survival and quality of life of HIV-1-infected patients; however, cART must be continued indefinitely to prevent viral rebound and associated disease progression. Inducing HIV-1 specific immune responses with a therapeutic immunization has been proposed to control viral replication after discontinuation of cART as an alternative to “cART for life”. The therapeutic vaccine use autologous monocyte-derived dendritic cells (MD-DCs) pulsed with autologous heat-inactivated whole HIV.

5.6.1 Description of the design

Thirty-six antiretroviral-treated chronic HIV-1 infected patients were randomized to receive three immunizations. One patient in the DC-control group was excluded from the analysis because of consent withdrawal before receiving any immunization. The 35 patients were followed up to 48 weeks after the first immunization. Week 0 was considered the day of second interruption of cART (2nd stop). Group 1 received immunizations at weeks -4 , -2 , and 0 (12 patients) and Group 2 at weeks 0 , 2 , and 4 (12 patients). These two different schedules were selected to assess whether cART could have any influence

in the response to immunizations. DC-control group patients (group 3) received placebo injections at weeks -4 , -2 , and 0 (Figure 5.2). Moreover, mRNA and miRNA were assessed at week -1 and week 3 , as shown in Figure 5.3.

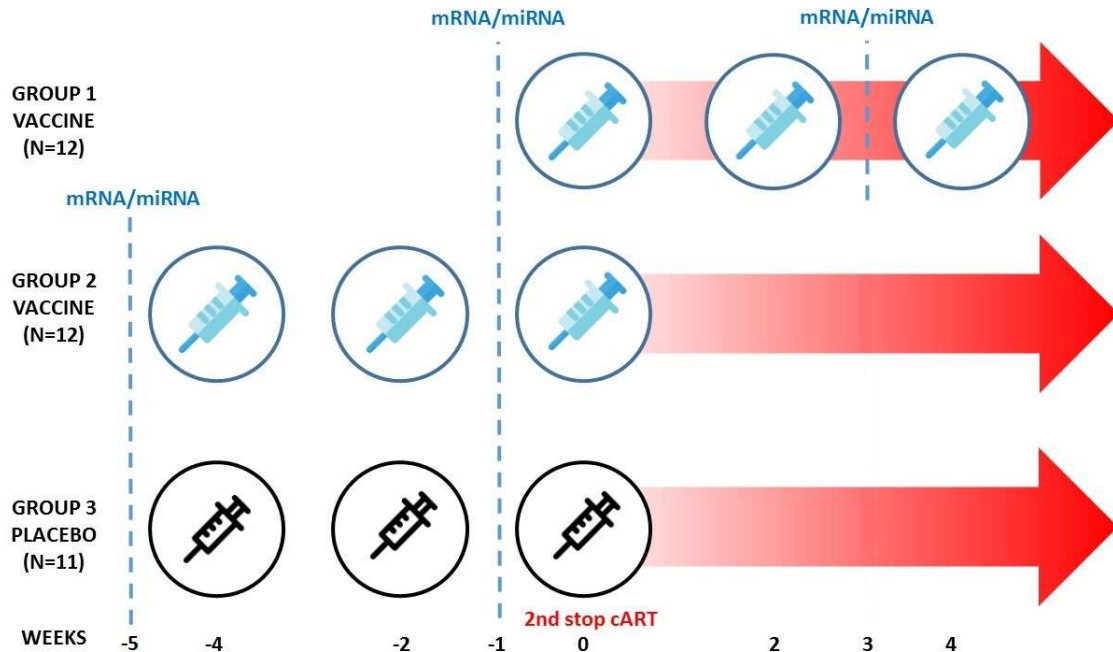


Figure 5.3: DCV2 clinical trial design. Vertical dashed lines indicates mRNA and miRNAs assessments. Red arrows show the time of cART initiation.

5.6.2 Baseline clinical parameters

Chronically HIV-1 infected patients on cART with baseline $CD4^+$ T lymphocytes above 450 cells/mm^3 , nadir $CD4^+$ T cell count above 350 cells/mm^3 , and undetectable viral load ($VL < 37 \text{ copies/ml}$) were enrolled. Patients had been on cART for at least the last 2 years before enrollment.

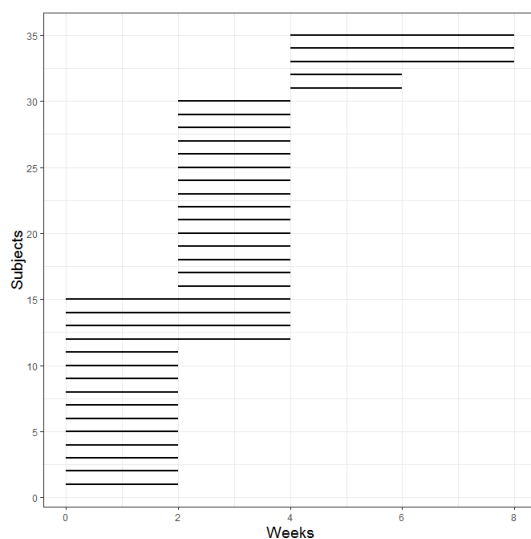
From Table 5.2 we observe most of the patients are males, and that the median age in each group is around 40 years. Regarding the consumption of tobacco, group 1 and group 2 present 7 smokers each and only 3 smokers for group 3. The median of HIV duration, that is, the time since the patient has been diagnosed with HIV (in years), is lower for group 2 than group 1 and 3, this is 8.5 years. Covariates baseline pre-cART VL and VL at stop 1 are similar in each of the three groups.

5.6.3 Viral rebound of HIV-infected patients in DCV2 trial

In this clinical trial, the time (weeks) to viral rebound is interval censored (see Chapter 4 for more information), as is presented in Figure 5.4. All the R code for this section is available at <http://doi.org/10.5281/zenodo.4678278>.

Table 5.2: Description of clinical covariates of the DCV2 study.

	Group 1 Vaccine	Group 2 Vaccine	Group 3 Placebo
n	12	12	11
Gender (male)	11	8	8
Age, median (IQR)	43 (40.75-49.75)	40.5 (38.5-46.5)	40 (34.5-45)
Tobacco (yes)	7	7	3
Risk HIV MSM	8	7	7
HIV duration, median (IQR)	12 (7.75-13.25)	8.5 (7-11.25)	13 (6.5-15.50)
Baseline pre-cART VL [log10 mean (SE) copies/ml]	4.92 (0.49)	4.83 (0.56)	4.64 (0.57)
VL stop 1 [log10 mean (SE) copies/ml]	4.91 (0.61)	5 (0.55)	4.81 (0.72)

**Figure 5.4:** Lengths of the ordered interval-censored times (weeks) until viral rebound of the DCV2 dataset.

Notice that the average length of the censoring intervals is 2.4 weeks. We are as well interested in the behaviour of these intervals per each group of intervention, according to the DCV2 trial, and this can be seen in Figure 5.5

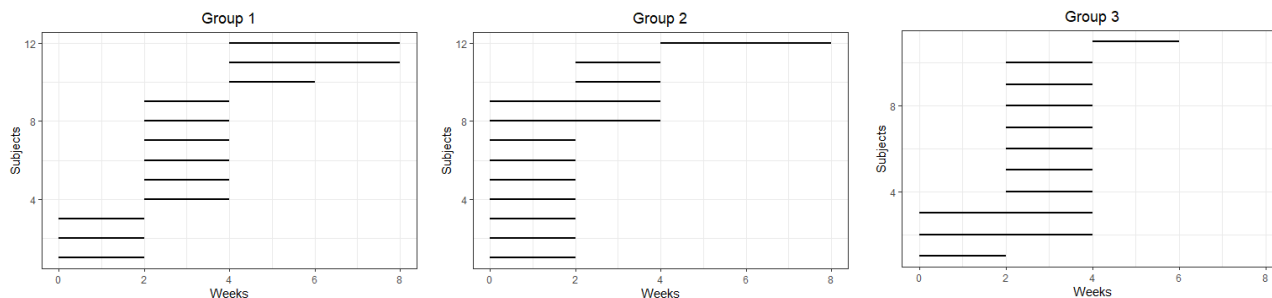


Figure 5.5: Lengths of the time (weeks) until viral rebound per intervention group.

We have applied the Fleming-Harrington class of test to check if the distribution of the time to viral rebound depends on the vaccine scheme. We have used the `FHtest` R package (Oller & Langohr, 2017), specifically the `FHtestics` function, which performs a test for interval-censored data based on the value of the score function. It uses the $G_{p,\lambda}$ family of statistics (being $\lambda = 0$) for testing the differences of two or more survival curves. We have found there is no statistical significant difference between them ($p = 0.399$). We graphically confirm this using the `interval` R package (see Figure 5.6).

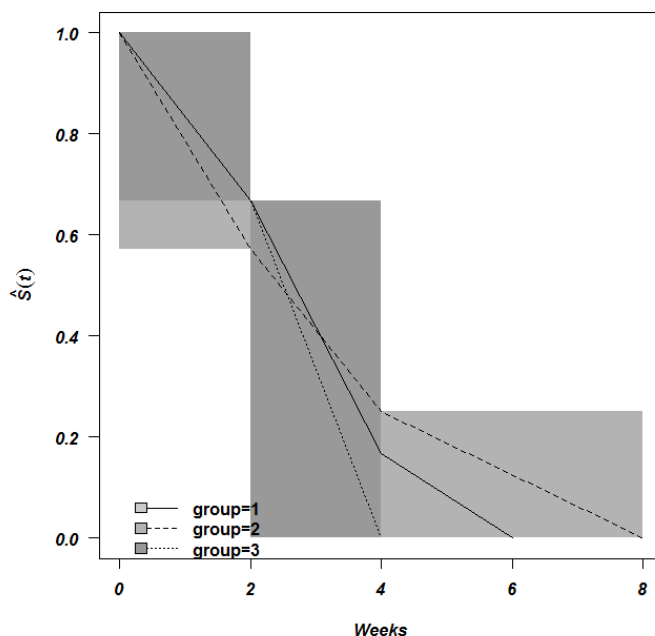


Figure 5.6: Turnbull's estimations of survival functions of times to viral rebound.

5.6.4 Time to viral rebound analysis using midpoint imputation

We present the results of the analysis of the elastic-net approach with the proportional hazards model (See Section 5.4). We have used midpoint imputation for the interval-censored time to viral rebound and the analyses were performed with the R package called `glmnet`. We considered leave-one-out cross

validation (LOOCV) to obtain the optimum value for λ , in this case is the λ_{\min} , i.e., the value of λ that gives minimum mean cross-validated error. The LOOCV process can be represented in the Figure 5.7 for the three main groups.

The graphics include the cross-validation curve (red dotted line), and upper and lower standard deviation curves along the λ sequence. In the upper part of the graph the number of variables (in this case mRNAs) selected are shown. Two λ values are indicated by the vertical dotted lines. The left vertical line in our plot shows where the CV-error curve hits its minimum as well as the number of variables selected according to that λ value. The right vertical line shows the most regularized model with CV-error within 1 standard deviation of the minimum. We decide to work and extract the λ_{\min} as we mentioned before.

The R code can be accessed at <http://doi.org/10.5281/zenodo.4678278>. The results can be seen in the Table 5.3, which contains the λ_{\min} , the symbol of the selected mRNAs, and its corresponding coefficient, for each group of treatment.

Table 5.3: Coefficients of the selected mRNAs for each group of treatment and overall.

Symbol	Group 1	Group 2	Group 3	All subjects
	$\alpha = 0.5, \lambda_{\min} = 1.10$	$\alpha = 0.5, \lambda_{\min} = 0.81$	$\alpha = 0.5, \lambda_{\min} = 0.86$	$\alpha = 0.5, \lambda_{\min} = 0.67$
CPA3		0.2193		
FECH		0.1049		
TNFRSF13B		-0.0598		
NUDT7		0.0382		
GSTM2		0.0008		
HLA-DPB2		0.0073		
MCL1		-0.1652		
MYCN		0.2811		
PCDH9		0.2643		
BAD		0.0872		
TDRD9			-0.0352	
NUDT17			0.0837	
LRP5			0.3296	
NT5C3A			0.0612	
RTN1			-0.0690	
CD16			-0.0264	
LOC100505915				0.0104
CENPBD1P1				-0.1894

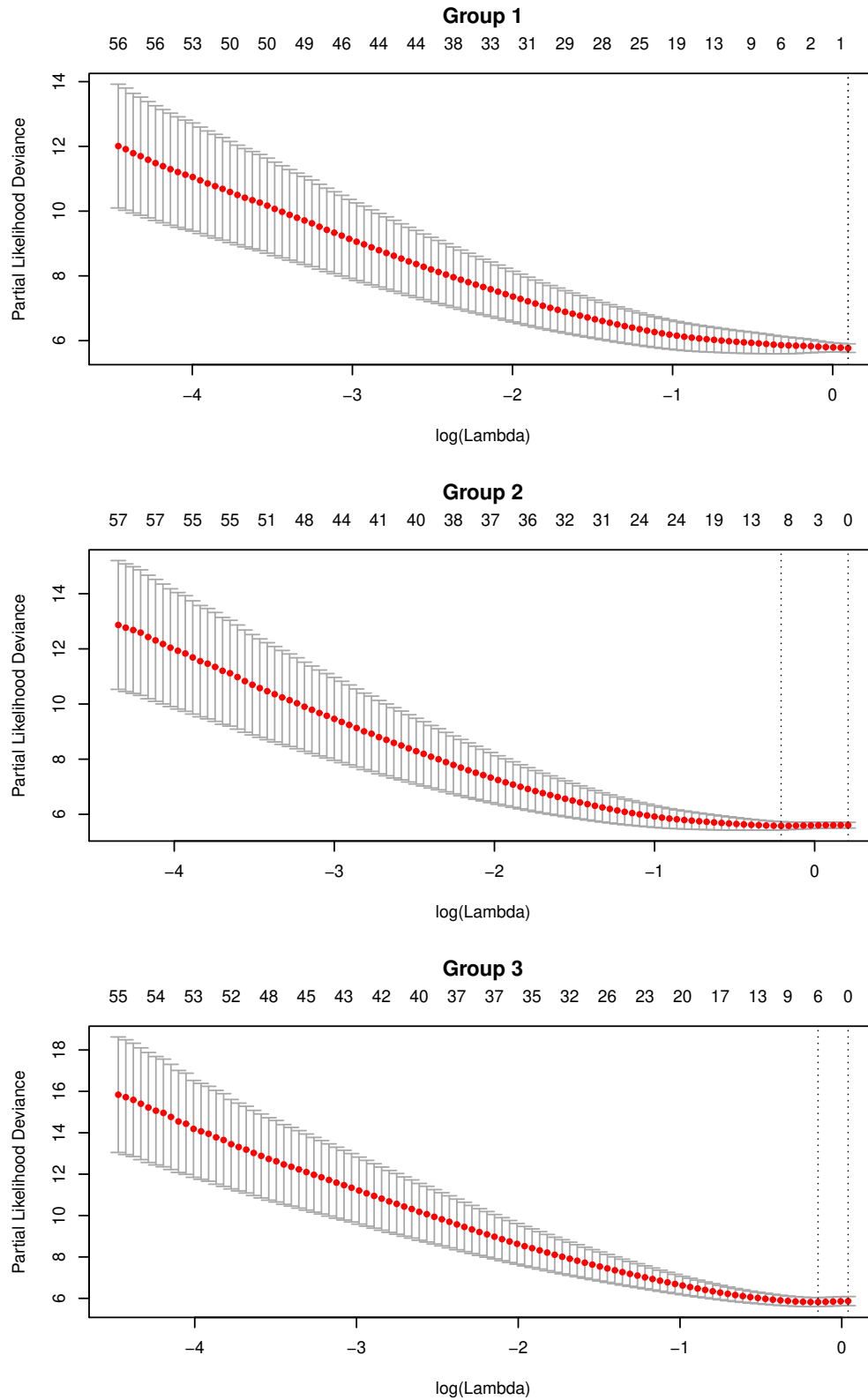


Figure 5.7: Cross validated error plots for the three groups of DCV2. Left dashed vertical line equals the minimum error, whereas the right dashed vertical line shows the cross-validated error within 1 standard error of the minimum.

When a coefficient is higher than zero it means that there is a higher risk of viral rebound. If the coefficient is lower than zero there is a lower risk of viral rebound. No mRNA was selected as predictor of viral rebound for group 1 (vaccine, early stop of treatment), for group 2 (vaccine, late stop of treatment) 10 mRNAs were selected, and for the group 3 (placebo) 6 mRNAs were selected. Moreover, as we can observe none mRNA was selected in more than one group. The official full name of every mRNA can be seen in the Table 5.4.

In the previous section we showed that there is no statistically significant difference among survival curves per group using the Fleming-Harrington class of test. This is why we are analyzing all the DCV2 trial participants together and running the `glmnet` package using the midpoint for the interval-censored times to viral rebound and the elastic net approach.

As illustrated in the Figure 5.8, the $\lambda_{\min} = 0.6711$ only allow us to select 2 mRNAs out of more than five thousand features, as we mentioned before, the selection is shown by the left vertical line in the graph. The selected mRNAs for all the subjects are (see Table 5.3): LOC100505915 ($\beta_1=0.0104>0$) and CENPBD1P1 ($\beta_2=-0.1894$), indicating higher and lower risk of viral rebound per every unit increase in each mRNA, respectively. The LOC100505915 mRNA was not previously related to the HIV, according to the literature, meanwhile the CENPBD1P1 mRNA was previously related to the K111 provirus, linked to HIV infection (Contreras-Galindo et al., 2013). Moreover, none of these mRNAs were previously selected by group.

Our findings suggest that these selected mRNAs (0, 10, and 6 per each treatment group and 2 for the overall set of participants) could correspond to potential biomarkers for HIV viral rebound and may be considered to determine the effectiveness of the dendritic cell-based vaccine.

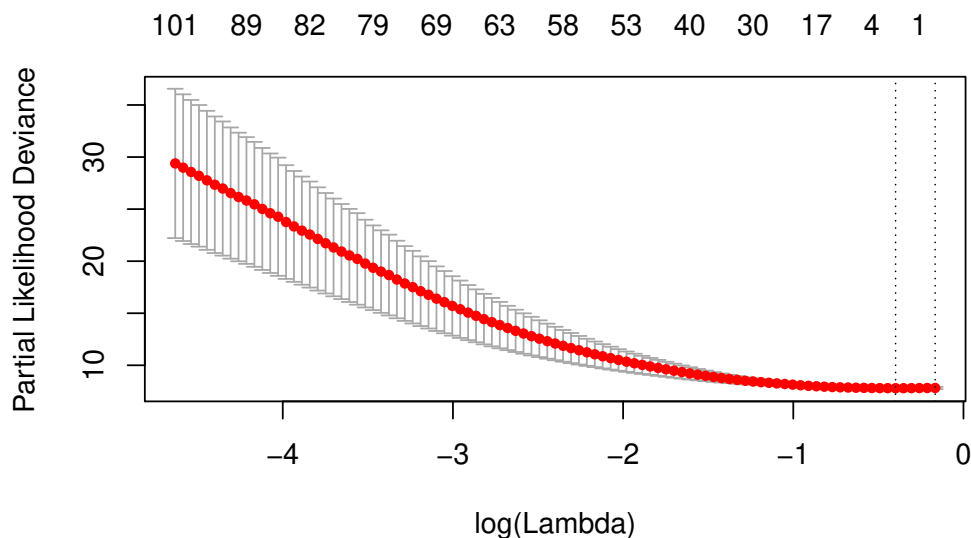


Figure 5.8: Cross validated error plot for the overall DCV2 set.

Table 5.4: Official full name of the selected mRNAs in each group of treatment.

Symbol	Description	Category
Selected for Group 2		
CPA3	Carboxypeptidase A3	Protein coding
FECH	Ferrochelatase	Protein coding
TNFRSF13B	TNF receptor superfamily member 13B	Protein coding
NUDT7	Nudix hydrolase 7	Protein coding
GSTM2	Glutathione S-transferase Mu2	Protein coding
HLA-DPB2	Major histocompatibility complex, class II, DP beta 2	Pseudogene
MCL1	MCL1, BCL2 family apoptosis regulator	Protein coding
MYCN	MYCN proto-oncogene BHLH transcription factor	Protein coding
PCDH9	Protocadherin 9	Protein coding
BAD	BCL2 associated agonist of cell death	Protein coding
Selected for Group 3		
TDRD9	Tudor domain containing 9	Protein coding
NUDT17	Nudix hydrolase 17	Protein coding
LRP5	LDL receptor related protein 5	Protein coding
NT5C3A	5'-nucleotidase, cytosolic IIIA	Protein coding
RTN1	Reticulon 1	Protein coding
CD16	Cell division cycle 16	Protein coding
Selected for all subjects		
LOC100505915	Uncharacterized LOC100505915	Protein coding
CENPBD1P1	CENPB DNA-binding domains containing 1 pseudogene 1	Protein coding

5.6.5 Fit of the AFT model by means of ad-hoc methods

We have maximize the log-likelihood function corresponding to the AFT model with Weibull distribution for interval-censored times to viral rebound as:

$$\ell = \frac{1}{n} \sum_{i=1}^n \log \left(\exp \left(-e^{\frac{1}{\sigma}(\log(t_{Li}) - \mu - \beta z)} \right) - \exp \left(-e^{\frac{1}{\sigma}(\log(t_{Ri}) - \mu - \beta z)} \right) \right) - \lambda P_{\alpha}(\beta), \quad (5.18)$$

where

$$P_{\alpha}(\beta) = \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1}{2}(1 - \alpha) \sum_{j=1}^p \beta_j^2 \right).$$

In our case, $n = 35$, t_{Li} and T_{Ri} , $i = 1, \dots, 35$ correspond to the lower and upper boundary of the interval times to viral rebound. The order of the z matrix is 35×5047 and it contains the mRNAs information for every subject. We have fixed $\alpha = 0.5$ but α can be defined as any value in $[0, 1]$. The parameters to be estimated correspond to λ , μ , σ and β_j , $j = 1, \dots, 5047$.

We ran the methods explained in Section 5.5.4 for the entire group of participants in the DCV2 trial, since, as we explained in Section 5.6.3, we did not find statistically significant differences between groups.

Approach A

The application of the first approach, which consider the use of the `mle2` function with the Nelder-Mead method, has selected the variables shown in Table 5.5 and obtained the parameter estimates shown therein. The full official name of each selected mRNA can be found in the Table 5.6.

Table 5.5: Coefficients of the selected mRNAs for each method (Approach A).

Symbol	Coefficient		
	Our algorithm	Without penalization	Univariate
PPP1R9A	-0.3672	-0.5214	
LOC100509457	0.0416	0.0447	
IL21R	-0.1625	-0.1249	
CYP1B1	0.1354	0.1946	
DUSP4	0.1120	0.1764	
CYGB			-0.532
TGIF2			-1.244
RHOF			-0.274
CASP1			0.339
RBM38			0.555

As shown in the Table 5.5 five mRNAs were selected by our implemented algorithm out of more than five thousand. As we mentioned before, a positive sign means that risk of viral rebound is higher, and thus the prognosis worse, for subjects with higher values of that mRNA. Thus, from Table 5.5, PPP1R9A and IL21R are associated with lower risk of viral rebound, whereas LOC100509457, CYP1B1, DUSP4 are associated with higher risk of viral rebound. It is important to notice that the value of each coefficient when compared with the corresponding value of the model fit that included the five mRNAs selected using the `survreg` function (without penalization) keeps the same sign and it has lower absolute value. The values obtained using the elastic-net penalization are closer to zero, as expected.

Table 5.6: Official full name of the selected mRNAs with each method (Approach A).

Symbol	Description
PPP1R9A	Protein phosphatase 1 regulatory subunit 9A
LOC100509457	HLA class II histocompatibility antigen, DQ alpha 1 chain-like
IL21R	Interleukin 21 receptor
CYP1B1	Cytochrome P450 family 1 subfamily B member 1
DUSP4	Dual specificity phosphatase 4
CYGB	Cytoglobin
TGIF2	TGFB induced factor homeobox 2
RHOF	Ras homolog family member F, filopodia associated
CASP1	Caspase 1
RBM38	RNA binding motif protein 38

We have also compared our method with the Newton-Raphson algorithm performed by the `survreg` function considering the previously selected mRNAs by our implemented algorithm and with the corresponding univariate models. The `survreg` function does not consider the penalization, that is, the parameters are obtained without using any penalization. When we run all the univariate models, the mRNAs selected corresponding to the most statistically significant models were completely different.

The main disadvantage of our algorithm is that it is computationally demanding, it takes approximately 2 weeks to completely achieve the results (for Intel(R) Pentium(R) CPU 3825U 1.90 GHz, RAM: 8 GB, and operating system of 64 bits). Regarding the biological meaning of our results, up to our knowledge, LOC100509457 was not previously related as a regulatory mRNA for the HIV. The other four selected mRNAs were related: PPP1R9A is a neuronal protein that was selected as a biomarker of cognitive impairment in HIV infection and Alzheimer's disease (Pulliam et al., 2019), the cellular immune responses and upregulation of IL21R was different in HIV patients when testing the response to a specific H1N1 vaccine (Pallikkuth et al., 2011), CYP1B1 has been used to develop new interventions for HIV positive smokers (Rao & Kumar, 2015), while some authors mentioned that it is possible that a high response of this mRNA is an early indicator of chronic obstructive pulmonary disease in HIV positive smokers (Logue et al., 2019), and DUSP4 was related to the CD4 T cells (Bignon et al., 2015). However, we cannot claim that these mRNAs are the best predictors of time to viral rebound, mainly because we could not confirm these results with other methods or data sets.

Approach B

When using the `iregnet` package for the entire set of mRNAs to explain the interval-censoring time to viral rebound, the coordinate descent algorithm that `iregnet` uses does not converge. The program instead says that we need to add more data. In this way, `iregnet` package cannot deal with interval-censored data in the setting when $p \gg n$. In our case, considering $n = 35$, the `iregnet` package converges until approximately 1000 covariates.

We have also replaced the interval-censored times by the midpoint imputation in each subject. We have ran the `iregnet` function and select the λ value closest to the one given by the `cv.glmnet` function. In this case the function select 2 mRNAs: CLUHP3 and ITGB2-AS1. The coefficients are shown in Table 5.7 and the official fullname in Table 5.8.

As considering the entire dataset of mRNAs the function `iregnet` does not converge, we decide approaching the problem by dividing the set into 5 disjoint sets of size 1 thousand mRNAs each (approximately) and executing the previous procedure on each subset. The model in each case selected 4, 4, 0, 106, and 0 mRNAs. Following, the sum of mRNAs selected (104 mRNAs) as predictors in each case were considered for a final model. Twelve predictors emerge from this final model (see Table 5.7). Unfortunately, in the scientific literature, we have not found a relationship between any of them with HIV, which leads us to intuitively think that by making this random allocation of the set of mRNAs, we are disassembling pathways between them and affecting their biological correlation structure.

Table 5.7: Coefficients of the selected mRNAs for each method (Approach B).

Symbol	Coefficient	
	iregnet with midpoint	iregnet 5 groups interval-censored
CLUHP3	0.0307	
ITGB2-AS1	0.0028	
FTL		-0.0015
ZNF658		0.0005
EML4		-0.0010
PHPT1		-0.0002
LRP12		0.0008
HMOX1		-0.0004
SIPA1L2		0.0014
UBE2L3		-0.0007
NPL		0.0011
TNIP1		0.0022
SLC16A10		-0.0008
COX8A		-0.0005

Table 5.8: Official full name of the selected mRNAs with each method (Approach B).

Symbol	Description
CLUHP3	Clustered mitochondria homolog pseudogene 3
ITGB2-AS1	ITGB2 antisense RNA 1
FTL	Ferritin light chain
ZNF658	Zinc finger protein 658
EML4	EMAP like 4
PHPT1	Phosphohistidine phosphatase 1
LRP12	LDL receptor related protein 12
HMOX1	Heme oxygenase 1
SIPA1L2	Signal induced proliferation associated 1 like 2
UBE2L3	Ubiquitin conjugating enzyme E2 L3
NPL	N-acetylneuraminase pyruvate lyase
TNIP1	TNFAIP3 interacting protein 1
SLC16A10	Solute carrier family 16 member 10
COX8A	Cytochrome c oxidase subunit 8A

To corroborate the previous results we have decided to consider exactly the same five previous subsets, but instead of working with interval-censored times to viral rebound, we have worked with the midpoint imputation. We have expected similar results but each group selected 171, 203, 175, 159, and 171 mRNAs, respectively. For the final model, 70 predictors have come out (the complete list is on the Appendix B). Furthermore, none of these predictors is repeated in each case considered in this approach. All of the above leads us to think that we are facing an identifiability problem, because $p \gg n$, we have many degrees of freedom and many possible combinations to maximize our function.

Other methods and further research are discussed in the next section.

5.7 Discussion

The availability of multiomic data sets has increased recently and this trend is expected to continue. Modern omic data sets can have a large number of individuals, be high dimensional, i.e., each subject can have information on hundreds of thousands of variables, and have a multilayer structure, this means data may involve clinical information, demographics, lifestyle, and multiple omics.

Our main idea in this chapter was to develop a model that uses a specific type of omics, the mRNAs, as possible biomarkers that help to predict the time until viral rebound of HIV-infected patients. This idea was motivated by the DCV2 clinical trial, which studies mRNAs and survival times. We have proposed the use of the elastic-net penalization since this technique allowed us to work with correlated high-dimensional data, which is the case of the mRNAs, that is, there is a structure of correlation among them that can be called “group effect”. The group effect means that some of them work as a group when regulating some cell activities in the human body. Another important aspect of elastic-net penalization is that it can perform well when the number of individuals n is smaller than the number of predictors p . In our particular case, we have a huge number of possible predictors (5047 mRNAs) and a low number of subjects ($n = 35$). This is a critical point in our application, because we can have an identifiability problem, since $p \gg n$.

Considering the PH model, we have shown an application of the elastic-net penalization using the midpoint of the interval during which the viral rebound occurred. When using the midpoint imputation we reduce the problem of dealing with interval-censored data to exact and right-censored data. Moreover, the use of midpoint imputation is straightforward because it is well known and described in the literature. To study the maximization of the log-likelihood function of the PH model considering an elastic-net penalization and interval-censored data as in Equation (5.10) is of further research. The use of all the complete information, by using the interval-censored times to viral rebound, could lead to better results.

Regarding the AFT model, we have described and implemented an ad-hoc algorithm to deal with interval-censored data and elastic-net penalization. The main advantage of our approach is that it deals with the complete high-dimensional dataset. Using the complete dataset, we considered the “grouping effect” or the correlation structure among the different mRNAs due to its biological structure. The major disadvantage of our method, with this particular dataset, which considers 35 patients and 5047 mRNAs, is that the Nelder-Mead algorithm took almost 2 weeks (for Intel(R) Pentium(R) CPU 3825U 1.90 GHz, RAM: 8 GB, and operating system of 64 bits) to select the significant mRNAs. The algorithm works better with lower numbers of predictors, for example, when we considered a random subset of 300 mRNAs, the results were obtained in less than 3 minutes. We have also used the Conjugate Gradient optimization method included in the `mle2` function using two options. In the first option, we have considered the analytical expression of the gradient as described in (5.17)

and, in the second option, we did not specify the expression for the gradient. The `mle2` function uses finite differences to find the gradient, if we don't include this argument. In each case, with the analytical expression of the gradient or by finite differences, the algorithm does not converge. When using the `iregnet` package, considering the whole dataset and the Weibull distribution, the coordinate descent algorithm implemented, did not converge neither.

We expected that the Cox model using the midpoint imputation (fitted using `glmnet` package) and the AFT model using the Weibull distribution would identify the same variables. However, with the AFT model, we identified 5 mRNAs (PPP1R9, LOC100509457, IL21R, CYP1B1, and DUSP4) using the approach A and the Nelder-Mead optimization method, and using the midpoint imputation with the PH model we identified 2 mRNAs (LOC100505915 and CENPBD1P1). In both cases we did not find any match between these two sets of mRNAs selected. This can probably be explained by different reasons: the first is the possible identifiability problem that we described previously; another possibility is that the time to viral rebound may not follow a Weibull distribution; finally, in the PH model we used a midpoint imputation approach and with the AFT model we considered the interval-censored times to viral rebound. Additionally, an important limitation of this approach that could be another reason to explain this difference, is the use of an ad-hoc method in the case of the AFT model. In this ad-hoc approach we have used an arbitrary threshold for the parameters to be equal to zero, we have obtained the value of the optimal λ in each iteration using LOOCV with the midpoint imputation.

Moreover, we have tried estimating the parameters by splitting the dataset into 5 subsets by using, for each subset, the `iregnet` package and the AFT model, considering the Weibull distribution, and two different scenarios: considering the interval-censored times to viral rebound and using the midpoint of the intervals. The selected predictors in each case were different and also different to the previously selected in the case of the use of our implemented algorithm, or by using the `glmnet` package. In this setting, the problem could be that we are disassembling the biological structure of the data without taking into account the real "grouping effect".

As a future research, we would like to continue working in improving the algorithm in R and its time of performance. Besides, we are interested in adapting other maximization methods such as the coordinate descent method, which is implemented in the `glmnet` and `iregnet` packages. We are also interested in using other software such as AMPL (Fourer et al., 2003) together with the R package `rneos` (Pfaff & Pfaff, 2020) which enables solving the optimization problems using the NEOS solvers (Server, 2016) and retrieve the results within R. Another possible alternative we are exploring is the use of Matlab for optimization and a correct solver such as `fmincon` or `fminunc`.

Summing up, a first step has been done to tackle the complex problem of fitting survival models with plenty of variables and few observations, but more research is needed to find a complete satisfactory solution to this problem.


MULTIPLE IMPUTATION APPROACH FOR INTERVAL-CENSORED TIME TO HIV RNA VIRAL REBOUND WITHIN A MIXED EFFECTS COX MODEL

We present a method to fit a mixed effects Cox model with interval-censored data. Our proposal is based on a multiple imputation approach that uses the truncated Weibull distribution to replace the interval-censored data by imputed survival times and then uses established mixed effects Cox methods for right-censored data. Interval-censored data were encountered in a database corresponding to a recompilation of retrospective data from eight Analytical Treatment Interruption (ATI) studies in 158 HIV-positive combination antiretroviral treatment (cART)-suppressed individuals. The main variable of interest is the time to viral rebound, which is defined as the increase of serum viral load to detectable levels in a patient with previously undetectable viral load, as a consequence of the interruption of cART. Another aspect of interest of the analysis is to consider the fact that the data come from different studies based on different grounds and that we have several assessments on the same patient. In order to handle this extra variability, we frame the problem into a mixed effects Cox model that considers a random intercept per subject as well as correlated random intercept and slope for pre-cART viral load per study. Our procedure has been implemented in R using two packages: `truncdist` and `coxme`, and can be applied to any data set that presents both interval-censored survival times and a grouped data structure that could be treated as a random effect in a regression model. The properties of the parameter estimators obtained with our proposed method are addressed through a simulation study.

The contents of this chapter have been published in:

Alarcón-Soto, Y., Langohr, K., Fehér, C., García, F., & Gómez, G. (2019). Multiple imputation approach for interval-censored time to HIV RNA viral rebound within a mixed effects Cox model. *Biometrical Journal*, 61, 299 - 318.

This chapter is based on the above paper after excluding part of the Notation and Preliminaries (section 6.2 because its content has been detailed in Chapter 4. The notation used here is coherent with the rest of this thesis, and differs slightly from the one used in the paper.

Received: 30 November 2017	Revised: 21 September 2018	Accepted: 24 September 2018
DOI: 10.1002/bimj.201700291		
RESEARCH PAPER	Biometrical Journal →	
Multiple imputation approach for interval-censored time to HIV RNA viral rebound within a mixed effects Cox model		
Yovaninna Alarcón-Soto ¹  Klaus Langohr ¹ Csaba Fehér ^{2,3} Felipe García ^{2,4} Guadalupe Gómez ¹		

6.1 Introduction

Most statistical methods developed for the analysis of survival data assume that the event that defines time origin is known and allow the event of interest \mathcal{E} that determines failure and, hence, the survival time, to be right-censored. In many situations, however, the event of interest \mathcal{E} cannot be observed and it is only known to have occurred within two random times, say L and R . In this set-up, we say that the time to \mathcal{E} , T , is interval-censored.

Interval-censored data often arises in medical or health studies that entail periodic follow-ups, and many clinical trials and longitudinal studies fall into this category (Sun, 2007). In such situations, interval-censored data may arise in several ways. For instance, an Human Immunodeficiency Virus (HIV)-infected patient is examined weekly to check if his/her viral load exceeds a certain threshold. Suppose that in a first measurement, at time t_1 , it does not exceed the threshold and it does in a second measurement at time t_2 . Hence, all that is known is that the viral load exceeded the threshold within the interval $(t_1, t_2]$, but the exact time of viral rebound is unknown.

We have encountered interval-censored data while studying the immunological response of HIV positive patients by means of different parameters of viral rebound dynamics within eight different ATI studies (Leal et al., 2017).

We start introducing the main concepts related to HIV-ATI studies to facilitate reading and understanding of the study that motivates this paper. The HIV is a virus that infects cells of the immune system, destroying or impairing their function. Infection with the virus results in a progressive deterioration of the immune system, leading to immune deficiency, and the immune system is considered deficient if it is no longer able to fulfil its role of fighting infection and disease (World Health Organization, 2017). HIV-infected patients are generally treated with a combination of antiretroviral drugs known as combination Antiretroviral Therapy (cART) in order to maximally suppress the HIV virus and stop the progression to the Acquired Immune Deficiency Syndrome (AIDS).

The ATI is a controlled interruption of the cART in HIV-positive patients and appears in different interventional or observational studies in this field. The objective of this interruption is the evaluation of the immunological response of the patients, described by different parameters related to the viral rebound dynamics (Treasure et al., 2016). In the case of our data set, depending on the study, a single ATI episode or several episodes are studied. Studies including more than one ATI episode are based on the “autovaccination” theory, according to which repeated encounters of the immune system with the antigenic stimulus (the virus) will be able to increase the specific immune response, leading to a major control of the posterior viral load (Graziani & Angel, 2015). The endpoint of interest in these studies is the time until viral rebound, which is defined as the first time that an HIV-infected patient, with previously undetectable serum viral load, surpasses the threshold of 20 copies/mL (or 1.30 in \log_{10} scale). Viral load values lower than 20 copies/mL were considered undetectable when calculating time to rebound.

Previous to the start of cART, most of the patients in our data set had a high viral load, referred as pre-cART viral load. As soon as they start cART, the viral load drops down to undetectable levels. In our case, at the beginning of the first ATI episode (week 0), all patients presented undetectable levels of viral load. From that moment, the viral load was measured once every week and the corresponding ATI episode was stopped as soon as the viral load was detectable again; see Figure 6.1.

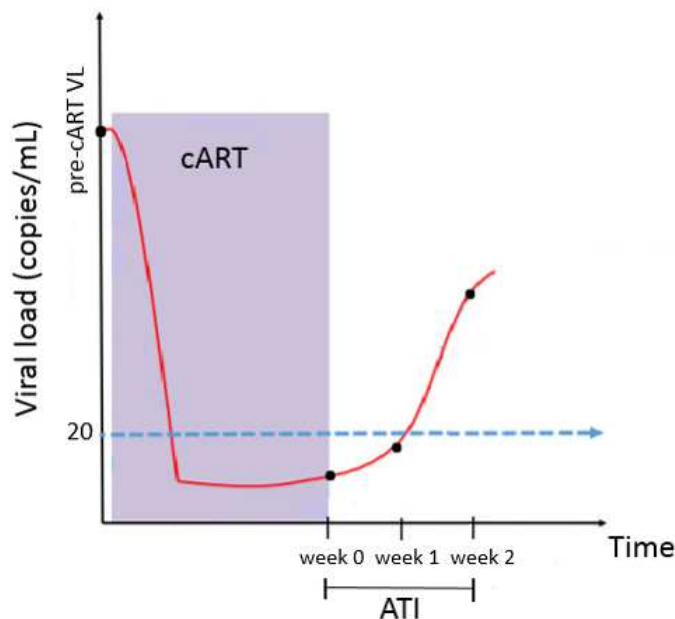


Figure 6.1: Viral load dynamics during cART and the first ATI episode. The last measure before starting the cART is denoted by pre-cART viral load (VL). In this case, the ATI stopped at week 2 because of detection of viral rebound.

To study viral rebound dynamics, we were interested in characterizing the percentiles of the time to viral load rebound based on the corresponding distribution function, in establishing differences between male and female patients, in determining the importance of pre-cART VL in suffering a viral rebound, and in assessing relative risks by means of hazard ratios based on a Cox proportional hazards model (Cox, 1972). To provide rigorous answers to all these issues, several nonparametric and parametric methods for interval-censored data have been developed and the literature is already abundant, see Gómez et al. (2009) for a thorough overview. Finkelstein (1986) was among the first authors to adapt the Cox model to interval-censored data. Nowadays, parameter estimation in the presence of interval-censored data can be carried out in R (R Core Team, 2020) by means of the `icenReg` package (Anderson-Bergman, 2017).

However, the ATI data set not only has interval-censored times to viral rebound but these are also quite heterogeneous because they correspond to eight different studies that are based on different grounds. Moreover, several patients underwent more than one treatment interruption. A proper data analysis should take this into account. Since the variability of measurements in different individuals

is usually larger than the variability between measurements in the same individual (Bland & Altman, 1994), a possible way to account for this extra variability could be to frame the problem into a mixed effects Cox model where a grouping variable corresponding to the subject would be added as random effect.

In the case of data coming from different studies, we might have two different scenarios. If there is homogeneity of the observed effect for some particular covariate, a single measure would be adequate to describe the general results. If, on the contrary, heterogeneity of effects is found, we should add a random effect of this covariate per study to capture this heterogeneity and carefully interpret the results. Our ATI data set falls in the second situation because of the different treatments used (combination of different drugs or different vaccines) and the different recruitment criteria for patients (such as early or late stage of HIV infection, CD4 count, viral load threshold, among others). Therefore, it is very likely to have some degree of variation (heterogeneity) among these studies, and for this reason, a random intercept and slope for pre-cART VL per study is considered.

Some other reasons to consider the inclusion of random effects are similar to those considered by Yamaguchi et al. (2002) and Senn (1998), who in the analyses of multicenter trials used random effects to model the center's effect variability. As Senn discusses, if we are interested in making inferences about patients from a given study, the fixed effect approach leaves little alternative but to use the results from that study only. The random effects approach will allow us to combine information with the given study with information from all the studies in a way which is more appealing and useful.

In the context of survival analysis, a mixed (fixed and random) effects Cox model with right-censored data has been presented by T. M. Therneau & Grambsch (2000) and its fit is accomplished with the `coxme` package of the same author (T. M. Therneau, 2018b). However, to the best of our knowledge, a mixed effects Cox model with interval-censored data has not been studied. Hence, our objective consists of developing a mixed effects Cox model with interval-censored data in order to correctly model the heterogeneity in our data set attributable to the repeated measures per patient and the different inclusion criteria per study.

A natural solution to the difficulties of direct estimation based on interval-censored data is to use an algorithm based on treating the interval-censored observations as missing data and imputing values for them, thus creating right-censored and exact data (Bebchuk & Betensky, 2000). Interval-censored data are actually incomplete data, not missing data, because the observed interval provides some information about the variable of interest (Sun, 2007). Nevertheless, we can still treat the underlying, unobserved true interval-censored failure times as missing and replace them by imputed times conditional on the observed information. Using the methodology and software already developed for right-censored data (T. M. Therneau, 2018b), our proposal is based on a multiple imputation approach using the truncated Weibull distribution to replace the censoring intervals by imputed survival times. Our idea is similar to the one of Satten et al. (1998) in the case of the Cox model without random effects, who used imputation methods to replace the interval-censored survival times by imputed values. The authors propose the use of a parametric model for the baseline hazard in order to generate imputed

failure times; following, a rank-based procedure based on the imputed failure times is used to estimate the regression coefficients.

Multiple imputation is a statistical technique to handle missing data that takes advantage of the flexibility in modern computing. With it, each missing value is replaced by two or more imputed values in order to represent the uncertainty about which value to impute (Rubin, 2004). According to the method for “repeated imputation” inference, each of the simulated complete data sets is analysed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Multiple imputation methods for related censoring problems regarding HIV data have been developed by Muñoz et al. (1989) and Taylor et al. (1990). Dorey et al. (1993) applied multiple imputation to interval-censored data corresponding to threshold-crossing time in some trials. Threshold-crossing times, somehow similar to viral load rebound, are common in medicine when patients move to a new risk category after crossing a threshold on some prognostic variable and because patient’s examinations occur only periodically, the exact time of crossing the threshold is only known to fall within a specified interval.

The article is organized as follows: in Section 6.2, we present the relevant notation and preliminaries used throughout the paper. Following, in Section 6.3, we present our multiple imputation-based approach to fit a mixed effects Cox model in the presence of interval-censored data. In Section 6.4, the methodology is applied to the ATI data set, followed by the section dedicated to the study of the properties of the fixed parameter estimators under different settings via a simulation study (Section 6.5). Finally, the main findings of this work are summarized in Section 6.6 and, so far, unresolved topics are discussed. Information on the implementation of the parameter estimation in R is presented in Appendix A and more details on the main features of each study in ATI data set are given in Appendix C.

6.2 Notation and preliminaries

Let T be the time until the event of interest, \mathcal{E} , which in the ATI studies corresponds to viral rebound. T is a non-negative random variable whose distribution function at time t , $F(t) = P(T \leq t)$, corresponds to the cumulative probability of reaching viral rebound before time t . The hazard function defined by $\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t)$, represents the instantaneous risk of viral rebounding and it is the function on which the model we propose will be based on.

The observable data, based on a sample of n patients, consists, for the i th individual ($i = 1, \dots, n$), of the random intervals $(L_i, R_i]$ during which the viral rebound occurred and the vector of covariates $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ir})'$, including the study label, gender, and pre-cART viral load. Interval-censored data include right-censored times as a particular case with $R_i = \infty$, which in our data base corresponds to patients whose viral load has not rebounded by the end of the study.

The Cox proportional hazards model, as described in (4.2), can be enhanced through the incorporation of random effect terms to account for within-cluster homogeneity in outcomes. This model,

called mixed effects Cox model, was developed by [T. M. Therneau & Grambsch \(2000\)](#) and implemented for right-censored data. The hazard function of a mixed effects Cox model for an individual i is given by

$$\lambda_i(t; \mathbf{X}, \mathbf{Z}) = \lambda_0(\mathbf{t}) \exp(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}) \quad \text{with } \mathbf{b} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \quad (6.1)$$

where \mathbf{X}_i and \mathbf{Z}_i are the i th rows of the design matrices corresponding to the fixed and random effects, and $\boldsymbol{\beta}$ and \mathbf{b} are the vectors of the fixed and random effects coefficients, respectively. In addition, as above, λ_0 is the unspecified baseline hazard function and the distribution of the random effects is assumed to follow a Gaussian distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, which in turn depends on a vector of parameters $\boldsymbol{\theta}$.

The partial likelihood function corresponding to the mixed effects Cox model is similar to that of the partial likelihood of the standard Cox model given in (4.3) and the expression of its logarithm for any fixed values of $\boldsymbol{\beta}$ and \mathbf{b} , is

$$\text{LPL}(\boldsymbol{\beta}, \mathbf{b}) = \log\{\text{PL}(\boldsymbol{\beta}, \mathbf{b})\} = \sum_{i=1}^n \int_0^{\infty} \left[Y_i(\mathbf{t}) \eta_i - \log \left\{ \sum_j Y_j(\mathbf{t}) \exp(\eta_j) \right\} \right], \quad (6.2)$$

where $\eta_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}$ is the linear score for subject i and $Y_i(t)$ is an indicator variable that takes value 1, if individual i is at risk at time t and 0 otherwise; for further details, see [T. M. Therneau \(2018b\)](#).

A straightforward adaptation of the partial likelihood function (6.2) to allow for the presence of interval-censored data is not possible, because with such data, it is not feasible to identify the exact ranking of the failure times and, consequently, the indicator variables $Y_i(t)$ cannot be determined for all t . Our proposal to overcome that problem is to use multiple imputation to replace the interval-censored survival times by exact and right-censored imputed values. The details are presented in the next section.

6.3 Parameter estimation in the mixed effects Cox model

The basic idea of our proposal consists of replacing the censoring intervals $(L, R]$ that contain the unknown survival times by imputed values based on a truncated Weibull distribution. The mixed effects Cox model in (6.1) can then be fitted to the imputed exact and right-censored data. These steps will be repeated several times in order to account for the uncertainty of the imputation step, which is ignored in the case of single imputation. Our proposal can be summarized by the following three steps.

Step 1 Imputation of interval-censored survival times.

Censoring intervals are replaced by imputed times following two steps. First, an accelerated failure time model is fitted to the whole data set, considering the covariates of interest, and the maximum likelihood estimators of the corresponding parameters are derived. Second, for each individual's censoring interval, its corresponding truncated Weibull distribution is obtained, and a random value is generated from that distribution. In the case of right-censored observations,

no imputation is performed. Hence, the resulting data set consists of uncensored (imputed) and right-censored survival times.

Step 2 Fit of the mixed effects Cox model and analysis.

The mixed effects Cox model (6.1) described in the previous section is fitted with the data resulting from Step 1 and the parameter estimates of interest are obtained.

Step 3 Pooling the results.

Given a pre-specified integer M , Steps 1 and 2 are repeated M times. Following, the parameters of interest are estimated as the average values of the estimates obtained in each of the M repetitions of Step 2.

Below we explain in more detail each of these steps.

6.3.1 Imputation of interval-censored survival times

The first step of our proposal consists of replacing the censoring intervals $(L, R]$ by imputed values based on a truncated Weibull distribution, where the truncation is induced by the respective intervals of the individuals. The motivation to use the Weibull distribution is twofold: on one hand, it is a common parametric model for survival data due the simplicity of the survival function and the flexibility of the hazard function, which is either constant, monotonically increasing or monotonically decreasing. On the other hand, it shares the assumption of proportional hazards with the Cox model. For that reason, given the Cox model in (4.2) and assuming that the baseline hazard function follows a Weibull distribution with cumulative distribution function $G(t) = 1 - \exp(-\lambda t^\alpha)$, survival times can be generated using the following expression (Bender et al., 2005):

$$T = \left(-\frac{\log(U)}{\lambda \exp(\boldsymbol{\beta}' \mathbf{x})} \right)^{1/\alpha}, \quad (6.3)$$

where U follows a uniform distribution on the interval from 0 to 1.

An estimation of $\boldsymbol{\beta}$, λ , and α in (6.3) are obtained after fitting the following accelerated failure time model –equivalent to the Cox model under the Weibull assumption–:

$$Y = \log T = \mu + \boldsymbol{\gamma}' \mathbf{X} + \sigma W,$$

where W follows the extreme value distribution. Standard software is used to obtain maximum likelihood estimators of μ , $\boldsymbol{\gamma}$ and σ and, from these, maximum likelihood estimators of the parameters in (6.3) are obtained by means of the following transformations:

$$\hat{\lambda} = \exp(-\hat{\mu}/\hat{\sigma}), \quad \hat{\alpha} = 1/\hat{\sigma}, \quad \hat{\boldsymbol{\beta}} = -\hat{\boldsymbol{\gamma}}/\hat{\sigma}.$$

For each individual's censoring interval an uncensored imputed survival time is randomly generated using its corresponding truncated Weibull distribution with parameters $\hat{\lambda}$, $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$. No imputation is needed in the case of right-censored observations.

Regarding truncated distributions and according to [Nadarajah & Kotz \(2006\)](#), let X be a random variable representing the truncated version of some distribution function $G(\cdot)$, over the interval $(L_i, R_i]$ of each individual with $0 < L_i < R_i < \infty$. It is straightforward to check that the distribution function of X is given by

$$F_{X_i}(x) = \frac{G(\max(\min(x, R_i), L_i) | \hat{\boldsymbol{\theta}}) - G(L_i | \hat{\boldsymbol{\theta}})}{G(R_i | \hat{\boldsymbol{\theta}}) - G(L_i | \hat{\boldsymbol{\theta}})},$$

and its corresponding inverse by

$$F_{X_i}^{-1}(p) = G^{-1}(G(L_i | \hat{\boldsymbol{\theta}}) + p(G(R_i | \hat{\boldsymbol{\theta}}) - G(L_i | \hat{\boldsymbol{\theta}}))).$$

Hence, to obtain a random value of X , for every censoring interval $(L_i, R_i]$, a random uniform number u_i is generated and an imputed value $x_i = F_X^{-1}(u_i)$ is derived.

Given an original sample of n individuals with $n-r$ interval-censored and r right-censored survival times, step 1 yields a sample of size n with $n-r$ imputed uncensored and the r right-censored survival times. In the next step, a mixed effects Cox model can be fitted to this sample as we explain below.

6.3.2 Fit of the mixed effects Cox model

In this step, the mixed effects Cox model (6.1) explained in Section 6.2 is fitted using the previously imputed values. The objective here is to estimate the vector of fixed effects regression parameters $\boldsymbol{\beta}$ and the vector of parameters $\boldsymbol{\theta}$ for the covariance matrix $\boldsymbol{\Sigma}$ of the random effects. In what follows, we sketch the main ideas of Therneau's method ([T. M. Therneau, 2018a](#)).

The MLE for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is based on an integrated penalized partial likelihood (IPL)

$$\text{IPL}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{q/2} |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{1/2}} \int \text{PL}(\boldsymbol{\beta}, \mathbf{b}) \exp\{-\mathbf{b}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{b} / 2\} d\mathbf{b}, \quad (6.4)$$

where \mathbf{b} is the vector of random effects coefficients, as presented in (6.1) and q corresponds to the number of random effects. When the variance of the random effect is zero, this collapses to the ordinary Cox partial likelihood.

Since expression (6.4) is not a tractable integral and in order to perform computations under this likelihood, we rewrite the logarithm of the integrand of equation (6.4), that is, $\text{LPPL}(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}) = \text{LPL}(\boldsymbol{\beta}, \mathbf{b}) - (1/2) \mathbf{b}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{b}$, as a second-order Taylor series about $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}})$ as follows

$$\text{LPPL}(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}) \approx \text{LPPL}(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\mathbf{b}}(\boldsymbol{\theta})) - (1/2) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \mathbf{b} - \hat{\mathbf{b}})' H(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \mathbf{b} - \hat{\mathbf{b}}),$$

where the Hessian H is evaluated at $(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\mathbf{b}}(\boldsymbol{\theta}))$. When $\boldsymbol{\theta}$ and hence $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are fixed, the relevant values of $\boldsymbol{\beta}$ and \mathbf{b} that maximize $\text{LPPL}(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta})$ are easily obtained using essentially the same methods as an ordinary Cox model.

As we are only interested in the values at $\hat{\boldsymbol{\beta}}$, the last term collapses to $(\mathbf{0}, \mathbf{b} - \hat{\mathbf{b}})' H(\mathbf{0}, \mathbf{b} - \hat{\mathbf{b}}) = (\mathbf{b} - \hat{\mathbf{b}})' H_{bb}(\mathbf{b} - \hat{\mathbf{b}})$, where H_{bb} is the portion of the Hessian corresponding to the random effects. When we replace the body of the integral in (6.4) with this approximation, then the result is an integral that can be solved in closed form. For further details see [T. M. Therneau \(2018a\)](#).

Basically, in this second step by fitting the mixed effects Cox model presented in (6.1), we obtain an estimation of the parameter vectors $\boldsymbol{\beta}$ and \mathbf{b} as well as of the corresponding covariance matrices.

6.3.3 Pooling the results

We need to repeat the previous steps (Steps 1 and 2) M times, where M is a pre-specified integer larger than 1. [Rubin \(2004\)](#) stated that multiple imputation using modest M , say between 2 and 10, is designed for situations with a modest fraction of missing information.

The M complete-data analyses corresponding to the M imputations under the mixed effects Cox model result in M repeated complete-data statistics, and these are combined to form one repeated-imputation inference that appropriately adjusts for interval-censored data under the model used to create the repeated imputations.

The estimated parameter vectors and their corresponding covariance matrices obtained in the M repetitions of steps 1 and 2 are denoted by $\hat{\boldsymbol{\beta}}_m, \hat{\mathbf{b}}_m, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_m}, \hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})_m, m = 1, \dots, M$.

Given a particular fixed effects parameter β_i , the repeated-imputation estimate $\hat{\beta}_{i,MI}$ is the average over the M estimates of this parameter, that is, $\hat{\beta}_{i,MI} = (\sum_{m=1}^M \hat{\beta}_{i,m})/M$. In addition, multiple imputation also provides a simple formula to estimate the variance of $\hat{\beta}_{i,MI}$ ([Rubin, 2004](#)), namely,

$$\widehat{\text{Var}}(\hat{\beta}_{i,MI}) = U_{i,MI} + \left(1 + \frac{1}{M}\right) B_{i,MI},$$

where $U_{i,MI} = (\sum_{m=1}^M \widehat{\text{Var}}(\hat{\beta}_{i,m}))/M$ is the within-imputation variance and B_{MI} is the between-imputation variance given by $B_{i,MI} = (\sum_{m=1}^M (\hat{\beta}_{i,m} - \hat{\beta}_{i,MI})^2)/(M-1)$. The B_{MI} term is inflated by a factor $1/M$ to take into account the finite number of imputations.

The same procedure is applied for the estimation of the random effects as well as for the elements of the covariance matrices.

6.3.4 Software issues

We have accomplished the imputation process in R using the `truncdist` contributed package ([Novomestky & Nadarajah, 2016](#)). This package includes the function `rttrunc` to impute the values per subjects. This function generates n random deviates that are drawn from the specified truncated distribution.

To fit the mixed effects Cox model, we used the R package called `coxme` ([T. M. Therneau, 2018b](#)). The central computational strategy implemented in this package is an outer and an inner loop. The outer loop searches over the parameters $\boldsymbol{\theta}$ of the variance matrix for a maximum of the IPL (6.4) and does it in 3 steps. For each trial value of $\boldsymbol{\theta}$ in this search, the first step is to calculate $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$.

The second step is to solve the penalized Cox model $LPL(\boldsymbol{\beta}, \mathbf{b}) - (1/2)\mathbf{b}'\boldsymbol{\Sigma}^{-1}\mathbf{b}$ to get the solution vector $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}})$. The iterative Newton-Raphson solution to this problem is the inner loop. The third step is to use the Laplace approximation to compute the log IPL, using the results of step 2. The implemented R code to follow the algorithm is presented in Appendix C.

6.4 Effect of gender and pre-cART VL on the time to HIV RNA viral rebound considering multiple random effects

The ATI data set (Leal et al., 2017) that motivated this study corresponds to a recompilation from 229 ATI episodes belonging to 158 different patients. The main virologic outcome of interest in these studies is the time to viral rebound, defined as the time between treatment interruption and the first detectable serum viral load (VL), which is measured as the number of HIV copies in a millilitre of blood. For the purpose of statistical analyses, it is commonly log-transformed because of its right-skewed distribution.

Following, we present a description of some of the variables in the data set. Therein, we refer to the eight studies by Study 1 to Study 8. More detailed information on the studies, including the inclusion criteria and the interventions, is provided in Appendix D. In addition, we will refer to the last viral load before the first initiation of cART by pre-cART VL.

6.4.1 Descriptive analysis of the ATI dataset

Table 6.1 presents the gender distribution and a numeric description of the log pre-cART VL separately for each of the eight studies as well as overall. Therein, N_{ATI} denotes the number of ATI periods per study and N the number of patients. As shown, 158 patients were involved in the 8 studies with a total of 229 ATI periods. Fifty nine patients were exposed to an immunomodulating intervention of some kind and 61% of the patients were men. Notice that in the case of both Study 6 and 7, we did not obtain the information on the gender of 17 patients. Concerning the pre-cART VL of the patients, the overall median of the log base 10-transformed pre-cART VL was 4.37 with similar values in all but one study (Study 4), within which the median log pre-cART VL was clearly lower (3.19). The reason for this resides in the inclusion criteria for Study 4 as explained in Appendix D.

Concerning the time until viral rebound, all but one of the 158 patients suffered a viral rebound during the respective follow-up times of the ATI studies and since VL was determined weekly, all these times were interval-censored. The exception corresponds to a right-censored observation of a patient in Study 5, whose viral load was below 20 copies/mL after four weeks when it was measured the last time. A graphical representation of all censoring intervals is shown in Figure 6.2, wherein, it can be observed that all interval lengths are multipliers of one week. This is due to the fact that the medical follow-up visits in all included studies were programmed with exact multiples of seven days following a strict protocol.

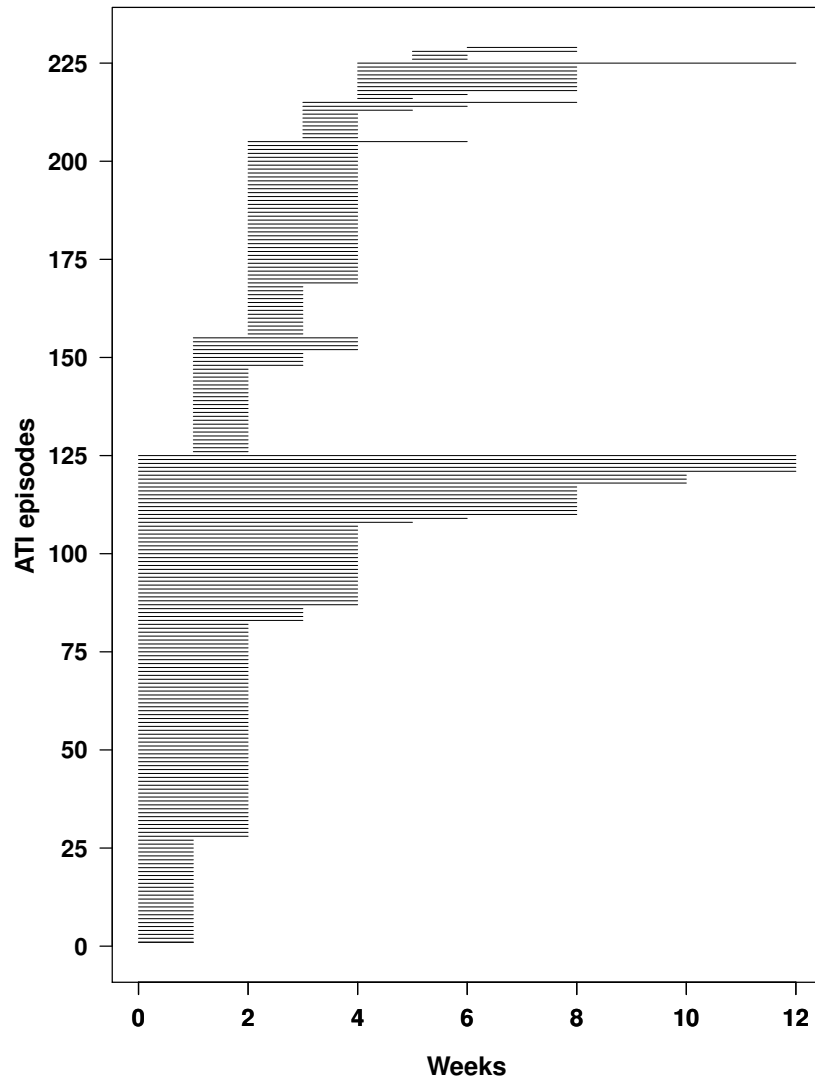


Figure 6.2: Lengths of the ordered interval-censored times (weeks) until viral rebound of the 229 ATI episodes. The average length of the censoring intervals is 2.6 weeks.

Table 6.1: Description of the eight studies in ATI data set.

Study	n _{ATI}	n	Gender		Log pre-cART VL			
			Male (Fem)	Missing	Mean (SD)	Median (IQR)	Min Max	Missing
Overall	229	158 ^a	96 (28)	34	4.37 (0.7)	4.31 (3.98 - 4.86)	2.33 - 6.00	3
Study 1 (García et al., 2005)	32	16	14 (2)	0	4.14 (0.64)	4.16 (3.88 - 4.57)	3.10 - 5.17	0
Study 2 (García et al., 2013)	70	35	27 (8)	0	4.78 (0.56)	4.69 (4.38 - 5.16)	3.20 - 5.74	0
Study 3 (ClinicalTrials.gov Identifier: NCT 02767193)	18	18	18 (0)	0	4.12 (0.57)	4.19 (3.95 - 4.48)	2.33 - 4.76	0
Study 4 (García et al., 2004)	11	11	5 (6)	0	3.21 (0.37)	3.19 (3.08 - 3.40)	2.61 - 3.80	0
Study 5 (García et al., 2003)	20	20	15 (5)	0	4.50 (0.49)	4.48 (4.11 - 4.89)	3.80 - 5.40	0
Study 6 (Mothe et al., 2015)	28	28	10 (1)	17	4.46 (0.77)	4.28 (3.96 - 4.87)	3.33 - 6.00	3
Study 7 (Fagard et al., 2003)	33	33	11 (5)	17	4.43 (0.52)	4.45 (4.08 - 4.70)	2.84 - 6.00	0
Study 8 (García et al., 1999, 2001)	17	10	7 (3)	0	4.59 (0.56)	4.40 (4.31 - 4.89)	3.89 - 5.70	0

^aThe sum of the column is not equal to 158 because of patients belonging to more than one study.

Additionally, in Figure 7.2, we provide the Turnbull estimates of the distribution functions of the time until viral rebound, obtained from Formula (4.1) by $1 - S_n(t)$. Therein, we can observe that the estimated probabilities of a viral rebound within the first two and four weeks are close to 0.6 and 0.9, respectively. Moreover, the separate Turnbull estimates of $F(t)$ in Studies 1 through 8 reflect a notable heterogeneity among the ATI studies: for example, the estimated probabilities of a viral rebound within the first two weeks vary from less than 0.2 (Study 4) to more than 0.8 (Study 1).

6.4.2 Fit of the mixed effects Cox model

For the purpose of studying the possible effect of gender and log pre-cART VL on the time until viral rebound taking into account the within-subject and within-study correlation, we fitted the mixed effects model presented in (6.1):

$$\lambda_i(t; \mathbf{X}, \mathbf{Z}) = \lambda_0(t) \exp(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}) \quad \text{with } \mathbf{b} \sim \mathbf{N}(\mathbf{0}, \Sigma), \quad (6.5)$$

where $\mathbf{X} \in \mathbb{R}^{229 \times 2}$ is the design matrix of the fixed effects Gender and log pre-cART VL and $\boldsymbol{\beta} \in \mathbb{R}^2$ is the fixed effects parameter vector. In the case of the log pre-cART VL, we decided to subtract 4, which

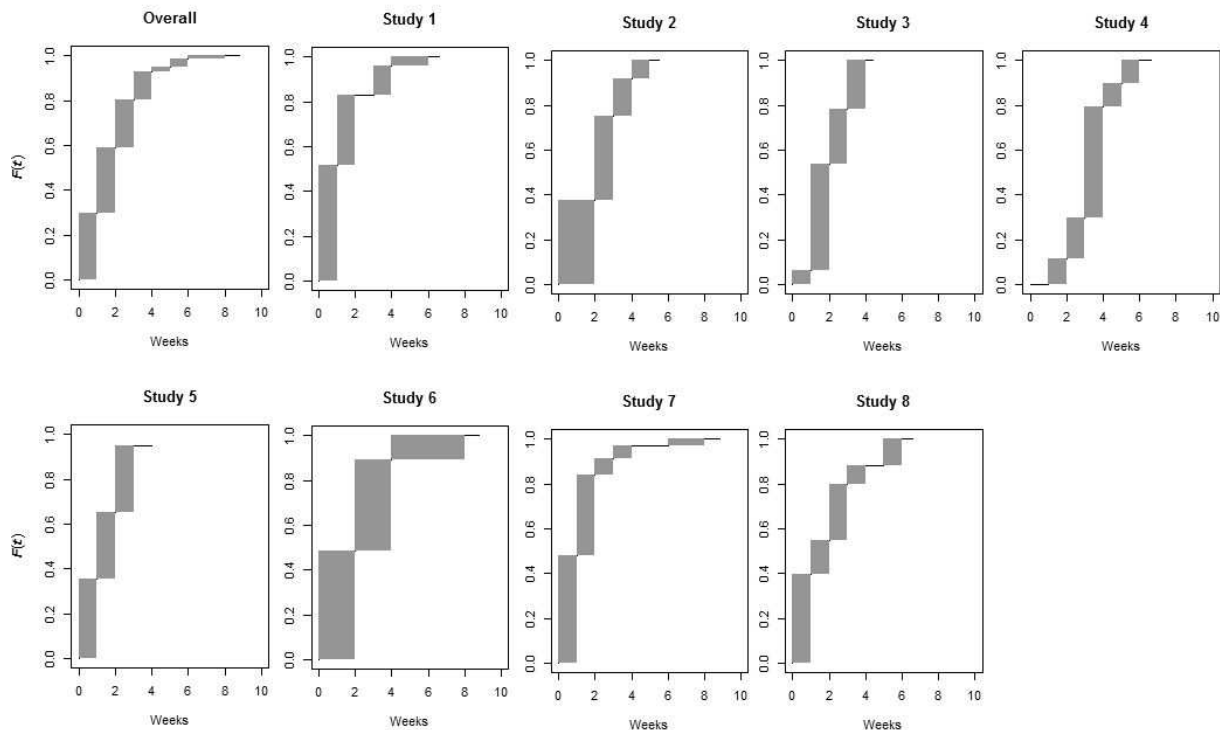


Figure 6.3: Non-parametric estimation of the distribution function of the time until viral rebound. The first graph (upper left corner) shows the estimation based on the pooled sample, the remaining graphs correspond to Studies 1 through 8.

is equivalent to a VL of 10000 HIV copies in a millilitre of blood and close to the overall median (4.31), since the value 0 was outside the range of the variable (Table 6.1). Concerning the random effects, we considered a random intercept per patient as well as correlated random intercept and slope for log pre-cART VL within each study. Hence, $\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_3)' \in \mathbb{R}^{158+8+8}$ and $\mathbf{Z} \in \mathbb{R}^{229 \times 174}$. In addition, the covariance matrix of \mathbf{b} is given by

$$\Sigma = \begin{pmatrix} \sigma_{\mathbf{b}_1}^2 & 0 & 0 \\ 0 & \sigma_{\mathbf{b}_2}^2 & \sigma_{\mathbf{b}_2, \mathbf{b}_3} \\ 0 & \sigma_{\mathbf{b}_2, \mathbf{b}_3} & \sigma_{\mathbf{b}_3}^2 \end{pmatrix}.$$

Notice that the inclusion of a random intercept per study implies study-specific baseline hazard functions $\lambda_{0,1}(t), \dots, \lambda_{0,8}(t)$.

For the model fit, we used the 3-step algorithm proposed in Section 6.3. In the first step of the algorithm, we replaced the interval-censored times until viral rebound by imputed times obtained randomly from the truncated Weibull distribution, and in Step 2, Model (6.1) was fitted. Steps 1 and 2 were repeated $M = 15$ times providing the parameter estimates corresponding to the fixed effects presented in Table 6.2.

Table 6.2: Estimation of the fixed effects parameters and the standard deviation of the random effects of Model (6.1) using the three-step imputation method.

	$\hat{\beta}$	$se(\hat{\beta})$	\widehat{HR}	95%CI
Gender (Female vs Male)	0.50	0.26	1.65	[0.99, 2.74]
Log pre-cART VL	0.60	0.23	1.83	[1.16, 2.86]
Random effects	$\hat{\sigma}_{b_1}$	$\hat{\sigma}_{b_2}$	$\hat{\sigma}_{b_3}$	$\widehat{Corr}_{b_2, b_3}$
	0.35	0.34	0.34	-0.47

According to the results obtained, the instantaneous risk of viral rebound among female patients is 1.65 times larger than the instantaneous risk among male patients with the same pre-cART VL. However, since the standard error is relatively large, the corresponding 95% confidence interval of the hazard ratio does include 1 and, hence, we actually cannot claim that women are at larger risk for viral rebound than men among the population of HIV-infected persons at a confidence level of 0.95. Concerning pre-cART VL, the results obtained are clear: the larger the pre-cART VL, the larger the risk of suffering a viral rebound. The adjusted hazard ratio of 1.83 implies that a unit increase of the log pre-cART VL, that is, a 10-fold increase of the VL, increases the instantaneous risk of a viral rebound by factor 1.83.

Regarding the random effects, the estimated standard deviation of the random intercept per subject, $\hat{\sigma}_{b_1}$ can be interpreted as the unexplained variation between individuals after controlling for the explanatory variables in the model. The value of the estimated standard deviation of the random intercept per study, $\hat{\sigma}_{b_2}$, reflects the heterogeneity among the eight studies with respect to the inclusion criteria. The standard deviation $\hat{\sigma}_{b_3}$ quantifies the variability of the slopes of pre-cART VL among the eight studies. The study-specific hazard ratios associated to pre-cART VL varies from 1.28 (Study 2) to 2.49 (Study 8; values not shown). Moreover, the negative value of the estimated correlation (-0.47) between random intercept and random slope of pre-cART VL, implies that the smaller the baseline hazard function per study, the larger the effect of the pre-cART VL.

6.5 Simulation study

The objective of the following simulation study was to explore the properties of the estimation method presented in Section 6.3 in terms of bias and mean squared error of the fixed effects estimator $\hat{\beta}$ in a setting very similar to the one of the data set at hand.

6.5.1 Simulation settings and data generation

Data sets were generated based on Model (6.5) under a total of 36 different scenarios shown in Table 6.3.

Table 6.3: Settings of the simulation study.

Sample size	$n \in \{100, 200, 300\}$
Assessment probability	$p \in \{0.2, 0.33, 0.5\}$
Number of imputations	$M = 15$
Distribution of T	Weibull, Gompertz
Percentage of right-censored observations	0 and 10
Distribution of X_1	$\text{Bin}(1, 0.23)$
Distribution of X_2	$N(0, 0.7)$
$(\beta_1, \beta_2)'$	$(0.5, 0.6)'$
Distribution of $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$	$MVN(0, \Sigma)$ (see (6.6))

Common to all scenarios, which were motivated by the data set presented in Section 6.4, were the distributions of the fixed effects variables X_1 (Binomial(1, 0.23)) and X_2 ($N(0, 0.7)$) as well as the parameter values $\beta_1 = 0.5$ and $\beta_2 = 0.6$. Regarding the random effects, the variable Study was generated from a multinomial distribution with equal probabilities for each of the eight studies, i.e., $p = 1/8$. In the case of Studies 1, 2, and 8, we duplicated the ATI episodes per patient, in order to reproduce the scenario of the real data set. The values of the random effects \mathbf{b}_1 , \mathbf{b}_2 , and \mathbf{b}_3 were generated from a multivariate normal distribution with mean 0 and

$$\Sigma = \begin{pmatrix} 0.35^2 & 0 & 0 \\ 0 & 0.35^2 & -0.47 \cdot 0.35 \cdot 0.35 \\ 0 & -0.47 \cdot 0.35 \cdot 0.35 & 0.35^2 \end{pmatrix}. \quad (6.6)$$

To generate survival times from the mixed effects proportional hazards model in (6.5), we used the inverse probability method described by Bender et al. (2005). According to the authors, a random survival time from Model (6.5) can be generated using the following equation:

$$T = \Lambda_0^{-1} \left(- \frac{\log(U)}{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})} \right),$$

where U follows a uniform distribution in the interval (0, 1) and Λ_0 is the cumulative baseline hazard function. Survival times can be generated assuming the conditional distribution of T given \mathbf{X} and \mathbf{Z} follows a Weibull or Gompertz distribution since both share the assumption of proportional hazards (Bender et al., 2005).

In the case of the Weibull distribution with shape and scale parameters $\kappa > 0$ and $\rho > 0$, the baseline hazard function is $\lambda_0 = \kappa \rho (\rho t)^{\kappa-1}$. Hence, $\Lambda_0(t) = (\rho t)^\kappa$, and following the inverse probability method, a realization of T is obtained by computing

$$t = \frac{\{-\log(u) \cdot \exp(-\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})\}^{1/\kappa}}{\rho},$$

where u is a realization of U .

Regarding the Gompertz distribution with shape and scale parameters $\alpha \in (-\infty, \infty)$ and $\rho > 0$, the baseline hazard function is $\lambda_0 = \rho \exp\{\alpha t\}$. Hence, $\Lambda_0(t) = (\rho/\alpha) \cdot (\exp(\alpha t) - 1)$, and following the inverse probability method, a realization of T is obtained by computing

$$t = \frac{1}{\alpha} \cdot \log\left(1 - \frac{\alpha \log(u)}{\rho \exp(-\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})}\right).$$

Notice that in the case of $\alpha < 0$ there is a point mass at infinity since the argument of the log function could then be negative. However, this possibility was ruled out in the simulation study as $\alpha = 1.5$ was used to generate the data.

Given all survival times $T_i, i = 1, \dots, n$, per data set, the censoring intervals $(L_i, R_i]$ were generated assuming noninformative censoring and following the procedure described in Gómez et al. (2009). All integers from 1 through 12, the maximum upper limit of the censoring intervals in our data set (Figure 6.2) were considered possible assessment times using three different assessment probabilities for each of the 12 values: $p \in \{0.2, 0.33, 0.5\}$. That is, the assessment times were generated according to a Bernoulli distribution with parameter p and for each survival time T_i generated, the interval $(L_i, R_i]$ was obtained as the smallest interval of assessment times among all those including T_i .

In addition to the underlying distribution and the assessment probabilities, the different scenarios were determined by the sample size per data set ($n \in \{100, 200, 300\}$) and the percentage of right-censored observations (none and 10%, respectively). The number of imputations was kept the same for all settings ($M = 15$).

6.5.2 Evaluation criteria

Given any of the 36 simulation settings, we generated $D = 1000$ data sets, for each of which β_1 and β_2 were estimated by means of the method presented in Section 6.3. Thus, we obtained $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_{1,d}, \hat{\beta}_{2,d})'$ for $d = 1, \dots, D$. Based on these estimations and given the true parameter vector $\boldsymbol{\beta}_0 = (0.5, 0.6)'$, we calculated the mean, variance, bias, and mean squared error (MSE) of $\hat{\beta}_1$, and $\hat{\beta}_2$ as follows:

$$\begin{aligned} \bar{\hat{\beta}}_i &= \frac{1}{D} \sum_{d=1}^D \hat{\beta}_{i,d}, \quad i = 1, 2, \\ \widehat{\text{Var}}(\hat{\beta}_i) &= \frac{1}{D-1} \sum_{d=1}^D (\hat{\beta}_{i,d} - \bar{\hat{\beta}}_i)^2, \quad i = 1, 2, \\ \widehat{\text{Bias}}(\hat{\beta}_i) &= \bar{\hat{\beta}}_i - \beta_{0,i}, \quad i = 1, 2, \\ \widehat{\text{MSE}}(\hat{\beta}_i) &= \widehat{\text{Var}}(\hat{\beta}_i) + \widehat{\text{Bias}}(\hat{\beta}_i)^2, \quad i = 1, 2. \end{aligned} \tag{6.7}$$

The bias is a measure for the accuracy of the estimators, whereas the MSE can be used as a measure for the precision.

Concerning the estimation of the variance, $\text{Var}(\hat{\beta}_{i,d})$ could also be estimated within each of the D runs and their mean could serve as well as an estimator of $\text{Var}(\hat{\beta}_i)$, $i = 1, 2$. However, Formula (6.7) is generally more appropriate, since the mean of the D variance estimates usually underestimates the true variance $\text{Var}(\hat{\beta}_i)$.

The whole simulation process was programmed in R (R Core Team, 2020) using the function `multimp` (see Appendix C) and is included in the Supporting Information.

6.5.3 Simulation results

The results of the simulations in terms of bias and MSE are shown in Tables 6.4 (no right-censored data) and 6.5 (10% of right-censored observations). In general, for the settings under study, it can be observed that our proposed method captures the real parameters in a proper way. Irrespective of the conditional distribution of T or the percentage of right-censored observations, the estimated bias of both $\hat{\beta}_1$ and $\hat{\beta}_2$ can be considered small.

Concerning the conditional distribution of the survival times, hardly any difference is observed between the Weibull and the Gompertz distribution with respect to the bias even though the former is always used for the imputation step of our method. Contrary to that, the MSE is generally larger in case of the Gompertz distribution indicating a somewhat smaller precision in that case. Only slight differences are observed comparing the settings without and with 10% of right-censored data. In the case of no right-censored observations, the bias is generally a bit lower, whereas the MSE is slightly bigger in most cases.

The sample size does have the expected impact on the precision of $\hat{\beta}_1$ and $\hat{\beta}_2$: the MSE decreases as n increases no matter the width of the intervals or the conditional distribution of the survival times. In terms of bias, generally, no big differences are observed between the different sample sizes. However, with a sample size of $n = 100$, the bias seems to depend on the assessment probability: the larger the assessment probability and, hence, the smaller the censoring intervals, the more accurate the estimator. This tendency is, generally, not observed for $n = 200$ and $n = 300$. Moreover, the bias, even though generally small, is almost always negative in the case of $n = 200$ and $n = 300$ indicating, hence, a slight underestimation of both parameters. By contrast, with a sample size of 100, the bias is most often positive.

6.6 Discussion

Our collaboration as data scientists with clinicians and virologists from Hospital Clínic of Barcelona, IRB Barcelona, and University of Barcelona led to the analysis of the eight different Analytical Treatment Interruption studies in chronic HIV-positive combination Antiretroviral Therapy (cART)-suppressed individuals. The main clinical question to be studied was whether gender and pre-cART VL were risk factors on the time until viral rebound in HIV-infected patients with previously undetectable viral load, taking into account the heterogeneity between the different studies and the fact that different patients

Table 6.4: Fixed parameters estimators without right-censored observations and 15 imputations ($\beta_1 = 0.5, \beta_2 = 0.6$).

	Weibull		Gompertz	
	Bias	MSE	Bias	MSE
$n = 100$				
<i>Assessment probability: $p = 0.2$</i>				
$\hat{\beta}_1$	0.060	0.246	0.045	0.293
$\hat{\beta}_2$	-0.001	0.078	-0.022	0.089
<i>Assessment probability: $p = 0.33$</i>				
$\hat{\beta}_1$	0.035	0.132	-0.007	0.135
$\hat{\beta}_2$	0.005	0.054	-0.030	0.064
<i>Assessment probability: $p = 0.5$</i>				
$\hat{\beta}_1$	0.001	0.090	-0.004	0.092
$\hat{\beta}_2$	-0.009	0.051	-0.021	0.049
$n = 200$				
<i>Assessment probability: $p = 0.2$</i>				
$\hat{\beta}_1$	-0.023	0.073	-0.028	0.092
$\hat{\beta}_2$	-0.029	0.039	-0.045	0.042
<i>Assessment probability: $p = 0.33$</i>				
$\hat{\beta}_1$	-0.034	0.049	-0.035	0.061
$\hat{\beta}_2$	-0.036	0.032	-0.044	0.038
<i>Assessment probability: $p = 0.5$</i>				
$\hat{\beta}_1$	-0.023	0.038	-0.032	0.041
$\hat{\beta}_2$	-0.032	0.030	-0.054	0.031
$n = 300$				
<i>Assessment probability: $p = 0.2$</i>				
$\hat{\beta}_1$	-0.038	0.046	-0.046	0.054
$\hat{\beta}_2$	-0.045	0.028	-0.064	0.036
<i>Assessment probability: $p = 0.33$</i>				
$\hat{\beta}_1$	-0.040	0.034	-0.051	0.039
$\hat{\beta}_2$	-0.045	0.027	-0.043	0.027
<i>Assessment probability: $p = 0.5$</i>				
$\hat{\beta}_1$	-0.040	0.027	-0.032	0.029
$\hat{\beta}_2$	-0.043	0.024	-0.052	0.027

Table 6.5: Fixed parameters estimators considering 10% right-censored observations and 15 imputations ($\beta_1 = 0.5, \beta_2 = 0.6$).

	Weibull		Gompertz	
	Bias	MSE	Bias	MSE
$n = 100$				
<i>Assessment probability: $p = 0.2$</i>				
$\hat{\beta}_1$	0.088	0.315	0.071	0.342
$\hat{\beta}_2$	0.027	0.090	-0.015	0.091
<i>Assessment probability: $p = 0.33$</i>				
$\hat{\beta}_1$	0.019	0.140	0.011	0.143
$\hat{\beta}_2$	0.019	0.061	-0.018	0.062
<i>Assessment probability: $p = 0.5$</i>				
$\hat{\beta}_1$	0.013	0.097	0.001	0.098
$\hat{\beta}_2$	0.007	0.055	-0.005	0.053
$n = 200$				
<i>Assessment probability: $p = 0.2$</i>				
$\hat{\beta}_1$	0.001	0.085	-0.022	0.091
$\hat{\beta}_2$	-0.024	0.045	-0.030	0.047
<i>Assessment probability: $p = 0.33$</i>				
$\hat{\beta}_1$	-0.017	0.052	-0.010	0.065
$\hat{\beta}_2$	-0.019	0.036	-0.028	0.035
<i>Assessment probability: $p = 0.5$</i>				
$\hat{\beta}_1$	-0.017	0.043	-0.015	0.046
$\hat{\beta}_2$	-0.023	0.030	-0.033	0.031
$n = 300$				
<i>Assessment probability: $p = 0.2$</i>				
$\hat{\beta}_1$	-0.030	0.049	-0.037	0.057
$\hat{\beta}_2$	-0.033	0.033	-0.054	0.036
<i>Assessment probability: $p = 0.33$</i>				
$\hat{\beta}_1$	-0.034	0.036	-0.030	0.037
$\hat{\beta}_2$	-0.036	0.027	-0.041	0.029
<i>Assessment probability: $p = 0.5$</i>				
$\hat{\beta}_1$	-0.027	0.028	-0.013	0.029
$\hat{\beta}_2$	-0.023	0.025	-0.031	0.026

had different number of ATI episodes. The first challenge we encountered was that the times until viral rebound were interval-censored. Other difficulties were that the studies had different inclusion criteria (for example, with respect to pre-cART VL values), and that some individuals participated in more than one study or were exposed to more than one ATI episode.

For these reasons, the analysis of this data set with the mixed effects Cox model seemed to be the natural choice. However, to the best of our knowledge, this model had not been studied with interval-censored data previously. Hence, our proposal is a first step to close the gap between the mixed effects Cox model (T. M. Therneau & Grambsch, 2000) and the Cox model with interval-censored data (Finkelstein, 1986).

The method proposed is based on multiple imputations in order to replace the censoring intervals by imputed values to simplify the data structure to uncensored and possibly right-censored survival times. For this step, we propose to generate random survival times from a truncated Weibull distribution. As an alternative to multiple imputation, single imputation methods imputation could have been applied, such as midpoint imputation replacing the censoring interval $(L_i, R_i]$ by its midpoint $(L_i + R_i)/2$. However, midpoint imputation is only reasonable when the time period between consecutive visits (or measurements) is short leading to approximately unbiased estimations (Law & Brookmeyer, 1992). But even in this case, the standard error of the estimator would be underestimated since single imputation methods ignore the imputation uncertainty (Kim, 2003) and do not take into account the variability of the censoring interval. In contrast, multiple imputation does not attempt to estimate each missing value through simulated values but rather to represent a random sample of the missing values. This process results in valid statistical inferences that properly reflect the uncertainty due to missing data (Yuan, 2010).

According to the results of the simulation study presented in Section 6.5, the estimation method proposed has desirable properties in terms of accuracy (small bias) and precision (low MSE) of the estimators of the fixed effects parameters. A relatively large MSE was only observed in the case of the smallest sample size ($n = 100$) and smallest assessment probabilities ($p = 0.2$) considered (see Tables 6.4 and 6.5). We have to admit, however, that the simulation study did only comprise 36 different scenarios determined by sample size, assessment probability, condition distribution of T given \mathbf{X} and \mathbf{Z} , and the percentage of right-censored observations. Further simulation studies would be desirable in order to explore the properties of our estimation method under additional settings or with mixed effects Cox models with a different number of fixed and random effects.

An apparent limitation of our estimation methods seems to be the fact that the imputation step is based exclusively on the (truncated) Weibull distribution. However, the Weibull distribution is a flexible distribution in the sense that its hazard function can have different shapes (constant, monotonically increasing, or monotonically decreasing). For this reason, its use must not necessarily be a limitation even though the underlying survival time distribution may be a different distribution. Actually, the simulation results in the case of the Gompertz distribution of T (small bias and low MSE) seem to confirm this supposition.

An important aspect of model fitting that has been beyond the scope of the present work is the comparison of nested mixed effect Cox models in the presence of interval-censored data. Nested mixed effects Cox models with right-censored data can be compared by means of the values of the log partial likelihood function after integrating out the random effects (T. M. Therneau, 2018b), whose difference multiplied by minus 2 follows a chi-squared distribution under the null hypothesis that the nested model does not improve the model fit. In our case, for example, we could be interested in testing, whether the inclusion of the different random effects actually improve the model fit. The application of the Likelihood Ratio Test proposed by T. M. Therneau (2018b) with our estimation method, however, is not straightforward because of the multiple imputations. One possible *ad hoc* solution to this question could be to compute, separately for the models to be compared, the corresponding values of the integrated log partial likelihood for each of the M model fits obtained with each imputation. Following, the mean difference over the M model fits could be calculated and compared to the quantiles of the corresponding chi-squared distribution. Further studies should address this topic in order to develop guidelines for researchers for how to compare nested mixed effects Cox models with interval-censored data.

Concerning the clinical results obtained, as could be expected, the higher the last viral load before the first initiation of cART (pre-cART VL), the larger the instantaneous risk of a viral rebound. In our data set, female patients were at larger instantaneous risk for viral rebound than male patients with the same pre-cART VL, however, using a 95% confidence level, we cannot claim that this is also valid among the population of HIV-infected patients. We did not study any further variables, but our R function `multimp` could be easily adapted to the estimation of more than two fixed parameters. The same could be done to consider more random effects.

We did also check whether the inclusion of the patient identifier and the (correlated) intercept and slope of pre-cART VL among studies as a random effects in the Cox model modifies the results obtained from a Cox model that ignored such random effects. To fit this model, we used the `icenReg` package (Anderson-Bergman, 2017), which enables the parameter estimation in the Cox model in the presence of interval-censored data. The differences observed were: $\hat{\beta}_1 = 0.51$ versus $\hat{\beta}_1 = 0.50$ in the case of variable Gender, and $\hat{\beta}_2 = 0.36$ versus $\hat{\beta}_2 = 0.60$ in the case of pre-cART VL. That is, ignoring the random effects ATI Study and patient identifier, the estimated hazard ratio associated with pre-cART VL would have been 1.43 and, hence, notably smaller than the estimated hazard ratio 1.83 reported in Section 6.4. These differences highlight the importance to consider random effects when fitting a Cox model in the presence of interval-censored data and a grouped data structure.

In Table 6.2, we also report the estimated standard deviations of the random effects. However, as Gleiss et al. (2018) point out, the values presented lack the direct comparability with the contribution of fixed effects. The authors are working on this issue addressing the explained variation in shared frailty models, which are a particular case of the mixed effects model, namely when a random intercept per random effect is considered.

Even though we have applied our method only to one data set, our method is valid for any data

set that presents both interval-censored survival times and a grouped data structure that could be treated as a random effect in a regression model. Nonetheless, a fit of the mixed effects Cox model with interval-censored data that did not require multiple imputations would be desirable. For this purpose, the expression of the likelihood function of the Cox model with interval-censored data presented by [Finkelstein \(1986\)](#) would need to be extended to consider random effects and tools of the area of optimization such as the ones presented by [Langohr & Gómez \(2005\)](#) could be useful to achieve this goal.

STATISTICAL METHODOLOGIES APPLIED TO BCN02 CLINICAL TRIAL

In this chapter, we present different analyses addressing the BCN02 clinical trial. We have divided the chapter in two main parts. In the first part, we start describing the clinical trial, its characteristics and goals. Following, we analyze the time until viral rebound. In this analysis, patients are classified in rebounders or controllers according to whether they keep the viral load controlled at week 12 after monitored antiretroviral pause (MAP). Based on the rebounder or controller classification we fit a log-binomial regression model. In the second part, we present the Enzyme-Linked Immunosorbent Spot (ELISpot) assays and the statistical methodologies used to work with this information. Finally, we present some conclusions regarding the methodologies applied in this clinical trial.

7.1 BCN02 clinical trial

The BCN02 study is a clinical trial to evaluate the safety and effect of a “kick and kill” strategy (for further information, see Chapter 3.10). The vaccine used is called MVA.HIVconsv and was used in combination with Romidepsin (RMD). The main idea was to study the behaviour of viral rebound after treatment interruption in early treated HIV-1 infected individuals. The study is a joint collaboration between the Hospital Universitari Germans Trias i Pujol, Badalona and in the Hospital Clínic, Barcelona. The trial is sponsored by IrsiCaixa AIDS Research Institute and the Coordinating Investigator is Beatriz Mothe Pujadas (more information at <https://clinicaltrials.gov/ct2/show/NCT02616874>).

As we mentioned before, this study is based on a combination of MVA.HIVconsv vaccine and a specific drug called Romidepsin. HIVconsv vaccine is defined as the most immunogenic candidate available so far and was supplied by the University of Oxford. RMD was clinically developed as an anti-

cancer drug and is approved for the treatment of cutaneous lymphoma. In the HIV field, RMD has been proposed as a potent HIV latent viral reservoir activator (for viral reservoir definition, see Chapter 3).

Design of the trial

The BCN02 clinical trial is an extension phase of the BCN01 clinical trial (NCT01612425) held in Barcelona in 2014 to evaluate the safety and immunogenicity of 2 different vaccines. All the patients of BCN01 who meet all eligibility criteria were invited to participate. The total number of patients in BCN02 trial is fifteen and the follow-up time corresponds to 32 weeks.

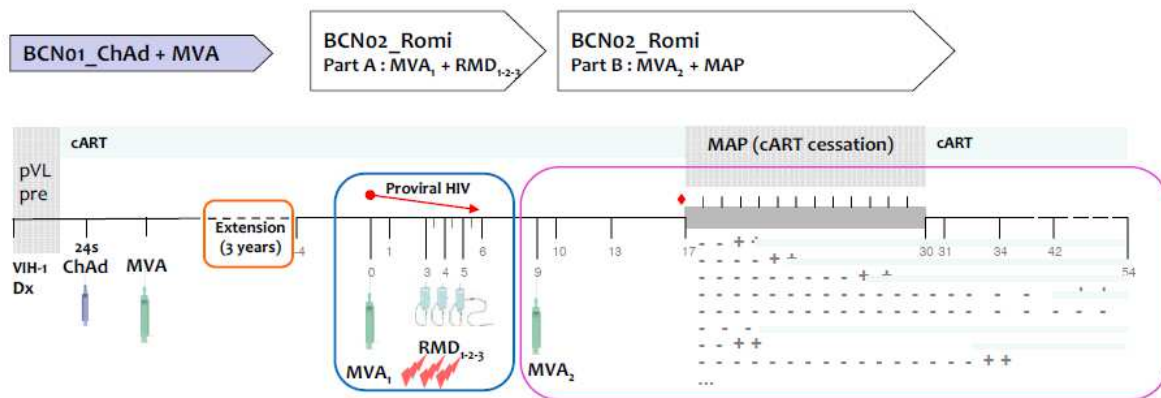


Figure 7.1: BCN02 diagram. Source: <https://www.flsidea.org/es/postcroi2017>.

In the Figure 7.1, we show the design of the BCN02 trial. The clinical trial was divided in two phases. In the first part of the trial, three doses of romidepsin were administered after a first vaccination with MVA.HIVconsv at weeks 3, 4 and 5 (Part A: $MVA_1 + RMD_{1-2-3}$, intervention phase). In the second part of the trial (Part B: $MVA_2 + MAP$), a second vaccination with MVA.HIVconsv was given 4 weeks after the last RMD infusion. After an interim analysis at week 13, a monitored antiretroviral pause (MAP) was offered to all participants at week 17. A closely monitored antiretroviral pause was planned to offer cART resumption as soon as viral rebound is detected.

7.2 Clinical and survival data of BCN02

In this section, we describe some clinical data at different stages of the BCN02 clinical trial. Moreover, we present the fitted survival models for the time until viral rebound, taking into account that this response is interval-censored. In the analyses, the profile of the patients is also important: controller or late-rebounder (from now on, rebounder). A controller patient is defined as a patient with viral load (VL) below 2,000 copies/ml at week 12 of MAP. A patient with VL greater than 2,000 copies/ml is considered rebounder.

Part of the analyses presented in this section are also presented in the manuscript “HIVconsv vaccines and romidepsin in early-treated HIV-1-infected individuals: Safety, immunogenicity and effect on the viral reservoir (study BCN02)” published in the journal *Frontiers in Immunology*, section Vaccines and Molecular Therapeutics (Mothe et al., 2020).

7.2.1 Descriptive analysis of clinical covariates

The Table 7.1 shows the demographic, clinical, and treatment characteristics of the fifteen study patients at BCN02 study entry. In the case of continuous variables, the median, the interquartile range, the minimum and maximum are shown for rebounder and controller patients in Table 7.2. In these analyses we did not consider the variable gender, because there was only one female patient.

Table 7.1: Demographic, clinical and treatment characteristics of study patients at study entry (n=15).

Variable	Median (range)*
Age (years)	43 (33 - 51)
Time since HIV acquisition to cART (days)	92 (28 - 164)
Pre cART log ₁₀ HIV RNA (copies/ml)	4.9 (3.2 - 5.8)
Time on cART (years)	3.23 (3.03 - 3.77)
cART regimen, n(%)	
TDF/FTC/RAL	11 (73)
ABC/3TC/RAL	2 (13)
ABC/3TC/DTG	2 (13)
CD4 ⁺ T-cell counts (cells/mm ³)	728 (416 - 1408)
Ratio CD4/CD8	1.37 (0.97- 1.93)

* Except when n(%) is specified

The median age of the patients is 43 years old, the median time since HIV acquisition to cART corresponds to 92 days. In the Table 7.1 we can observe the patients distribution to any of the three cART regimen. Eleven patients adhere to the first regimen, corresponding to the combination of Tenofovir Disoproxil Fumarate (TDF), Emtricitabine (FTC), and Raltegravir (RAL). Two patients adhere to the second combination, given by Abacavir (ABC), Lamivudine (3TC), and RAL. Also two patients receive the third combination that consists of ABC, 3TC, and Dolutegravir (DTG).

Table 7.2: Summary of continuous covariates for BCN02.

Variable	Rebounders (n=10)				Controllers (n=3)			
	Med	IQR	Min	Max	Med	IQR	Min	Max
Demos								
Age	42	38–45	33	48	32	31–36	30	40
At HIV-1 diagnosis								
Days HIV to cART	82	67–107	32	118	112	70–138	28	164
log ₁₀ (VL) at cART init	5.02	4.88–5.16	4.26	5.48	3.35	3.28–4.59	3.20	5.82
CD4 absolute v ₀	453	382–560	299	785	631	608.5–633	586	635
CD4/CD8 ratio v ₀	0.56	0.45–0.69	0.44	1.26	0.56	0.42–0.85	0.27	1.14
At BCN02 entry								
CD4 absolute	728	533–1331	416	1408	657	652.5–814	648	971
CD4/CD8 ratio	1.37	1.22–1.47	1.00	1.93	1.33	1.15–1.54	0.97	1.74
Total months on cART	38.88	37.56–40.20	36.36	41.04	41.64	39.30–42	36.96	42.36
Months on UD VL	35.52	35.40–35.88	31.32	40.32	36.36	36.18–39	36	41.64
At MAP								
CD4 absolute	735	484–1075	468	1269	854	675–902	496	950
CD4/CD8 ratio	1.30	1.25–1.48	0.87	1.75	1.52	1.14–1.57	0.76	1.61
Total months on cART	41.88	44.04–46.32	41.88	46.80	46.80	44.88–47.28	42.96	47.76
Months on UD VL	41.52	40.92–43.68	37.80	45.60	42.48	42.24–44.34	42	46.20
Vaccine Immunogenicity								
HIV convs Magnitude								
At BCN02 entry	160	0–287	0	2640	0	0–655	0	1310
At BCN02 peakimmunog	1965	1380–3940	530	6901	2480	1605–2668	730	2855
HIV convs pools breadth								
At BCN02 entry	1	0–2	0	2	0	0–1	0	2
At BCN02 peakimmunog	4	4–6	3	6	6	5.5–6	5	6
HIV convs immunodominance								
At BCN02 entry	6	0–8	0	37	0	0–38	0	76
At BCN02 peakimmunog	89	68–100	54	100	86	82–93	78	100
Responses OUTside HIVconsv								
At BCN02 entry	4525	3088–5385	1635	8945	3328	1872–3428	415	3528
At BCN02 peakimmunog	705	405–1150	130	5385	530	325–925	120	1320
Viral reservoir								
Week 0 BCN02	190	107–434	18	752	65	62.5–116.5	60	168
Week 3 BCN02	157	110–494	26	892	46	34.5–100	23	154
Week 6 BCN02	131	105–464	60	656	43	36.5–185.5	30	328
Week 17 BCN02	144	119–420	29	680	54	35–88	16	122

Table 7.2 shows the absolute frequencies of variables explored to explain the binary outcome defined as rebounders versus controllers. As can be seen, the table summarizes the results for 13 of the 15 patients participating in the study. This is because one of the patients did not stop the treatment for safety reasons and one patient was taking unauthorized drugs for the study. In the table it is possible to observe the median, the interquartile range (IQR), the minimum and the maximum for different demographic information, and variables regarding different stages of the study: at HIV-1 diagnosis, at BCN02 entry, at MAP. Regarding HIV-1 diagnosis, the median days on infection to cART was higher for controllers patients. In addition, in these patients the median amount of viral load ($\log_{10}(\text{VL})$ at cART init) was lower and the median CD4 counts greater than early-rebounders patients. At BCN02 entry the situation is more or less similar for every group. At MAP, a slightly difference can be appreciated, the median value for CD4 absolute and CD4/CD8 ratio is higher for controllers.

Table 7.2 shows as well information on the vaccine immunogenicity and viral reservoir at BCN02 entry and at BCN02 peak immunogenicity. Almost in every variable of these categories it is possible to observe that the response values are higher for rebounders compared with controllers. Supplementary information can be found at the Table E.1 in the Appendix E.

7.2.2 Survival models for time until viral rebound

Survival analysis is used to analyse data in which the outcome is the time until an event of interest (for more information, see Chapter 4). In the BCN02 clinical trial, the event of interest corresponds to viral rebound, and the time until viral rebound is interval-censored, since the moment when the viral load crosses the threshold cannot be observed exactly. The Figure 7.2 shows the Turnbull estimator of the survival function. As can be seen, approximately 30% of the patients do not present a viral rebound in the follow-up time of 32 weeks.

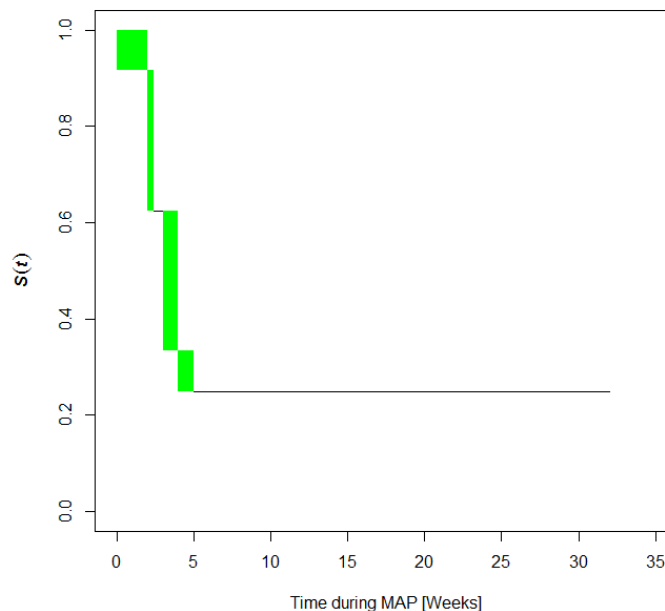


Figure 7.2: Estimation of the survival function of the time to viral rebound in the BCN02 clinical trial obtained with the Turnbull estimator.

Likewise, we fitted univariate Cox proportional hazards models as in Equation (4.2) (Cox, 1972) considering different covariates. To fit the models, we used the function `ic_sp` from the `icenReg` R package (Anderson-Bergman, 2017), since this function allows us to fit a semi-parametric model for interval-censored data. The covariance matrix of the regression coefficients is estimated via bootstrapping. Considering the 95% confidence intervals associated with each covariate, we found none that could explain the time until viral rebound of the patient, i.e., all the intervals include zero (information available at Appendix E, Table E.2).

7.3 Log-binomial regression model to study the patient profile

The log-binomial model, with binomial error and logarithm link function, is proposed as an alternative to the usual logistic regression. Logistic regression uses the odds ratio (OR) as a measure of association. However, when the frequency of the variable of interest is high (e.g., 10% or 20%) in the study population, it tends to increase the magnitude of the association (Szklo & Nieto, 2014). Furthermore, the OR often has a difficult or unintuitive interpretation. Due to this, it is useful to look for an alternative model. The log-binomial model is presented as a good alternative to overcome the pitfalls described. Its main advantage is that its effect size measure is the relative risk (RR) which has a more intuitive interpretation than the OR.

7.3.1 Log-binomial regression model

The log-binomial model (Wacholder, 1986) is a generalized lineal regression model for a binary outcome with the logarithm as a link function. Let Y be a binary outcome and $\mathbf{X} = (X_0, X_1, \dots, X_p)'$ be the set of covariates with $X_0 = 1$. The log-binomial model can be expressed as

$$P(Y = 1|\mathbf{X}) = e^{\beta' \mathbf{X}}, \quad (7.1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ are the parameters of the model.

The observable data set consists of a vector of n independent observations of the outcome, $\{y_1, \dots, y_n\}'$, and a $x \times (p + 1)$ matrix $\mathbf{X} = \{x_{ij}\}$ recording the explanatory-variable values, with x_{ij} being the value of the j th explanatory variable for observation i . A specific feature of the log-binomial model is that the model is consistent to the probability laws only if the parameters satisfy the constraints:

$$\mathbf{X}_i \boldsymbol{\beta} = \beta_0 x_{i0} + \dots + \beta_p x_{ip} \leq 0 \quad \text{for all } 1 \leq i \leq n. \quad (7.2)$$

To estimate the parameters of the model, we maximize the likelihood of the observed data (Savu et al., 2010),

$$L(\boldsymbol{\beta}; \mathbf{y}) = \prod_{i=1}^n e^{y_i \mathbf{X}_i \boldsymbol{\beta}} (1 - e^{\mathbf{X}_i \boldsymbol{\beta}})^{1-y_i}$$

over the restricted parameter space (7.2).

To fit log-binomial regression models, we use the R function `glm`, which fits generalized linear models, specifying the link function and the distribution for the error. This function uses an R process known as “step-halving”. This process controls the iteration, if the iteration tends to be outside the parameter space, it recalculates the adjusted value. It repeats this process over and over as long as it remains a positive value until the adjusted value is negative. By this, we mean that the linear predictor, $\hat{\alpha} + \hat{\beta}' X$, must always be negative, with $\alpha < 0$ always but

there may be $\beta > 0$. This ensures that even if the values do not converge they are always within the parameter space. If the positive values are at the beginning of the estimate, then the adjustment stops and prompts the user for better initial values (Williamson et al., 2013).

7.3.2 Fitted univariate log-binomial regression models

We fitted univariate log-binomial regression models. The binary outcome Y from the Equation (7.1) refers to the patient profile, rebounders or controllers. A patient with VL lower than 2,000 copies/ml at week 12 after MAP is considered controller. The estimated relative risks obtained from the log-binomial models for different covariates analyzed are shown in Figure 7.3. Univariate log-binomial regression models used to detect factors associated with virologic control during MAP revealed that VL before ART initiation (pre-ART VL) was the only factor associated with control of viral rebound after ART interruption, considering a 95% confidence interval. For each log increase of the pre-ART VL, the probability of becoming a controller decreased by 66% (RR: 0.34; 95% CI: 0.14, 0.79). The fitted models, their coefficients, the standard errors associated to each coefficient, and the risk ratio (\widehat{RR}) with its 95% confidence interval can be seen at Table E.3 in Appendix E.

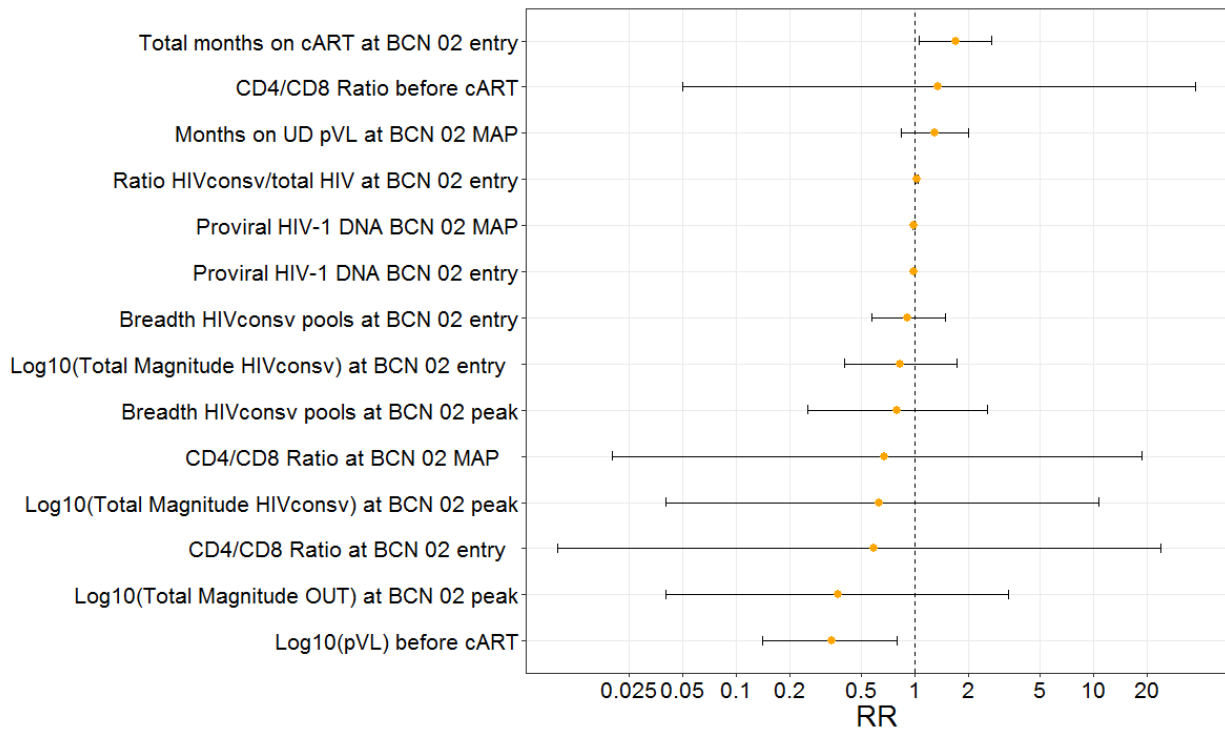


Figure 7.3: Estimate relative risks for a subset of clinically relevant covariates analysed in the univariate log-binomial regression model.

7.4 ELISpot assays for BCN02

In this section, we provide statistical methodologies to analyze Enzyme-Linked Immunosorbent Spot (ELISpot) assays data. Following, we present the definition, the main concepts associated with the ELISpot assays, and the methodology to analyze the data coming from these assays.

7.4.1 What is an ELISpot assay?

The ELISpot assay is an immunoassay that measures the frequency of cytokine-secreting cells at the single-cell level. In this assay, cells are cultured on a surface coated with a specific capture antibody in the presence or absence of stimuli. Proteins, such as cytokines, that are secreted by the cells will be captured by the specific antibodies on the surface. After an appropriate incubation time, cells are removed and the secreted molecule is detected using a detection antibody in a similar procedure to that employed by the Enzyme-Linked ImmunoSorbent Assay (ELISA). By using a substrate with a precipitating rather than a soluble product, the end result is visible spots on the surface (for more details see Figure 7.4). A single cell forms a coloured “footprint” (spot) on the bottom of the well representing its secretory activity.

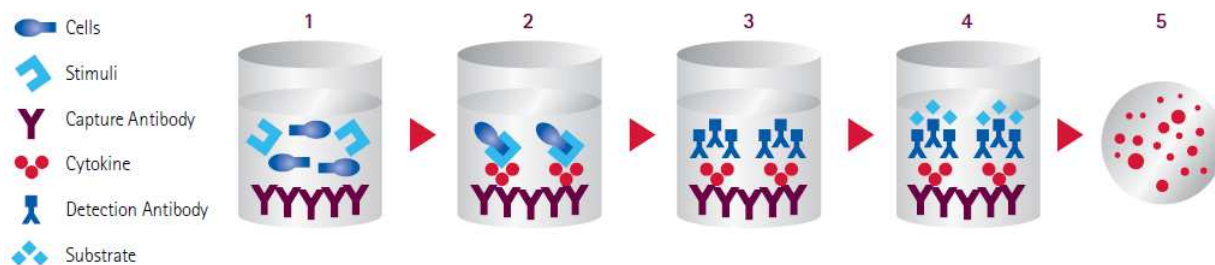


Figure 7.4: The ELISpot assay workflow. (1) Coat membrane with capture antibody. Add immune cells and stimulate. (2) Responding cells produce cytokines. The cytokine of interest binds to the capture antibody beneath the cell. (3) Wash to remove cells. Add a second cytokine-specific biotinylated antibody which binds to the cytokine-antibody complex. (4) Add streptavidin-enzyme conjugate. (5) Add enzyme substrate and develop. Within a well, each responding cell will result in the development of one spot. Source: <https://merckmillipore.com>, accessed: March, 2020).

The ELISpot assay is carried out in a 96-well plate (Figure 7.5 a)), and an automated ELISpot reader is used for the analysis (Figure 7.5 b)). The assay is therefore easy to perform and allows rapid lecture of a large number of samples. The output from the ELISpot reader is a set of images and its corresponding Excel files that includes spot sizes and spot count per well (see Figure 7.5 c1) and c2)).

While ELISpot assays allows one to directly visualize and count extremely low frequencies of cytokine secreting T cells amongst millions of non-secreting bystander cells, the interpretation of ELISpot data can become ambiguous when (a) spot numbers in antigen-containing wells are low, (b) spot counts in negative control wells are elevated, and particularly (c) when both of the above occur simultaneously. Thus, the primary task, before any statistical analysis, must be the optimization of the basic assay parameters and reagents such that the assay yields low background signal in the negative control wells, and the maximal number of antigen-induced spots in test wells, i.e., the signal to noise ratio is maximized (Dittrich & Lehmann, 2012).

For this analysis, our main variables of interest are spot counts and spot size. Spot count corresponds to cytokine-producing cell numbers. Due to diffusion properties, a true spot has a densely colored center which fades toward the edges; the size or color intensity of the spots is determined by the amount of cytokine released. Spot size refers to relative amounts of cytokine produced per cell.

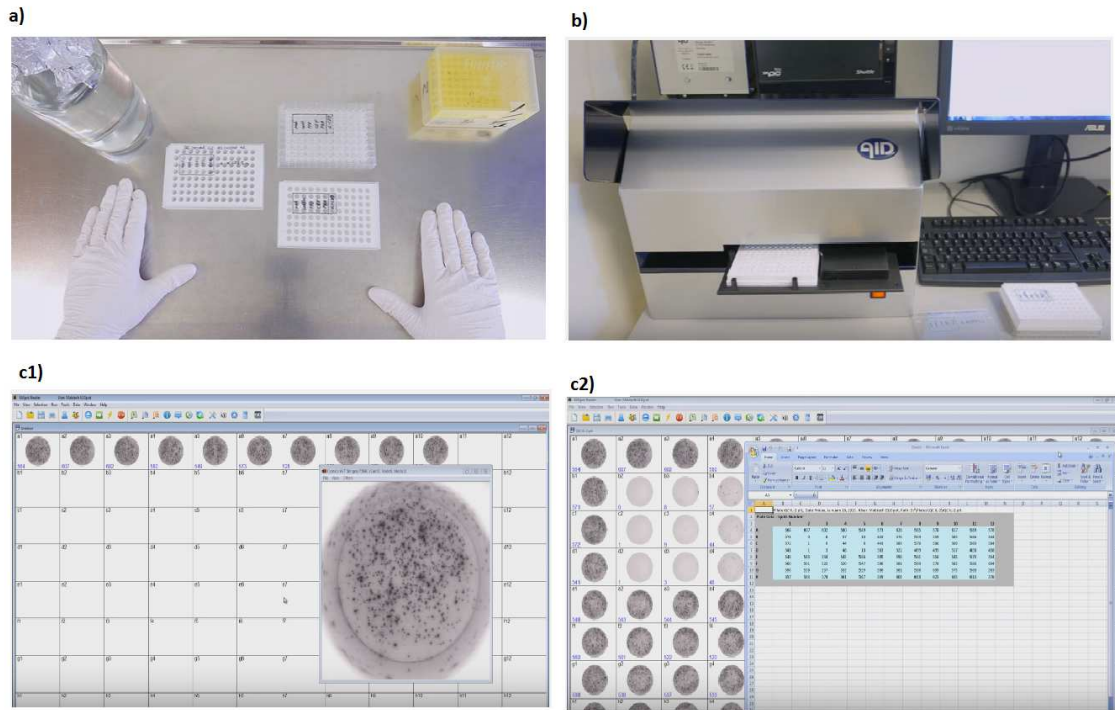


Figure 7.5: The ELISpot assay tools (Mabtech, 2015). (a) 96-well plate, b) ELISpot reader, c1) ELISpot image output and c2) ELISpot excel file output.

7.4.2 Plate organization and settings for BCN02

The plate is an array of 8 rows with 12 wells in each. Wells are organized from columns 1 to 12 and from rows A to H. The half of the 96-well plate disposition for this study can be seen in Figure 7.6. For this study, half a plate corresponds to a different time point. There are six different time points: MVA₁ (week 0), MVA₁+1 (week 1), MVA₁+3 (pre-RMD, week 3), MVA₂ (week 9), MVA₂+1 (week 10) and MVA₂+4 (week 13). Every well contains a pool of peptides, as can be seen in Figure 7.6.

In the plate (Figure 7.6), we observe the target of the most conserved regions of the HIV-1 proteome (HIV-consv), from A1 to A6 and its replicates from B1 to B6, which have the potential to enhance host immune control and facilitate clearance of the latent reservoir. Moreover, we observe other components such as OUT (the non-conserved regions of HIV-1, from C1 to C6 and E1 to E6 and its corresponding replicates from D1 to D6 and F1 to F6, respectively), HTI (cells G1 to G5), negative and positive controls. Negative controls consist of cells cultured without stimuli, whereas T-cell activator are commonly used as positive controls. The last ones are used to confirm cell and assay functionality, these often include phytohemagglutinin (PHA) and CEF (Cytomegalovirus, Epstein-Barr virus, influenza virus) peptide pools, that induce secretion of many common cytokines.

HIVconsv immunogen codes for the 14 most conserved regions of the consensus HIV-1 Gag, Pol, Vif, and Env proteins and can be seen in Figure 7.7. HIVconsv immunogen is vectored by modified vaccinia virus Ankara, MVA.HIVconsv (Mothe, 2016).

The description of peptide pools for HIVConsv, OUT and HTI can be observed in Tables E.4, E.5 and E.6 from Appendix E, respectively.

	1	2	3	4	5	6
A	p1	p2	p3	p4	p5	p6
B	p1	p2	p3	p4	p5	p6
C	Gag-1	Gag-2	Pol-1	Pol-2	Pol-3	V-T
D	Gag-1	Gag-2	Pol-1	Pol-2	Pol-3	V-T
E	Env-1	Env-2	Env-3	Env-4	Nef	Acc
F	Env-1	Env-2	Env-3	Env-4	Nef	Acc
G	Gag-p1	Gag-p2	Pol-p1	Pol-p2	Vif-Nef	CEF
H	NEG	NEG	NEG	NEG	PHA	CEF

HIVconsv
 OUT
 HTI
 Positive
 Negative

Figure 7.6: Pool of peptides organization in half a plate for BCN02.

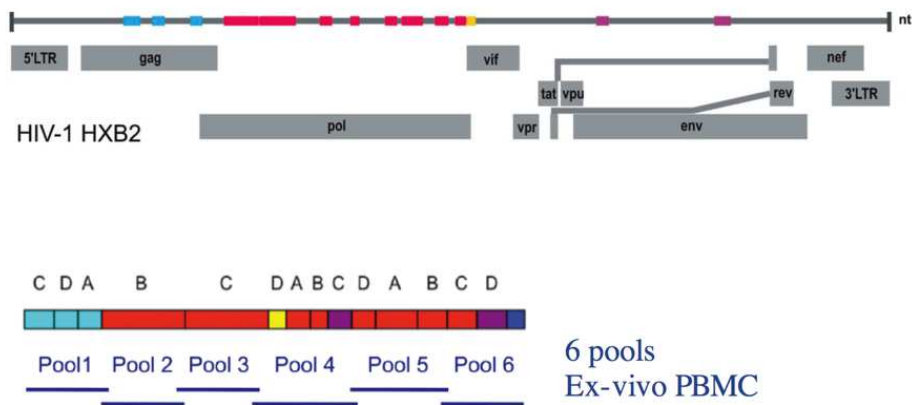


Figure 7.7: Schematic representation of the selected conserved regions in the HIV proteome from different HIV-1 clades included in the HIVconsv immunogen and distribution of the set of 6 peptide pools for ELISpot assay. Source: [Mothe \(2016\)](#).

On average, T-cell ELISpot counts show linearity for PBMCs in the range of 100,000- 800,000 cells. Where possible, cells should be serially diluted and plated in triplicate. However, given the restrictions of well size in 96-well plates, seeding more than 400,000 cells per well may result in overcrowding and cell stacking. In this study, HIVconsv and OUT pool of peptides were plated in two replicates. Most of the cells were seeding with 100,000 cells (see Table E.7 in Appendix E)

7.5 Statistical methodologies to work with spot counts and spot size from ELISpot assay data

When working with data from ELISpot assays, we have two main objectives. The first one is to build a routine in R to read and sort the data. This is because the ELISpot reader currently delivers outputs in Excel that are then manually manipulated. The second goal is to study the distribution of spot size and spot count over time, its variability, and its relationship. At this point, it is necessary to mention that, as far as we know, there are not studies that analyze the spot size or its relationship with the spot counts.

7.5.1 Data management from BCN02 ELISpot assay

ELISpot data for this specific assay was processed using an ELISpot reader (Software version: ImmunoSpot 5.1.36). The software of this reader allows for data extraction via Excel, as we mentioned before. Every patient and every complete plate (that considers two different timepoints) are represented by a unique Excel sheet. The data is organized in different outputs within the same Excel sheet, as can be seen in Table 7.3.

Table 7.3: Type of outputs from the ImmunoSpot reader.

Type of output	Description	Dimension
Spot counts	Absolute frequency of spot counts	8×12
Mean spot sizes	Area of each spot (in 10^3 mm^2)	8×12
Histogram distribution	Distribution according to its size (in $\log \text{ mm}^2$). The histogram considers from -3 to 2 by 0.2 units	96×26
Well areas covered	Percentage of the well covered by spot. Not used for this study.	8×12
Total intensity of all foreground object per well	Total intensity foreground after removing noise. Not used for this study.	8×12

At this moment, to work with these type of data, clinicians need to summarize the information from every Excel sheet manually. To help in this work, and considering this is a typical type of data in clinical assays similar to BCN02, we create a routine to read all the Excel sheets and then perform the analyses with R. The raw data is available at <http://doi.org/10.5281/zenodo.3870744> and the R script and its environment can be accessed in <http://doi.org/10.5281/zenodo.3870750>.

We need to mention that we decide to work with the data coming from the histogram distribution, since this output shows the most raw data version. In these Excel sheets it is common to find a discrepancy in the counts

coming from the histogram distribution and the one from the spot counts output. This is due to the fact that when obtaining the counts, the ImmunoSpot reader selects only the 90% percent of the well to count. The reason to use this percentage is to avoid the accumulation of material on the edges. Once the spot count is obtained in the selected area, the software extrapolates to have the 100%. We decide to work with the 90% without any extrapolation, in this way we know the counts and the distribution of each count regarding its size.

Furthermore, once we have the spot counts, this data need to meet the previously established positivity criteria based on the background definition. The background corresponds to the average of the negative control counts. The positivity criteria is defined as the maximum of a) greater than 3 times the background, b) greater than background plus 3 standard deviations or c) greater than 5 spots. If the positivity criteria is not met, the count is considered to be zero as well as the corresponding size.

7.5.2 Distribution of spot size and spot count over time and its variability

To study the relation between spot counts and size, it is necessary to consider the corresponding replicates (if any) in different scenarios. The six different scenarios are given by the combination of the HIV region (HIVconsv, OUT and HTI) and the corresponding variable of interest (spot count or spot size). We obtained the mean and standard deviation in each scenario across the 6 different timepoints previously explained. Depending on each case we will compute the mean and standard deviation.

The standard deviation deserves a especial mention, since depending on the situation, we must calculate it in different ways. In the case of spot count and any region (IN/OUT/HTI) we obtain the standard deviation in the usual way. In the case of spot size and HTI region (without replicates), to obtain the standard deviation we need to use an unbalanced one-way random effects ANOVA model, that consider the effect of each patient. Finally, in the case of spot size and IN/OUT region (with replicates), to obtain the standard deviation we used unbalanced two-way random effects ANOVA models, that consider the effect of the patient and the well. We explain these scenarios in greater detail below.

Usual mean and standard deviation for spot count

Considering the covariate spot count, we have two situations, as can be seen in Figure 7.8. The first situation is without replica, as in the case of HTI region in the ELISpot plate. In the figure, we observe the case of three specific patients. The summary for a specific timepoint and pool of peptides is the frequency or the number of spots (spot count). The second situation is with replica, for the HIVconsv (IN) and OUT region. In each of these situations, we can obtain the mean and standard deviation for spot counts in the usual way.

Unbalanced one-way random effects ANOVA model

The second scenario in the ELISpot from the BCN02 trial is the case of the study of the variability of the spot size variable considering the HTI region, without replicates, as represented by the Figure 7.9. In the figure we present the hypothetical situation for three different controller patients. In this case, each well has many spots, and we are interested in obtaining the variability of the spot size for a specific timepoint and pool of peptides. In this case, it is clear that we need to control for the patient. The way of doing this is considering the unbalanced one-way random effects ANOVA model.

This section follows the Chapter 11 of Analysis of variance for random models, volume II: Unbalanced data (Sahai & Ojeda, 2004). The random effects model for the unbalanced one-way classification is given by

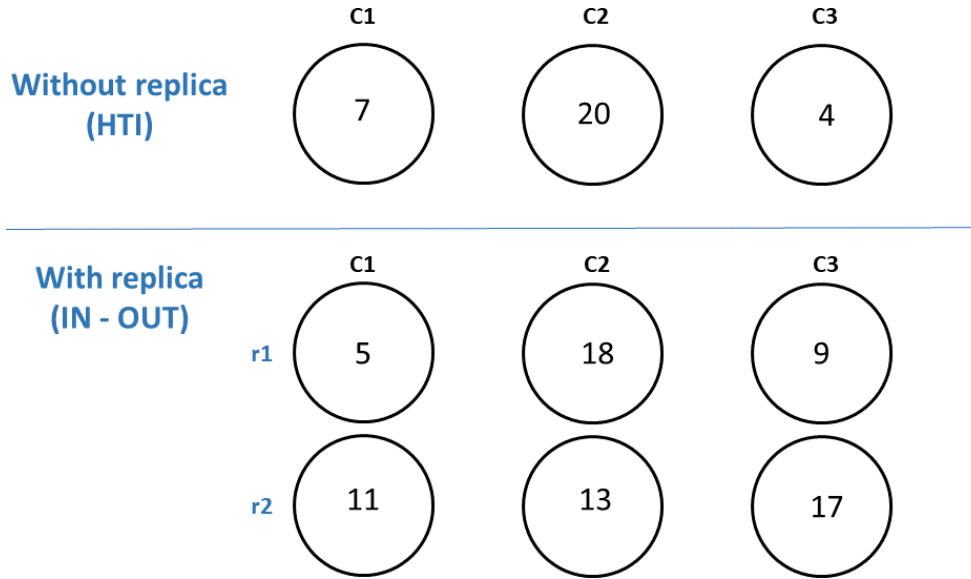


Figure 7.8: Scenario 1: spot count with or without replicate.

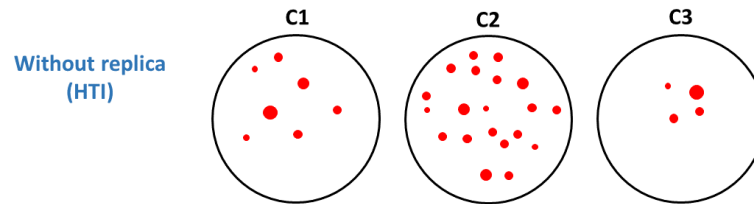


Figure 7.9: Scenario 2: spot size without replicate.

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, a; \quad j = 1, \dots, n_i, \quad (7.3)$$

where x_{ij} is the j th observation in the i patient, μ is the overall mean, α_i is the random effect of the i th patient factor and ϵ_{ij} is the error term. It is assumed that $-\infty < \mu < \infty$ is a constant, and α_i s and ϵ_{ij} s are mutually and completely uncorrelated random variables with zero means and variances σ_α^2 and σ_ϵ^2 , respectively. Here, σ_α^2 and σ_ϵ^2 are known as the components of variance.

Analysis of variance estimators

The analysis of variance (ANOVA) method of estimating variance components σ_ϵ^2 and σ_α^2 consists of equating observed values of the mean squares MS_B and MS_W to their expected values, and solving the resulting equations for σ_ϵ^2 and σ_α^2 . The estimators thus obtained are

$$\begin{aligned} \hat{\sigma}_{\epsilon, ANOV}^2 &= MS_W \\ \hat{\sigma}_{\alpha, ANOV}^2 &= \frac{MS_B - MS_W}{n_0}, \end{aligned} \quad (7.4)$$

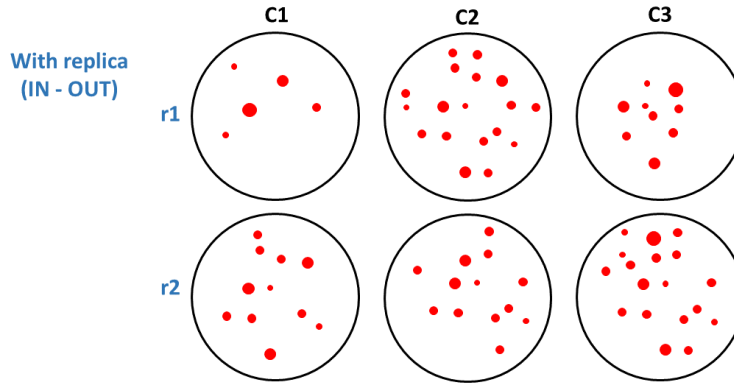


Figure 7.10: Scenario 3: spot size with replicate.

where $MS_W = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$, $MS_B = \sum_{i=1}^a n_i (\bar{x}_i - \bar{x})^2$, and $n_0 = (N^2 - \sum_{i=1}^a n_i^2) / N(a-1)$.

Unbalanced two-way random effects ANOVA model

The third scenario corresponds to the study of the variability of the spot size variable considering the HIVconsV (IN) and OUT region, considering the replicates, as can be seen in the Figure 7.10. In the figure, we present the hypothetical situation for three different controller patients. In this case, each well has many spots but also each well has its replica, and we are interested in obtaining the variability of the spot size for a specific timepoint and pool of peptides. As can be seen we need to control by patient and also by replicate, so in this case, we apply an unbalanced two-way random effects ANOVA model.

This section follows the Chapter 12 of Analysis of variance for random models, volume II: Unbalanced data (Sahai & Ojeda, 2004). In this section, we consider the random effects model involving two factors, the patient (A) and the replica (B), in a factorial arrangement where the numbers of observations in each well are different. We further assume that the model does not involve any interaction terms. Consider the two factors A and B and let there be $n_{ij} \geq 0$ observations corresponding to the (i, j) th cell. The model is given by

$$x_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}; \quad i = 1, \dots, a; \quad j = 1, \dots, b; \quad k = 0, \dots, n_{ij}, \quad (7.5)$$

where x_{ijk} is the k th observation corresponding to the i th level of factor A (the patient) and the j th level of factor B (the replica), μ is the overall mean, α_i s and β_j s are random effects, i.e., α_i is the effect of the i th level of factor A , β_j is the effect of the j th level of factor B , and ϵ_{ijk} is the error term. It is assumed that $-\infty < \mu < \infty$ is a constant and α_i s, β_j s, and ϵ_{ijk} s are mutually and completely uncorrelated random variables with means zero and variances σ_α^2 , σ_β^2 , and σ_ϵ^2 , respectively. The parameters σ_α^2 , σ_β^2 , and σ_ϵ^2 are known as the variance components.

For the model in (7.5) there is no unique analysis of variance. The conventional ANOVA obtained by an analogy with the corresponding balanced design is given in Table 7.4

Analysis of variance estimators

The analysis of variance or Henderson's Method I (Henderson, 1953) for estimating variance components is to equate the sums of squares or mean squares in Table 7.4 to their respective expected values. The resulting

Table 7.4: Analysis of variance for the model in (7.5).

Source of variation	Degrees of freedom	Sum of squares	Mean square	Expected square mean
Factor A	$a - 1$	SS_A	MS_A	$\sigma_\epsilon^2 + r_5\sigma_\beta^2 + r_6\sigma_\alpha^2$
Factor B	$b - 1$	SS_B	MS_B	$\sigma_\epsilon^2 + r_3\sigma_\beta^2 + r_4\sigma_\alpha^2$
Error	$N - a - b + 1$	SS_E	MS_E	$\sigma_\epsilon^2 + r_1\sigma_\beta^2 + r_2\sigma_\alpha^2$

equations are

$$\begin{aligned}
SS_A &= (N - k_1)\sigma_\alpha^2 + (k_3 - k_2)\sigma_\beta^2 + (a - 1)\sigma_\epsilon^2 \\
SS_B &= (k_4 - k_1)\sigma_\alpha^2 + (N - k_2)\sigma_\beta^2 + (b - 1)\sigma_\epsilon^2 \\
SS_E &= (k_1 - k_4)\sigma_\alpha^2 + (k_2 - k_3)\sigma_\beta^2 + (N - a - b + 1)\sigma_\epsilon^2.
\end{aligned} \tag{7.6}$$

The variance components estimators are obtained by solving the equations in (7.6) for σ_α^2 , σ_β^2 and σ_ϵ^2 . The estimators thus obtained are given by

$$M = \begin{bmatrix} \hat{\sigma}_{\alpha, ANOV}^2 \\ \hat{\sigma}_{\beta, ANOV}^2 \\ \hat{\sigma}_{\epsilon, ANOV}^2 \end{bmatrix} = \begin{bmatrix} N - k_1 & k_3 - k_2 & a - 1 \\ k_4 - k_1 & N - k_2 & b - 1 \\ k_1 - k_4 & k_2 - k_3 & N - a - b + 1 \end{bmatrix}^{-1} \begin{bmatrix} SS_A \\ SS_B \\ SS_E \end{bmatrix}. \tag{7.7}$$

Further simplification of (7.7) yields

$$\begin{aligned}
\hat{\sigma}_{\epsilon, ANOV}^2 &= \frac{\theta_1(SS_E + SS_A) + \theta_2(SS_E + SS_B) - (SS_E + SS_B + SS_A)}{\theta_1(N - b) + \theta_2(N - a) - (N - 1)} \\
\hat{\sigma}_{\beta, ANOV}^2 &= \frac{SS_E + SS_B - (N - a)\hat{\sigma}_{\epsilon, ANOV}^2}{N - k_3} \\
\hat{\sigma}_{\alpha, ANOV}^2 &= \frac{SS_E + SS_A - (N - b)\hat{\sigma}_{\epsilon, ANOV}^2}{N - k_4},
\end{aligned} \tag{7.8}$$

where

$$\theta_1 = \frac{N - k_1}{N - k_4} \quad \text{and} \quad \theta_2 = \frac{N - k_2}{N - k_3}.$$

7.5.3 Results

In this section, we present the results for the HIVconsV (IN) region. The results for the OUT and HTI region can be found in Appendix E.2. First, we present the profiles of spot counts for the different pool of peptides (from p1 to p6) across time.

In the Figure 7.11 we observe some differences between controller (C) and rebounders (R) patients. In the case of the pool of peptides p1, the profile is superior for controllers. The opposite situation is observed for the pool p3. This mean that the pool p1 elicits higher response for controller patients and the pool p3 generates higher response for rebounder patients.

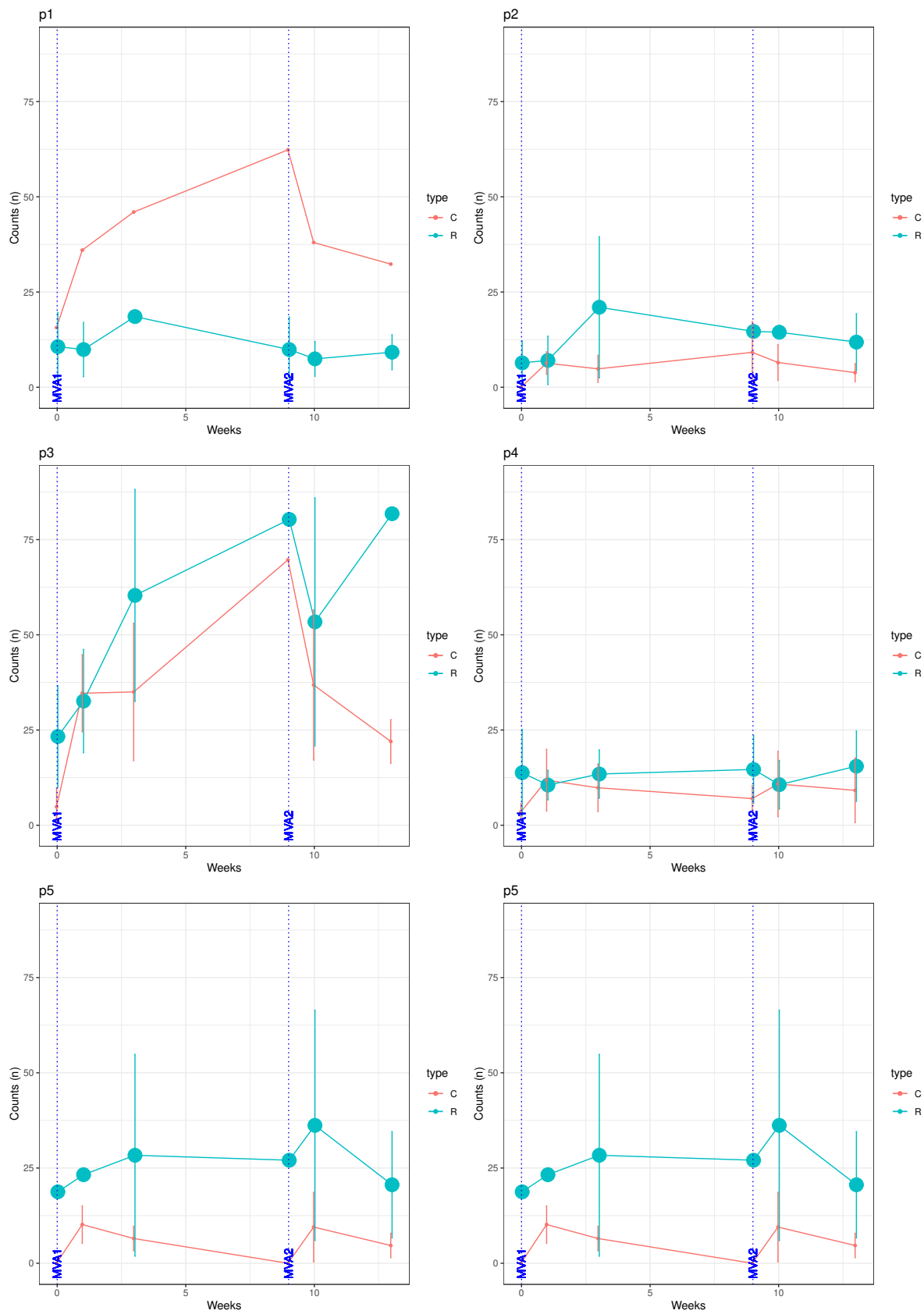


Figure 7.11: Mean and 95% confidence interval for spot counts in HIVconsv region for each pool of peptides and each patient profile (C: Controller; R: Rebounder).

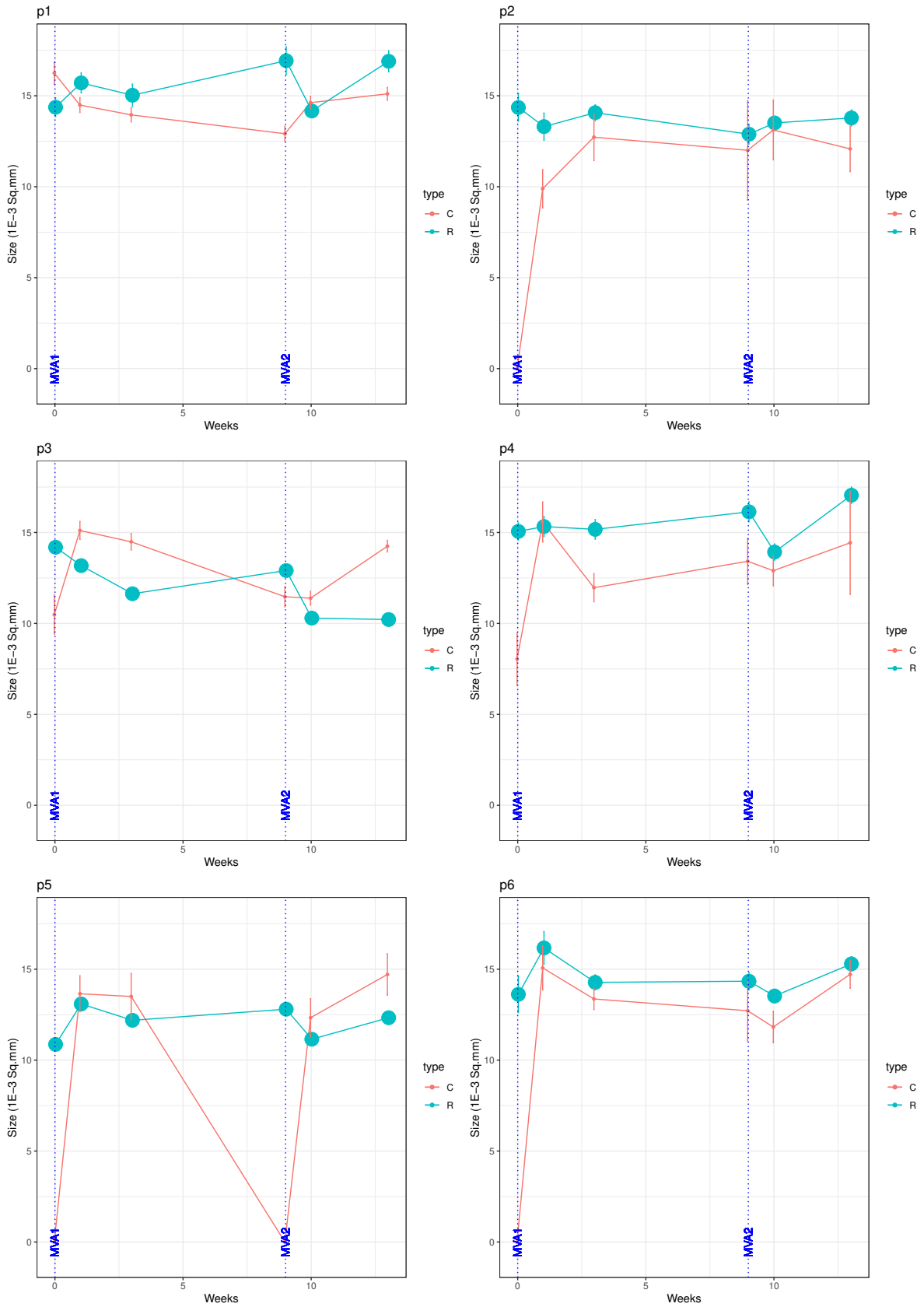


Figure 7.12: Mean and 95% confidence interval for spot size in HIVconsv region for each pool of peptides and each patient profile (C: Controller; R: Rebounder).

Looking at the Figure 7.12, we cannot distinguish a clear difference between the profile of the patients, controllers and rebounders across time in any of the pool of peptides for the HIVconv region. It is important then to have a descriptive approach for the two covariates, spot count and spot size, without considering the patient profile and to observe if the different curves across time are different or not.

From the Figure 7.13, we observe the two variables do not present the same profile across time. This is probably, because these two variables are not given the same information about the patients across time. It is important then to study the possible correlation between spot count and spot size. As can be seen in Figure 7.14 the Spearman correlation between these variables is weak $\rho = 0.11$ (p-value=0.0008) and there is no clear difference between different patient profile (controller and rebounder).

7.6 Discussion

In this chapter, we have presented the different nature of the different types of data in a clinical trial, and in this way, the different statistical methodologies applied to obtain clinical conclusions. The clinical trial presented, BCN02, continues in the line with all the work presented so far, considering data coming from HIV-1-infected patients.

We have presented the clinical trial and the description of their patients, according to different covariates of interest. We have applied univariate models for the interval-censored time until viral rebound and we did not find any covariate to explain this outcome.

To study the patient profile, since this is a binary variable we have fitted univariate log-binomial regression model. We have presented this model as an alternative to the logistic regression model. The log-binomial model is useful when we want to estimate the relative risk in a study with common results in the population. We believe that there is still a lot of new information to contribute about the log-binomial regression model, for example, regarding the convergence problems that can appear. However, this model offers a good solution to obtain estimates of the relative risk in a cohort study and the prevalence ratio in a cross-sectional study, and a easier interpretable measurement, compared with the OR from the usual logistic regression model.

For the survival analysis and for the log-binomial fitted models we have considered only univariate approaches because of the low sample size for this study. Only 15 patients, 3 of them classified as Controllers, 10 Rebounders, one that never stop the antiretroviral treatment and one that was consuming unauthorized drugs. An extension of this study will be conducted, the BCN03, and the sample size will be bigger, this will allow for considering more than one covariate in the survival analysis as in the log-binomial regression models.

Another important aspect we have addressed in this clinical trial was working with data from ELISpot assays. We developed an R routine to read and sort the data coming from the Excel sheets from the ImmunoSpot reader. In these assays, we worked with two main variables: the spot count (already considered in previous studies) and the spot size (a variable, as far as we know, never considered). An interesting conclusion from the results is that these two variables do not present the same type of information since their profiles across time are different. However, the clinical information provided by the spot size is still unknown. At this moment, clinicians are considering to work with rapid ELISpot, in which a big percentage of the spots appear in the first hours of incubation, so that it is less time-consuming. Another path of research, that can be addressed from a statistical point of view is to create a Shiny App to read and sort the data from the ImmunoSpot reader. This would make posterior analyses easier, avoiding manually manipulation of the Excel sheets.

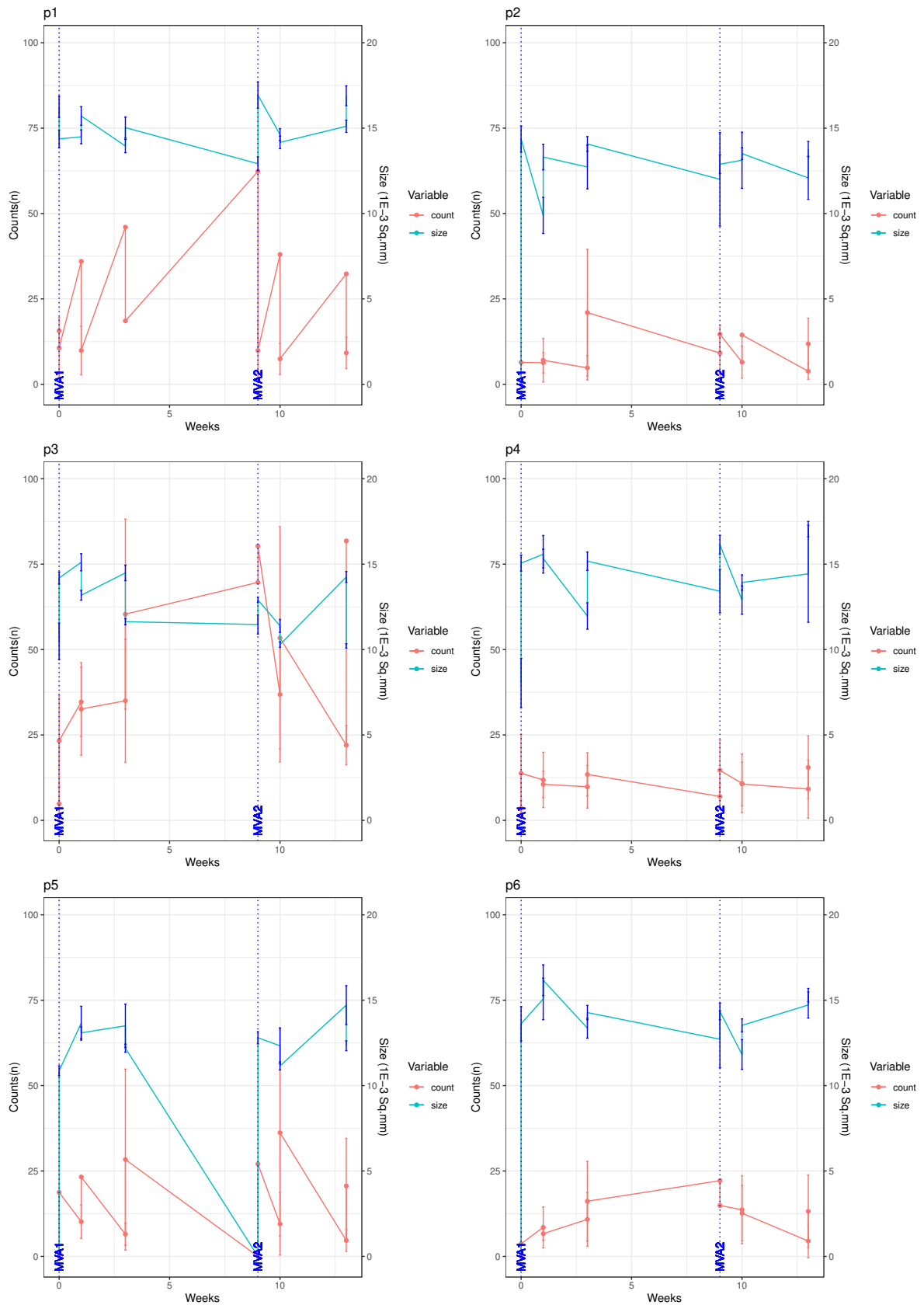


Figure 7.13: Mean and 95% confidence interval for spot size and spot count in HIVconsv region for each pool of peptides.

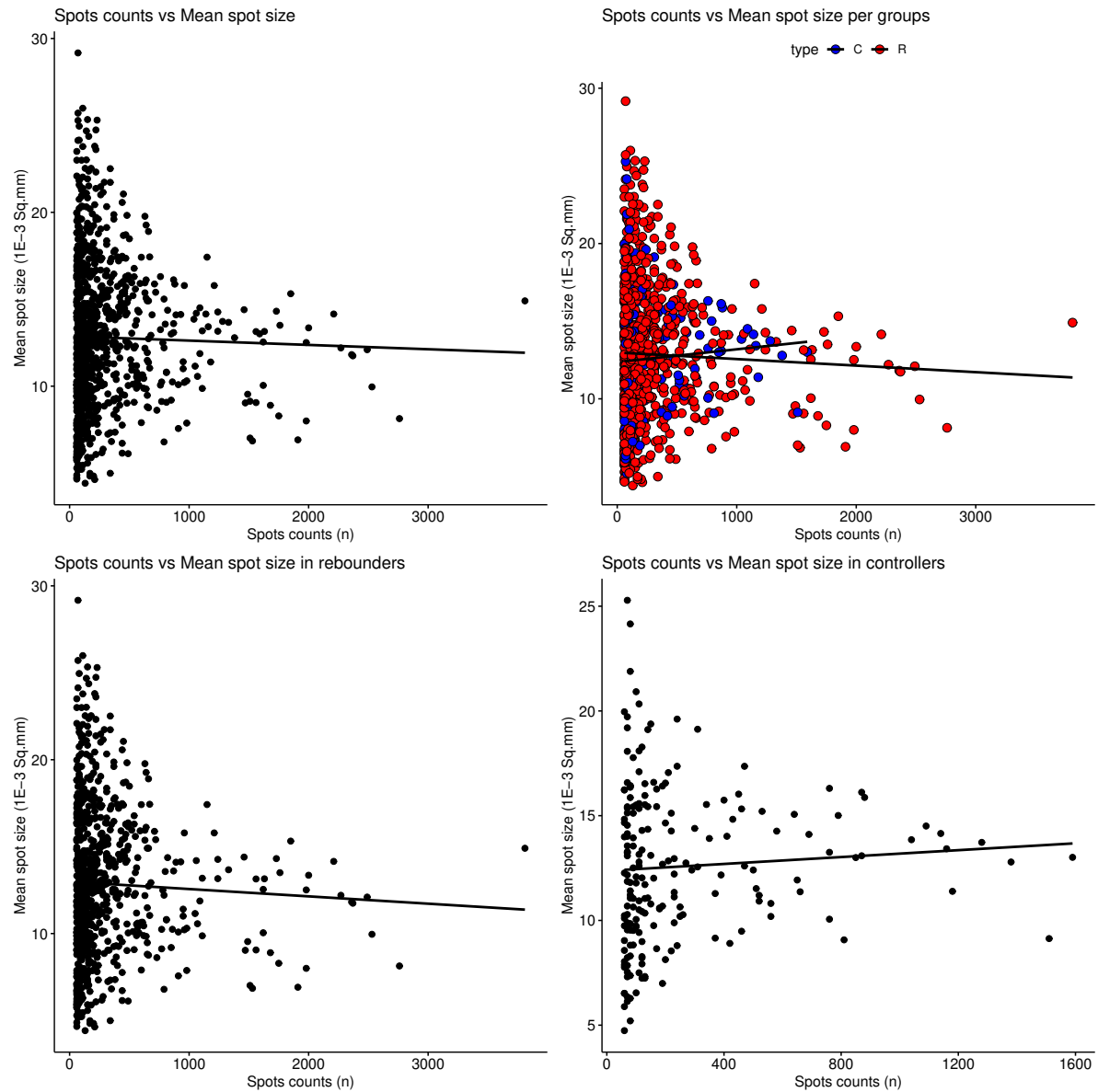


Figure 7.14: Spearman correlations between spot size and spot count in HIVconsV region considering all the pool of peptides and all the timepoints. C: controller, R: rebounder.

CONCLUSIONS AND FURTHER RESEARCH

This thesis presents different statistical approaches to handle diverse biological and clinical questions that arise in the development of a therapeutic vaccine for HIV. It has been essential for us to study these problems from the perspective of Data Science, working hand in hand with a multidisciplinary team.

As Data Science is a relatively new discipline, there is a lack of definitions directly linked to Biomedicine. However, we found that “Data Science” is an increasingly searched term (Chapter 2). For this reason, we propose to establish a proper definition of Data Science in Biomedicine. One key aspect of Data Science in Biomedicine is that multidisciplinary teams involved in a clinical trial need to use a common language that can be understood by all members of the team. In Chapter 4, we have presented the vocabulary used in HIV studies. We have also explained some common methods used along this work, such as survival and omics data analysis (Chapter 5).

In this thesis we have presented different statistical approaches to address the questions that arise from three independent clinical trials based on different type of data. The questions related to the first two clinical trials gave rise to the development of two important methodological contributions of this thesis: the application of the elastic-net penalization to the accelerated failure time (AFT) model and the fit of a mixed effects Cox model with interval-censored data. The first question was to identify biomarkers (from mRNAs) of the viral rebound in the DCV2 clinical trial (Chapter 5). In these studies with high dimensional data it is common that the number of covariates is greater than the number of subjects in the study. Besides, in omic layers such as transcriptome (mRNA) there is a correlation structure that need to be considered. To address this problem, we have proposed the use of an elastic-net approach for the accelerated failure time model, which allows us to work with high-dimensional data and consider the grouping effect. The second main clinical question we studied was whether gender and pre-cART viral load were risk factors on the time until viral rebound in the Analytical Treatment Interruption (ATI) studies (Chapter 6). In order to handle the variability due to the fact that some patients had more than one assessment, we have proposed a method to estimate the parameters of a mixed effects Cox model with interval-censored data. The third question was related to improving the statistical analysis used in the BCN02 clinical trial (Chapter 7) in two scenarios: when using ELISpot data and when defining a dichotomous variable for responders and controllers. Here, we did not develop a new methodology but applied existing techniques

that were not used in this area. Following, we present our main findings and methodological contributions, as well as our limitations and further research.

Identifying mRNA biomarkers related to the time to viral rebound is of critical importance. The knowledge on whether some biomarkers are associated with higher or lower risk of viral rebound can help the development of the therapeutic HIV vaccine. For this purpose, we proposed the use of an accelerated failure times (AFT) model using an elastic-net penalization approach to estimate the parameters that considers the correlation structure among mRNAs and its high-dimensionality nature. We have derived the expression of the penalized log-likelihood and maximized it using two ad-hoc methods under the assumption that time to viral rebound follows a Weibull distribution. When applying these ad-hoc methods to the DCV2 study, we found different sets of predictors according to the method used. With the first approach (Approach A, Section 5.5.4), we identified 5 mRNAs (PPP1R9A, LOC100509457, IL21R, CYP1B1, and DUSP4) as possible predictors of the time to viral rebound. According to the literature, four of these (all except the LOC100509457) are already related to some diseases presented in HIV-infected patients. This work, however, presents several limitations. First, the computational time of our first approach algorithm, two weeks, is highly demanding and needs to be improved (for Intel(R) Pentium(R) CPU 3825U 1.90 GHz, RAM: 8 GB, and operating system of 64 bits). Moreover, a simulation study is also needed to describe the properties of our ad-hoc method, but at this point is unfeasible for the time of computation. To achieve less time-consuming results, the use of a more powerful computer or the parallelization of this problem should be explored; these points, however, are beyond of the scope of this thesis.

For the second approach (Approach B, Section 5.5.4), we have used the `iregnet` package considering the interval-censored times to viral rebound and using the midpoint of these intervals. For the complete dataset, the coordinate descent method did not converge, this is why we splitted the data into 5 disjoint randomly assigned subsets and selected the best predictors in each model to finally fit a new model that considered the previously selected mRNAs. The set of final predictors in each model, considering interval censoring or midpoint imputation, are different between them and also different from the previously selected with our implemented ad-hoc algorithm using Nelder-Mead optimization. Moreover, none of these selected predictors, based on the scientific literature, seems to have a connection with the HIV field. The intuition in this line says that splitting the whole set of mRNAs in disjoint subsets we are ignoring the biological structure.

This chapter opens the door to a range of research lines. It would be of interest to study the maximization of the likelihood function of the proportional hazards model using the elastic-net penalization. We could add to this the study of other optimization algorithms, such as the coordinate descent algorithm, which is implemented in the `glmnet` and `iregnet` packages. Moreover, the possibility to take into account other regularization techniques such as adaptive lasso, adaptive elastic-net, and smoothly clipped absolute deviation (SCAD), among others, could be object of further investigation. Another critical point that can be seen in Figure 5.4 is that we have dealt with intervals that have length equal to two weeks or four weeks. An application that considers more random interval-censored times to viral rebound could help to better study the performance of our methods. Finally, another aspect to be addressed corresponds to the incorporation of micro RNAs (miRNAs) to the model used. This addition is not trivial because of the correlation structure between mRNAs and miRNAs. The relationship between them is not bijective, that is, one miRNA can regulate more than one mRNA and, at the same time, one mRNA can be regulated by more than one miRNA.

As a first step to close the gap between the mixed effects Cox model using right-censored data and the Cox model that considers interval censoring, we developed a multiple imputation approach for interval-censored time to HIV RNA viral rebound within a mixed effects Cox model. After that, we conducted a simulation study considering 36 different scenarios to examine the properties of the obtained estimators. The results of this simu-

lation showed that our method has desirable properties such as low bias and low minimum squared error (MSE). We are planning new simulation studies in order to explore the estimators properties under additional scenarios with different fixed and random effects. Machine learning is a complementary different approach that could be used for multiple imputation (for more information, see [Brownlee \(2020\)](#)). As a limitation we based our imputation step on the truncated Weibull distribution, however this distribution is flexible enough and allow its hazard function have different shapes. Another limitation is that we cannot compare nested mixed Cox models when using interval-censored data. To overcome this we suggest the development of guidelines for the comparison of nested mixed Cox models. Apart from this, our main future goal is the parameter estimation of a mixed effects Cox model with interval-censored data without resorting to multiple imputation.

In our last research contribution we have presented different statistical methodologies applied to the BCN02 clinical trial. For our first goal, the identification of variables that explain the profile of the patient, we have proposed a log-binomial regression model. The main advantage of this model as compared to the classical logistic regression model is the more intuitive interpretation of the risk ratio rather than the odds ratio, which tends to exaggerate the magnitude of the association between exposure and outcome. The main limitation is related to the small sample size that allowed us to fit only univariate models. In the next study, BCN03, we will have a larger sample size to study the combination of different variables to explain the response. For our second goal, the analysis of ELISpot assays data, we have used an unbalanced one/two way ANOVA to obtain the correct variance estimation for the spot size and count, the main variables from these assays. The lack of prior research studies on the topic can be seen as a limitation. We believe that the development of an R Shiny App could help read and sort the data as well as to avoid the manual manipulation of the Immunospot Excel sheet outputs. Another future aspect of study could be the identification of new clinical information from the spot size, since now we know that it provides different information than spot counts, using rapid ELISpot.

To sum up, in this thesis, we have both developed new statistical methodology and applied standard statistical methods to answer important questions related to the therapeutic vaccine for HIV. The results obtained are important and useful, according to the clinicians involved in the HIV studies described and their multidisciplinary teams. As can be seen in this thesis, Data Science plays and will continue playing a major role in the path to the cure of HIV.

BIBLIOGRAPHY

- Alarcón-Soto, Y., Langohr, K., Fehér, C., García, F., & Gómez, G. (2019). Multiple imputation approach for interval-censored time to HIV RNA viral rebound within a mixed effects cox model. *Biometrical Journal*, *61*(2), 299–318.
- Allers, K., Hütter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E., & Schneider, T. (2011). Evidence for the cure of HIV infection by CCR5 Δ 32/ Δ 32 stem cell transplantation. *Blood*, *117*(10), 2791–2799.
- Altman, L. K. (1981). Rare cancer seen in 41 homosexuals. *New York Times*, *3*, A20.
- Amaro, R. E. (2016). Drug discovery gets a boost from data science. *Structure*, *24*, 1225–1226.
- Anderson-Bergman, C. (2017). icenReg: Regression models for interval censored data in R. *Journal of Statistical Software*, *81*, 1–23.
- Barré-Sinoussi, F., Chermann, J.-C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., ... Montagnier, L. (1983). Isolation of a t-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, *220*(4599), 868–871.
- Bebchuk, J. D., & Betensky, R. A. (2000). Multiple imputation for simple estimation of the hazard function based on interval censored data. *Statistics in Medicine*, *19*, 405–419.
- Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*, 1713–1723.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.
- Bhavnani, S. P., Muñoz, D., & Bagai, A. (2016). Data science in healthcare: implications for early career investigators. *Circulation: Cardiovascular Quality and Outcomes*, *9*, 683–687.
- Bignon, A., Régent, A., Klipfel, L., Desnoyer, A., de la Grange, P., Martinez, V., ... Balabanian, K. (2015). DUSP4-mediated accelerated T-cell senescence in idiopathic CD4 lymphopenia. *Blood, The Journal of the American Society of Hematology*, *125*(16), 2507–2518.

- Bland, J. M., & Altman, D. G. (1994). Correlation, regression, and repeated data. *British Medical Journal*, 308, 896.
- Bolker, B., & R Development Core Team. (2020). *bbmle: Tools for general maximum likelihood estimation* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=bbmle> (R package version 1.0.23.1)
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63, 1059–1078.
- Broström, G. (2019). *eha: Event history analysis* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=eha> (R package version 2.8.0)
- Brownlee, J. (2020). *Data preparation for machine learning: Data cleaning, feature selection, and data transforms in python*. Machine Learning Mastery.
- Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA Journal of Applied Mathematics*, 6(3), 222–231.
- Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19), 2363–7.
- Chase, J. A. D., & Vega, A. (2016). Examining health disparities using data science. *Research in Gerontological Nursing*, 9, 106–107.
- Chen, D.-G., Chen, J., Lu, X., Grace, Y. Y., & Yu, H. (2016). *Advanced statistical methods in data science*. Springer.
- Cleveland, W. S. (2014). Data science: An action plan for expanding the technical areas of the field of statistics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7, 414–417.
- Clumeck, N., Mascart-Lemone, F., De Maubeuge, J., Brenez, D., & Marcelis, L. (1983). Acquired immune deficiency syndrome in black africans. *Acquired immune deficiency syndrome in black Africans*, 1.
- Cohen, M., Chen, Y., McCauley, M., Gamble, T., Bollinger, R., & Bryson, Y. (2011). Antiretroviral treatment to prevent the sexual transmission of HIV-1: results from the HPTN 052 multinational randomized controlled trial. In *6th international AIDS society conference on HIV pathogenesis, treatment and prevention, Rome, abstract MOAX0102*.
- Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N., ... Fleming, T. R. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 365(6), 493–505.
- Contreras-Galindo, R., Kaplan, M. H., He, S., Contreras-Galindo, A. C., Gonzalez-Hernandez, M. J., Kappes, F., ... others (2013). HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Research*, 23(9), 1505–1513.

- Conway, D. (2010). The Data Science Venn Diagram. *Drew Conway*, 10.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.
- Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review*, 1.
- Davenport, T. H., & Patil, D. (2012). Data scientist. *Harvard Business Review*, 90, 70–76.
- De Bin, R., Sauerbrei, W., & Boulesteix, A.-L. (2014). Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in Medicine*, 33(30), 5310–5329.
- Deeks, S. G. (2012). Shock and kill. *Nature*, 487(7408), 439–440.
- Deeks, S. G., Autran, B., Berkhout, B., Benkirane, M., Cairns, S., Chomont, N., ... Barré-Sinoussi, F. (2012). Towards an HIV cure: a global scientific strategy. *Nature Reviews Immunology*, 12(8), 607.
- De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? a consensual definition and a review of key research topics. In *AIP Conference Proceedings* (Vol. 1644, pp. 97–104).
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.
- Dittrich, M., & Lehmann, P. V. (2012). Statistical analysis of ELISPOT assays. *Handbook of ELISPOT: Methods and Protocols*, 173–183.
- Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26, 745–766.
- Dorey, F. J., Little, R. J., & Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, 12, 1589–1603.
- Eisinger, R. W., Dieffenbach, C. W., & Fauci, A. S. (2019). HIV viral load and transmissibility of HIV infection: Undetectable equals untransmittable. *The Journal of the American Medical Association*.
- Espasandín-Domínguez, J., Benítez-Estévez, A. J., Cadarso-Suárez, C., Kneib, T., Barreiro-Martínez, T., Casas-Méndez, B., & Gude, F. (2018). Geographical differences in blood potassium detected using a structured additive distributional regression model. *Spatial Statistics*, 24, 1–13.
- Fagard, C., Oxenius, A., Günthard, H., Garcia, F., Le Braz, M., Mestre, G., ... Hirschel, B. (2003). A prospective trial of structured treatment interruptions in human immunodeficiency virus infection. *Archives of Internal Medicine*, 163, 1220–1226.
- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

- Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., ... Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205), 56–61.
- Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 845–854.
- Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3), 317–322.
- Fourer, R., Gay, D. M., & Kernighan, B. W. (2003). *AMPL. A modeling language for mathematical programming*. Thomson/Brooks/Cole.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer Series in Statistics New York.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1.
- Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1), 101–113.
- García, F., Climent, N., Guardo, A. C., Gil, C., León, A., Autran, B., ... Gallart, T. (2013). A dendritic cell-based vaccine elicits T cell responses associated with control of HIV-1 replication. *Science Translational Medicine*, 5, 166ra2–166ra2.
- García, F., Lejeune, M., Climent, N., Gil, C., Alcamí, J., Morente, V., ... Gallart, T. (2005). Therapeutic immunization with dendritic cells loaded with heat-inactivated autologous HIV-1 in patients with chronic HIV-1 infection. *Journal of Infectious Diseases*, 191, 1680–1685.
- García, F., Plana, M., Arnedo, M., Brunet, M., Castro, P., Gil, C., ... Martorell, J. (2004). Effect of mycophenolate mofetil on immune response and plasma and lymphatic tissue viral load during and after interruption of highly active antiretroviral therapy for patients with chronic HIV infection: a randomized pilot study. *Journal of Acquired Immune Deficiency Syndromes*, 36, 823–830.
- García, F., Plana, M., Arnedo, M., Ortiz, G. M., Miró, J. M., Lopalco, L., ... Gatell, J. M. (2003). A cytostatic drug improves control of HIV-1 replication during structured treatment interruptions: a randomized study. *AIDS*, 17, 43–51.
- García, F., Plana, M., Ortiz, G. M., Bonhoeffer, S., Soriano, A., Vidal, C., ... Pantaleo, G. (2001). The virological and immunological consequences of structured treatment interruptions in chronic HIV-1 infection. *AIDS*, 15, F29–F40.
- García, F., Plana, M., Vidal, C., Cruceta, A., Pantaleo, G., Pumarola, T., ... Gatell, J. M. (1999). Dynamics of viral load rebound and immunological changes after stopping effective antiretroviral therapy. *AIDS*, 13, F79–F86.

- Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy-analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3), 307–315.
- Gleiss, A., Gnant, M., & Schemper, M. (2018). Explained variation in shared frailty models. *Statistics in Medicine*, 37, 1482–1490.
- Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109), 23–26.
- Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 215–223.
- Gómez, G., & Cadarso-Suárez, C. (2017). El modelo de riesgos proporcionales de Cox y sus extensiones. Impacto en estadística y biomedicina. *Gaceta de la Real Sociedad Matemática Española*, 20, 513–538.
- Gómez, G., Calle, M. L., Oller, R., & Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, 9, 259–297.
- Gómez, G., Julià, O., & Langohr, K. (2015). Análisis de Supervivencia [Computer software manual].
- Gómez-Mateu, M., Lorenzo-Arribas, A., Bofill, M., Vilor-Tejedor, N., Barrio, I., Espasandín-Domínguez, J., ... Pérez-Álvarez, N. (2016). Big data in biomedical research. Perspectives from the Biostatnet-CRM Workshop. *BEIO*, 257–277.
- Gottlieb, M. S., Schanker, H. M., Fan, P. T., Saxon, A., Weisman, J. D., & Pozalski, I. (1981). Pneumocystis pneumonia—Los Angeles. *Morbidity and Mortality Weekly Report*, 30(21), 250–2.
- Graziani, G. M., & Angel, J. B. (2015). Evaluating the efficacy of therapeutic HIV vaccines through analytical treatment interruptions. *Journal of the International AIDS Society*, 18, 20497.
- Greenhouse, J. (2013). Statistical thinking: the bedrock of data science. *The Huffington Post*.
- Gupta, R. K., Peppas, D., Hill, A. L., Gálvez, C., Salgado, M., Pace, M., ... Olavarria, E. (2020). Evidence for HIV-1 cure after CCR5Δ32/Δ32 allogeneic haemopoietic stem-cell transplantation 30 months post analytical treatment interruption: a case report. *The Lancet HIV*.
- Hackstadt, A. J., & Hess, A. M. (2009). Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, 10(1), 11.
- Hare, W., Nutini, J., & Tesfamariam, S. (2013). A survey of non-gradient optimization methods in structural engineering. *Advances in Engineering Software*, 59, 19–28.
- Hastie, T., & Qian, J. (2014). Glmnet vignette. Retrieved June, 9(2016), 1–30.

- Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2), 226–252.
- HIV i-Base. (2017). *HIV treatment information and advocacy*. <http://i-base.info/guides/art-in-pictures/the-hiv-lifecycle-in-detail>.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müßig, A., Allers, K., ... Thiel, E. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *New England Journal of Medicine*, 360(7), 692–698.
- Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating ‘omics’ data sets. *Nature Reviews Molecular Cell Biology*, 7(3), 198–210.
- Kane, M. J. (2014). Cleveland’s action plan and the development of data science over the last 12 years. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7, 423–424.
- Kauffmann, A., Gentleman, R., & Huber, W. (2008). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3), 415–416.
- Khan, H. R., & Shaw, J. E. H. (2015). AdapEnetClass-package: A Class of Adaptive Elastic Net Methods for Censored Data [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=AdapEnetClass> (R package version 1.2)
- Khan, H. R., & Shaw, J. E. H. (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing*, 26(3), 725–741.
- Khan, H. R., & Shaw, J. E. H. (2019). Variable selection for accelerated lifetime models with synthesized estimation techniques. *Statistical Methods in Medical Research*, 28(3), 937–952.
- Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 489–502.
- Kirch, W. (2008). *Encyclopedia of Public Health* (Vol. 1).
- Kızılersü, A., Kreer, M., & Thomas, A. W. (2018). *The weibull distribution*. Wiley Online Library.
- Klein, J. P., & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.
- Klein, N., Kneib, T., Lang, S., & Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, 9, 1024–1052.

- Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1), 112–147.
- Langohr, K. (2004). *Regression models with an interval-censored covariate*. Universitat Politècnica de Catalunya.
- Langohr, K., & Gómez, G. (2005). Likelihood maximization using web-based optimization tools. *The American Statistician*, 59, 192–202.
- Law, C. G., & Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, 11(12), 1569–1578.
- Leal, L., Lucero, C., Plan, M., Climent, N., Martinez, E., Castro, P., ... Garcia, F. (2017). Viral outcomes after treatment interruptions to evaluate a functional cure [Abstract] [Computer software manual]. CROI Foundation/IAS-USA. (In CROI 2017. Conference on Retroviruses and Opportunistic Infections)
- Lewin, S. R., & Rasmussen, T. A. (2020). Kick and kill for HIV latency. *The Lancet*, 395(10227), 844–846.
- Logue, E. C., Neff, C. P., Mack, D. G., Martin, A. K., Fiorillo, S., Lavelle, J., ... Fontenot, A. P. (2019). Upregulation of chitinase 1 in alveolar macrophages of HIV-infected smokers. *The Journal of Immunology*, 202(5), 1363–1372.
- Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., ... Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, 18, 1039–1065.
- Mabtech. (2015, May 14). *ELISpot tutorial - how to count spots using an AID ELISpot reader*. <https://www.youtube.com/watch?v=SPuFjr44U-I>.
- Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(suppl_1), D54–D58.
- Massicotte, P., & Eddelbuettel, D. (2019). *gtrendsr: Perform and display google trends queries* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=gtrendsR> (R package version 1.4.4)
- McIntosh, A. M., Stewart, R., John, A., Smith, D. J., Davis, K., Sudlow, C., ... Porteous, D. J. (2016). Data science for mental health: a UK perspective on a global challenge. *The Lancet Psychiatry*, 3, 993–998.
- Mell, P., & Grance, T. (2009). The nist definition of cloud computing. *National Institute of Standards and Technology*, 53, 50.
- Mothe, B. (2016). Shaping CTL immunodominance with conserved HIV vaccines after early treatment (BCN01). *DNA*, 4, 1.

- Mothe, B., Climent, N., Plana, M., Rosàs, M., Jiménez, J. L., Muñoz-Fernández, M. Á., ... León, A. (2015). Safety and immunogenicity of a modified vaccinia ankara-based HIV-1 vaccine (MVA-B) in HIV-1-infected patients alone or in combination with a drug to reactivate latent HIV-1. *Journal of Antimicrobial Chemotherapy*, 70, 1833–1842.
- Mothe, B., Rosàs-Umbert, M., Coll, P., Manzardo, C., Puertas, M. C., Morón-López, S., ... Moltó, J. (2020). HIVcons vaccine and romidepsin in early-treated HIV-1-infected individuals: Safety, immunogenicity and effect on the viral reservoir (study BCN02). *Frontiers in Immunology*.
- Muñoz, A., Wang, M.-C., Bass, S., Taylor, J. M., Kingsley, L. A., Chmiel, J. S., & Polk, B. F. (1989). Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men. *American Journal of Epidemiology*, 130, 530–539.
- Nadarajah, S., & Kotz, S. (2006). R programs for truncated distributions. *Journal of Statistical Software*, 16, 1–8.
- NAM Publications. (2017). *HIV & AIDS information*. <http://www.aidsmap.com/CD4-cell-counts/page/1044596/>.
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4), 308–313.
- Novomestky, F., & Nadarajah, S. (2016). *truncdist: Truncated random variables [Computer software manual]*. (R package version 1.0-2)
- Oller, R., Gómez, G., & Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics*, 32(3), 315–326.
- Oller, R., & Langohr, K. (2017). *FHtest: An R package for the comparison of survival curves with censored data*. *Journal of Statistical Software*, 81(15), 1–25.
- Opportunistic Infections Project Team of the Collaboration of Observational HIV Epidemiological Research in Europe (COHERE) in EuroCoord. (2012). CD4 cell count and the risk of AIDS or death in HIV-infected adults on combination antiretroviral therapy with a suppressed viral load: a longitudinal cohort study from COHERE. *PLoS Medicine*, 9(3), e1001194.
- Oxford University Press (Ed.). (2008). *Oxford English Dictionary* (Vol. 30).
- Pallikkuth, S., Kanthikeel, S. P., Silva, S. Y., Fischl, M., Pahwa, R., & Pahwa, S. (2011). Upregulation of IL-21 receptor on B cells and IL-21 secretion distinguishes novel 2009 H1N1 vaccine responders from nonresponders among HIV-infected persons on combination antiretroviral therapy. *The Journal of Immunology*, 186(11), 6173–6181.
- Pfaff, B., & Pfaff, M. B. (2020). Package *rneos*.

- Pinheiro, D., Hamad, F., Cadeiras, M., Menezes, R., & Nezamoddini-Kachouie, N. (2016). A data science approach for quantifying spatio-temporal effects to graft failures in organ transplantation. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the IEEE* (pp. 3433–3436).
- Pitchenik, A. E., Fischl, M. A., Dickinson, G. M., Becker, D. M., Fournier, A. M., O'Connell, M. T., ... Spira, T. J. (1983). Opportunistic infections and Kaposi's sarcoma among haitians: evidence of a new acquired immunodeficiency state. *Annals of Internal Medicine*, 98(3), 277–284.
- Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one*, 2, e308.
- Polak, E. (2012). *Optimization: algorithms and consistent approximations* (Vol. 124). Springer Science & Business Media.
- Project, T. B. (2015). hgu133plus2cdf: hgu133plus2cdf [Computer software manual]. (R package version 2.18.0)
- Pulliam, L., Sun, B., Mustapic, M., Chawla, S., & Kapogiannis, D. (2019). Plasma neuronal exosomes serve as biomarkers of cognitive impairment in HIV infection and Alzheimer's disease. *Journal of Neurovirology*, 25(5), 702–709.
- R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rao, P., & Kumar, S. (2015). Polycyclic aromatic hydrocarbons and cytochrome P450 in HIV pathogenesis. *Frontiers in Microbiology*, 6, 550.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47.
- Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67, 819–829.
- Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.
- Rodger, A. J., Cambiano, V., Bruun, T., Vernazza, P., Collins, S., Degen, O., ... Lundgreen, J. (2019). Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy (PARTNER): final results of a multicentre, prospective, observational study. *The Lancet*.
- Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

- Ruiz-Riol, M., & Brander, C. (2019). *Can we just kick-and-kill HIV: possible challenges posed by the epigenetically controlled interplay between HIV and host immunity*. Future Medicine.
- Sahai, H., & Ojeda, M. M. (2004). *Analysis of variance for random models, volume 2: Unbalanced data: Theory, methods, applications, and data analysis* (Vol. 2). Springer Science & Business Media.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3, 210–229.
- Sánchez, A., Fernández-Real, J., Vegas, E., Carmona, F., Amar, J., Burcelin, R., ... Reverter, F. (2012). Multivariate methods for the integration and visualization of omics data. In *Bioinformatics for Personalized Medicine* (pp. 29–41). Springer.
- Sanz, R. G., & Sánchez-Pla, A. (2019). Statistical analysis of microarray data. In *Microarray Bioinformatics* (pp. 87–121). Springer.
- Satten, G. A., Datta, S., & Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association*, 93, 318–327.
- Savu, A., Liu, Q., & Yasui, Y. (2010). Estimation of relative risk and prevalence ratio. *Statistics in Medicine*, 29(22), 2269–2281.
- Schutt, R., & O’Neil, C. (2013). *Doing data science: Straight talk from the frontline*. O’Reilly Media, Inc.
- Sengupta, S., & Siliciano, R. F. (2018). Targeting the latent reservoir for HIV-1. *Immunity*, 48(5), 872–895.
- Senn, S. (1998). Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine*, 17, 1753–1765.
- Server, N. (2016). State-of-the-art solvers for numerical optimization. *Wisconsin Institute for Discovery* Available in: <https://neos-server.org/neos/>.
- Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111), 647–656.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 486–494.
- Sharp, P. M., & Hahn, B. H. (2011). Origins of hiv and the aids pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1(1), a006841.
- Shewchuk, J. R. (1994). *An introduction to the conjugate gradient method without the agonizing pain*. Carnegie-Mellon University. Department of Computer Science.

- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1.
- Smith, M. T., Vermeulen, R., Li, G., Zhang, L., Lan, Q., Hubbard, A. E., ... Rothman, N. (2005). Use of omics technologies to study humans exposed to benzene. *Chemico-Biological Interactions*, 153, 123–127.
- The stages of HIV infection*. (n.d.). <https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/46/the-stages-of-hiv-infection>. (Accessed: 2020-04)
- Stigler, S. M. (2016). *The seven pillars of statistical wisdom*. Harvard University Press.
- Study to Evaluate the Safety and Effect of HIVconsv Vaccines in Combination With Histone Deacetylase Inhibitor Romidepsin on the Viral Rebound Kinetic After Treatment Interruption in Early Treated HIV-1 Infected Individuals*. *ClinicalTrials.gov Identifier: NCT02616874*. (2018). <https://clinicaltrials.gov/ct2/show/NCT02616874>. (Accessed: 2020-04)
- Stute, W., & Wang, J.-L. (1994). The jackknife estimate of a Kaplan-Meier integral. *Biometrika*, 81(3), 602–606.
- Sun, J. (2007). *The Statistical Analysis of Interval-censored Failure Time Data*. Springer Science & Business Media.
- Szklo, M., & Nieto, F. J. (2014). *Epidemiology: beyond the basics*. Jones & Bartlett Publishers.
- Tanase, C., OGREZEANU, I., & BADIU, C. (2011). *Molecular pathology of pituitary adenomas*. Elsevier.
- Taylor, J. M., Muñoz, A., Bass, S. M., Saah, A. J., Chmiel, J. S., & Kingsley, L. A. (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statistics in Medicine*, 9, 505–514.
- The Chartered Society of Physiotherapy . (2008). *Encyclopedia of Public Health* (L. Breslow, Ed.). New York: Macmillan Reference USA. Retrieved from <http://www.csp.org.uk/topics/public-health>
- Therneau, T. (2015). *A package for survival analysis in S. version 2.38*.
- Therneau, T. M. (2018a). *Coxme and the Laplace Approximation*. Retrieved from <http://cran.r-project.org/web/packages/coxme/vignettes/laplace.pdf>
- Therneau, T. M. (2018b). *coxme: Mixed effects Cox models [Computer software manual]*. (R package version 2.2-10)
- Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag.

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Treasure, G. C., Aga, E., Bosch, R. J., Mellors, J. W., Kuritzkes, D. R., Para, M., . . . Li, J. Z. (2016). Relationship among viral load outcomes in HIV treatment interruption trials. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 72, 310.
- Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33, 1–67.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–295.
- U.S. Department of Health and Human Services. (2019). *AIDS info*. <https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/46/the-stages-of-hiv-infection>.
- Vanable, P. A., Ostrow, D. G., McKirnan, D. J., Taywaditep, K. J., & Hope, B. A. (2000). Impact of combination therapies on HIV risk perceptions and sexual risk among HIV-positive and HIV-negative gay and bisexual men. *Health Psychology*, 19(2), 134.
- Van Dyk, A. C. (2010). *HIVAIDS care and counselling: a multidisciplinary approach*. Pearson South Africa.
- Wacholder, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences. *American Journal of Epidemiology*, 123(1), 174–184.
- Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *The Lancet*, 377, 537–539.
- Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15), 1871–1879.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.
- Williamson, T., Eliasziw, M., & Fick, G. H. (2013). Log-binomial models: exploring failed convergence. *Emerging Themes in Epidemiology*, 10(1), 14.
- Wilson, E. B. (1927). What is statistics? *Science*, 65, 581–587.
- World Health Organization. (2017). *HIVAIDS*. <http://www.who.int/features/qa/71/en/>. (Accessed: 2018-09-20)
- Wu, Y., & Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics*, 71(3), 782–791.

- Yamaguchi, T., Ohashi, Y., & Matsuyama, Y. (2002). Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Statistical Methods in Medical Research*, *11*, 221–236.
- Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 49, 1–11.
- Zapf, A., Huebner, M., Rauch, G., & Kieser, M. (2018). What makes a biostatistician? *Statistics in Medicine*, 1–7.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.
- Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*(4), 1733.



SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1: Number of publications associated with the topics “Data Science” (denoted by DS), “Big Data” (denoted by BD) and “Cloud Computing” (denoted by CC) in different countries from 2004 to 2019.

Year	Topic	USA	UK	Japan	Germany	Australia	Spain	Italy	India	China
2004	DS	14,025	3,514	1,555	2,848	1,408	982	1,957	404	920
	BD	130	29	18	43	10	19	11	2	21
	CC	28	3	0	4	0	1	4	0	1
2005	DS	16,184	4,284	1,769	3,562	1,565	1,256	2,371	528	1,234
	BD	161	37	19	45	16	20	10	3	29
	CC	30	5	2	3	0	0	3	1	2
2006	DS	16,319	4,359	1,673	3,600	1,717	1,322	2,374	554	1,410
	BD	116	52	14	36	17	26	21	4	48
	CC	24	2	3	9	0	1	2	2	2
2007	DS	16,211	4,420	1,703	3,446	1,715	1,299	2,562	590	1,656
	BD	164	42	25	47	17	22	20	9	60
	CC	26	3	3	2	0	3	5	2	13
2008	DS	17,766	4,791	1,791	3,763	1,917	1,540	2,583	713	2,083
	BD	190	49	20	57	28	29	22	5	62
	CC	35	5	2	9	1	4	1	3	12
2009	DS	19,142	5,317	2,045	4,310	2,250	1,868	3,075	850	2,667
	BD	167	58	18	51	31	29	28	11	75
	CC	37	9	3	15	5	4	3	1	11
2010	DS	21,075	5,889	2,145	4,744	2,674	2,116	3,295	1,041	3,368
	BD	202	66	25	74	28	17	31	14	102

Year	Topic	USA	UK	Japan	Germany	Australia	Spain	Italy	India	China
	CC	68	9	9	11	6	7	4	7	17
2011	DS	23,402	6,568	2,401	5,185	2,961	2,352	3,650	1,163	4,343
	BD	242	56	23	65	43	32	38	11	130
	CC	112	7	11	27	6	4	9	7	56
2012	DS	25,647	7,662	2,778	5,796	3,425	2,706	4,002	1,379	5,902
	BD	238	66	12	78	39	37	44	16	112
	CC	133	16	13	23	13	15	9	13	50
2013	DS	27,087	7,957	2,815	6,054	3,922	2,962	4,514	1,583	7,019
	BD	349	95	36	101	61	33	41	20	170
	CC	143	37	17	44	18	24	13	7	81
2014	DS	27,676	7,826	2,783	6,164	4,066	3,105	4,655	1,737	8,646
	BD	504	126	45	116	81	55	61	24	216
	CC	171	37	12	27	26	23	26	17	118
2015	DS	29,007	8,699	2,926	6,319	4,503	3,179	4,857	1,875	10,329
	BD	686	188	47	146	101	72	66	44	302
	CC	172	42	15	44	25	38	32	31	111
2016	DS	28,770	8,664	2,899	6,585	4,758	3,174	4,865	2,034	10,716
	BD	817	230	47	188	124	84	103	70	326
	CC	210	50	21	48	31	36	34	65	137
2017	DS	29,268	8,665	2,888	6,525	4,628	3,346	4,946	1,998	11,831
	BD	900	283	64	211	132	95	92	71	486
	CC	229	49	16	47	45	56	51	48	214
2018	DS	31,619	9,469	3,329	6,907	5,218	3,609	5,444	2,308	13,312
	BD	1,033	296	65	221	154	127	150	113	567
	CC	294	57	25	55	38	52	49	103	242
2019	DS	29,775	9,394	3,351	6,831	5,120	3,665	5,369	2,004	14,561
	BD	1,019	345	69	264	186	153	182	120	564
	CC	215	47	18	39	40	35	51	75	236

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table B.1: Set of predictors for the 5 subsets defined using iregnet and midpoint imputation.

Symbol	Coef	Description
KLRG1	0.0393	killer cell lectin like receptor G1
ECI2	0.0138	enoyl-CoA delta isomerase 2
NAMPT	0.0009	nicotinamide phosphoribosyltransferase
PLXNC1	0.0092	plexin C1
LPAR6	-0.0027	lysophosphatidic acid receptor 6
SLC25A13	-0.0801	solute carrier family 25 member 13
TSHZ1	-0.0013	teashirt zinc finger homeobox 1
HTATIP2	-0.0085	HIV-1 Tat interactive protein 2
CCT2	-0.0008	chaperonin containing TCP1 subunit 2
MAN1A2	0.0466	mannosidase alpha class 1A member 2
CCR5	-0.0285	C-C motif chemokine receptor 5 (gene/pseudogene)
CR1	0.0806	complement component 3b/4b receptor 1 (Knops blood group)
PRUNE2	-0.0116	prune homolog 2
CLECL1	0.0959	C-type lectin like 1
EIF4G1	0.0735	eukaryotic translation initiation factor 4 gamma 1
ERCC1	-0.1148	ERCC excision repair 1, endonuclease non-catalytic subunit
F3	-0.0019	coagulation factor III, tissue factor
FCER1A	-0.0265	Fc fragment of IgE receptor 1a
SEL1L3	-0.6494	SEL1L family member 3
FUT8	0.7262	fucosyltransferase 8
EPHX4	-0.1602	epoxide hydrolase 4
ZNF658	0.0030	zinc finger protein 658
GLRX	-0.0411	glutaredoxin

Symbol	Coef	Description
CYP4V2	-0.2422	cytochrome P450 family 4 subfamily V member 2
PHPT1	-0.0122	phosphohistidine phosphatase 1
LRP12	-0.0028	LDL receptor related protein 12
HSP90AB1	-0.0029	heat shock protein 90 alpha family class B member 1
IFNGR2	0.1090	interferon gamma receptor 2 (interferon gamma transducer 1)
IL13RA1	-0.0011	interleukin 13 receptor subunit alpha 1
ITGA2B	0.0158	integrin subunit alpha 2b
ITPR3	0.2600	inositol 1,4,5-trisphosphate receptor type 3
JARID2	0.4977	jumonji and AT-rich interaction domain containing 2
LGALS3	0.0473	lectin, galactoside binding soluble 3
LSS	0.0770	lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)
MYCN	-0.0061	v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog
NCK1	-0.0020	NCK adaptor protein 1
NCL	-0.2835	nucleolin
RNF165	0.0437	ring finger protein 165
SCL22A18	-0.1193	solute carrier family 22 member 18
MRPL27	-0.0606	mitochondrial ribosomal protein L27
SERPINE2	-0.2815	serpin family E member 2
P3H2	-0.5630	prolyl 3-hydroxylase 2
GPALPP1	-0.0775	GPALPP motifs containing 1
HEPACAM	-0.0030	hepatic and glial cell adhesion molecule
LPAR5	0.3626	lysophosphatidic acid receptor 5
SIPA1L2	-0.0353	signal induced proliferation associated 1 like 2
TRAPPC1	-0.3007	trafficking protein particle complex 1
PLEKHA1	0.0957	pleckstrin homology domain containing A1
RBM4	0.0032	RNA binding motif protein 4
LOC642852	0.0191	uncharacterized LOC642852
CRNDE	0.1033	colorectal neoplasia differentially expressed (non-protein coding)
C11orf1	-0.0548	chromosome 11 open reading frame 1
SCARNA17	0.3755	small Cajal body-specific RNA 17
TTK	0.0146	TTK protein kinase
ZNF736	-0.0588	zinc finger protein 736
DNAJB14	0.2250	DnaJ heat shock protein family (Hsp40) member B14
CHD9	0.0036	chromodomain helicase DNA binding protein 9
NPL	-0.0033	N-acetylneuraminatase pyruvate lyase
ACOX2	0.4263	acyl-CoA oxidase 2
NIFK	0.3283	nucleolar protein interacting with the FHA domain of MKI67
SUPT3H	-0.0595	SPT3 homolog, SAGA and STAGA complex component
CDC42BPA	-0.0098	CDC42 binding protein kinase alpha
TOX2	-0.0193	TOX high mobility group box family member 2
CDK5R1	0.0950	cyclin dependent kinase 5 regulatory subunit 1
CD1D	0.1286	CD1d molecule

Symbol	Coef	Description
SOCS5	-0.2084	suppressor of cytokine signaling 5
DLGAP5	-0.0114	DLG associated protein 5
TNIP1	0.1719	TNFAIP3 interacting protein 1
ERP29	0.1034	endoplasmic reticulum protein 29
EIF3E	0.0830	eukaryotic translation initiation factor 3 subunit E

R CODE USED TO FIT THE MIXED EFFECTS COX MODEL

```
multimp <- function(data, Tl, Tu, dist = "weibull", nimp, uncens,
                    var1, var2, rand1, rand2){
  require(truncdist)
  require(coxme)
  require(actuar)

  aft <- survreg(Surv(t1, t2, cens, type="interval") ~ var1 + var2, data)
  lambda <- exp(-aft$coef[[1]]/aft$scale)
  alpha <- 1/aft$scale

  mat <- matrix(0, ncol = nimp, nrow = nrow(data))
  colnames(mat) <- as.vector(paste("M", 1:nimp, sep = ""))

  for(i in 1:nrow(data)){
    mat[i,] <- with(data,
                    rtrunc(nimp, spec = dist, a = Tl[i], b = Tu[i],
                          shape = alpha,
                          scale = (lambda * exp(-(aft$coef[[2]] *
data$var1[i]/sca
+ aft$coef[[3]] *
data$var2[i]/sca))^(1/alpha)))
  }
  mat[data$Tu == Inf, ] <- rep(data$Tl[data$Tu == Inf], nimp)
  dat <- data.frame(data, mat)
```



```

S <- vector("list", nimp)
mfit <- vector("list", nimp)
beta1 <- rep(0, nimp)
varbeta1 <- rep(0, nimp)
beta2 <- rep(0, nimp)
varbeta2 <- rep(0, nimp)
std1 <- rep(0, nimp)
std2 <- rep(0, nimp)
std3 <- rep(0, nimp)
corr1 <- rep(0, nimp)

for(i in 1:nimp){
  S[[i]] <- with(dat, Surv(dat[, (ncol(dat) - nimp + i)],
                           uncens))
  mfit[[i]] <- coxme(S[[i]] ~ var1 + var2 + (1|rand1)
                    + (1 + var2 | rand2), dat)

beta1[i] <- as.numeric(fixef(mfit[[i]])[1])
varbeta1[i] <- as.numeric(vcov(mfit[[i]])[1])
beta2[i] <- as.numeric(fixef(mfit[[i]])[2])
varbeta2[i] <- as.numeric(vcov(mfit[[i]])[4])

  std1[i] <- sqrt(mfit[[i]]$vcoef[[1]])
  std2[i] <- sqrt(mfit[[i]]$vcoef[[2]][1])
  std3[i] <- sqrt(mfit[[i]]$vcoef[[2]][4])
  corr1[i] <- mfit[[i]]$vcoef[[2]][2]
}

b1 <- round(mean(beta1), 3)
wi1 <- mean(varbeta1)
bi1 <- (sum((beta1 - b1)^2)) / (nimp - 1)
varb1 <- wi1 + (1 + (1/nimp)) * bi1
sb1 <- round(sqrt(varb1), 4)
hr1 <- round(exp(b1), 2)
ci1 <- round(hr1 * exp(qnorm(c(0.025, 0.975)) * sb1), 2)

b2 <- round(mean(beta2), 3)
wi2 <- mean(varbeta2)
bi2 <- (sum((beta2 - b2)^2)) / (nimp - 1)
varb2 <- wi2 + (1 + (1/nimp)) * bi2
sb2 <- round(sqrt(varb2), 4)
hr2 <- round(exp(b2), 2)
ci2 <- round(hr2 * exp(qnorm(c(0.025, 0.975)) * sb2), 2)

sd1 <- round(mean(std1), 2)

```

```
sd2 <- round(mean(std2), 2)
sd3 <- round(mean(std3), 2)
cor <- round(mean(corr1), 2)

vars <- c(b1, sb1, hr1, ci1, b2, sb2, hr2, ci2, sd1, sd2, sd3, cor)
names(vars) <- c('Coef', 'se(coef)', 'HR', 'Lower 95%', 'Upper 95%',
                'Coef', 'se(coef)', 'HR', 'Lower 95%', 'Upper 95%',
                'sd1', 'sd2', 'sd3', 'Corr')

return(vars)

}
```


ADDITIONAL INFORMATION ON THE ATI DATA SET

The eight ATI studies differ from each other with respect to the types of intervention defined by different therapeutic vaccines or drug combinations during the combination Antiretroviral Therapy (cART) and the inclusion criteria, which depended on the CD4 cell count, the plasma viral load (VL), previous years during cART, or the stage of infection; see Table D.1. In addition, the number of patients varied from one study to the other as well as the geographical recruitment area. Following, some more detailed information is provided on each study.

In Study 1 (García et al., 2005), the first ATI episode was done to harvest autologous HIV virus to create the therapeutic vaccine. The therapeutic vaccine was administered just before the second ATI. Patients are coming from a previous study with non-advanced chronic HIV-1 infection.

Study 2 (García et al., 2013) has a similar design as that of Study 1. In this study, there are 2 ATI episodes, the first one is to harvest the autologous virus and the second is the post-vaccine stop.

Study 3 (<https://clinicaltrials.gov/ct2/show/NCT02767193>) corresponds to the ongoing new version of the dendritic cell-based vaccine trial. The first ATI episode was done, as in the previous studies, to harvest autologous HIV virus to create the therapeutic vaccine, which was administered just before the second ATI. Patients in this trial had to be on stable cART for at least one year and the average of all measurements of CD4 cells during the year before starting cART.

Study 4 (García et al., 2004) was carried out to evaluate the effect of mycophenolate mofetil (MMF) on the immunologic control during the ATI. In particular, the first ATI episode was done to evaluate the effect of MMF over the viral load dynamics. MMF is a well characterized drug widely used in renal transplantation because of its ability to selectively inhibit lymphocyte division and it may inhibit HIV replication. Patients were chronic HIV-1 infected persons at very early stages and were treated with cART for 12 months.

Part of the objectives of Study 5 (García et al., 2003) was to study the effect of concomitant hydroxyurea treatment (HU), which was administered to a group of patients but not in the first ATI episode. The study group were patients with chronic HIV infection from the Spanish EARTH-2 study. This study 5 was a designed study to explore the effect of controlled and repeated interruptions over the immune response against the virus. They were based on the 'autovaccination' theory, as explained in the introduction. The first ATI episode does not have

any specific ‘importance’, but the joint set of ATI episodes (which corresponds to the studied intervention in this case).

Study 6 ([Mothe et al., 2015](#)) corresponds to patients participating in a randomized phase I HIV vaccine trial with recombinant modified vaccinia Ankara-based and Gag-Pol-Nef polyprotein with or without a drug to reactivate latent HIV in 3 centers. The only ATI episode of this study was done 8 weeks after the last dose of MVA-B and the viral rebound dynamics were assessed during the first 12 weeks after cART interruption. The ATI was done to evaluate the effect of the vaccine in the control of viral load during the absence of cART. Participants were chronically HIV-infected individuals recruited at three HIV units in Barcelona and Madrid (Spain).

In Study 7 ([Fagard et al., 2003](#)), cART was interrupted for 2 weeks, restarted for 8 weeks. After 4 such cycles, treatment was indefinitely suspended 40 weeks after study entry. The ATI rationale in this study is to try to prove the ‘autovaccination hypothesis’, that is that reexposure to HIV during treatment interruptions may stimulate the HIV-specific immune response and lead to low viremia after withdrawal of cART.

Study 8 ([García et al., 1999, 2001](#)) involved patients with chronic HIV-1 infection in very early stages who started a twice daily three-drug regimen. cART was discontinued after one year of treatment and effective virologic response. The ATI rationale in this study is the same as in the Study 5.

Table D.1: Inclusion criteria and treatment for the different studies of the ATI data set.

Study	Patients	Treatment (apart from cART)	CD4 counts (cells/mm³)	Viral Load (copies/mL)
Study 1	16	- Dendritic cell-based HIV-vaccine (n=12) - Placebo (n=4)	>500 (pre-cART)	>5,000 (pre-cART) <20 for at least 104 weeks while receiving cART
Study 2	35	- Dendritic cell-based HIV-vaccine (n=24) - Placebo (n=11)	>450 (baseline)	<37 (enrollment)
Study 3	18		>350 (previous years) >450 (at enrollment)	Undetectable at least 6 months before inclusion
Study 4	11	- Received MMF (n=7) - Did not (n=4)	>500	200-5,000 (baseline)
Study 5	20		>500 (pre-cART)	>5,000 (pre-cART)
Study 6	28	- Ankara-based vaccine and Gag-Pol-Nef polyprotein with a drug to reactive latent HIV (n=19) - Placebo (n=9)	>450	Not specified
Study 7	33		>740	Undetectable for a median of 21 months
Study 8	10		>500 (last 3 months)	>10,000

MORE INFORMATION OF BCN02 CLINICAL TRIAL

E.1 Additional tables of BCN02 clinical trial

Table E.1: Summary of continuous covariates of BCN02.

Variable	Rebounders (n=10)				Controllers (n=3)			
	Med	IQR	Min	Max	Med	IQR	Min	Max
At HIV-1 Diagnosis								
First log ₁₀ (VL)	5.11	4.88–5.24	4.26	5.49	4.77	4.06–5.29	3.34	5.82
Response to cART init								
Weeks to UD VL	12	4–12	4	36	4	4–14	4	24
log ₁₀ (VL) at week ₄	1.63	1.60–2.20	1.56	3.35	1.60	1.58–2.76	1.56	3.93
log ₁₀ (VL) at week ₁₂	1.60	1.56–1.60	1.56	1.61	1.60	1.58–2.11	1.56	2.63
log ₁₀ (VL) at week ₂₄	1.60	1.56–1.60	1.56	1.64	1.56	1.56–1.58	1.56	1.60
CD4 absolute week ₂₄	593	498–826	388	1,297	734	621–786	508	838
CD4/CD8 ratio week ₂₄	1.05	0.84–1.41	0.63	1.52	1	0.95–1.15	0.89	1.30
At MAP								
Weeks since last VAX	13	8–19	8	25	12	10.5–14	9	16
Vaccine Immunogenicity								
Total HIV magnitude								
At BCN02 entry	5,163	3,375–6,980	1,635	8,945	3,328	2,526–3,428	1,725	3,528
At BCN02 peakimmunog	2,880	2,535–4,790	720	8,355	2,600	1,898–3,388	1,195	4,175
HTI responses								
At BCN02 entry	907	560–1,710	170	2,990	1,270	748.5–1,445.5	227	1,621
At BCN02 peakimmunog	200	70–390	0	550	340	170–770	0	1,200

Variable	Rebounders (n=10)				Controllers (n=3)			
	Med	IQR	Min	Max	Med	IQR	Min	Max
Viral reservoir								
Week 24 BCN01	824	494–1,391	134	3,826	177	142.5–339	108	501
Week 60 BCN01	384	170–623	79	1,382	78	67–233.5	56	389
MAP	146	144–550	29	829	34	25–81	16	128
RMD-PK								
AUC ₁	392.9	386–427.1	314	439.5	473.5	432.5–548.9	391.5	624.3
Reservoir CA RNA								
AUC RMD	54.7	14.1–175.2	3.1	183.5	18.14	18.14–18.14	18.14	18.14

Table E.2: Univariate fitted survival models of BCN02.

Variables	$\hat{\beta}$	s.e.($\hat{\beta}$)	95% CI	\widehat{HR}
Site (HUGTIP)	0.51	1.66	-2.74, 3.77	1.67
Age	0.24	1.73	-3.14, 3.62	1.27
First log ₁₀ (VL)	0.43	10.82	-20.78, 21.6	1.53
Days HIV to cART	-0.01	0.08	-0.17, 0.16	0.99
log ₁₀ (BSL VL at cART initiation)	0.53	22.96	-44.47, 45.5	1.70
CD4 absolute v ₀	-0.01	0.12	-0.25, 0.24	0.99
CD4/CD8 ratio v ₀	-0.22	3.92	-7.90, 7.47	0.81
Weeks to UD VL	0.02	0.19	-0.36, 0.39	1.02
log ₁₀ (VL) at w ₄	-0.27	11.53	-22.87, 22.3	0.76
CD4 absolute w ₂₄	-0.001	0.004	-0.01, 0.01	1.00
CD4/CD8 ratio w ₂₄	-0.22	1.63	-3.42, 2.98	0.80
CD4 absolute BCN02 entry	-0.00	0.002	0.00, 0.00	1.00
CD4/CD8 ratio BCN02 entry	0.35	1.82	-3.21, 3.92	1.42
Total months on cART BCN02 entry	-0.07	0.47	-0.99, 0.86	0.94
Months on UD pVL BCN02 entry	-0.13	5.36	-10.64, 10.4	0.88
CD4 absolute MAP	-0.001	0.03	-0.07, 0.06	1.00
CD4/CD8 ratio MAP	0.06	14.72	-28.79, 28.9	1.07
Total months on cART MAP	-0.05	1.76	-3.51, 3.40	0.95
Months on UD VL MAP	-0.12	0.18	-0.46, 0.23	0.89
Weeks since last Vax to MAP	0.002	0.71	-1.38, 1.39	1.03
log ₁₀ (Mag HIVconsv peak)	0.25	3.20	-6.02, 6.53	2.27
log ₁₀ (Total HIV mag BSL)	2.89	2.36	-1.74, 7.52	24.25
log ₁₀ (Total HIV mag peak)	0.24	1.48	-2.66, 3.14	2.51
log ₁₀ (Ratio HIVconsv/total BSL)*	-0.03	1.96	-3.86, 3.80	0.02
log ₁₀ (Total out BSL)	2.92	3.66	-4.25, 10.1	27.42
log ₁₀ (Total out peak)	0.39	5.03	-9.46, 10.24	2.08
log ₁₀ (HTI responses BSL)	0.88	15.86	-30.21, 32	3.70

Variables	$\hat{\beta}$	s.e.($\hat{\beta}$)	95% CI	\widehat{HR}
\log_{10} (HTI responses peak)*	-0.01	1.42	-2.80, 2.79	0.91
Breadth total HIVconsv (6) BSL	0.11	0.43	-0.72, 0.96	1.13
Breadth total HIVconsv (6) peak	-0.68	2.14	-1.36, 1.15	0.50
Breadth total HIVconsv (18) BSL	0.24	0.27	-0.10, 0.71	1.27
Breadth total HIVconsv (18) peak	-0.20	1.97	-3.27, 3.24	0.82
Breadth total HTI (5) BSL	0.57	1.88	-1.88, 3.33	1.77
Breadth total HTI (5) peak	0.04	0.45	-0.81, 0.82	1.04
HIV DNA w24 BCN01	0.001	0.01	-0.01, 0.01	1.00
HIV DNA w60 BCN01	0.002	0.02	0.00, 0.01	1.00
HIV DNA w0 BCN02	0.002	0.003	-0.01, 0.01	1.00
HIV DNA w3 BCN02	0.001	0.004	-0.01, 0.01	1.00
HIV DNA w6 BCN02	0.002	0.08	-0.08, 0.08	1.00
HIV DNA w17 BCN02	0.004	0.02	-0.01, 0.02	1.00
AUC ₁	0.002	0.15	-0.05, 0.02	1.00
AUC RMD	-0.01	0.01	-0.04, 0.04	0.99

* $\log_{10}(x + 1)$ transformation used.

Table E.3: Univariate log-binomial regression models for patient profile.

Variables	$\hat{\beta}$	s.e.($\hat{\beta}$)	\widehat{RR}	95% CI
Intercept	-0.92	0.55		
Site (HUGTIP)	-1.16	1.08	0.31	0.04, 2.62
Intercept	-1.39	0.61		
Vaccine ARM BCN01 (B)	-0.22	1.08	0.80	0.10, 6.70
Intercept	3.39	1.24		
\log_{10} (First VL)	-1.02	0.37	0.36	0.17, 0.75
Intercept	-3.43	1.37		
Days HIV to cART	0.02	0.01	1.02	1.00, 1.04
Intercept	3.69	1.42		
\log_{10} (BSL VL at cART init)	-1.15	0.44	0.32	0.13, 0.75
Intercept	-1.44	1.31		
CD4/CD8 ratio v0	-0.04	1.84	0.96	0.03, 35.55
Intercept	-1.29	0.78		
Weeks to UD VL	-0.02	0.06	0.98	0.89, 1.10
Intercept	-2.91	1.30		
\log_{10} (VL) week 4	0.66	0.41	1.94	0.87, 4.32

Variables	$\hat{\beta}$	s.e.($\hat{\beta}$)	\widehat{RR}	95% CI
Intercept	-1.34	1.49		
CD4 absolute w ₂₄	-0.00	0.002	1.00	1.00, 1.00
Intercept	-0.98	2.00		
CD4 absolute w ₂₄	-0.45	1.83	0.64	0.02, 23.08
Intercept	-0.95	1.37		
CD4 absolute BCN02 entry	-0.00	0.00	1.00	1.00, 1.00
Intercept	-0.21	2.52		
CD4/CD8 ratio BCN02 entry	-0.99	1.88	0.40	0.01, 15.98
Intercept	-5.23	7.46		
Total months on cART BCN02 entry	0.09	0.18	1.10	0.77, 1.58
Intercept	-4.66	5.17		
Months on UD VL BCN02 Entry	0.09	0.13	1.09	0.84, 1.42
Intercept	-1.44	1.47		
CD4 absolute MAP	-0.00	0.00	1.00	1.00, 1.00
Intercept	-0.47	2.10		
CD4/CD8 ratio MAP	0.77	1.63	0.46	0.02, 11.31
Intercept	-3.04	9.99		
Total months on cART MAP	0.03	0.22	1.04	0.68, 1.59
Intercept	-3.91	6.71		
Months on UD VL MAP	0.06	0.15	1.06	0.78, 1.43
Intercept	-0.88	1.37		
Weeks since last Vax to MAP	-0.04	0.10	0.96	0.78, 1.17
Intercept	-1.15	0.60		
log ₁₀ (Mag HIVconsv BSL)*	-0.25	0.37	0.78	0.37, 1.62
Intercept	0.73	4.54		
log ₁₀ (Mag HIVconsv peak)	-0.67	1.40	0.51	0.03, 7.98
Intercept	1.02	5.13		
log ₁₀ (Total HIV mag peak)	-0.73	1.52	0.48	0.02, 9.51
Intercept	-1.37	0.68		
log ₁₀ (Ratio HIVconsv/total BSL)*	-0.15	0.75	0.86	0.20, 3.77
Intercept	-1.80	3.10		
log ₁₀ (Ratio HIVconsv/total peak)	0.00	0.03	1.00	0.94, 1.07
Intercept	0.68	3.13		
log ₁₀ (Total out peak)	-0.79	1.20	0.45	0.04, 4.71

Variables	$\hat{\beta}$	s.e.($\hat{\beta}$)	\widehat{RR}	95% CI
Intercept	-0.07	3.82		
\log_{10} (HTI responses BSL)	-0.48	1.32	0.62	0.05, 8.29
Intercept	-0.94	0.85		
\log_{10} (HTI responses peak)*	-0.26	0.41	0.77	0.34, 1.73
Intercept	-1.18	0.59		
Breadth total HIVconsv (6) BSL	-0.35	0.56	0.70	0.23, 2.12
Intercept	-5.53	4.17		
Breadth total HIVconsv (6) peak	0.76	0.72	2.14	0.52, 8.82
Intercept	-3.95	3.14		
Breadth total HIVconsv (18) peak	0.29	0.33	1.33	0.70, 2.54
Intercept	-1.34	1.16		
Breadth total HTI (5) BSL	-0.06	0.52	0.94	0.34, 2.63
Intercept	-0.68	0.64		
Breadth total HTI (5) peak	-0.61	0.53	0.54	0.19, 1.54
Intercept	0.09	0.56		
HIV DNA w24 BCN01	-0.00	0.00	1.00	0.99, 1.00
Intercept	-0.20	0.62		
HIV DNA w60 BCN01	-0.01	0.00	0.99	0.99, 1.00
Intercept	-0.56	0.71		
HIV DNA w0 BCN02	-0.01	0.01	0.99	0.98, 1.01
Intercept	-0.45	0.62		
HIV DNA w3 BCN02	-0.01	0.01	0.99	0.98, 1.01
Intercept	-0.74	0.67		
HIV DNA w6 BCN02	-0.00	0.00	1.00	0.99, 1.00
Intercept	-0.01	0.50		
HIV DNA w17 BCN02	-0.01	0.01	0.99	0.97, 1.00
Intercept	-0.04	0.44		
HIV DNA MAP	-0.01	0.01	0.99	0.97, 1.00
Intercept	-1.34	1.34		
HIV DNA MAP	-0.03	0.05	0.97	0.87, 1.08

* $\log_{10}(x + 1)$ transformation used.

Table E.4: Pool of peptides for HIVconsv in ELISpot assay of BCN02.

Pool #	Protein coverage	Number of peptides
p1	Gag Clade C, D and A	28
p2	Pol Clade B	30
p3	Pol Clade C	30
p4	Vif Clade D, Pol Clade A, Env Clade C	24
p5	Pol Clade A, B and D	31
p6	Pol Clade C, Env Clade D, Mouse & Macaque	23

Table E.5: Pool of peptides for OUT in ELISpot assay of BCN02.

Pool #	Protein coverage	Number of peptides
OUT Gag-1	Gag clade B (p17-p24)	50
OUT Gag-2	Gag clade B (p24-p15)	39
OUT Pol-1	Gag/pol TF, Prot, RT	39
OUT Pol-2	RT	39
OUT Pol-3	RT, Int	40
OUT V-T	Vif (39), Tat(19)	58
OUT Env-1	Gp 120	46
OUT Env-2	Gp 120	46
OUT Env-3	Gp 120, gp41	47
OUT Env-4	Gp 41	47
OUT Nef	Nef	49
OUT Acc	Vpu (19), Vpr(22), Rev(26)	67

Table E.6: Pool of peptides for HTI in ELISpot assay of BCN02.

Pool #	WITHOUT AAA-containing peptides	Number of peptides
Gag-p1	Seg 1-2-3	23
Gag-p2	Seg 4-5-6-7	20
Pol-p1	Seg 8-9-10	29
Pol-p2	Seg 11-12-13	14
Vif-Nef	Seg 14-15-16	9

Table E.7: Cells per well in each timepoint for every subject in BCN02.

Patient	w0	w1	w3	w9	w10	w13
A12	100,000	100,000	100,000	100,000	100,000	100,000
A13	100,000	72,000	100,000	100,000	100,000	100,000
A04	100,000	100,000	100,000	100,000	86,000	92,000
A14	100,000	100,000	100,000	100,000	100,000	100,000
B03	100,000	100,000	100,000	100,000	100,000	100,000
B05	100,000	100,000	100,000	145,000	100,000	100,000
B06	100,000	100,000	100,000	100,000	100,000	100,000
A15	100,000	100,000	100,000	100,000	100,000	100,000
B13	100,000	100,000	100,000	100,000	100,000	100,000
B07	100,000	100,000	100,000	100,000	100,000	100,000
A05	100,000	100,000	100,000	100,000	100,000	100,000
A09	100,000	100,000	100,000	100,000	100,000	100,000
A02	100,000	100,000	100,000	100,000	100,000	100,000
B10	100,000	100,000	100,000	100,000	100,000	100,000
B14	100,000	100,000	100,000	100,000	100,000	100,000

E.2 Results for OUT and HTI region

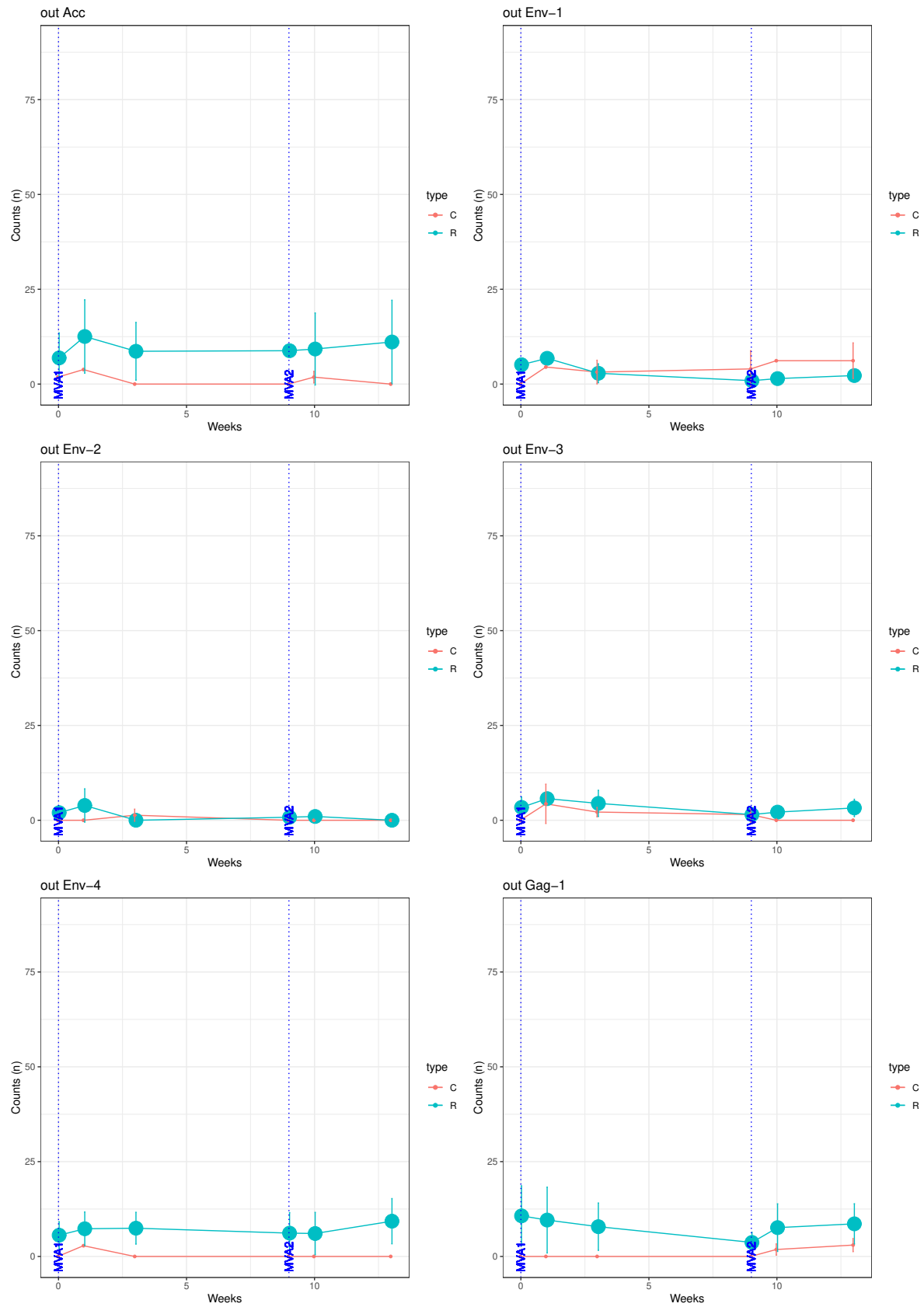


Figure E.1: Mean and 95% confidence interval for spot counts in OUT region for each pool of peptides and each patient profile (first part).

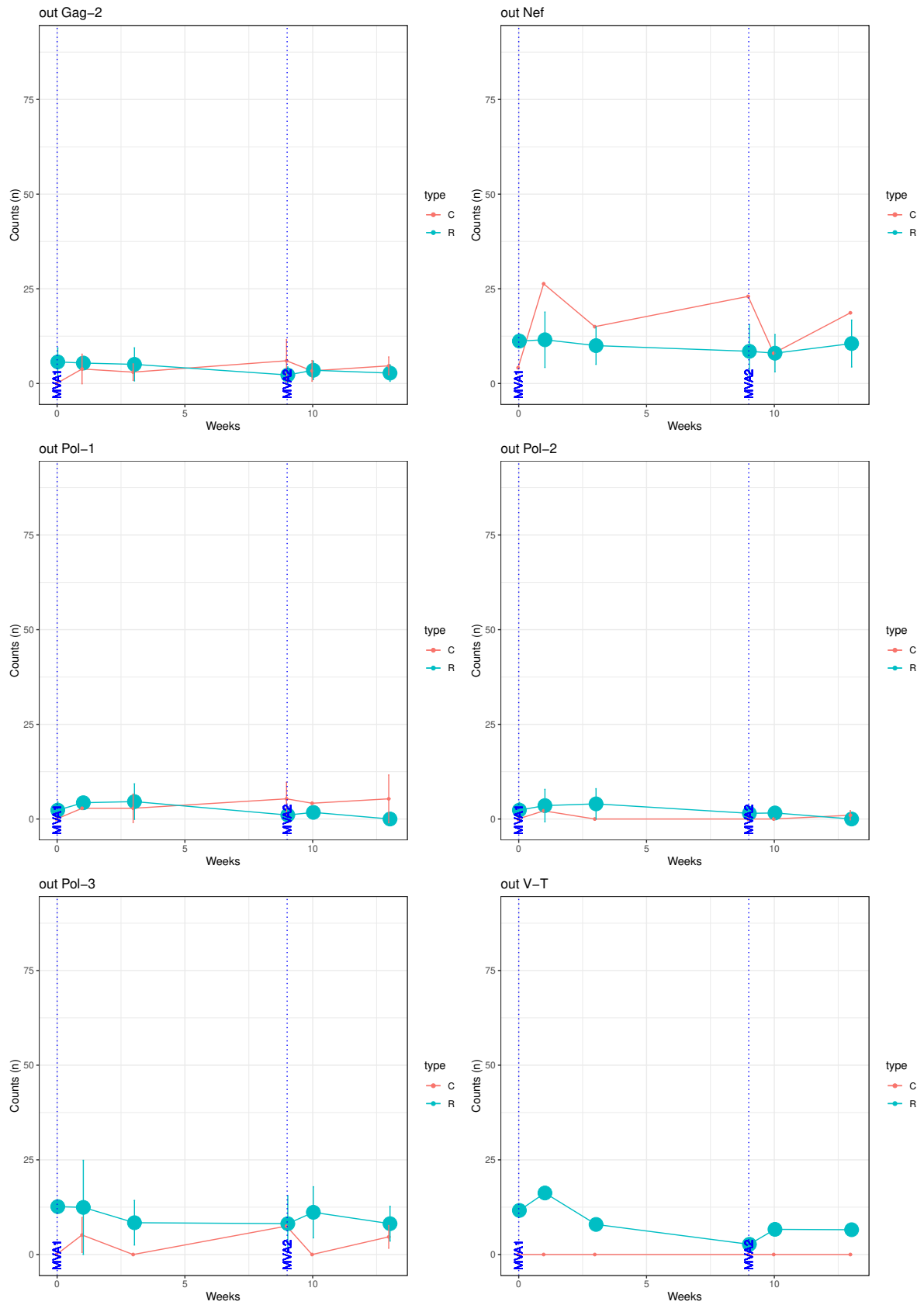


Figure E.2: Mean and 95% confidence interval for spot counts in HIVconsv region for each pool of peptides and each patient profile (second part).

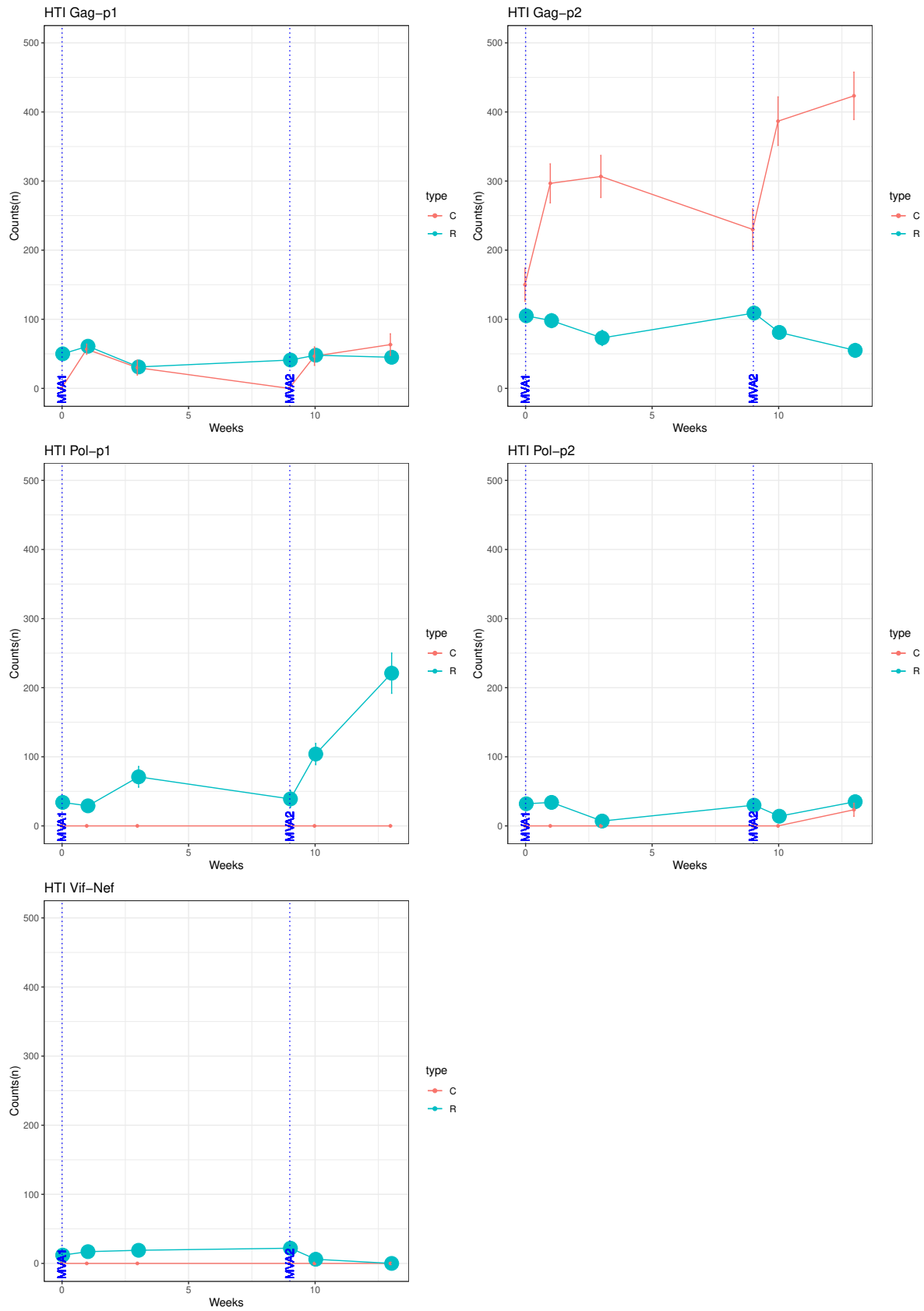


Figure E.3: Mean and 95% confidence interval for spot counts in HTI region for each pool of peptides and each patient profile.

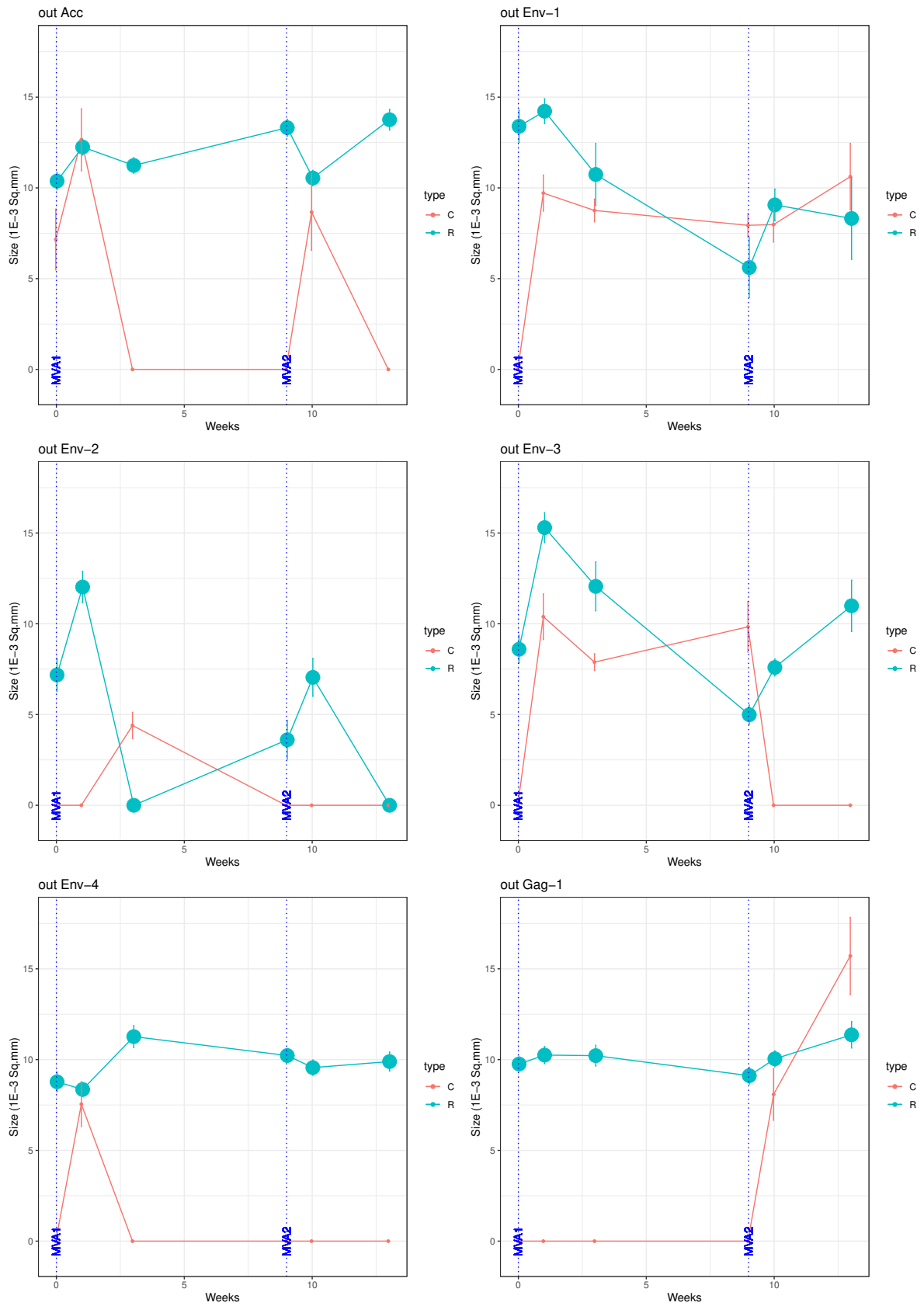


Figure E.4: Mean and 95% confidence interval for spot size in OUT region for each pool of peptides and each patient profile (first part).

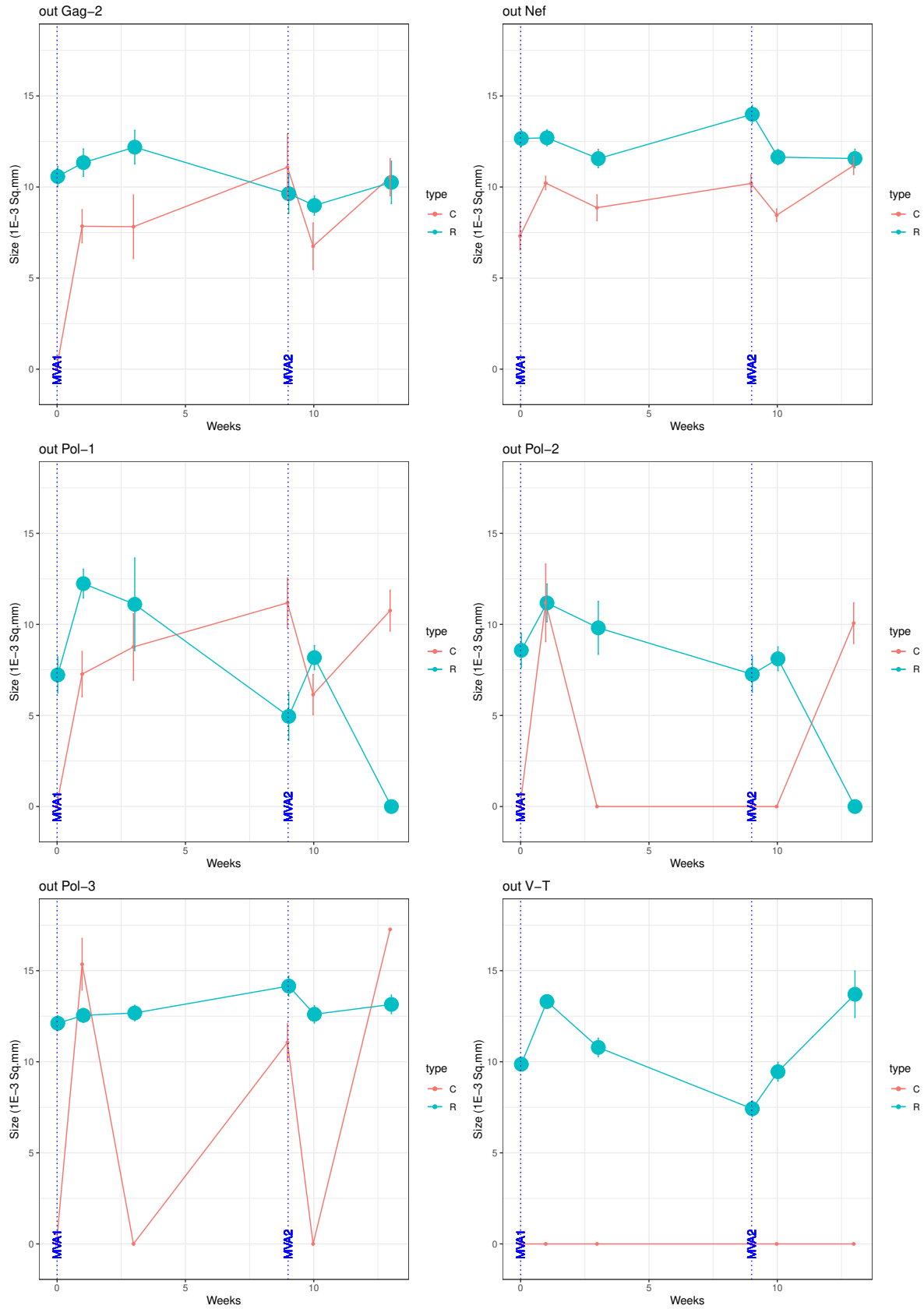


Figure E.5: Mean and 95% confidence interval for spot size in OUT region for each pool of peptides and each patient profile (second part).

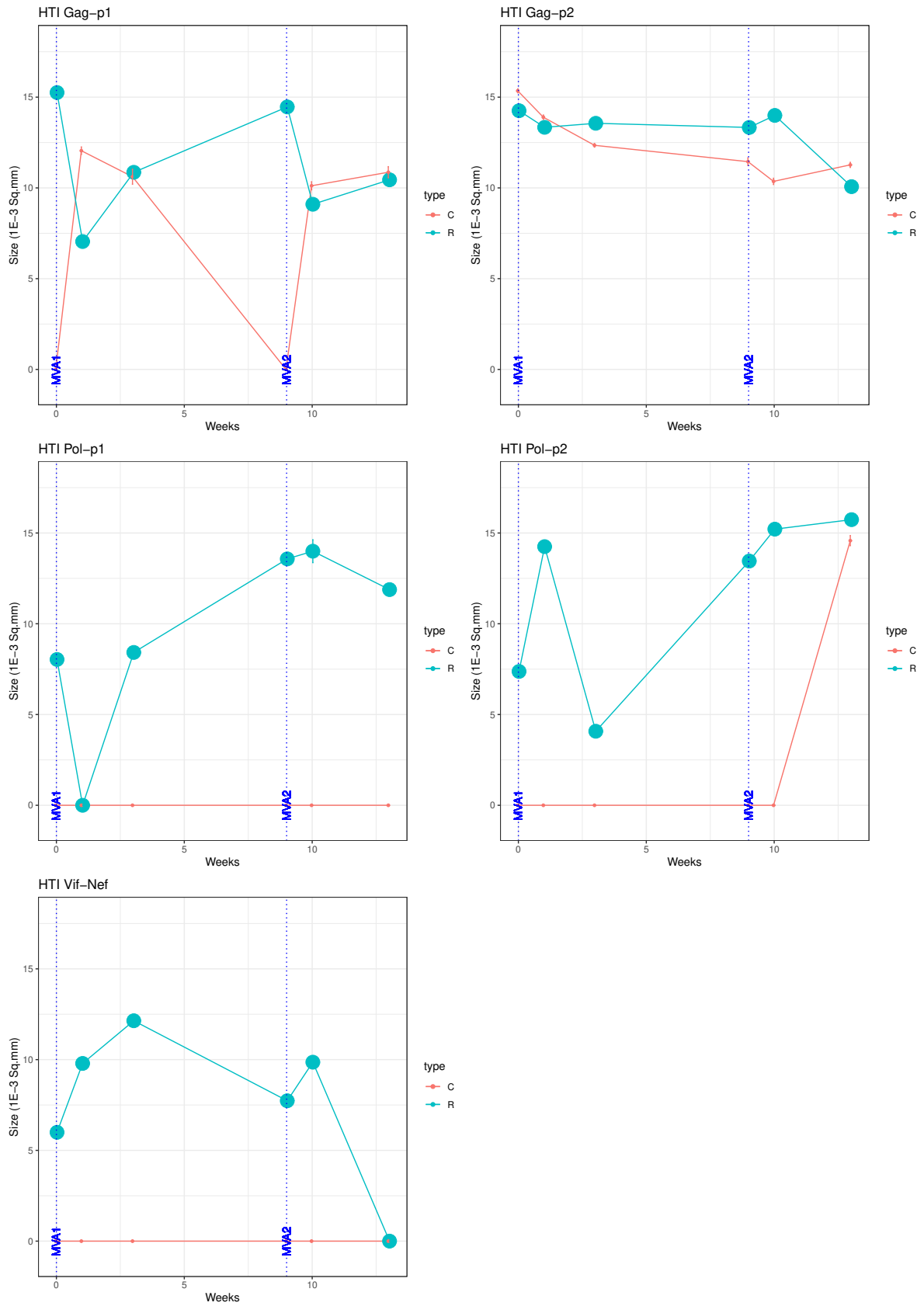


Figure E.6: Mean and 95% confidence interval for spot size in HTI region for each pool of peptides and each patient profile.

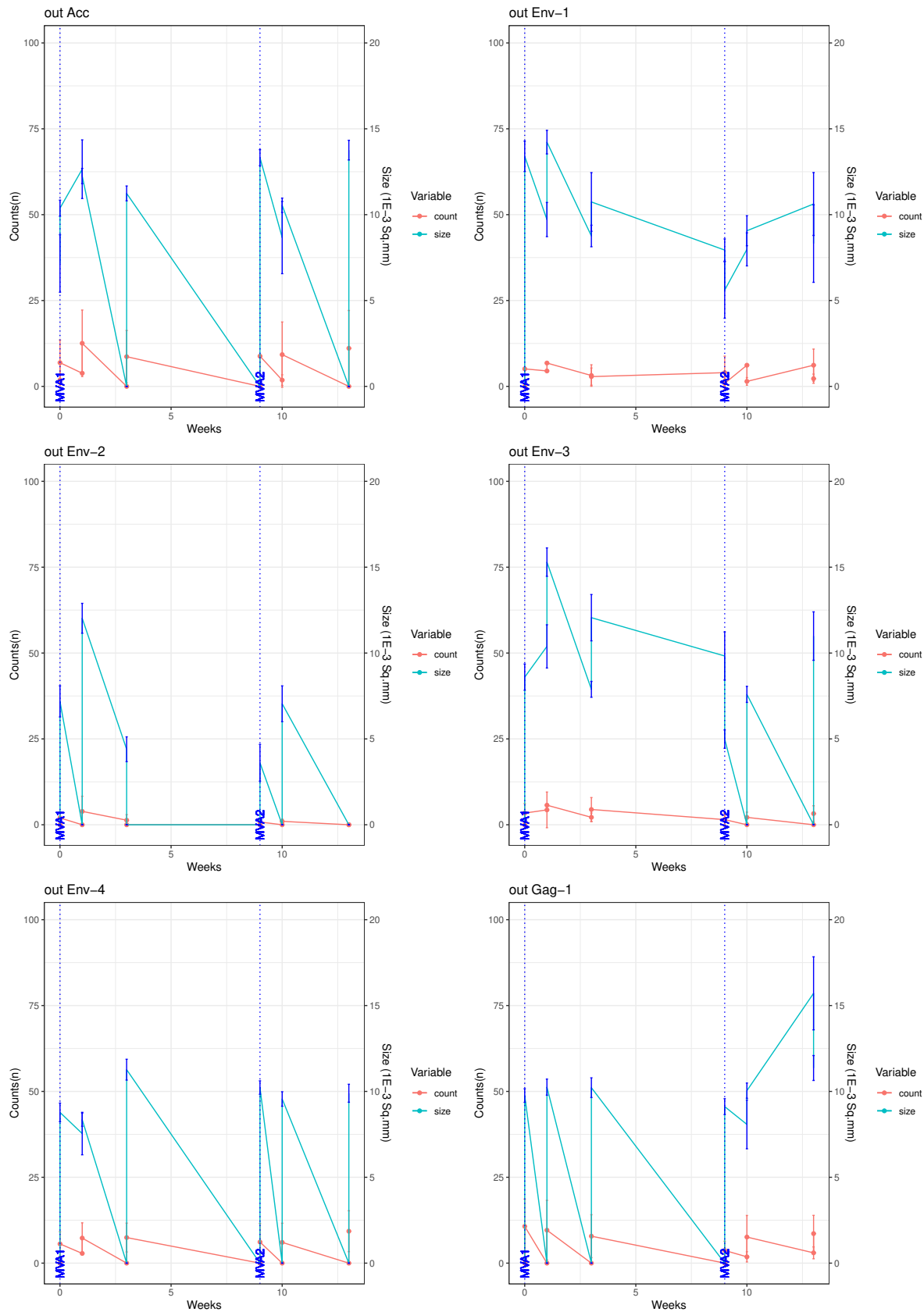


Figure E.7: Mean and 95%confidence interval for spot size and spot count in OUT region for each pool of peptides (first part).

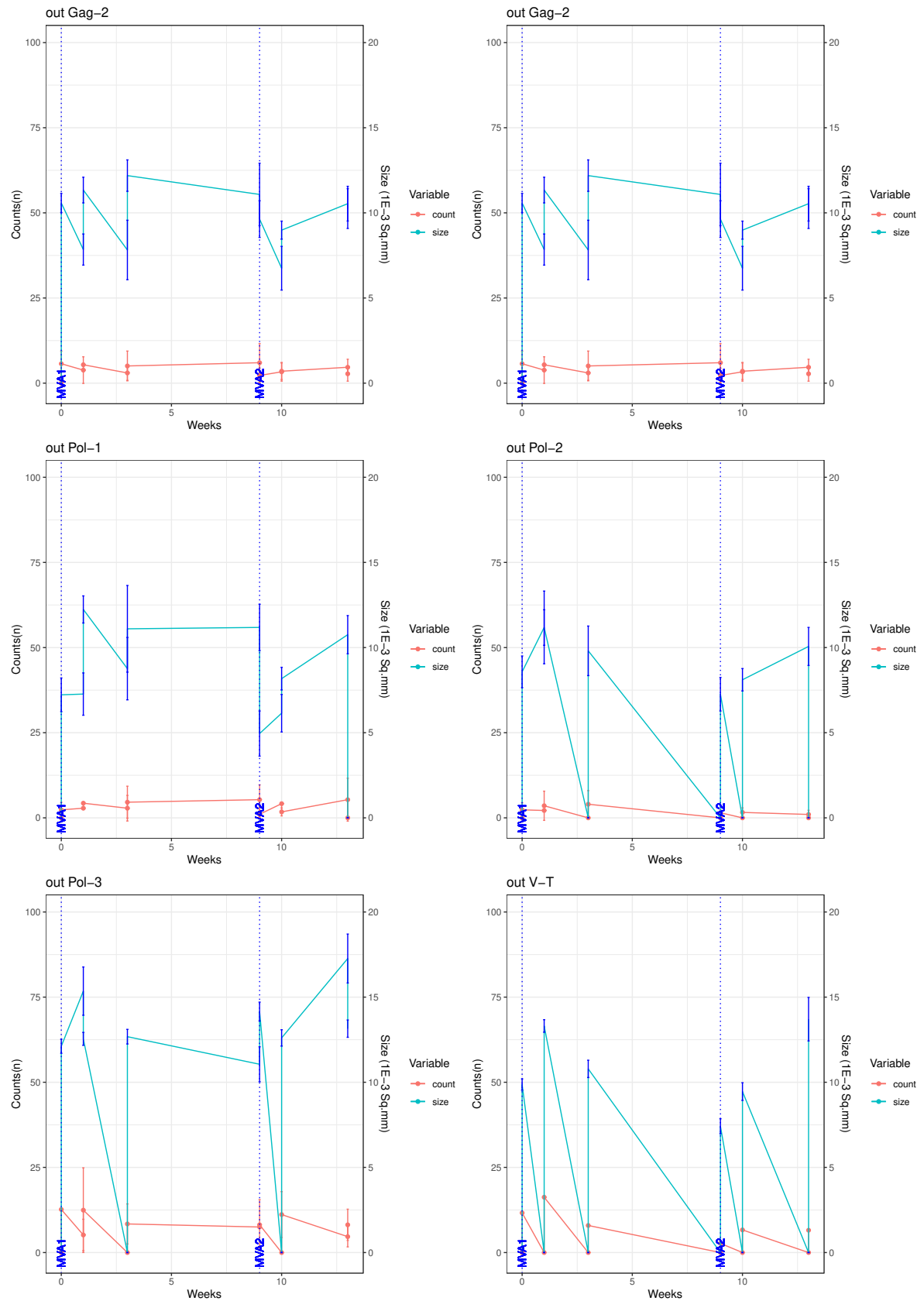


Figure E.8: Mean and 95% confidence interval for spot size and spot count in OUT region for each pool of peptides (second part).

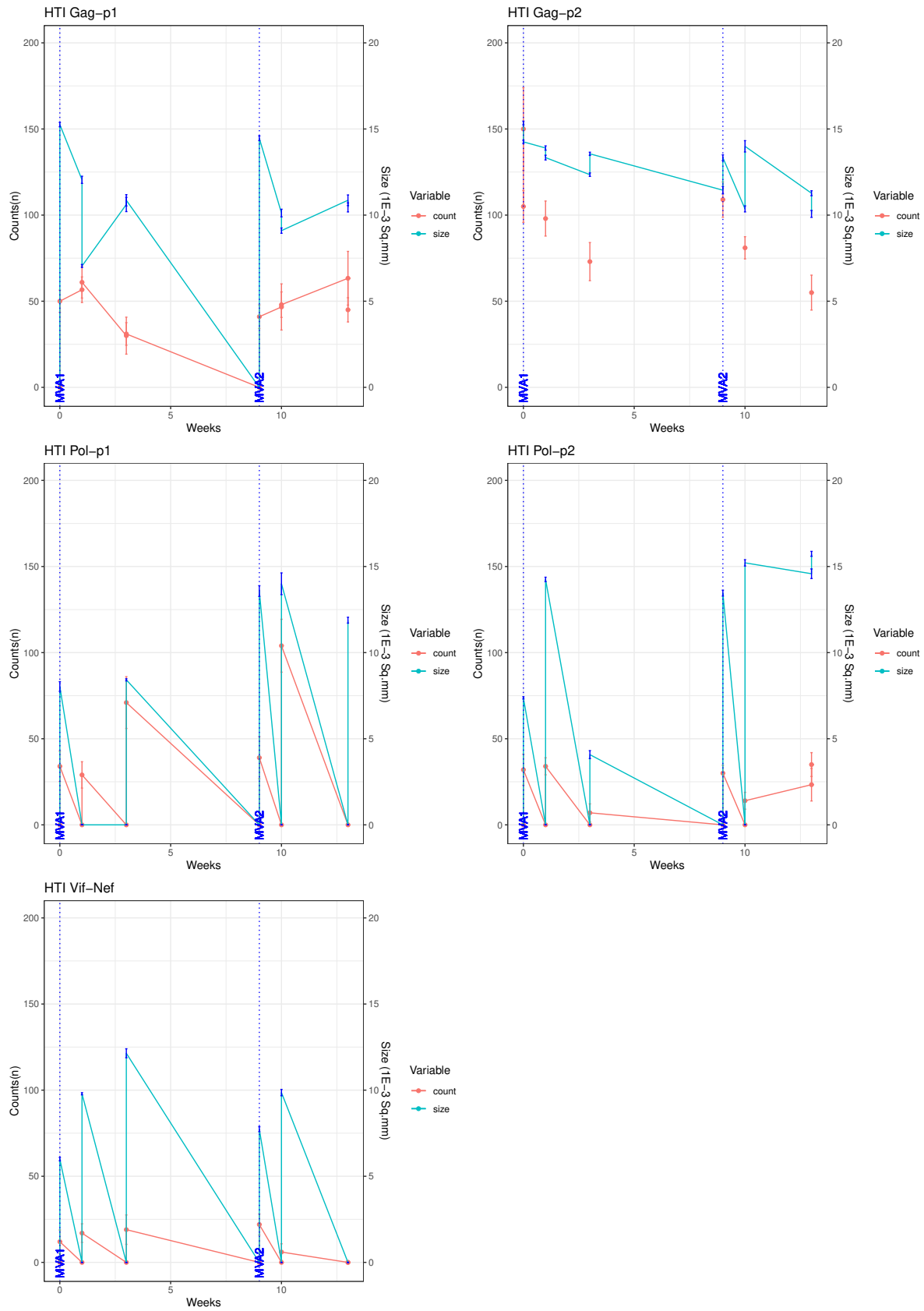


Figure E.9: Mean and 95% confidence interval for spot size and spot count in HTI region for each pool of peptides.

GLOSSARY

This Glossary is intended to explain briefly the main biological terms presented in this research. These definitions are based mainly on those presented by the World Health Organization ([World Health Organization, 2017](#)) and in two Public Health Encyclopedias ([The Chartered Society of Physiotherapy, 2008](#); [Kirch, 2008](#)).

Acquired Immunodeficiency Syndrome is a term which applies to the most advanced stages of HIV infection. It is defined by the occurrence of any of more than 20 opportunistic infections or HIV-related cancers.

Antibody is the generic name for any immunoglobulin produced, no matter how this occurs. Humans can produce many specific antibodies. This may be an active process by a healthy host in response to the challenge of exposure to a foreign antigen transmitted via the placenta or in maternal milk from mother to offspring, or it may be artificially induced by immunization with live attenuated organisms, killed organisms, or a protein derivative. The antibody is the basic ingredient of the host's defenses against infection. By measuring the concentration of specific antibodies in individuals and populations it is possible to determine levels of susceptibility and resistance to infection by specific pathogens. At the population level, this is called "sero-epidemiology".

Antigen is a substance that is capable of inducing a specific immune response in the host into which it is introduced. The immune response is mediated via an immunoglobulin (protein) molecule, called an antibody, which is formed by B-lymphocytes and T-helper cells that are the basic ingredients of the host's immune system. An antigen is an organic compound- a protein, polysaccharide or glycolipid. Sometimes it is an entire organ or tissue that has been transplanted into the host, which rejects it and attempts to destroy it. An antibody has the capacity to bind specifically to the (foreign) antigen and thereby neutralize it so it can be destroyed by the host's phagocytes.

Antiretroviral Therapy consists of the combination of antiretroviral drugs to maximally suppress the HIV virus and stop the progression of HIV disease. ART also prevents onward transmission of HIV. Huge reductions have been seen in rates of death and infections when use is made of a potent ARV regimen, particularly in early stages of the disease. The WHO recommends ART for all people with HIV as soon as possible after

diagnosis without any restrictions of CD4 counts. It also recommends offer of pre-exposure prophylaxis to people at substantial risk of HIV infection as an additional prevention choice as part of comprehensive prevention. Countries are now following to adapt and implement these recommendations within own epidemiological settings.

CD4 is a type of lymphocyte. CD4 T lymphocytes (CD4 cells) help to coordinate the immune response by stimulating other immune cells, such as macrophages, B lymphocytes (B cells), and CD8 T lymphocytes (CD8 cells), to fight infection. HIV weakens the immune system by destroying CD4 cells.

CD8 is a T cell with CD8 receptor recognizes antigens on the surface of a virus-infected cell and binds to the infected cell and kill it.

Cytokine The term cytokine is derived from a combination of two Greek words - “cyto” meaning cell and “kynos” meaning movement. Cytokines are cell signalling molecules that aid cell to cell communication in immune responses and stimulate the movement of cells towards sites of inflammation, infection and trauma. Cytokines exist in peptide, protein and glycoprotein (proteins with a sugar attached) forms. The cytokines are a large family of molecules that are classified in various different ways due to an absence of a unified classification system. Examples of cytokines include the agents interleukin and the interferon which are involved in regulating the immune system’s response to inflammation and infection.

Dendritic cells are a type of antigen-presenting cells that form an important role in the adaptive immune system. The main function of dendritic cells is to present antigens. In addition, only the dendritic cells have the capacity to induce a primary immune response in the inactive or resting naive T lymphocytes. To do this, the dendritic cells capture the antigens from invading bodies, which they process and then present on their cell surface, along with the necessary accessory or co-stimulation molecules. Dendritic cells also contribute to the function of B cells and help to maintain their immune memory. Dendritic producing cytokines and other factors that promote B cell activation and differentiation.

Enzyme-Linked Immunosorbent Assays are the most widely used type of assay. They have evolved from viral lysate tests to tests containing recombinant protein and synthetic peptide antigens. They have high sensitivity and specificity. ELISAs are designed specifically for screening large numbers of specimens at a time, making them suitable for use in surveillance and centralized blood transfusion services. As ELISAs require sophisticated equipment and skilled technicians to perform the tests, their use is limited to certain circumstances.

Enzyme-Linked Immunospot assay is a highly sensitive immunoassay that measures the frequency of cytokine-secreting cells at the single-cell level. In this assay, cells are cultured on a surface coated with a specific capture antibody in the presence or absence of stimuli. Proteins, such as cytokines, that are secreted by the cells will be captured by the specific antibodies on the surface. After an appropriate incubation time, cells are removed and the secreted molecule is detected using a detection antibody in a similar procedure to that employed by the ELISA. The detection antibody is either biotinylated and followed by a streptavidin-enzyme conjugate or the antibody is directly conjugated to an enzyme. By using a substrate with a precipitating rather than a soluble product, the end results are visible spots on the surface. Each spot corresponds to an individual cytokine-secreting cell.

Env is a viral gene that encodes the protein forming the viral envelope. The expression of the Env gene enables retroviruses to target and attach to specific cell types, and to infiltrate the target cell membrane.

Gag (Group-specific antigen) is the genetic material that codes for the core structural proteins of a retrovirus.

Human Immunodeficiency Virus is a virus that infects cells of the immune system, destroying or impairing their function. Infection with the virus results in progressive deterioration of the immune system, leading to “immune deficiency”. The immune system is considered deficient if it is no longer able to fulfill its role of fighting infection and disease. Infections associated with severe immunodeficiency are known as “opportunistic infections”, because they take advantage of a weakened immune system.

Interferon gamma is a cytokine that is critical for innate and adaptive immunity against viral, some bacterial and protozoal infections. Aberrant IFN- γ expression is associated with a number of autoinflammatory and autoimmune diseases. The importance of IFN- γ in the immune system stems in part from its ability to inhibit viral replication directly, and most importantly from its immunostimulatory and immunomodulatory effects. IFN- γ is produced predominantly by natural killer and natural killer T cells as part of the innate immune response, and by CD4 Th1 and CD8 cytotoxic T lymphocyte effector T cells once antigen-specific immunity develops.

Immunogenicity is the ability of a particular substance, such as an antigen or epitope, to provoke an immune response in the body.

Latency Reversing Agents are small pharmacological molecules that could help uncover where HIV is hiding in the cells of HIV-positive individuals whose viral load has been suppressed below the level of treatment by effective ART.

Latent HIV reservoir Resting CD4 cells (or other cells) that are infected with HIV but not actively producing HIV. Latent HIV reservoirs are established during the earliest stage of HIV infection. Although ART can reduce the level of HIV in the blood to an undetectable level, latent reservoirs of HIV continue to survive. When a latently infected cell is reactivated, the cell begins to produce HIV again. For this reason, ART cannot cure HIV infection.

Lymphocyte is one of the subtypes of white blood cell in a vertebrate’s immune system. Lymphocytes include natural killer cells, T cells and B cells. They are the main types of cells found in lymph, which prompted the name lymphocyte.

Nef is a small protein encoded by primate lentiviruses. These include HIV-1, HIV-2 and SIV. Nef localizes primarily to the cytoplasm but also partially to the Plasma Membrane and is one of many pathogen-expressed proteins, known as virulence factors, which function to manipulate the host’s cellular machinery and thus allow infection, survival or replication of the pathogen. Nef stands for “Negative Factor” and although it is often considered dispensable for HIV-1 replication, in infected hosts the viral protein markedly elevates viral titers.

PBMC (Peripheral Blood Mononuclear Cell) is any peripheral blood cell having a round nucleus. These cells consist of lymphocytes (T cells, B cells, NK cells) and monocytes. These cells can be extracted from whole blood using ficoll, a hydrophilic polysaccharide that separates layers of blood, and gradient centrifugation, which will separate the blood into a top layer of plasma, followed by a layer of PBMCs and a bottom fraction of polymorphonuclear cells and erythrocytes.

Peptides are chemical agents belonging to the protein family. A peptide is composed of a mixture of several amino acids. These agents are involved in the composition of a large number of substances produced by the body, in particular hormones that regulate body functions, enzymes that carry out chemical reactions, transmitter molecules, neurotransmitters that carry nerve impulses, and so on.

Seroconversion is the time period during which a specific antibody develops and becomes detectable in the blood. After seroconversion has occurred, the disease can be detected in blood tests for the antibody. During an infection or immunization, antigens enter the blood, and the immune system begins to produce antibodies in response. Before seroconversion, the antigen itself may or may not be detectable, but the antibody is, by definition, absent. During seroconversion, the antibody is present but not yet detectable. Any time after seroconversion, the antibodies can be detected in the blood, indicating a prior or current infection.

SIV (Simian immunodeficiency viruses) are retroviruses that cause persistent infections in at least 45 species of African non-human primates.

Spot Within an ELISPOT well, spot is the “footprint” of a single cell that has released a relatively high amount of cytokines. True spots have a dense center with a light outer ring caused by the diffusion of the cytokine from the producing cell. The color depth or the size of spots depends on the amount of secreted cytokines.

Spot size is the relative amounts of cytokine produced per cell.

T cells or T lymphocyte, is a type of lymphocyte (a subtype of white blood cell) that plays a central role in cell-mediated immunity. T cells can be distinguished from other lymphocytes, such as B cells and natural killer cells, by the presence of a T-cell receptor on the cell surface. They are called T cells because they mature in the thymus from thymocytes.

Therapeutic vaccine When most people hear the word vaccine, they think of a way to prevent disease. However, therapeutic vaccines are not used for prevention. Instead, they are used as a method of treatment. Just like a regular vaccine, therapeutic vaccines are used to stimulate the immune system to target an infection or a type of diseased cell. In other words, they help teach the body how to do a better job of protecting itself in order to control, or get rid of, an otherwise difficult to treat condition.

Viral rebound When a person on ART has persistent, detectable levels of HIV in the blood after a period of undetectable levels. Causes of viral rebound can include drug resistance or poor adherence to an HIV treatment regimen.

Viral load HIV-1 viral load refers to the number of viral particles found in each milliliter. The more HIV-1 viral particles in the blood, the faster the CD4⁺ T-cells are likely destroyed and the faster the progress toward AIDS.