UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Finite Mixture Models based on Scale Mixtures of Skew-Normal distributions applied to serological data

”Documento Definitivo”

Doutoramento em Estatística e Investigação Operacional

Especialidade de Bioestatística e Bioinformática

Tiago Miguel Dias Domingues

Tese orientada por:

Nuno Henriques dos Santos de Sepúlveda

Maria Helena Mouriño Silva Nunes

Documento especialmente elaborado para a obtenção do grau de doutor

2021

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Finite Mixture Models based on Scale Mixtures of Skew-Normal distributions applied to serological data

Doutoramento em Estatística e Investigação Operacional

Especialidade de Bioestatística e Bioinformática

Tiago Miguel Dias Domingues

Tese orientada por:

Nuno Henriques dos Santos de Sepúlveda

Maria Helena Mouriño Silva Nunes

Júri:

Presidente:

- Doutora Maria Teresa dos Santos Hall de Agorreta de Alpuim, Professora Catedrática e Presidente do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa

Vogais:

- Doutor Nuno Henriques dos Santos de Sepúlveda, Investigador da *Politechnika Warszawska* (Orientador)
- Doutora Arminda Manuela Andrade Pereira Gonçalves, Professora Auxiliar da Escola de Ciências da Universidade do Minho
- Doutor Tiago Matias Machado dos Santos Seara Paixão, Investigador Auxiliar do Instituto Gulbenkian de Ciência
- Doutor Ruy Miguel Sousa Soeiro de Figueiredo Ribeiro, Professor Associado com Agregação da Faculdade de Medicina da Universidade de Lisboa
- Doutora Maria Ivette Leal de Carvalho Gomes, Professora Catedrática Aposentada da Faculdade de Ciências da Universidade de Lisboa
- Doutora Luísa da Conceição dos Santos do Canto e Castro de Loura, Professora Associada da Faculdade de Ciências da Universidade de Lisboa

Documento especialmente elaborado para a obtenção do grau de doutor

2021

Agradecimentos

Grande parte do conteúdo deste trabalho de doutoramento foi realizado durante um longo período de confinamento devido à pandemia que mudou as nossas vidas.

Gostaria de começar por agradecer aos meus orientadores, Professora Doutora Helena Mouriño e Professor Doutor Nuno Sepúlveda, por todos os ensinamentos durante a realização deste (duro) trabalho. Por terem acreditado que eu seria capaz de superar tal prova. Saio uma pessoa mais madura a nível pessoal e intelectual. Com referências que me vão acompanhar o resto da vida. A eles, o meu profundo agradecimento.

Ao Centro de Estatística e Aplicações (CEAUL) por toda a ajuda no financiamento deste projeto. À FCUL pela cedência de espaços para a realização dos nossos trabalhos.

À ação COST EUROMENE (European Network on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome) pelo financiamento da minha missão a Londres onde tive a oportunidade de trabalhar o primeiro conjunto de dados sobre fadiga crónica que será explorado ao longo deste trabalho. À Doutora Eliana Lacerda e ao Doutor Luis Nacul pela cedência dos dados. Aos investigadores da London School of Hygiene and Tropical Medicine com quem tive contacto e que me enriqueceram a todos os níveis.

À Professora Doutora Solange Gil, por me ter dado a possibilidade de aplicar as minhas metodologias num conjunto de dados gentilmente cedido por ela e pela sua equipa que levaram à publicação de um artigo.

À Escola Superior de Gestão e Tecnologia de Santarém (ESGTS) e ao Instituto Superior de Engenharia de Lisboa (ISEL) pelo apoio institucional. Em particular, agradecer aos coordenadores de área Professor Jorge Honório e Professora Isabel Duarte da ESGTS e ao Professor José Leonel Rocha do ISEL por acreditarem em mim.

A todos os meus colegas das referidas instituições pelas palavras de apoio e por tentarem aliviar a minha carga de trabalho durante a realização deste trabalho. Um agradecimento especial ao Ricardo.

À minha querida colega e amiga Jessica Lomba. Por termos sido o apoio um do outro. Por me ter dado a mão sempre que precisei. A sua generosidade e amabilidade são impagáveis. Obrigado por tudo.

Agradeço também aos meus restantes colegas de doutoramento: Ivo Ferreira, Marcos Roberto e Mafalda Ponte por todas as conversas e partilha de bons momentos.

À minha família, por serem o meu suporte emocional. Ao meu querido sobrinho, Tomás Domingues Fernandes, que sem saber faz com que os meus dias sejam mais felizes, aliviando os momentos mais stressantes da vida. Dedico ainda este trabalho a uma pessoa muito especial e que sei que muita se orgulha de assistir ao término desta etapa da minha vida: meu querido avô António Domingues.

Ao Gonçalo, por tudo o que representa. Por toda a paciência e apoio. Por me fazer perceber que a vida pode ser muito mais do que imaginamos.

Agradecer à Madalena Vaz da Silva. Por todos os dias dar significado à palavra ”amigo”. Não tenho palavras para descrever o quanto me ensina e me acrescenta enquanto pessoa. Agradeço também a todos os amigos que me apoiaram e continuam a apoiar.

Por último, aos meus alunos que me ensinam todos os dias.

A todos o meu obrigado.

Abstract

Serological data can be described as a mixture of distributions, with each mixture component representing a serological population (e.g. seronegative and seropositive population). In seroepidemiological studies of infectious diseases, mixture models with Normal distribution are mostly used, which implies that the components that make up the mixture are approximately symmetric. However, it has been observed that, especially in seropositive populations, it is possible to observe skewness to the left, leading to the violation of the assumption of normality underlying the data. Thus, and in order to capture the possible skewness in serological data, the family of Scale Mixtures of Skew-Normal (SMSN) distributions is used, of which the Skew-Normal distribution and the Skew-t distribution are particular cases. In the case of the Skew-t distribution, being a heavy-tailed distribution, it allows capturing the possible existence of outliers.

In addition to the models used to describe the behavior of the serological data, the issue of estimating the cutoff point for classifying an individual as seropositive is explored. In this sense, two perspectives on the problem are presented: one in which the true state of the disease is unknown; another in which this state is known a priori.

The generalization of the use of a cutoff point without statistical methodology to support the estimation of this point may have consequences in the seroprevalence of a population, that is, in the proportion of seropositive individuals. Thus, three methods based on mixture models are proposed in this work for estimating the cutoff point when the true infection status is unknown.

Keywords: serology; finite mixture models; skew-normal distribution; skew-t distribution; cutoff point.

Resumo

A serologia é a área científica que se dedica ao estudo do soro sanguíneo, nomeadamente à identificação de anticorpos no sangue, permitindo reconstruir o historial de infeção de um indivíduo.

Os dados serológicos podem ser modelados por uma mistura finita de distribuições, sendo que cada componente da mistura representa uma população serológica (p.ex. população seronegativa e seropositiva).

Os modelos de mistura finitos são modelos bastante flexíveis, utilizados para modelar dados de populações heterogéneas, permitindo captar características como multimodalidade, assimetria e curtose. O racional por detrás destes modelos baseia-se na consideração de subpopulações em número finito e em diferentes proporções.

O método da máxima verosimilhança com recurso ao algoritmo de expectação-maximização (EM) é o método usual para a estimação dos parâmetros do modelo.

Em estudos seroepidemiológicos de doenças infecciosas são utilizados maioritariamente modelos de mistura de componentes com distribuição normal, o que implica que as mesmas sejam simétricas. No entanto, tem sido possível observar assimetrias à esquerda, sobretudo nas populações seropositivas, levando a que o pressuposto da normalidade dos modelos de mistura gaussianos seja violado.

Desta forma, e de modo a captar o possível enviesamento em dados serológicos, consideramos aqui a família de distribuições baseadas em misturas do parâmetro de escala da distribuição normal-assimétrica, das quais a distribuição normal-assimétrica e a distribuição t de Student assimétrica são casos particulares. Esta família de distribuições foi inicialmente discutida por Andrews e Mallows (1974), tendo o seu trabalho sido estendido por Branco e Dey (2001). Desta forma, considerando modelos de mistura baseados em distribuições da família de misturas do parâmetro de escala da distribuição normal-assimétrica, tem-se que para a estimação dos parâmetros é utilizada uma variante do algoritmo EM, designado por algoritmo

de expectação-maximização condicional (ECM). Esta versão do método consiste em particionar o passo M em sub-passos, condicional aos valores dos parâmetros do passo anterior. Trata-se de uma alternativa que simplifica o passo M do tradicional algoritmo EM quando é necessário proceder à estimação de muitos parâmetros, permitindo reduzir o tempo computacional para a produção das estimativas dos mesmos.

Tal como foi referido anteriormente, a distribuição t de Student assimétrica é um caso particular da família de misturas do parâmetro de escala da distribuição normal-assimétrica, sendo esta uma distribuição de caudas pesadas, que permite captar a possível existência de *outliers* e modelar convenientemente possíveis observações extremas. Até ao momento este é um trabalho pioneiro na aplicação desta família de distribuições a dados serológicos.

Para ilustrar a aplicação destes modelos foram utilizados dados de anticorpos contra sete herpesvírus (citomegalovirus (CMV), Epstein-Barr antigénio EBNA1 (EBV-EBNA1), Epstein-Barr antigénio VCA (EBV-VCA), vírus do herpes humano tipo 6 (HHV-6), herpes simplex tipo 1 (HSV1), herpes simplex tipo 2 (HSV2) e vírus da varicela zoster (VZV)), disponíveis no banco de dados de síndrome de fadiga crónica do Reino Unido. Foram ainda analisados dados de anticorpos contra o vírus SARS-CoV-2 (*Severe acute respiratory syndrome coronavirus 2*) resultantes da análise de quatro antigénios (RBD – *glycoprotein receptor-binding domain*; S^{tri} — *S trimeric spike protein*; S1 — *spike glycoprotein S1 domain*; S2 – *SARS-CoV-2 spike glycoprotein S2 domain*), disponibilizados por Rosado et al., 2020.

No caso do primeiro conjunto de dados, pretendeu-se averiguar a possível existência de relação entre a exposição a um herpesvírus e a manifestação da condição, uma vez que até ao momento a etiologia da síndrome de fadiga crónica é desconhecida. Numa primeira fase, foram ajustados aos dados de anticorpos os modelos baseados nas distribuições normal-assimétrica e t de Student assimétrica, bem como os tradicionais modelos simétricos (Normal e t de Student) para comparação. Observou-se então que os modelos baseados em distribuições assimétricas apresentam melhor qualidade de ajustamento. Além disso, como a exposição do indivíduo ao vírus não era conhecida *a priori*, foi possível testar o número adequado de componentes do modelo de mistura através de *bootstrap* paramétrico. Relativamente aos vírus para os quais as infeções ocorrem maioritariamente na infância, observou-se apenas uma população serológica (seropositiva), como é o caso dos vírus HHV-6 e VZV. Para o vírus HSV1, verificou-se que o modelo que melhor se ajusta aos dados é o modelo baseado na distribuição normal-assimétrica com três componentes. Neste caso, a componente intermédia foi interpretada como sendo correspondente a uma subpopulação seronegativa, tendo em conta que a distribuição é assimétrica à direita e sendo esta a interpretação dada a populações seronegativas, segundo estudos anteriores. Para os restantes vírus, foram consideradas duas componentes serológicas (i.e, populações seronegativas e seropositivas).

No caso dos dados relativos ao vírus SARS-CoV-2, dado que o verdadeiro estado de infeção é conhecido *a priori*, considerou-se o ajustamento de um modelo de mistura com duas componentes (seropositiva e seronegativa, respetivamente).

Além dos modelos utilizados para descrever o comportamentos deste tipo de dados, é explorada a questão da estimação do ponto de corte para proceder à classificação do estado serológico de um indivíduo, isto é, proceder à categorização da concentração de anticorpos. Note-se que, no caso em que o verdadeiro estado de infeção não é conhecido (caso do conjunto de dados de síndrome de fadiga crónica), não é possível utilizar métodos de determinação do ponto de corte, como, por exemplo, métodos baseados na curva ROC. Nesse sentido, são apresentadas duas perspetivas sobre o problema: uma em que o verdadeiro estado de infeção é desconhecido; outra em que esse estado é conhecido *a priori*.

No primeiro cenário, o procedimento usual (tendo em conta que na maioria dos casos é assumida a normalidade dos dados), é utilizar a regra dos 3-σ, isto é, selecciona-se uma população de controlo (seronegativa) e procede-se ao cálculo da média mais 3 desvios padrão. A generalização da utilização de um ponto de corte sem metodologia estatística que sustente a sua estimação pode ter consequências nas conclusões acerca da seroprevalência de uma população, isto é, da proporção de indíviduos seropositivos. Assim, são propostos neste trabalho três métodos baseados em modelos de mistura para a estimação do ponto de corte, aplicados quando o verdadeiro estado de infeção é desconhecido: 1) determinação do quantil de probabilidade 99.9% da população seronegativa; 2) determinação do mínimo das densidades do modelo de mistura; 3) determinação do ponto para o qual é fixada *a priori* em 90% a probabilidade condicional de classificação do indivíduo como seropositivo dada a quantidade de anticorpo. Os métodos propostos foram validados através da sua aplicação ao conjunto de dados de anticorpos contra o vírus SARS-CoV-2, onde o verdadeiro estado de infeção é conhecido, procedendo-se ao cálculo da sensibilidade, especificidade e acurácia. Foi ainda realizado um estudo de simulação aplicado aos dados de anticorpos contra o vírus SARS-CoV-2, com o objetivo de avaliar a capacidade dos modelos para identificar as respetivas populações serológicas, variando a dimensão da amostra ($100, 500$ e 1000) e a proporção de indivíduos seronegativos

(30%, 60% e 90%) e seropositivos (70%, 40% e 10%). Os resultados obtidos revelam algumas fragilidades dos modelos na identificação das respetivas populações serológicas, considerando pequenas amostras e valores extremos da proporção de indivíduos seronegativos e seropositivos (p.ex. 30% e 90%), alertando para a possível existência de falsos positivos e/ou falsos negativos nestas situações. Além deste aspeto, estes resultados também permitem avaliar a qualidade dos programas de vacinação aplicados, bem como o tempo de imunização da população.

Com este trabalho pretende-se fornecer novas ferramentas para a análise e tratamento de dados serológicos, sendo uma alternativa a metodologias *standard*, pouco específicas, utilizadas até aqui para modelação e estimação de ponto de corte.

Palavras-chave: serologia; modelos de mistura finitos; distribuição normal-assimétrica; distribuição t de Student assimétrica; ponto de corte.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AIC: Akaike's Information Criterion

BIC: Bayesian Information Criterion

CMV: Cytomegalovirus

EBV: Epstein-Barr virus

EBNA1: Epstein-Barr nuclear antigen 1

ECM: Expectation-conditional-maximization algorithm

ELISA: Enzyme-linked immunoabsorbent assays

EM: Expectation-maximization algorithm

HC: Healthy controls

HHV-6: Human herpes virus 6

HSV1: Herpes simplex virus 1

HSV2: Herpes simplex virus 2

IgA: Immunoglobulin type A

IgD: Immunoglobulin type D

IgE: Immunoglobulin type E

IgG: Immunoglobulin type G

IgM: Immunoglobulin type M

LRT: Likelihood Ratio Test

ME/CFS: Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

ME-M: Mild/moderate symptoms of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

ME-S: Severely affected patients with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

MLE: Maximum likelihood estimator

MS: Multiple sclerosis

OD: Optical density

RBD: glycoprotein receptor-binding domain

ROC: Receiver Operating Characteristic

RT-qPCR: Reverse transcription quantitative PCR

S1: spike glycoprotein S1 domain

S2: spike glycoprotein S2 domain

S^{tri}: S trimeric spike protein

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

SMSN: Scale Mixtures of Skew-Normal distributions

SN: Skew-Normal distribution

ST: Skew-t distribution

VCA: Viral Capsid Antigen

VZV: Varicella-Zoster virus

Scientific contributions of this Ph.D work

- **Published Work**

- **Dias Domingues, T.**, Grabowska, A., Lee, Ji-Sook, Ameijeiras-Alonso, J., Westermeier, F., Scheibenbogen, C., Cliff, J., Nacul, L. Lacerda, E., Mouriño, H., Sepúlveda, N. (2021). ”Herpesviruses serology distinguishes different subgroups of patients from the United Kingdom Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Biobank”. In: *Frontiers in Medicine* (*main reference for chapter 4*)

- **Dias Domingues, T.**, Mouriño, H. and Sepúlveda, N. (2021). “Analysis of antibody data using Finite Mixture Models based on Scale Mixtures of Skew-Normal distributions”. In: medRxiv. DOI: https://doi.org/10.1101/2021.03.08.21252807. (*main reference for chapter 3*)

- **Dias Domingues, T.**, Mouriño, H. and Sepúlveda, N. (2020). “A statistical analysis of serological data from the UK myalgic encephalomyelitis/chronic fatigue syndrome biobank”. In:AIP Conference Proceedings 2293.1. DOI: https://doi.org/10.1063/5.0026633 (*for more details see appendix B*)

- **Additional Work**

- Moreira da Silva, J., Prata, S., **Dias Domingues, T.**, Leal, R. O., Nunes, T., Tavares, L., Almeida, V., Sepúlveda, N., Gil, S. (2020). “Detection and modeling of anti-Leptospira IgG prevalence in cats from Lisbon area and its correlation to retroviral infections, lifestyle, clinical andhematologic changes”. In:Veterinary and Animal Science, 10, p. 100144. DOI: https://doi.org/10.1016/j.vas.2020.100144 (*for more details see appendix C*)

Chapter 1

Introduction

To introduce the theme of this doctoral thesis, it is inevitable to mention the year 2020, when the country and the world were ravaged by the COVID-19 pandemic that changed the lives of millions of people since the subject of this doctoral thesis is the analysis of serological data. The words serology, antibodies, seropositive, seroprevalence, among others, have taken over everyday conversations. The scientific community had to mobilize to provide answers on how to act against a virus that was totally unknown until then. Once again, Science put itself at the service of society, revealing its power in knowing the unknown.

When this doctoral work began we were far from imagining what the future would bring, more specifically in the application of this doctoral work in a pandemic scenario.

The main goal of this work was to use mixture models based on distributions from the Scale Mixtures of Skew-Normal (SMSN) distributions family to model serological data. Also, we aimed to answer the problem of estimating the cutoff point for classifying an individual as seropositive. Currently, and according to the knowledge we have so far, many laboratories manufacture serological tests using a general criterion for estimating the cutoff point, the most used being the mean plus three standard deviations. Thus, alternative methods to estimate the cutoff point are proposed throughout this work.

In order for the reader to become familiar with the type of data, some basic concepts on serological data, essential for understanding the following chapters, will be presented throughout this chapter. More specifically, the concept of antibody, its types and functions in the immune system are introduced. Some techniques for obtaining serological data are also presented, namely laboratory techniques such as enzyme-linked immunoabsorbent assays (ELISA) that are the most common in this type of analysis.

In chapter 2 we present a brief description of the statistical methodology used in this work. It starts with a characterization of the Skew-Normal and Skew-t distributions that are the distributions used in this work. This distributions are part of the SMSN distributions having the particularity of incorporating an additional parameter to the traditional symmetric distributions that allows to control the skewness of the data. Next, some basic concepts of mixture models are introduced. The basic assumption of these models is that data is composed of different latent populations, each one representing a distinct serological state.

Chapters 3, 4 and 5 concern applications to real data of the methodologies used throughout this work. In particular, chapters 3 and 4 are already under peer review for possible publication. In chapter 3 we used mixture models based on SMSN distributions to a data set related to antibodies against 6 different common herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) from the UK biobank. In addition to the previously mentioned models we explore the definition of an individual's serological status. An individual is classified as seropositive by performing a serological test for which an antibody value greater than a certain value c causes them to be classified as seropositive, otherwise they are classified as seronegative. The usual interpretation is that individuals are seropositive after a recent infection whilst seronegative individuals reflect either absence of any infection or an infection that occurred a long time ago. For a given infectious agent, there are several criteria to define a cutoff for seropositivity.

In the most pragmatic approach, one uses a general cutoff advised by the manufacturer of the lab reagents used for antibody quantification. However, it is unclear whether such cutoff holds true in general or whether its determination followed any statistical rationale.

In order to circumvent this scourge, three methods are proposed for estimating the cutoff point for defining an individual's serological status. Note that the motivation to define a method for estimating the cutoff point arises in cases where the true infection status is not known. Otherwise, there are several methods in the literature to determine an optimal cutoff point, the best known being the Receiver Operating Characteristic (ROC) curve.

Thus, in the application chapter 3 we also explore the cutoff point estimation problem for the case where the true infection status of the individuals is unknown, since ME/CFS is not even recognized as a disease by the World Health Organization (WHO).

Whereas viruses can be disease-causing (note that the human immunodeficiency virus (HIV) causes the acquired immunodeficiency syndrome (AIDS)), and the ME/CFS is a condition with unknown etiology, in chapter 4 we conducted an association study in order to understand if is possible to establish an association between ME/CFS and chronic herpesviruses infections. Basically, it is intended to evaluate the impact of the choice of cutoff point on the presence of a clinical condition.

In chapter 5 we explore the definition of an individual serological status from the perspective of knowing the true infection status. Antibody data against Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was used for this purpose and are freely avaliable at https://github.com/MWhite-InstitutPasteur/SARSCoV2SeroDXphase2. Note that since we know the true infection status of the individual it is possible to evaluate the quality of the methods established in chapter 3 though sensitivity, specificity, and accuracy, as well as compared to other methods that are already well described in the literature, such as the ROC curve-based method.

Chapter 6 presents the general conclusions of this work.

Finally, the bibliography and some appendices are presented, namely some supplementary material from chapter 4, papers presented in conferences and collaborations with other institutions where the methods developed here were applied.

1.1 Basic concepts in Human serology[1]

Human serology is the area dedicated to the study of serum or blood plasma. More specifically, serology is used to identify antibodies against different microorganisms that invade the human body, providing concrete answers on the epidemiology of specific diseases, as well as allowing the evaluation of the effectiveness of immunization / vaccination programs. Basically,

[1] Main reference: The Immunological Basis for Immunization Series. Module 1: General Immunology. Global Programme for Vaccines and Immunization. World Health Organization

serology allows the reconstruction of an individual's infection history.

Daily, the human being comes into contact with pathogens responsible for diseases such as viruses, bacteria, fungi and parasites, and the human immune system is responsible for protection against these pathogens. The defense organism's function is immunity, which is classified as innate or non-specific and adaptive.

Innate immunity is always present in the body, acting in a similar way to any exposure to a foreign agent. Adaptive immunity, as the name implies, is specific / adapted / personalized to a given pathogen.

The human immune system is a very complex system in its own right, and it is well known that the most important cellular component in this system is leukocyte cells, also called leukocytes or white blood cells. Leukocytes are produced in bone marrow and are subsequently released into the blood and transported throughout the body. We can say that these are the soldiers who are on duty to protect against the enemy.

From the moment a certain pathogen invades the organism, we enter the first phase of the immune system's action through innate immunity. This includes the skin, mucous membranes, hair and other mechanical factors, which prevent the entry of the pathogen or remove it from the body surface. If the pathogen crosses these barriers, then you will encounter another biochemical one where there are chemical mediators such as saliva, sweat, tears and hydrochloric acid in the stomach that try to prevent the pathogen from entering the cells. Once inside the organism and in the last line of defense of the innate system there are several effector cells and mechanisms that try to hinder or destroy the pathogen, including cytokines and cells such as natural killers, macrophages and neutrophils.

When the innate responses are not sufficient to remove the pathogen, the adaptive immunity is then activated. This is the stage where the production of antibodies by lymphocytes occurs. The activity of the adaptive immune system is ensured by two types of lymphocytes: B lymphocytes or B cells and T lymphocytes or T cells. B cells are responsible for the production of antibodies, and T cells are divided into two main sub-groups: CD8+ T cells which kill cells infected with a pathogen; and CD4+ T cells which help coordinate the response of both CD8+ T cells and B cells.

The antibodies produced by B cells are nothing more than glycoproteins[2], also called im-

[2] Glycoproteins are proteins containing glycans attached to amino acid side chains. Glycans are oligosaccharide chains; which are saccharide polymers, that can attach to either lipids (glycolipids) or amino acids (glycoproteins). Typically, these bonds are formed through a process called glycosylation. For more information see https://www.news-medical.net/health/What-is-a-Glycoprotein.aspx

munoglobulins (Ig) and their function is to neutralize and prevent the normal functioning of pathogens. Each pathogen contains antigens to which antibodies bind via receptors. These receptors are part of the natural structure of the antibody and their shape is adapted to the antigenic determinant or epitope[3] of the antigen. That is why it is often said that antibodies specific to a particular antigen are produced. Basically, it is possible to make the analogy that the relationship between antibody and antigen is like a key and lock.

Schematically,

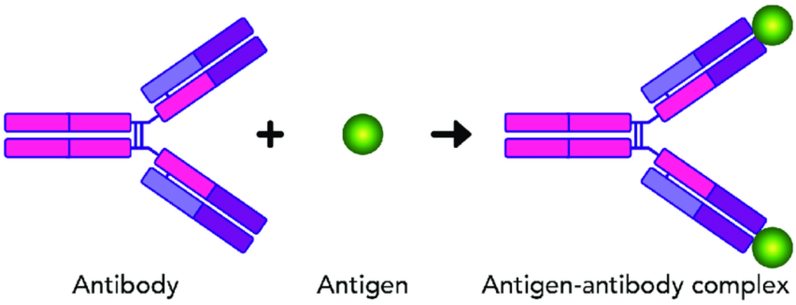


Figure 1.1: Antibody-antigen structure (Salazar et al., 2017)

After the first contact with a particular antigen, the adaptive immune system learns to deal with that pathogen by memorizing it. This property of the system means that successive infections caused by the same pathogen are treated more quickly and efficiently by the adaptive immune system.

For the destruction of an antigen to be effective, there must be a great affinity between the antibody that binds to the different epitopes that make up an antigen. In this way, it is necessary that the B cells producing the antibodies expand in order to guarantee a sufficient number to fight the respective antigen. Due to the mass production of antibodies, there are those that will serve to eliminate the antigen and there will be others that will continue to circulate in the body acting on future infections.

Depending on their function, antibodies can belong to five classes, namely IgA, IgD, IgE, IgG and IgM.

IgG are the main immunoglobulins present in the bloodstream and represent about 80 % of the total immunoglobulins. These are responsible for containment of viruses and bacteria,

[3] An epitope, also known as antigenic determinant, is the part of an antigen that is recognized by the immune system, specifically by antibodies, B cells, or T cells. The epitope is the specific piece of the antigen to which an antibody binds.

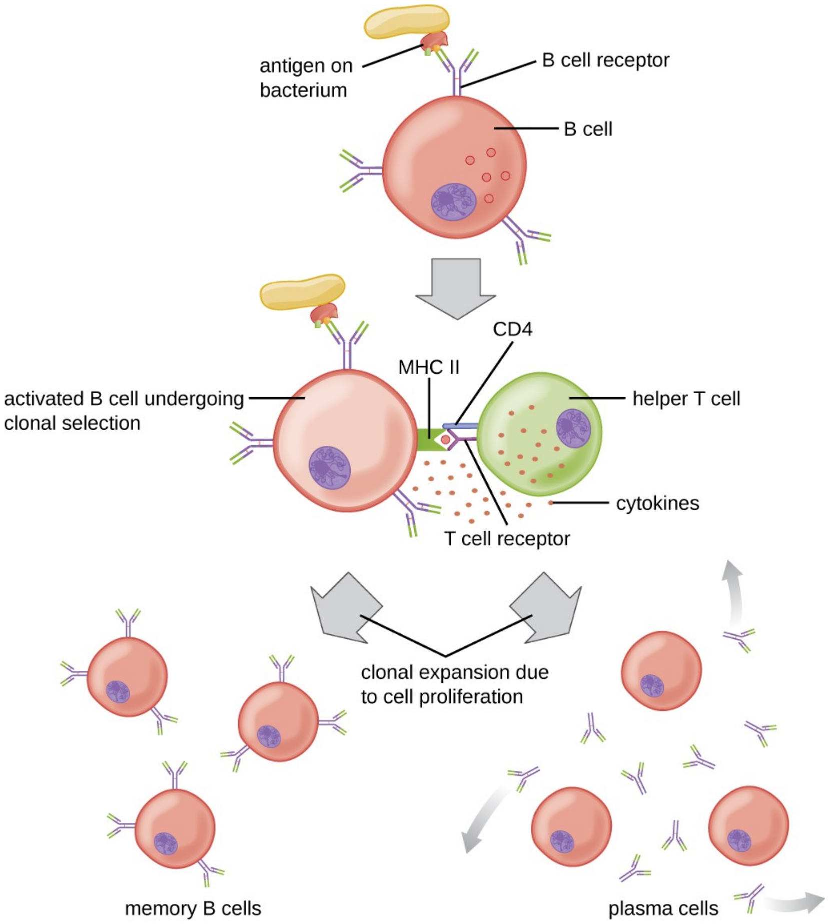


Figure 1.2: Antibody mediated-immunity (Source: https://courses.lumenlearning.com/cuny-kbcc-microbiologyhd/chapter/b-lymphocytes-and-humoral-immunity/)

helping in phagocytosis and bacterial lysis[4]. IgGs are also able to pass through the maternal placenta. IgMs are confined to the bloodstream and are unable to pass through the maternal placenta. IgA is the second most abundant immunoglobulin in serum. It is present in the secretions of the gastrointestinal and respiratory tracts as well as in colostrum and breast milk. Regarding IgD, its function is not yet fully known.

In the next section we explore the behavior of the most important immunoglobulins in the immune response, IgG and IgM depending on the type of immune response.

1.2 Primary and secondary immune response

As mentioned above, the most important immunoglobulins in the immune response are IgG and IgM.

[4] Lysis is the breaking down of the membrane of a cell, often by viral, enzymic, or osmotic mechanisms that compromise its integrity.

When a pathogen invades the human body, the production of antibodies against pathogen-derived antigens takes up to 10 days. During this phase, also called the ”lag phase” the lymphocytes find the antigen replicating itself and only then start the production of antibodies. It is at this stage of the initial contact with the antigen, also called the primary response, that the concentration of IgM and IgG antibodies increases considerably, reaching a plateau and then decreasing as the response is produced and the pathogen is eliminated.

Considering that there is a second encounter with the same antigen, that is, a secondary response, the immune response is much faster in the sense that antibodies are produced more quickly and it is possible to observe more persistent levels of them. This latter property is largely due to IgGs, also called ”memory antibodies” that remain in circulation even after the antigen has been eliminated. Figure 1.3 shows the relationship between IgM and IgG antibodies taking into account a primary and secondary response.

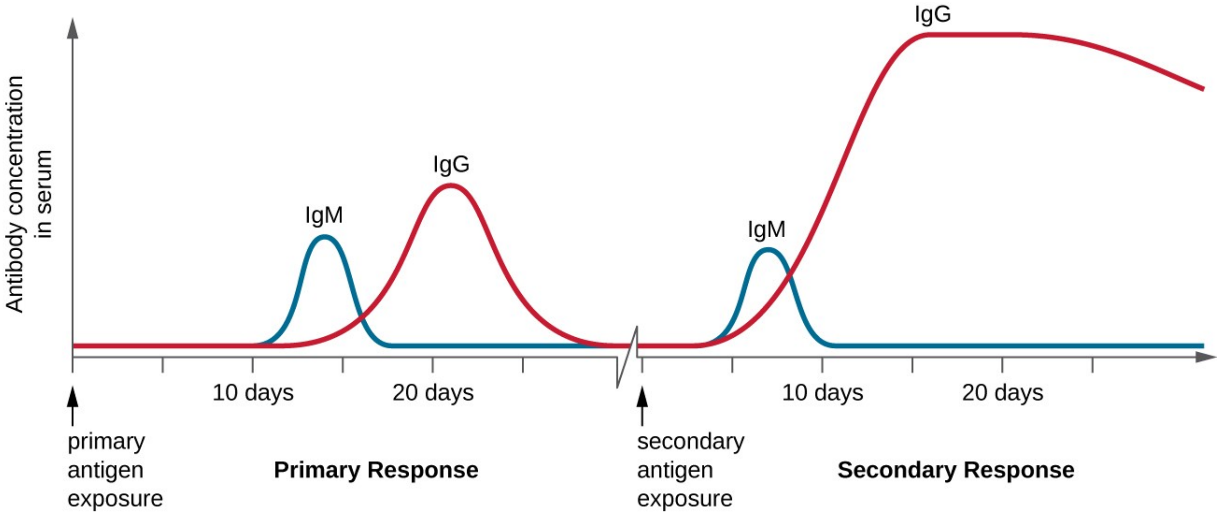


Figure 1.3: Primary immune response versus secondary immune response (Source: https://courses.lumenlearning.com/cuny-kbcc-microbiologyhd/chapter/b-lymphocytes-and-humoral-immunity/)

Next, we discuss one of the most used techniques for the quantification of antibodies and antigens, the ELISA assays.

1.3 Quantification of antibody concentration

The detection of antibodies in serum or blood plasma can be done through different techniques, the most used being ELISA assays. ELISA assays emerged in 1971 when two groups

of researchers introduced an immunoenzymatic method for the detection and quantification of specific antigens or antibodies against a given pathogen. They observed that proteins could be immobilized on a solid polystyrene surface and that the antibody-antigen reaction could be revealed by the formation of colored products formed by the enzyme-substrate reaction, in the presence of an electron donor component called chromogen (Hipólito, 2017).
ELISA assays can be classified into direct, indirect, sandwich, competitive and non-competitive, with each method defined by the research method used, that is, whether the research relates to antibodies or antigens.

The direct method is used to search for antigens in the serum that will react with antibodies specific to that antigen producing an intensity of colored light called optical density (OD). This light intensity reveals the amount of antibodies that have bound to the antigen. Staining is produced by a protein called an enzyme conjugate, the most used proteins being alkaline phosphatase and peroxidase.

The indirect method is used to search for primary antibodies against a given antigen by binding those (unconjugated) antibodies to conjugated secondary antibodies. The color is developed when the enzyme substrate is added. The color intensity is directly proportional to the concentration of antigens present in the solution sample.

The sandwich method considers a primary (unconjugated) antibody immobilized under the plate that captures the antigen. Subsequently, the conjugated secondary antibody is added that will bind to the antigen captured by the immobilized primary antibody, thus forming a sandwich. It is through the conjugated secondary antibody that the intensity of light is then produced, its intensity being proportional to the concentration of antigen in the sample (Figure 1.4).

In competitive tests, based on a microplate with an inert surface with wells, the antigens or antibodies that are thus bound to the plate are placed in it. If antigens are placed, then an amount of antibody will be added that will bind to the antigen. It turns out that the higher the concentration of antigen, the lower the ability of the antibody to establish the binding, hence the term ”competition”.

As mentioned earlier, in ELISA assays, whenever binding of an antibody (analyte) to an antigen is established, a coloration is produced, and the more intense it is, the greater the amount of antibody that has bound to the antigen. The light intensity is translated in terms of OD values and the OD values are compared with known concentrations of antigen (Figure 1.5). Next the

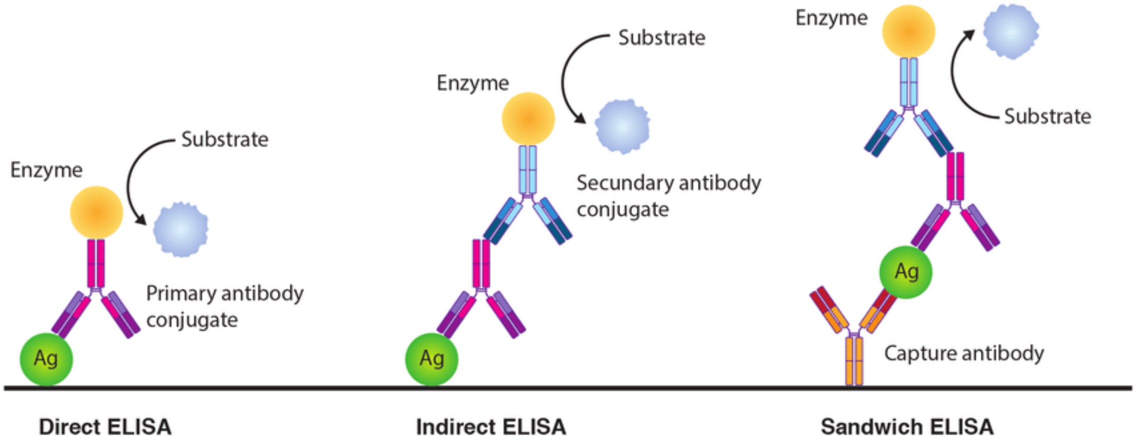


Figure 1.4: Direct, indirect and sandwich ELISA assays (Salazar et al., 2017)

most common procedure is to fit a 4 parameter logistic model (4-PL) (Hoogen et al., 2020), that is

$$y_j = a + \frac{(d-a)}{1+(x_j/c)^b},$$

where y_j is the response at concentration x_j, a is the upper asymptote, d is the lower asymptote, c is the concentration at the inflection point of the curve and b is the growth factor (Azadeh et al., 2017).

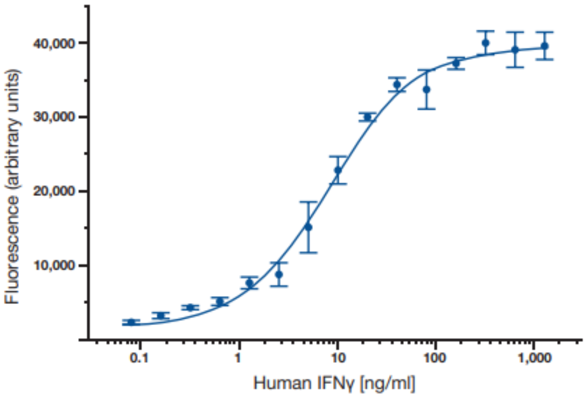


Figure 1.5: ELISA standard curve (Source: https://www.bio-rad-antibodies.com/elisa-results-quantitative-semi-qualitative.html). x axis: known concentrion of antigen; y axis: OD values (light intensity).

Note that the antibody concentration tends to be very low for OD values very close to the lower asymptote while OD values very close to the upper asymptote correspond to higher antibody concentration. During the whole process of constructing the calibration curve, there are several statistical procedures that should be taken into account when obtaining antibody data, namely the minimum detectable concentration, the reliable detection limit and the limit of quantitation which are well described in O’Connell et al., 1993.

Chapter 2

Statistical methodology

After the reader has had a first contact with what serological data are, we introduce in this chapter the statistical methodology that will be used to analyze this type of data. In particular, the reader's attention is drawn to the fact that given the organization of this document, in each of the application chapters the statistical methodology used is described in great detail, so we will only give a brief introduction in this chapter. Throughout the text we will refer to the application chapters that explore each of the methodologies used.

When analyzing data of the antibody responses against a specific virus, it is usually assumed the existence of two or more latent, unobservable populations representing different serological states (e.g., seronegative and seropositive). In this scenario, data is typically described by a mixture of two or more probability distributions (Dias Domingues et al., 2020). We start this chapter with some basic concepts in finite mixture models.

2.1 Finite mixture models

Finite mixture models are very flexible models, used to model data from heterogeneous populations, allowing to capture population characteristics such as multimodality, skewness and kurtosis (Lachos Dávila et al., 2018). The rationale behind these types of models is that, given a population, it is possible to consider subpopulations in a finite number, in different proportions, with each subpopulation characterized by a probability density function and a parameter space. Naturally, in the case in which infinite subpopulations are considered, we are in the presence of the so-called infinite mixture models (which we do not address in this work).

The development of finite mixture models deserved the attention of Everitt and Hand (1981), Lindsay (1995), Böhning (2000), McLachlan and Peel (2000) due to the mathematical treatability of these types of models, but largely due to the flexibility of their use in several areas such as biology, medicine and economics. The first important work on mixture distributions is due to Karl Pearson (1894), in which he adjusted a mixing model with normal components to data that represented the distance from the forehead to the body length of 100 crabs from Bay of Naples (G. McLachlan and Peel, 2000). Throughout this chapter we review the properties of these models, considering their application in the general case and specifying their application to serological data in chapters 3, 4 and 5.

2.1.1 Formulation of a finite mixture model

In general, let $Z_1,...,Z_n$ be the identical and independent random variables for a sample of size n, $G_1,...,G_g$ be the partition from a superpopulation G (sample space) and $\pi_1,...,\pi_g$ the probabilities of sampling an observation belonging to each latent population (with the usual restriction of $\sum_{k=1}^{g}\pi_k = 1$ and $0 < \pi_k \leq 1$). The probability density function (pdf) of a mixture of distributions is then given by

$$f(z) = \sum_{k=1}^{g} \pi_k f_k(z;\theta_k), \tag{2.1}$$

where $f_k(z;\theta_k)$ is the mixing probability density function associated with $k-$th latent population and parameterized by a vector θ_k. The quantities π_k are the proportions or weights of the mixture. The number of components g, can be a known value or a parameter to be estimated from a sample.

As previously mentioned, the density of a mixture of distributions depends on parameters that need to be estimated, so they become part of a parametric family, and the density of the mixture can be written in the form

$$f(z;\Theta) = \sum_{k=1}^{g} \pi_k f_k(z;\theta_k), \tag{2.2}$$

being Θ the vector that contains all the unknown parameters of the mixture model, being defined by $\Theta = (\pi_1,...,\pi_{g-1},\theta_1,...,\theta_g)$.

2.1.2 Identifiability of Mixture Models

When a mixture model is fitted to a set of observations in which it is unknown which component each observation belongs to, it is important to ensure the model identifiability, i.e, there is a unique characterization for any of the mixing distributions considered (G. McLachlan and Peel, 2000). This fact is important for the parameter estimation.

Definition 1 *A mixture of distributions with pdf given in (2.2) is said to be identifiable if and only if*

$$\begin{aligned} \sum_{k=1}^{g} \pi_k f_k(z;\theta_k) &= \sum_{j=1}^{\tilde{g}} \tilde{\pi}_j f_j(z;\tilde{\theta}_j) \Rightarrow \\ &\Rightarrow \quad g = \tilde{g} \wedge (\forall k = 1,...,g, \exists j = 1,...,\tilde{g} : \pi_k = \tilde{\pi}_j \wedge \theta_k = \tilde{\theta}_j). \end{aligned} \tag{2.3}$$

In addition to characterizing a mixture model through its density function and its distribution function, it is necessary to characterize it in terms of location and scale.

2.1.3 Moments of mixture models

Definition 2 *Let Z a random variable with pdf given by (2.2). The moments of order n of Z are*

$$E(Z^n) = \sum_{k=1}^{g} \pi_k E(Z_k^n), \tag{2.4}$$

where $E(Z_k^n)$ is the moment of order n of a random variable with pdf $f_k(z;\theta_k), k = 1,...,g$.

Definition 3 *The variance of a random variable with pdf given by (2.2) is given by*

$$\begin{aligned} V(Z) &= E(Z^2) - E^2(Z) \\ &= \sum_{k=1}^{g} \pi_k E(Z_k^2) - E^2(Z) \\ &= \sum_{k=1}^{g} \pi_k (V(Z_k) + E^2(Z_k)) - E^2(Z) \\ &= \sum_{k=1}^{g} \pi_k V(Z_k) + \sum_{k=1}^{g} \pi_k (E(Z_k) - E(Z))^2, \end{aligned} \tag{2.5}$$

where $V(Z_k)$ is the variance of a random variable with pdf $f_k(z;\theta_k), k = 1, ..., g$.

2.1.4 Maximum Likelihood method in mixture models

Regarding the estimation of the parameters of a mixture model, several methods have been proposed, namely the maximum likelihood method, graphic methods, minimum distance method and Bayesian methods (G. McLachlan and Peel, 2000).

In the context of serological data, the serological status of an individual can be a latent variable leading to a problem of incomplete data. In this sense we explore the maximum likelihood method via EM algorithm to estimate the model parameters.

2.1.4.1 EM algorithm

The EM algorithm is an iterative method used to calculate the maximum likelihood estimators in the context of incomplete data whenever the maximum likelihood method does not produce analytical solutions.

Consider $(z_1, ..., z_n)$ an observed random sample of dimension n from a mixture of g components with pdf given by

$$f(z_i;\Theta) = \sum_{k=1}^{g} \pi_k f_k(z_i;\theta_k), \tag{2.6}$$

being Θ the vector that contains all the unknown parameters of the mixture model, i.e., the vector to be estimated by the maximum likelihood method. The correspondent log likelihood function is given by

$$\log L(\Theta) = \sum_{i=1}^{n} \log \big(\sum_{k=1}^{g} \pi_k f_k(z_i;\theta_k) \big). \tag{2.7}$$

For the algorithm to be applied, it is necessary to constitute the complete sample, that is, the sample that contains the information of the unobservable latent variable, Y, which for a sample of dimension n we will consider as being defined by $(Y_1, ..., Y_n)$ with $Y_i = (Y_{i1}, ..., Y_{ig})$, where the element k of Y_i, designated by y_{ik} is defined as follows (G. McLachlan and Peel, 2000)

$$y_{ik} = \begin{cases} 1 & \text{, if } z_i \text{ comes from the } k^{th} \text{ component} \\ 0 & \text{, otherwise} \end{cases} \tag{2.8}$$

In this way, the complete sample is then formed by the pair (z_n, y_n) independent and identically distributed in which $Y_1, ..., Y_n$ are independent realizations of a multinomial distribution, i.e,

$$Y_1, ..., Y_n \frown Multinomial(1, \pi_1, ..., \pi_g).$$

The pdf is given by

$$f(y_i; \Theta) = \prod_{k=1}^{g} \pi_k^{y_{ik}}, \tag{2.9}$$

that is, $f(y_i; \Theta)$ is the marginal probability density of Y. Considering that Y is the indicator variable of the component from which z_i comes from, we can obtain the conditional probability of Z_i given $Y_i = y_i$, i.e,

$$f_{Z_i|Y_i=y_i}(z_i; \Theta) = \prod_{k=1}^{g} f_k(z_i, \theta_k)^{y_{ik}}. \tag{2.10}$$

By the Total Probability Theorem, we obtain the pdf of the complete data, that is the joint probability density,

$$f((z_i, y_i); \Theta) = \prod_{k=1}^{g} [\pi_k f_k(z_i, \theta_k)]^{y_{ik}}. \tag{2.11}$$

and the log-likelihood function is given by

$$\log L(\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{g} y_{ik} \log\{\pi_k f_k(z_i; \theta_k)\}. \tag{2.12}$$

As we can see by the expression (2.12), we can't obtain the maximum likelihood estimators (MLE) for the parameters since $Y_i = y_i$ is a random variable. Then we need a iterative process to obtain the MLE for the parameters.

2.1.4.2 The iterative process

The EM algorithm works through two steps, an E (expectation) and an M (maximization) steps.

In step E and in the iteration $p+1$ of the algorithm, the conditional expected value of the log likelihood function defined by the equation (2.12) given the incomplete sample is calculated, using the value for Θ the value in the previous iteration, $\Theta^{(p)}$. That is,

$$Q(\Theta,\Theta^{(p)}) = E_{\Theta^{(p)}}\{\log L(\Theta)|z_i\}. \tag{2.13}$$

Note that function (2.12) is linear in y_{ik}, so with function (2.13) we intend to calculate the probability that a given observation, z_i, belongs to the $k-th$ component. In fact,

$$Q(\Theta,\Theta^{(p)}) = \sum_{i=1}^{n}\sum_{k=1}^{g} E_{\Theta^{(p)}}\{Y_{ik}|z_i\}\log\{\pi_k f_k(z_i\dot{;}\theta_k)\}. \tag{2.14}$$

How,

$$\begin{aligned} E_{\Theta^{(p)}}\{Y_{ik}|z_i\} &= P_{\Theta^{(p)}}\{Y_{ik}=1|z_i\} \\ &= \frac{P_{\Theta^{(p)}}\{Y_{ik}=1,z_i\}}{P_{\Theta^{(p)}}(z_i)} \\ &= \frac{\pi_k^{(p)} f_k(z_i;\theta_k^{(p)})}{\sum_{h=1}^{g}\pi_h^{(p)} f_h(z_i;\theta_h^{(p)})} \\ &= w_{ik}^{(p+1)} (i=1,...,n;k=1,...,g), \end{aligned} \tag{2.15}$$

Thus, given the result in (2.15), the expression (2.14) can be written in the form

$$Q(\Theta,\Theta^{(p)}) = \sum_{i=1}^{n}\sum_{k=1}^{g} w_{ik}^{(p+1)}\log\{\pi_k f_k(z_i;\theta_k)\}. \tag{2.16}$$

In the $(p+1)$ interaction of the step M is calculated the new value of Θ that maximizes the expression (2.16), that is, the updated maximum likelihood estimates of the parameters are determined, $\Theta^{(p+1)}$. Dempster et al., 1977 demonstrated that

$$L(\Theta^{(p+1)}) \geq L(\Theta^{(p)}), p = 0, 1, ... \tag{2.17}$$

which implies that L converges to some L^* by a previously limited sequence of values.
The algorithm stops as soon as the stopping criterion is established, that is, whenever the value of the criterion becomes less than a given constant (G. McLachlan and Peel, 2000).

The serological populations that make up a mixture can exhibit distinct behaviors given the particularities of antibody data, in particular IgG data.

Parker et al., 1990 states that the distribution of the seronegative population is Normal, noting that in the case of the seropositive population there is a skew of the antibody data to the left, because there is a decrease over time. This fact motivated us to use distributions that capture this skewness such as the Skew-Normal and Skew-t distributions. In the section 2.2 a brief characterization of these distributions is presented.

2.2 Scale Mixtures of Skew-Normal distributions as mixing distributions

The study of class of distributions used in this work to model serological data begins with the most widely known and used probability distribution, the Normal distribution.

Due to its good properties, namely the fact that it is symmetrical around the mean, its support is the real line, has several applications due to the central limit theorem, the normal distribution also ends up being used to model serological data.

In many studies it is assumed that the components that make up a mixture are well modelled by a normal distribution with mean value μ and standard deviation σ. It turns out that the serological data have some peculiarities, namely the fact that they can be asymmetric in each component, presenting heavy tails, which ends up violating the assumption of symmetry present in the normal distribution. Thus, it becomes necessary to control the skewness inherent in this type of data.

Azzalini, 1985 dedicated to the study of skewed distributions, namely in a generalization of the normal distribution including a skewness parameter which he called the Skew-Normal distribution. The Skew-Normal distribution is a particular case of a class of distributions called Scale Mixtures of Skew-Normal (SMSN) distributions, of which the Skew-t distribution is also a part and which will also be used throughout this work (Basso et al., 2010; Dias Domingues et al., 2021). The following is a brief characterization of these distributions.

2.2.1 Special case: Skew-Normal distribution

Definition 4 *A random variable W has a Skew-Normal distribution with location parameter μ, scale parameter σ^2 and skewness parameter α (denoted as $W \frown SN(\mu, \sigma^2, \alpha)$) if its probability density function (pdf) can be written as*

$$\begin{aligned} f_W(w) &= 2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(w-\mu)^2}{2\sigma^2}} \times \int_{-\infty}^{\alpha\frac{(w-\mu)}{\sigma}} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\ &= 2\phi(w;\mu,\sigma^2)\Phi\left(\frac{\alpha(w-\mu)}{\sigma}\right), \; w,\mu,\alpha \in \mathbb{R}, \; \sigma^2 \in \mathbb{R}^+ \end{aligned} \tag{2.18}$$

where $\phi(.;\mu,\sigma^2)$ denotes the pdf of the Normal distribution with mean μ and variance σ^2; $\Phi(.)$ denotes the the cumulative distribution function of the standard Normal distribution.

For the sake of simplicity, considering the standard Skew-Normal distribution, $SN(\alpha)$, we have that when $\alpha = 0$, the $SN(\alpha)$ reduces to the $N(0,1)$ and when $\alpha \to \infty$, the $SN(\alpha)$ converges to the half-Normal distribution (Figueiredo et al., 2013).

The mean and variance of the Skew-Normal distribution are respectively given by,

$$E(W) = \mu + \sigma\sqrt{\frac{2}{\pi}}\frac{\alpha}{\sqrt{1+\alpha^2}}, \quad V(W) = \left(1 - \left(\frac{2}{\pi}\frac{\alpha}{\sqrt{1+\alpha^2}}\right)^2\right)\sigma^2. \tag{2.19}$$

2.2.2 Special case: Skew-t distribution

Definition 5 *A random variable Z has a Skew-t distribution with location parameter μ, scale parameter σ^2, skewness parameter α and v degrees of freedom, i.e., $Z \frown ST(\mu, \sigma^2, \alpha, v)$, if its pdf is given by*

$$f_Z(z) = 2\,t(z;\mu,\sigma,v+1)\;T\left(A\sqrt{\frac{v+1}{d+v}};v+1\right). \tag{2.20}$$

where $t(.;\mu,\sigma,v+1)$ denotes the probability density function of a Generalized Student-t

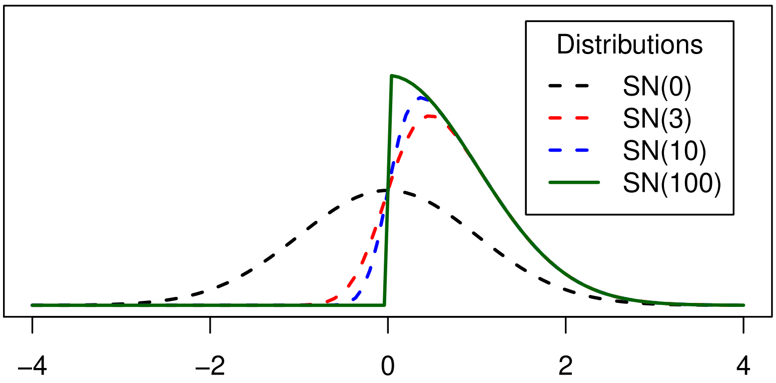


Figure 2.1: Density functions of standard Skew-Normal distribution varying the skewness parameter

distribution[1] *with location parameter* μ, *scale parameter* σ *and* $v+1$ *degrees of freedom;* $T(.;v+1)$ *represents the cumulative distribution function of a standard Student-t distribution with* $v+1$ *degrees of freedom.*

For the sake of simplicity, considering the standard Skew-t distribution, $ST(\alpha)$, we have that when $\alpha = 0$, the $ST(\alpha)$ reduces to the standard Student's t distribution, when $\alpha \to \infty$, the $ST(\alpha)$ converges to the half-t distribution and when $v \to \infty$ converges to the Skew-Normal distribution (Azzalini, 2014).

The mean and variance of the Skew-t distribution are respectively given by,

$$E(Z) = \mu + \sigma b_v \delta, \text{ if } v > 1, \quad V(Z) = \sigma^2 \left[\frac{v}{v-2} - \left(b_v \delta \right)^2 \right] \text{ if } v > 2, \tag{2.21}$$

where $b_v = \frac{\sqrt{v}\,\Gamma(\frac{1}{2}(v-1))}{\sqrt{\pi}\,\Gamma(\frac{1}{2}v)}$ and $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$.

The full derivation of the Skew-Normal and Skew-t distributions as particular cases of the SMSN family can be found in the application chapter 3, sub-section 3.3.1.

[1] We consider the Generalized Student's t-distribution as the non-standardized Student's t-distribution (Jackman, 2009)

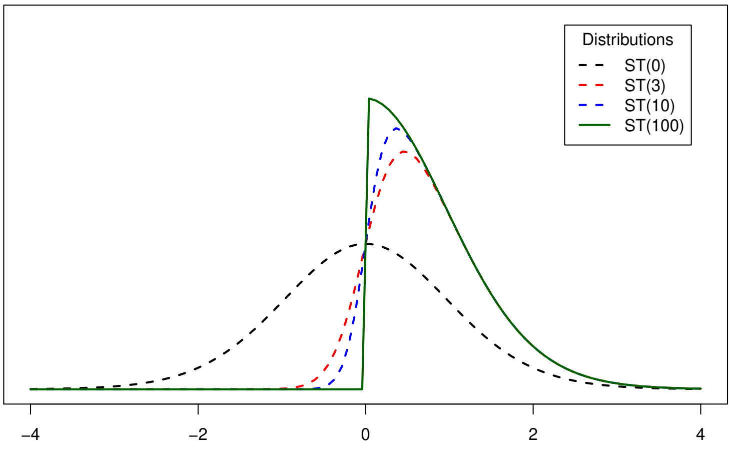


Figure 2.2: Density functions of standard Skew-t distribution varying the skewness parameter

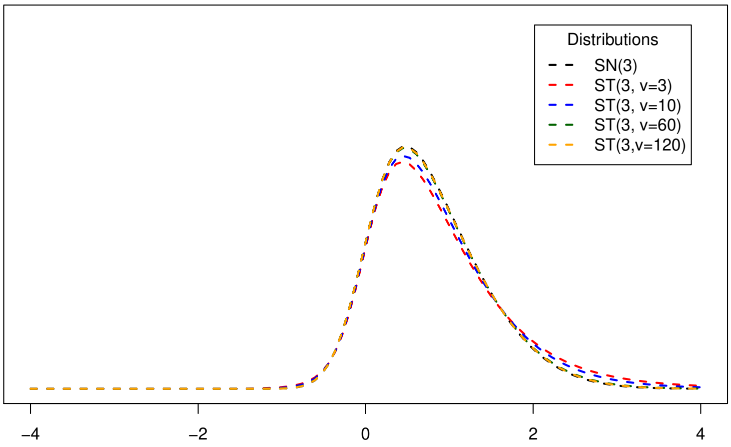


Figure 2.3: Density functions of Skew-t distribution with $\alpha = 3$ and with degrees of freedom $v = 3, 10, 60, 120$

2.2.3 Maximum likelihood estimators of the Skew-Normal and Skew-t distributions

Regarding the estimation of the parameters of the Skew-Normal distribution and the Skew-t distribution, it is not possible to obtain closed expressions for the parameter estimators by the maximum likelihood method as proved below (Yalçınkaya et al., 2018):

2.2.3.1 Case of the Skew-Normal distribution

For a random sample of size n, the likelihood function is given by

$$\begin{aligned} L(\Theta;\mathbf{w}) &= \prod_{i=1}^{n} f_W(w_i) = \prod_{i=1}^{n} \frac{2}{\sigma}.\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{w_i-\mu}{\sigma})^2}.\Phi\Big(\alpha(\frac{w_i-\mu}{\sigma})\Big) = \\ &= \prod_{i=1}^{n} \frac{2}{\sigma}.\prod_{i=1}^{n}(2\pi)^{-1/2}.e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{w_i-\mu}{\sigma})^2}.\prod_{i=1}^{n}\Phi\Big(\alpha(\frac{w_i-\mu}{\sigma})\Big). \end{aligned} \tag{2.22}$$

Then, the log-likelihood is given by

$$\log(L(\Theta;\mathbf{w})) = n\log(\frac{2}{\sigma}) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{w_i-\mu}{\sigma}\right)^2 + \sum_{i=1}^{n}\log\left(\Phi\Big(\alpha(\frac{w_i-\mu}{\sigma})\Big)\right). \tag{2.23}$$

ML for μ:

$$\begin{aligned} \frac{\partial \log(L(\Theta;\mathbf{w}))}{\partial\mu} &= -\frac{1}{2}\sum_{i=1}^{n} 2.\Big(\frac{w_i-\mu}{\sigma}\Big).\Big(-\frac{1}{\sigma}\Big) + \sum_{i=1}^{n}\frac{\phi\Big(\alpha.(\frac{w_i-\mu}{\sigma})\Big)}{\Phi\Big(\alpha.(\frac{w_i-\mu}{\sigma})\Big)}.\Big(\alpha.(-\frac{1}{\sigma})\Big) = \\ &= \frac{1}{\sigma}\sum_{i=1}^{n}\Big(\frac{w_i-\mu}{\sigma}\Big) - \sum_{i=1}^{n}\frac{\phi\Big(\alpha.(\frac{w_i-\mu}{\sigma})\Big)}{\Phi\Big(\alpha.(\frac{w_i-\mu}{\sigma})\Big)}.\Big(\frac{\alpha}{\sigma}\Big). \end{aligned} \tag{2.24}$$

ML for σ:

$$\begin{aligned}
\frac{\partial \log(L(\Theta;\mathbf{w}))}{\partial \sigma} &= -\frac{n}{\sigma}-\frac{1}{2}\sum_{i=1}^{n}-2.\left(\frac{w_i-\mu}{\sigma}\right)\left(\frac{w_i-\mu}{\sigma^2}\right)-\sum_{i=1}^{n}\frac{\phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}{\Phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}.\left(\alpha.(-\frac{w_i-\mu}{\sigma^2})\right)= \\
&= -\frac{n}{\sigma}+\sum_{i=1}^{n}\left(\frac{(w_i-\mu)^2}{\sigma^3}\right)+\sum_{i=1}^{n}\frac{\phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}{\Phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}.\left(\alpha.(\frac{w_i-\mu}{\sigma^2})\right)= \\
&= \frac{1}{\sigma}\left(-n+\sum_{i=1}^{n}\left(\frac{w_i-\mu}{\sigma}\right)^2+\sum_{i=1}^{n}\frac{\phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}{\Phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}.\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)\right) \qquad (2.25)
\end{aligned}$$

ML for α:

$$\frac{\partial \log(L(\Theta;\mathbf{w}))}{\partial \alpha} = \sum_{i=1}^{n}\frac{\phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}{\Phi\left(\alpha.(\frac{w_i-\mu}{\sigma})\right)}.\left(\frac{w_i-\mu}{\sigma}\right). \tag{2.26}$$

2.2.3.2 Case of the Skew-t distribution

For a random sample of size n, the likelihood function is given by

$$\begin{aligned}
L(\Theta;\mathbf{z}) &= \prod_{i=1}^{n} f_Z(z_i)=\prod_{i=1}^{n}2.\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}.\left(1+\frac{\left(\frac{z_i-\mu}{\sigma}\right)^2}{v}\right)^{-\frac{v+1}{2}}.F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}};v+1\right)= \\
&= \prod_{i=1}^{n}2.\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}.\prod_{i=1}^{n}\left(1+\frac{\left(\frac{z_i-\mu}{\sigma}\right)^2}{v}\right)^{-\frac{v+1}{2}}.\prod_{i=1}^{n}F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}};v+1\right).
\end{aligned} \tag{2.27}$$

We have then that the log-likelihood is given by,

$$
\begin{aligned}
\log\Bigg(L(\Theta;\mathbf{z})\Bigg) &= \log\left(\prod_{i=1}^{n} 2.\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\right) + \log\left(\prod_{i=1}^{n}\left(1+\frac{\left(\frac{z_i-\mu}{\sigma}\right)^2}{v}\right)^{-\frac{v+1}{2}}\right) + \\
&+ \log\left(\prod_{i=1}^{n} F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}};v+1\right)\right) = \\
&= n\Bigg(\log(2)+\log\left(\Gamma\left(\frac{v+1}{2}\right)\right)-\log(\sigma)-\frac{1}{2}\log(v\pi)-\log\left(\Gamma\left(\frac{v}{2}\right)\right)+ \\
&+ \left(\frac{v+1}{2}\right)\log(\sigma^2 v)\Bigg)+\left(-\frac{v+1}{2}\right)\sum_{i=1}^{n}\log\left(\sigma^2 v+(z_i-\mu)^2\right)+ \\
&+ \sum_{i=1}^{n}\log\left(F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}};v+1\right)\right). \qquad (2.28)
\end{aligned}
$$

ML for μ:

$$
\begin{aligned}
\frac{\partial\log(L(\Theta;\mathbf{z}))}{\partial\mu} &= \left(-\frac{v+1}{2}\right)\sum_{i=1}^{n}-\frac{2(z_i-\mu)}{\sigma^2 v+(z_i-\mu)^2}+\sum_{i=1}^{n}\frac{f_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}{F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}\times \\
&\times \left(\left(-\frac{\alpha}{\sigma}\right).\sqrt{\frac{v+1}{d(z_i)+v}}+A(z_i).\left(\frac{v+1}{d(z_i)+v}\right)^{-\frac{1}{2}}.\left(\frac{(\frac{z_i-\mu}{\sigma^2}).(v+1)}{(d(z_i)+v)^2}\right)\right) \\
&= (v+1)\sum_{i=1}^{n}\frac{(z_i-\mu)}{\sigma^2 v+(z_i-\mu)^2}+\sum_{i=1}^{n}\frac{f_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}{F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}\times \\
&\times \left(\left(-\frac{\alpha}{\sigma}\right).\sqrt{\frac{v+1}{d(z_i)+v}}+A(z_i).\left(\frac{v+1}{d(z_i)+v}\right)^{-\frac{1}{2}}.\left(\frac{(\frac{z_i-\mu}{\sigma^2}).(v+1)}{(d(z_i)+v)^2}\right)\right)
\end{aligned}
\qquad (2.29)
$$

ML for σ:

$$\begin{aligned}\frac{\partial \log(L(\Theta;\mathbf{z}))}{\partial \sigma} &= \frac{nv}{\sigma} + \left(-\frac{v+1}{2}\right)\sum_{i=1}^{n}\frac{2\sigma v}{\sigma^2 v+(z_i-\mu)^2} + \sum_{i=1}^{n}\frac{f_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}{F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)} \times \\ &\times \left(\left(-\frac{\alpha}{\sigma^2}\right).(z_i-\mu).\sqrt{\frac{v+1}{d(z_i)+v}} + A(z_i).\left(\frac{v+1}{d(z_i)+v}\right)^{-\frac{1}{2}}.\left(\frac{(\frac{(z_i-\mu)^2}{\sigma^3}).(v+1)}{(d(z_i)+v)^2}\right)\right)\end{aligned} \tag{2.30}$$

ML for α:

$$\frac{\partial \log(L(\Theta;\mathbf{z}))}{\partial \alpha} = \sum_{i=1}^{n}\frac{f_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}{F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)} \times \left(\left(\frac{z-\mu}{\sigma}\right).\sqrt{\frac{v+1}{d(z_i)+v}}\right). \tag{2.31}$$

ML for v:

$$\begin{aligned}\frac{\partial \log(L(\Theta;\mathbf{z}))}{\partial v} &= n\left(\frac{1}{2}\psi\left(\frac{v+1}{2}\right) - \frac{1}{2v} - \frac{1}{2}\psi\left(\frac{v}{2}\right) + \frac{v+1}{2v} + \frac{\log(\sigma^2 v)}{2}\right) + \\ &+ \left(-\frac{1}{2}\right)\sum_{i=1}^{n}\log\left(\sigma^2 v+(z_i-\mu)^2\right) + \\ &+ \left(-\frac{v+1}{2}\right)\sum_{i=1}^{n}\frac{\sigma^2}{\sigma^2 v+(z_i-\mu)^2} + \sum_{i=1}^{n}\frac{f_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)}{F_T\left(A(z_i).\sqrt{\frac{v+1}{d(z_i)+v}}\right)} \times \\ &\times \left(A(z_i).\frac{1}{2}.\left(\frac{v+1}{d(z_i)+v}\right)^{-\frac{1}{2}}.\left(\frac{d(z_i)-1}{(d(z_i)+v)^2}\right)\right),\end{aligned} \tag{2.32}$$

where $\psi(.)$ denotes the digamma function, i.e., $\psi(x) = \frac{\partial}{\partial x}\ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$.

Thus, it is necessary to use iterative methods for parameter estimation. There are several iterative methods for estimating the parameters, such as Newton-Raphson (NR), Nelder Mead (NM), and Iteratively Re-weighting Algorithm (IRA) (Yalçınkaya et al., 2018). Since we will be using mixtures of Skew-Normal and Skew-t distributions throughout our applications, we will use the EM algorithm for estimating the model parameters. In particular, in the case of distributions belonging to the SMSN family, we will use a variant of the EM algorithm called expectation-conditional-maximization (ECM) whose description of the algorithm is detailed in then application chapter 5, sub-section 5.3.2.

Chapter 3

Serological analysis of herpesviruses using data from the United Kingdom ME/CFS biobank

Finite mixture models have been widely used in antibody (or serological) data analysis in order to help classifying individuals into either antibody-positive or antibody-negative. The most popular models are the so-called Gaussian mixture models which assume a Normal distribution for each component of a mixture. In this work, we propose the use of finite mixture models based on a flexible class of scale mixtures of Skew-Normal distributions for serological data analysis. These distributions are sufficiently flexible to describe right and left asymmetry often observed in the distributions associated with hypothetical antibody-negative and antibody-positive individuals, respectively. We illustrate the advantage of these alternative mixture models with a data set of 406 individuals in which antibodies against six different human herpesviruses were measured in the context of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome.

3.1 Introduction

Antibodies are key immunological proteins produced by B cells upon molecular recognition of an antigen derived from an infectious agent. In general, they contribute to microbial clearance and, if maintained in the body over time, they comprise the basis of the so-called immunological memory, which translates into a quicker and more efficient immune response in the case of repeated exposure to the same infectious agent. In turn, autoantibodies are antibodies recognizing components from the body and they are usually present in autoimmunity diseases, such as multiple sclerosis or rheumatoid arthritis. In the laboratory, antibodies (or autoantibodies) against a specific antigen are usually quantified by the enzymatic-linked immunosorbent assays (ELISA) using serum samples; see ref. Wang et al., 2003; Vila Nova et al., 2018; Shikova et al., 2020 for some recent studies using these assays. The respective readout is a light intensity, also known as optical density, which can be converted into a concentration or a titre using a calibration curve of known antibody concentrations. In practice, these assays are easily standardized, widely available, and ideal for high-throughput analysis of antibodies against a single antigen (Wang et al., 2003). Such advantages make ELISA particularly suitable for large-scale sero-epidemiology surveys where one aims to estimate the prevalence of exposure to a given pathogen in the population (Wang et al., 2003; Cook et al., 2011; Hsiang et al., 2012). With the recent development of high-throughput technologies, antibody quantification is currently shifting from the traditional ELISA to more advance assays, such as microarray (Helb et al., 2015; Loebel et al., 2017), luminex (Lammie et al., 2012; Blomberg, Rizwan, et al., 2019), or cytometry bead assays (Sowa et al., 2017), where a large number of different antibodies can be evaluated in the same serum sample. However, some of these promising technologies still require some degree of optimization before their widely applicability in biomedical research (Hoogen et al., 2020; Wu et al., 2020).

Statistical analysis of antibody (or serological) data is usually carried out under the assumption that the antibody distribution consists of different latent populations, each one representing a distinct antibody state or different degrees of exposure to a given antigen. This assumption gives rise to a serological data analysis based on finite mixture models (G. McLachlan and Peel, 2000). These models can be more or less complex depending on the number of components and mixing distributions used to describe the data. Due to its conceptual simplicity and easy of interpretation, the most popular finite mixture model in routine serological applications invokes the existence of two components related to hypothetical seronegative and seropositive individuals or, equivalently, antibody-negative and antibody-positive individuals (Gay, 1996; Chis Ster, 2012; Rogier et al., 2015). Models comprising more than two components have also been found appropriate to describe data from some studies (Parker et al., 1990; Baughman et al., 2006; M. C. Rota et al., 2008; Nhat et al., 2017; Moreira da Silva et al., 2020), but they might bring some ambiguity when interpreting which components are associated with antibody positivity (Sepúlveda, Stresman, et al., 2015). In turn, the most popular choice for the mixing distributions is the Lognormal distribution in the original scale of the measurements or, equivalently, the Normal distribution after logarithmic transformation of the data (Chis Ster, 2012; Parker et al., 1990). Gamma and Weibull are other choices for the mixing distributions among textbook probability distributions (Rogier et al., 2015; Nhat et al., 2017). Alternatively, less trivial mixture models can be used in the analysis. This is the case of a mixture between two truncated Normal distributions describing the situation where observations might fall below the lower limit of detection or above the upper limit of detection of the assay (Baughman et al., 2006). Another interesting model is a mixture between a Normal and a combination of half-Normal distributions for the hypothetical seronegative and seropositive populations, respectively (Gay, 1996). The rationale behind this proposal is that the distribution of the seropositive population should be left skewed, because antibody levels tend to decrease over time (Parker et al., 1990). Notwithstanding their suitability to tackle specific characteristics of serological data, none of the above models would appear to provide sufficiently flexibility in terms of skewness and flatnesss of each mixing distribution that could be used as the basis of data analysis automation in high-throughput serological studies.

In this scenario, we propose the scale mixture of Skew-Normal distributions (SMSN) as a flexible mixing distribution for serological data analysis. The flexibility of this family is attributed to four parameters that control the location, the scale, the skewness and the flatness of the resulting distribution. In addition, SMSN includes the Normal distribution, the Generalized Student's t-distribution, and its skewed version as special cases (Basso et al., 2010). We illustrate the advantage of using these models by analysing a data set related to antibodies against 6 different common herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) (Cliff et al., 2019).

3.2 Data under analysis

ME/CFS is complex disease whose patients experience a long-lasting fatigue that cannot be alleviated by rest or suffer from post-exertional malaise upon minimal physical and mental activity (Fukuda et al., 1994; Carruthers et al., 2003). The aetiology of the disease remains unknown, but it is often linked with common viral infections, including common herpesviruses (Rasa et al., 2018).

To accelerate current knowledge on ME/CFS, it was created a large disease-specific biobank in the United Kingdom (E. M. Lacerda, Bowman, et al., 2017; E. M. Lacerda, Mudie, et al., 2018). The data set under analysis is part of this biobank and it was published in a recent study with the aim of investigating the immunological component of the disease (Cliff et al., 2019). In the data set, there is a total of 406 individuals, all adults, divided into three main groups: healthy controls (HC, $n = 107; 26.4\%$), patients with ME/CFS ($n = 250; 61.8\%$), and patients with multiple sclerosis (MS, $n = 49; 12.1\%$). The group of patients with ME/CFS was further divided into a subgroup of 196 patients with mild or moderate symptoms (ME-M) and another subgroup of 54 severely affected patients who are home- or even bed-bound (ME-S). A detailed description about the recruitment of study participants, inclusion/exclusion criteria, and ethics can be found in the original reference (Cliff et al., 2019).

The data set comprises six serological variables corresponding to the antibody concentration against the following common herpesviruses: human cytomegalovirus, CMV; Epstein-Barr virus, EBV; human herpesvirus-6, HHV-6; types 1 and 2 herpes simplex viruses, HSV-1 and HSV-2, respectively; and varicella-zoster virus, VZV. Note that the tested antibodies against EBV were specific to the viral-capsid antigen.

In each serum sample, the concentration of each vira-specific antibody was expressed in arbitrary units per ml (U/ml) according the corresponding optical density determined by commercial ELISA kits. According to the ELISA's kit manufacturers, samples with antibodies concentration ≤ 8 U/ml should be classified as seronegative and those with concentration ≥ 12 U/ml should be classified as seropositive for all antibodies with the exception of HHV-6. Samples with IgG concentration between 8 and 12 U/ml should be classified as equivocal. For antibodies against HHV-6, seronegative and seropositivity should be defined as ≤ 10.5 U/ml or ≥ 12.5 U/ml, respectively. Samples with concentrations between these two limits were considered equivocal.

3.3 Statistical analysis of serological data

3.3.1 Finite mixture models based on SMSN distributions

When analysing serological data related to the antibody responses against a specific antigen, it is usually assumed the existence of two or more latent, unobserved populations, which might represent different levels of exposure to that antigen. For simplicity, individuals that were never exposed to a given antigen are considered as seronegatives whilst individuals exposed to it are considered seropositives. In this scenario, the respective data from a specific antibody are typically described by a mixture of two or more probability distributions.

Let $G_1,...,G_g$ be the partition from a superpopulation G (sample space) and $\pi_1,...,\pi_g$ the probabilities of sampling an individual belonging to each latent population (with the usual restriction of $\sum_{k=1}^{g}\pi_k = 1$ and $0 \leq \pi_k \leq 1$). A random variable Z is a finite mixture of independent random variables $Z_1, Z_2, ..., Z_g$ if the probability density function (pdf) of Z is given by

$$f(z) = \sum_{k=1}^{g} \pi_k f_{Z_k}(z;\boldsymbol{\theta}_k), \tag{3.1}$$

where $f_{Z_k}(z;\boldsymbol{\theta}_k)$ is the mixing probability density function of Z_k associated with the k-th latent population and parameterized by the vector $\boldsymbol{\theta}_k = \left\{\theta_1,...,\theta_g\right\}$.

The most popular choice for the mixing distribution in serological analysis is the Normal distribution which is symmetric around the mean and it is a mesokurtic distribution (with a kurtosis of 3 irrespective of the mean and standard deviation of the distribution). However, serological data from populations on the brink of malaria elimination show long tails and marked right asymmetry (Rogier et al., 2015) in each latent population even after applying log-transformation. In such cases, one can use instead the Generalized Student t as the mixing distribution, because it has heavier tails than the Normal distribution. However, this distribution remains in the realm of the symmetric distributions. To incorporate asymmetry in the modelling, one can alternatively use the less-known Skew-Normal as the mixing distribution (Lin et al., 2007a).

A random variable W_k has a Skew-Normal distribution with location parameter μ_k, scale parameter σ_k^2 and skewness parameter α_k (denoted as $W_k \frown SN(\mu_k, \sigma_k^2, \alpha_k)$) if its pdf can be written as

$$\begin{aligned} f_{W_k}(w) &= 2\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{(w-\mu_k)^2}{2\sigma_k^2}} \times \int_{-\infty}^{\alpha_k\frac{(w-\mu_k)}{\sigma_k}} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\ &= 2\phi(w;\mu_k,\sigma_k^2)\Phi\left(\frac{\alpha_k(w-\mu_k)}{\sigma_k}\right),\ w,\mu_k,\alpha_k \in \mathbb{R},\ \sigma_k \in \mathbb{R}^+ \end{aligned} \tag{3.2}$$

where $\phi(.;\mu_k,\sigma_k^2)$ denotes the pdf of the Normal distribution with mean μ_k and variance σ_k^2; $\Phi(.)$ denotes the the cumulative distribution function of the standard Normal distribution (Azzalini, 1985; Azzalini and Capitanio, 2003; Basso et al., 2010). The mean and variance of the Skew-Normal distribution are respectively given by,

$$E(W_k) = \mu_k + \sigma_k\sqrt{\frac{2}{\pi}}\frac{\alpha_k}{\sqrt{1+\alpha_k^2}}, \quad V(W_k) = \left(1-\left(\frac{2}{\pi}\frac{\alpha_k}{\sqrt{1+\alpha_k^2}}\right)^2\right)\sigma_k^2. \tag{3.3}$$

Additionally, the Skew-Normal distribution can be used to construct a more general class of flexible distributions, the scale mixtures of Skew-Normal (SMSN) distributions.

The random variable Z in expression (3.1) belongs to the SMSN family with location parameter μ_k, scale parameter σ_k^2, skewness parameter α_k and mixing distribution $H_k(.;\boldsymbol{\nu}_k)$ parameterized by θ_k (denoted as $Z_k \frown SMSN(\mu_k,\sigma_k^2,\alpha_k)$) if it can be written in the following way:

$$Z_k = \mu_k + \frac{W_k}{\sqrt{U_k}}, \tag{3.4}$$

where U_k is a random variable with distribution function $H_k(.,\boldsymbol{\nu}_k)$ and pdf $h_k(.,\boldsymbol{\nu}_k)$; $\boldsymbol{\nu}_k$ is either a scalar or a vector of parameters indexing the distribution of U_k; and $W_k \frown SN(0,\sigma_k^2,\alpha_k)$ which is assumed to be independent of U_k (Basso et al., 2010; Lachos Dávila et al., 2018).

Based on expression (3.4), it is worth noting that the conditional distribution
$Z_k|U_k = u$ takes the form

$$\begin{aligned} F_{Z_k|U_k=u}(z) &= P(Z_k \leq z \mid U_k = u) = P(\mu_k + \frac{1}{\sqrt{u}}W_k \leq z) \\ &= P(W_k \leq \sqrt{u}(z-\mu_k)) = F_{W_k}(\sqrt{u}(z-\mu_k)),\ z \in \mathbb{R}. \end{aligned} \tag{3.5}$$

Thus,

$$
\begin{aligned}
f_{Z_k|U_k=u}(z) &= \frac{d}{dz}|_{U_k=u}\, F_{W_k}(z) = \sqrt{u} \times f_{W_k}(\sqrt{u}(z-\mu_k)) \qquad (3.6)\\
&= \sqrt{u}\frac{2}{\sigma_k}\phi\left(\frac{\sqrt{u}(z-\mu_k)}{\sigma_k}\right)\Phi\left(\frac{\alpha_k\sqrt{u}(z-\mu_k)}{\sigma_k}\right),\ z \in \mathbb{R},
\end{aligned}
$$

where $\phi(.)$ represents the pdf of the standard Normal distribution. Which is equivalent to, $f_{Z_k|U_k=u}(z) = 2\phi\left(z;\mu_k,\frac{\sigma_k^2}{u}\right)\Phi\left(\frac{\alpha_k(z-\mu_k)}{\sigma_k/\sqrt{u}}\right)$, $z \in \mathbb{R}$, where $\phi(.;\mu_k,\frac{\sigma_k^2}{u})$ denotes the pdf of the $N(\mu_k,\frac{\sigma_k^2}{u})$. Hence, $Z_k|U_k = u \frown SN(\mu_k,\frac{\sigma_k^2}{u},\alpha_k)$.

The marginal probability density distribution of Z_k is given by

$$
f_{Z_k}(z) = \int_0^{+\infty} 2\phi\left(z;\mu_k,\frac{\sigma_k^2}{u}\right)\Phi\left(\frac{\alpha_k(z-\mu_k)}{\sigma_k/\sqrt{u}}\right)dH(u;\boldsymbol{\nu}),\ z \in \mathbb{R}. \qquad (3.7)
$$

The name of this class of distributions relies on the fact that the density function of Z_k (3.4) involves an infinite mixture of Skew-Normal distributions.

To model different patterns arising from serological data, we rely on 4 particular cases of the SMSN family. The first one is the case of the Skew-Normal distribution itself. This happens when U_k is not a random variable but rather the scalar $u = 1$. Then, variable Z_k in expression (3.4) simplifies to $Z_k = \mu_k + W_k$. Hence,

$$
F_{Z_k}(z) = P(W_k \leq z-\mu_k) = F_{W_k}(z-\mu_k), z \in \mathbb{R}, \qquad (3.8)
$$

$$
f_{Z_k}(z) = f_{W_k}(z-\mu_k) = 2\phi(z-\mu_k;0,\sigma_k^2)\Phi\left(\alpha_k\left(\frac{z-\mu_k}{\sigma_k}\right)\right). \qquad (3.9)
$$

Therefore, $Z_k \frown SN(\mu_k,\sigma_k^2,\alpha_k)$.

The second case is a simplification of the previous one when $\alpha_k = 0$. In this case, the Skew-Normal distribution reduces to the usual (symmetric) Normal distribution. In fact, when $\alpha_k = 0$ we get

$$
f_{Z_k}(z) = 2\phi(z-\mu_k;0,\sigma_k^2)\Phi(0) = \phi(z-\mu_k;0,\sigma_k^2) = \phi(z;\mu_k,\sigma_k^2), z \in \mathbb{R}, \qquad (3.10)
$$

where $\phi(.;\mu_k,\sigma_k^2)$ represents the pdf of the $N(\mu_k,\sigma_k^2)$ distribution.

The third and fourth cases are the skew Generalized Student's t-distribution and its symmetric counterpart, hereafter referred to as Skew-t and Student's t-distributions for short, respectively. These distributions can be obtained as follows.

Let U_k be a Gamma distribution with shape and rate parameters $\frac{v}{2}$ and $\frac{v}{2}$, respectively, that is, $U_k \frown Gamma(\frac{v}{2},\frac{v}{2})$. The formulation is such that the mean of U_k is equal to one.

Note that $Z_k = \mu_k + \frac{W_k}{\sqrt{U_k}}$, where $W_k \frown SN(0,\sigma_k^2,\alpha_k), U_k \frown Gamma(\frac{v}{2},\frac{v}{2})$ are independent random variables, is equivalent to $Z_k = \mu_k + \frac{W_k}{\sqrt{\frac{R_k}{v}}}$ where R_k is a χ^2 distribution with v degrees of freedom.

The conditional cumulative distribution function and the corresponding pdf of $Z_k|U_k = u$ are given by the expressions (3.5) and (3.6), respectively. According to expression (3.7), the marginal probability density distribution of Z_k takes the form

$$\begin{aligned} f_{Z_k}(z) &= \int_0^{+\infty} f_{Z_k|U_k=u}(z) f_{U_k}(u)du \\ &= \int_0^{+\infty} 2\sqrt{u}\phi(\sqrt{u}(z-\mu_k);0,\sigma_k^2)\Phi\left(\frac{\alpha_k\sqrt{u}(z-\mu_k)}{\sigma_k}\right)\frac{(\frac{v_k}{2})^{\frac{v_k}{2}} u^{\frac{v_k}{2}-1} e^{-\frac{v_k}{2}-u}}{\Gamma(\frac{v_k}{2})}du \\ &= \frac{2v_k^{\frac{v_k}{2}}}{\sigma_k\sqrt{\pi}2^{\frac{v_k+1}{2}}\Gamma(\frac{v_k}{2})}\int_0^{+\infty}\Phi(\sqrt{u}A)u^{\frac{1}{2}(v_k-1)}e^{-\frac{1}{2}u(d+v_k)}du, \end{aligned} \tag{3.11}$$

with $A = \frac{\alpha_k(z-\mu_k)}{\sigma_k}, d = \left(\frac{z-\mu_k}{\sigma_k}\right)^2$.

Integrating expression (3.11) by substitution of the variable $s = \frac{1}{2}u(d+v_k)$, we obtain

$$
\begin{aligned}
f_{Z_k}(z) &= \frac{2}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} \int_0^{+\infty} \Phi\left(A\sqrt{\frac{2s}{d+v_k}}\right) s^{\frac{1}{2}(v_k-1)} e^{-s} ds \\
&= \frac{2\,\Gamma(\frac{v_k+1}{2})}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} \int_0^{+\infty} \Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right) \frac{1}{\Gamma(\frac{v_k+1}{2})} s^{\frac{1}{2}(v_k-1)} e^{-s} ds \\
&= \frac{2\,\Gamma(\frac{v_k+1}{2})}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} \times \\
&\qquad \times \int_0^{+\infty} P\left(Z \le A\sqrt{\frac{2}{d+v_k}}\,\sqrt{s}|S=s\right) \frac{1}{\Gamma(\frac{v_k+1}{2})} s^{\frac{1}{2}(v_k-1)} e^{-s} ds. \qquad (3.12)
\end{aligned}
$$

It is important to notice the following Lemma.

Lemma (Azzalini, 2014): Suppose that $Z \frown N(0,1), Y \frown Gamma(m,1), R \frown t_{2m}$, $m > 0$. It can be proved that

$$
E\left(\Phi(c\sqrt{Y})\right) = \int_0^{+\infty} P(Z \le c\sqrt{y}|Y=y) f_Y(y) dy = P(R \le c\sqrt{m}), c \in \mathbb{R}. \qquad (3.13)
$$

The application of this Lemma to expression (3.12) leads to

$$
\begin{aligned}
f_{Z_k}(z) &= \frac{2\,\Gamma(\frac{v_k+1}{2})}{\sigma_k\sqrt{\pi v_k}\,\Gamma(\frac{v_k}{2})}\left(1+\frac{d}{v_k}\right)^{-\frac{1}{2}(v_k+1)} E\left(\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\right) \\
&= 2\,t(z;\mu_k,\sigma_k,v_k+1)\, E\left(\Phi\left(A\sqrt{\frac{2}{d+v_k}}\sqrt{s}\right)\right) \\
&= 2\,t(z;\mu_k,\sigma_k,v_k+1)\, P\left(T \le A\sqrt{\frac{v_k+1}{d+v_k}};v_k+1\right) \\
&= 2\,t(z;\mu_k,\sigma_k,v_k+1)\, T\left(A\sqrt{\frac{v_k+1}{d+v_k}};v_k+1\right),
\end{aligned}
\qquad (3.14)
$$

where $t(.;\mu_k,\sigma_k,v_k+1)$ denotes the probability density function of a Generalized Student-t distribution with location parameter μ_k, scale parameter σ_k and v_k+1 degrees of freedom; $T(.;v_k+1)$ represents the cumulative distribution function of a standard Student-t distribution with v_k+1 degrees of freedom.

In short, if $Z_k \frown ST(\mu_k, \sigma_k^2, \alpha_k, v_k)$, then its pdf is given by

$$f_{Z_k}(z) = 2\, t(z; \mu_k, \sigma_k, v_k + 1)\; T\left(A\sqrt{\frac{v_k+1}{d+v_k}}; v_k + 1\right). \tag{3.15}$$

It should be noted that when the skewness parameter is equal to zero, i.e., $\alpha_k = 0$, the quantity $A = \frac{\alpha_k(z-\mu_k)}{\sigma_k} = 0$, and the above expression takes the form

$$f_{Z_k}(z) = 2\, t(z; \mu_k, \sigma_k, v_k + 1)\; P(T \leq 0; v_k + 1) = t(z; \mu_k, \sigma_k, v_k + 1). \tag{3.16}$$

which corresponds to the probability density function of a Generalized Student-t distribution with location parameter μ_k, scale paramter σ_k and $v_k + 1$ degrees of freedom.

As the degrees of freedom tends to infinity, the Skew-t distribution converges to the Skew-Normal distribution (Azzalini, 1985; Azzalini and Capitanio, 2003; Basso et al., 2010).

The mean and variance of the Skew-t distribution are respectively given by,

$$E(Z_k) = \mu_k + \sigma_k b_{v_k}\delta_k, \text{ if } v_k > 1, \quad V(Z_k) = \sigma_k^2\left[\frac{v_k}{v_k - 2} - \left(b_{v_k}\delta_k\right)^2\right] \text{ if } v_k > 2, \tag{3.17}$$

where $b_{v_k} = \frac{\sqrt{v_k}\,\Gamma(\frac{1}{2}(v_k-1))}{\sqrt{\pi}\,\Gamma(\frac{1}{2}v_k)}$ and $\delta_k = \frac{\alpha_k}{\sqrt{1+\alpha_k^2}}$.

3.3.2 Estimation and model selection

Suppose that we have a random sample $X_1, ..., X_n$ representing the antibody levels of n individuals. In general, it is very difficult to determine the maximum likelihood (ML) estimates of the parameters of any given finite mixture model by direct maximization of the corresponding log-likelihood functions. One way to surpass this problem is to consider the Expectation-Maximization (EM) algorithm given that the latent serological status of each individual is unknown and, thus, we are in the presence of a problem of incomplete data.

A full derivation of an EM-type algorithm for fitting mixtures of SMSN can be found in Basso et al., 2010. In brief, the E-step is the same as in the traditional mixtures of

Normal distributions, which has been widely studied in the literature (Lin et al., 2007a; Basso et al., 2010; Lachos Dávila et al., 2018). Replacing the classical M-step with a sequence of conditional maximization steps (CM-steps), one obtains closed form expressions that simplify the implementation of the algorithm. Also, the observed information matrix can be derived analytically (Ferreira et al., 2011).

There are several methods to determine the optimal number of components that constitute the mixture, g (G. McLachlan and Peel, 2000; Oliveira-Brochado et al., 2005; Lukočienė et al., 2009). A simple way to do it is to use penalized forms of the log-likelihood function: the information criteria. They rely on the idea that an increase in the number of components in the mixture leads to a better fit of the data, thus, increasing the maximized likelihood function. Invoking the parsimony principle to determine the best model for the data, enhancement in model fitting is penalized by an increase in the number of parameters included in the model. In this framework, two of the most popular measures are Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Xie et al., 2013). In general the best model is the one that provides the lowest estimate of AIC or BIC value among all models tested. About the use of one or the other criterion, the BIC criterion demonstrates consistency in determining the number of components of the mixture models (Xie et al., 2013; Mehrjou et al., 2016). In addition, BIC tends to select simpler models (ideally with less number of components) than AIC, which simplifies data interpretation. Hence, this information criterion is preferred for model selection.

Another way to assess the number of components in a mixture model is to carry out a hypothesis testing, namely the Likelihood Ratio Test (LRT). However, the regularity conditions for the validity of classical asymptotic approximation of the test statistic are not met in the context of finite mixture models, because the null hypothesis associated with this hypothesis is specified in the boundary of the parameter space rather than its interior (G. McLachlan and Peel, 2000). In some cases, the true parameter is in a non-identifiable subset of the parameter space (Feng et al., 1996). As a consequence, there is no guarantee that, under the null hypothesis, the likelihood ratio statistic asymptotically follows a χ^2 distribution with the degrees of freedom given by the difference between the number of parameters under the alternative and the null hypothesis (G. McLachlan and Peel, 2000). To surpass this problem, a bootstrap approach can be carried out to estimate the p-value of this non-standard LRT (Feng et al., 1996; Yu et al., 2019).

Let us consider the test specified by $H_0 : g = g_0$ versus $H_1 : g = g_1$ where $g_0 < g_1$. Let us also denote $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ the vector of parameters of the mixture models under H_0 and H_1 hypotheses, respectively, $\boldsymbol{x} = (x_1,...,x_n)$ the observed data and $T(\boldsymbol{x};\boldsymbol{\psi}_0,\boldsymbol{\psi}_1)$ the test statistic of LRT. The bootstrap approach is given by the following algorithm (Yu et al., 2019):

1. Use the EM algorithm to estimate the $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ estimates under the H_0 and H_1 hypotheses, respectively. Calculate $T(\boldsymbol{x};\hat{\boldsymbol{\psi}}_0,\hat{\boldsymbol{\psi}}_1)$;

2. Simulate $N = 10,000$ independent samples $\boldsymbol{x}^*_1,...,\boldsymbol{x}^*_n$ using the mixture model under H_0 and parameterized by $\hat{\boldsymbol{\psi}}_0$;

3. For each bootstrap sample i, calculate $T(\boldsymbol{x}^*_i;\hat{\boldsymbol{\psi}}_{0_i},\hat{\boldsymbol{\psi}}_{1_i})$, where $\hat{\boldsymbol{\psi}}_{0_i}$ and $\hat{\boldsymbol{\psi}}_{1_i}$ are the estimated parameter vectors for the bootstrap sample i under the H_0 and H_1 hypotheses, respectively;

4. Estimate the p-value as $\frac{1}{N}\sum_{i=1}^{N} I\{T(\boldsymbol{x}^*_i;\hat{\boldsymbol{\psi}}_{0_i},\hat{\boldsymbol{\psi}}_{1_i}) > T(\boldsymbol{x};\hat{\boldsymbol{\psi}}_0,\hat{\boldsymbol{\psi}}_1)\}$, where $I\{.\}$ is the indicator function.

Another important statistical test in the context of mixtures based on SMSN is to address the significance of the asymmetry parameters of the mixing distributions composing the mixture model based on Skew-Normal or Skew-t distributions . To attain this goal, a LRT can also be carried out. Suppose that we have a mixture model with all g components given by either Skew-Normal or Skew-t distributions. In this test, the hypotheses under testing are the following:

$$H_0 : \alpha_k = 0,\ \forall_{k=1,...,g} \text{ versus } H_1 : \exists_{k=1,...,g},\ \alpha_k \neq 0.$$

The test statistic is given by $\Lambda = -2\ln\frac{\widehat{L}_0}{\widehat{L}_1}$, where $\widehat{L}_0$ and $\widehat{L}_1$ correspond to the likelihood functions of the mixture model evaluated at the maximum likelihood estimates under H_0 and H_1, respectively. In contrast with the previous test related to the number of components, the usual asymptotic approximation for the distribution of the LRT statistic under H_0 holds, that is, a χ^2 distribution with g degrees of freedom.

Finally, the quality of the fit of the estimated models should be assessed. For a matter of simplicity, the Pearson's χ^2 test for goodness-of-fit can be used. To apply this test, the data under analysis can be divided into bins according to the sampled 5%-quantiles or deciles (*i.e.*, 10%-quantiles).

3.3.3 Estimation of seropositivity

After determining the best finite mixture model for the data, the next step is usually to estimate the seroprevalence, that is, the prevalence of antibody-positive individuals in the population (or, the probability of an individual being antibody-positive). Seropositivity is usually defined by a cutoff, denoted by c, in the respective antibody distribution above which individuals would be considered seropositive. In the context of finite mixture models, cutoff determination requires the interpretation of each latent population in terms of seronegativity and seropositivity. To do that, one typically assumes the seronegative population as the one with lowest average value while the remaining components are interpreted as different levels of seropositivity upon recurrent infections. In this scenario, the seropositivity of i-th individual can be seen as resulting from a Bernoulli random variable $Y_i \frown Ber(p)$ where $p = P[X_i \geq c]$ and X_i $(i = 1, ..., n)$ represents the random variable representing the underlying antibody concentration. The probability p is also called seroprevalence and it embodies the probability of exposed individuals to a given antigen in the population. According to the maximum likelihood method, seroprevalence can be estimated as the proportion of seropositive individuals in the sample. Therefore, different estimates for the seroprevalence can be obtained according to the methods used to determine the cutoff.

In this work, we consider the following three different methods for determining the seropositivity cutoff:

- **Method 1:** It is based on the 99.9%-quantile associated with the estimated seronegative population. This method is the most popular in sero-epidemiology (Sepúlveda, Stresman, et al., 2015; Saraswati et al., 2019). It is often called as the 3σ rule, because the 99.9%-quantile is given by the mean plus 3 times the standard deviation of a normally distributed seronegative population;

- **Method 2:** It relies on the minimum of the density mixture functions. In the case of two latent populations, the cutoff corresponds to the absolute minimum, and in the case of three or more latent populations the cutoff corresponds to the lowest relative minimum. This point can be calculated using the Dekker's algorithm (Brent, 1973). It should be noted that the minimum of the mixing function is not expected to coincide with the point of intersection of the probability densities of each individual subpopulation;

- **Method 3:** It imposes a threshold in the the so-called conditional classification curves (Sepúlveda, Stresman, et al., 2015). Under the assumption that all components but the first one refer to seropositive individuals, the conditional classification curve of seropositive individuals given the antibody level x is defined as

$$p_{+|x} = \frac{\sum_{k=2}^{g} \pi_k f_k(x;\boldsymbol{\theta}_k)}{\sum_{k=1}^{g} \pi_k f_k(x;\boldsymbol{\theta}_k)}. \tag{3.18}$$

In turn, the classification curve of seronegative individuals is given by

$$p_{-|x} = 1 - p_{+|x}. \tag{3.19}$$

After calculating these curves, one can impose a minimum value for the classification of each individual. In this case, two cutoff values arise in the antibody distribution, one for the seronegative individuals and another for seropositive individuals. Mathematically, the classification rule is given as follows

$$C_i = \begin{cases} \text{seronegative} & \text{, if } \; x_i \leq c_- \\ \text{equivocal} & \text{, if } \; c_- < x_i < c_+ \\ \text{seropositive} & \text{, if } \; x_i \geq c_+ \end{cases} \tag{3.20}$$

where c_- and c_+ are the cutoff values in the antibody distribution that ensure a minimum classification probability, say 90%. To calculate these cutoff values in practice, one can use the bisection method providing an initial interval where they might be located (Sepúlveda, Stresman, et al., 2015).

3.3.4 R packages

We used the package `mixsmsn` to fit different mixture models based on SMSN (Prates et al., 2013). In particular, we used the function `smsn.mix` to estimate the model parameter via the

EM algorithm and the function `rmix` to generate random samples from a given mixture model in the bootstrap method. For fitting the Student's t-distribution, we considered the R package `extraDistr` (Wolodzko, 2020), namely, the function `dlst` to calculate their density and the function `plst` to define the cumulative distribution function. The fitting of the Skew-Normal distributions was performed with the package `sn` (Azzalini, 2020). The functions `dsn` and `psn` were used to calculate the probability density function and the cumulative distribution function of the Skew-Normal distribution, respectively. In the case of the Skew-t distribution, the functions `dst` and `pst` were used to calculate the probability density function and the cumulative distribution function, respectively.

3.4 Results

3.4.1 Analysis of serological data by finite mixture models based on SMSN

The statistical analysis was performed after applying the base 10 logarithmic transformation to the data. The number of components g in the mixture models was allowed to vary from 1 (single distribution) to 3 components. When fitting the mixtures of Skew-t distributions, the package `mixsmsn` only allowed to fit models with a common degree of freedom for all mixing distributions (*i.e.*, $v_1 = ... = v_g = v$).

Our results suggested that the 6 antibodies under evaluation could be divided into two major classes: (i) the first one including antibodies against HHV-6 and VZV where there was evidence for a single serological population (Table 3.1) and (ii) another one including the antibodies against the remaining four herpesviruses where there was evidence for the existence of more than one serological population in the respective data (Table 3.2).

According to BIC, the best models for the antibodies against HHV-6 and VZV were Skew-Normal and Skew-t distributions, respectively. The estimated distributions were both left skew (Figure 3.1A; $\alpha_{HHV6} = -1.82$ with 95% CI =(-2.44;-1.02) and $\alpha_{VZV} = -4.54$ with 95% CI =(-6.94;-2.14). They would appear to have a good fit to the data at the 5% significance level ($p_{gof} = 0.140$ and 0.076, respectively). In the case of antibodies against VZV, further

Table 3.1: Antibody data with evidence for a single serological population, where g represents the number of components in the mixture models, p is the number of parameters of the model, $\mathscr{L}_{\max}$ is the value of the maximized log-likelihood function, p_{gof} is the maximum p-value for the goodness-of-fit test when dividing data into deciles or 5%-quantiles, and p_{boot} is the bootstrap p-value for testing $H_0 : g = 1$ versus $H_1 : g = 2$.

Virus	**SMSN**	g	p	$\mathscr{L}_{\max}$	**BIC**	p_{gof}	p_{boot}
HHV-6	Normal	1	2	-129.46	270.94	0.064	0.064
		2	5	-116.97	263.97	0.169	
		3	8	-110.43	268.91	0.462	
	Skew-Normal	**1**	**3**	**-121.35**	**260.71**	**0.140**	**0.027**
		2	7	-117.35	276.75	0.084	
		3	11	-109.40	284.87	0.152	
	Student's t	1	3	-124.38	266.77	0.157	0.042
		2	6	-117.14	270.32	0.122	
		3	9	-105.36	264.78	0.254	
	Skew-t	1	4	-118.81	261.65	0.148	0.409
		2	8	-116.83	281.71	0.076	
		3	12	-104.00	586.83	0.001	
VZV	Normal	1	2	-108.76	229.53	< 0.001	< 0.001
		2	5	-7.28	44.60	0.159	
		3	8	-1.70	51.45	0.153	
	Skew-Normal	1	3	-23.94	65.90	< 0.001	0.180
		2	7	-0.11	42.27	0.406	
		3	11	0.10	65.87	0.068	
	Student's t	1	3	-61.90	141.80	< 0.001	< 0.001
		2	6	-7.41	50.86	0.082	
		3	9	-1.68	57.42	0.113	
	Skew-t	**1**	**4**	**-7.89**	**39.81**	**0.076**	**0.375**
		2	8	-0.05	48.16	0.211	
		3	12	5.47	62.14	0.134	

evidence was obtained for a single population when testing one Skew-t distribution against a mixture of two Skew-t distributions, respectively ($p_{boot} = 0.375$). However, when testing one Skew-Normal distribution against a mixture of two Skew-Normal distributions for the antibodies against HHV-6, the respective result was in the borderline of the 5% statistical significance ($p_{boot} = 0.027$).

In terms of serological classification, the evidence for a single population would appear to represent a putative seropositive population. This interpretation is consistent with the prior knowledge that HHV-6 and VZV are usually acquired during childhood and more than 95% of the adult populations typically shows evidence of antibody positivity against these viruses (Braun et al., 1997). In addition, the core values of these distributions are higher than the cutoff for seropositivity suggested by the lab protocol. Finally, a left skewness is also predicted for a hypothetical seropositive population, because the antibodies should decay over time (Parker et al., 1990).

It is worth noting that most of the mixture models under comparison could also fit data of these two antibodies well. This the case of the mixture of two or three Normal distributions ($p_{gof} = 0.169$ and 0.462 for antibodies against HHV-6 and $p_{gof} = 0.159$ and 0.153 for VZV), which are typically used in serological data analysis. Therefore, although not being the best models for HHV-6 and VZV-related antibodies, these models could have been used for subsequent serological analyses.

With respect to the antibodies against the remaining herpesviruses, the respective data analysis was not so straightforward, because the model with lowest BIC estimate could not fit the data well according to the Pearson's goodness-of-fit test at 5% significance level (Table 3.2). This is the case of the mixtures of two Skew-Normal distributions for the antibodies against CMV (BIC=509.69 and $p_{gof} = 0.038$), HSV-1 (BIC=563.52 and $p_{gof} = 0.003$), and HSV-2 (BIC=570.68 and $p_{gof} = 0.013$). For these antibodies, the best models were considered to be a mixture of two Skew-t distributions (BIC=511.45 and $p_{gof} = 0.072$), a mixture of three Skew-Normal distributions (BIC=570.70 and $p_{gof} = 0.104$), and a mixture of two Normal distributions (BIC=585.27 and $p_{gof} = 0.516$), respectively, because they were the best models ranked by BIC with a good fit for the data (Figure 3.1B). Interestingly, for the HSV-2-related antibody data, when the mixture of two Normal distributions was compared to the mixture of 2 Skew-Normal distribution by a likelihood ratio test, the first model was strongly rejected ($p < 0.0001$), which suggested the asymmetry of at least one of the components. This inconsistency between this test and the selected model can be explained by the unavailability of fitting a mixture of a Normal distribution and a Skew-Normal distribution in the package `smsn`. For the EBV-related antibody data, the best model according to BIC was a mixture of two Skew-t distributions, which also had a good fit for the data (BIC=299.32 and $p_{gof} = 0.248$; Figure 3.1B).

In terms of interpretation of each component, there was evidence of putative seronegative and seropositive populations for antibodies against CMV, EBV, and HSV-2 (Figure 3.1B). This interpretation was supported by the observation that the cutoff value suggested by the commercial kits lies between these hypothetical serological populations. In the case of antibodies against HSV-1, the respective interpretation was not so obvious, because (i) the best mixture model was composed of three components and (ii) the cutoff suggested by the commercial kits lies in the middle of the intermediate distribution, which shows right asymmetry. In theory, the distribution of a putative seronegative population tends to show right asymmetry (Parker et al., 1990) and, if so, this intermediate component should be interpreted accordingly. However, this interpretation opens the door for the presence of two sets of seronegative populations resulting from distinct background signals in absence of antibodies. In absence of additional information

Table 3.2: Antibody data with evidence for more than one serological population, where g represents the number of components in the mixture models, p is the number of parameters of the model, $\mathscr{L}_{\max}$ is the value of the maximized log-likelihood function, and p_{gof} is the maximum p-value for the goodness-of-fit test when dividing data in deciles or 5%-quantiles.

Virus	**SMSN**	g	p	$\mathscr{L}_{\max}$	**BIC**	p_{gof}
CMV	Normal	1	2	-409.11	830.24	< 0.001
		2	5	-245.75	521.54	0.016
		3	8	-233.70	515.45	0.018
	Skew-Normal	1	3	-357.61	733.23	< 0.001
		2	7	-233.82	509.69	0.038
		3	11	-226.64	519.35	0.146
	Student's t	1	3	-410.14	838.29	< 0.001
		2	6	-238.54	513.12	0.038
		3	9	-231.23	516.59	0.046
	Skew-t	1	4	-357.71	739.45	< 0.001
		2	**8**	**-231.55**	**511.45**	**0.072**
		3	12	-226.93	525.93	0.324
EBV	Normal	1	2	-342.30	696.62	< 0.001
		2	5	-152.66	335.36	< 0.001
		3	8	-129.30	306.65	0.173
	Skew-Normal	1	3	-226.42	470.86	< 0.001
		2	7	-130.57	303.17	0.084
		3	11	-128.02	322.10	0.054
	Student's t	1	3	-240.21	498.43	< 0.001
		2	6	-151.61	339.26	< 0.001
		3	9	-129.41	312.88	0.117
	Skew-t	1	4	-173.14	370.31	< 0.001
		2	**8**	**-125.63**	**299.32**	**0.248**
		3	12	-126.29	324.66	0.087
HSV-1	Normal	1	2	-442.27	896.56	< 0.001
		2	5	-291.59	613.22	< 0.001
		3	8	-264.94	577.94	0.003
	Skew-Normal	1	3	-394.55	807.11	< 0.001
		2	7	-260.74	563.52	0.003
		3	**11**	**-252.32**	**570.70**	**0.104**
	Student's t	1	3	-443.73	905.48	< 0.001
		2	7	-291.73	619.51	< 0.001
		3	9	-264.98	584.02	0.002
	Skew-t	1	4	-395.43	814.88	< 0.001
		2	8	-260.88	569.82	0.001
		3	12	-251.86	575.79	< 0.001
HSV-2	Normal	1	2	-427.29	866.59	< 0.001
		2	**5**	**-277.62**	**585.27**	**0.516**
		3	8	-269.24	586.54	0.007
	Skew-Normal	1	3	-337.36	692.74	< 0.001
		2	7	-264.32	570.68	0.013
		3	11	-257.19	580.45	0.003
	Student's t	1	3	-428.40	874.81	< 0.001
		2	6	-277.84	591.71	0.688
		3	9	-269.60	593.26	0.004
	Skew-t	1	4	-337.79	699.60	< 0.001
		2	8	-264.52	577.10	0.007
		3	12	-257.38	586.83	0.001

about the serological data, this intermediate component was considered to represent a putative seronegative population.

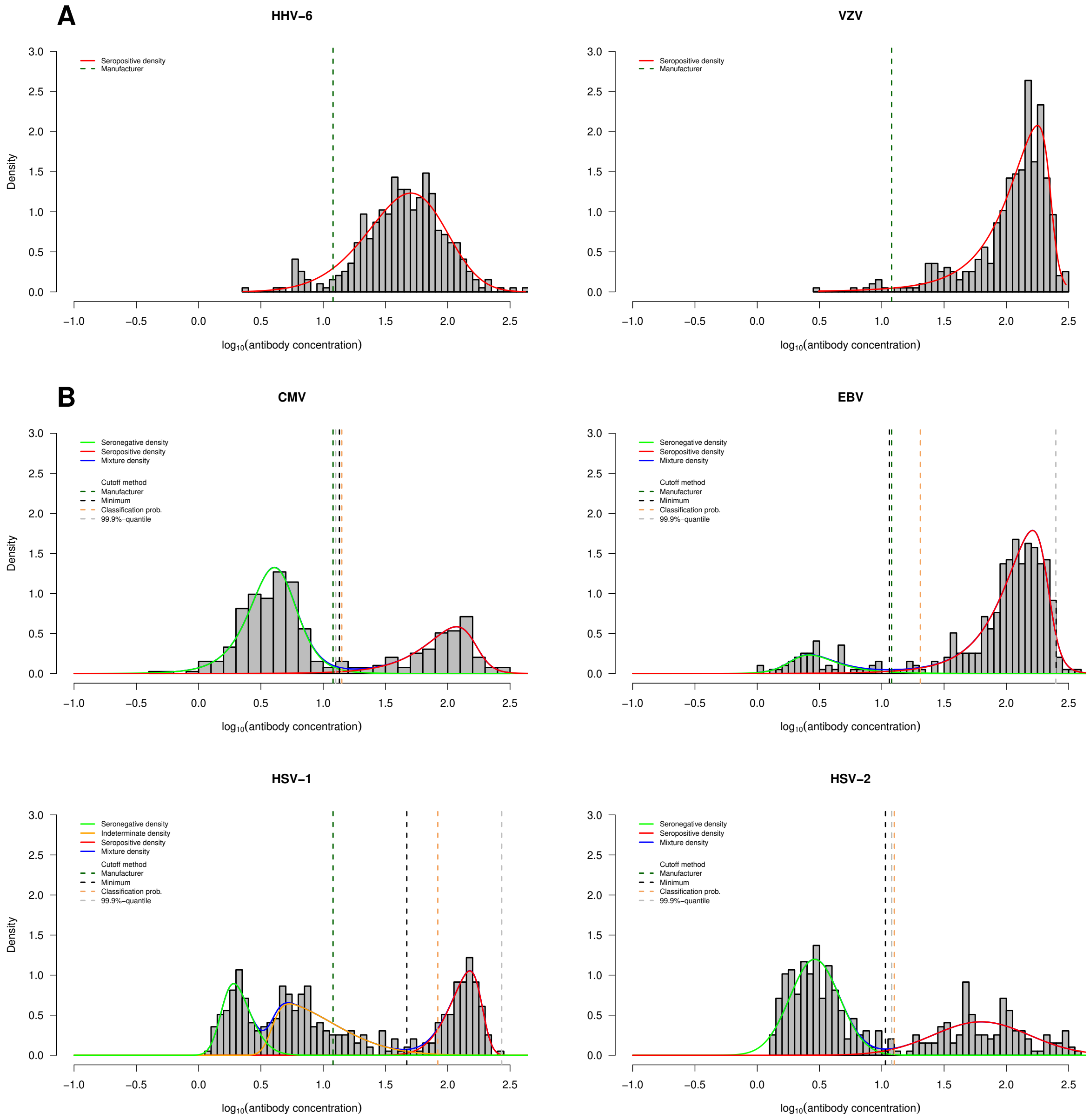


Figure 3.1: Best models for the data under analysis. **A.** Antibody distributions with evidence for a single serological population (HHV-6 and VZV). **B.** Antibody distributions with evidence for more than one serological population (CMV, EBV, HSV1, and HSV2). Antibody concentration in x axis is given in $\log_{10}$ units.

3.4.2 Estimation of cutoff for seropositivity

After fitting the mixture models to the data, the following step of the analysis was to estimate a cutoff value for seropositivity and the subsequent seroprevalence in the different study groups (Table 3.3).

For CMV and HSV-2 antibody data, the cutoff values did not vary substantially from one method to another. Interesting, the cutoff values estimated by method 1 (the 3σ rule) almost perfectly matched with the ones suggested by the commercial kits (12.6 U/ml and 12.0 U/ml for CMV and HSV-2 respectively versus 12.0). This good matching between estimates could be explained by a good approximation of the Normal distribution for the seronegative population (Figure 3.1B) and, therefore, we could infer that the cutoff value suggested by the commercial kits was derived from the 3σ rule; this information was absent from the original study (Cliff et al., 2019). Since the seronegative and seropositive populations were separated well in these antibody distributions, the estimates of seroprevalence across the different study groups were almost invariant with respect to cutoff value used.

With respect to the EBV antibody data, the hypothetical seronegative population is asymmetric to the right ($\alpha_1 = 1.74$; 95% CI=(-1.30; 4.80); Figure 3.1B) with heavy tails ($\nu = 4.52$; 95% CI=(0.79;8.26)). As a consequence, the cutoff value of 249.5 U/ml derived from method 1 was quite different from the one suggested by the commercial kit. However, this cutoff value was considered non-informative, because it was well located within the seropositive population and implied seroprevalence estimates close to zero for the different study groups. In contrast, the cutoff values from the remaining methods were in the same order of magnitude of the one suggested by the commercial kits. Therefore, the subsequent seroprevalence estimates of each study group did not differ substantially among these methods. Again, the consistency of the resulting seroprevalence estimates was due to the fact that the seronegative and seropositive populations were well separated in these data.

The largest differences in the cutoff values for seropositivity were observed for the HSV-1 antibody data. Coincidentally, this was the data set where the best mixture model was composed of three components. As discussed earlier in this paper, the intermediate component was considered a second hypothetical seronegative population, which resulted in a shift in the calculation of seropositivity towards higher values. As such, the cutoff seropositive based on the commercial kit led to the highest seroprevalence estimates for all study groups with a global estimate of 45.2% (95% CI=(40.2%;50.2%). As an extreme case, the 3σ rule produced again a too-high cutoff value due to the right asymmetry of both seronegative populations. Such unrealistic cutoff value led a zero seroprevalence estimates and rendered the respective analysis useless.

Finally, although not being the main objective of this study, the comparison of the four study groups suggested that, given a method for determining seropositivity and antibody under analysis, the seroprevalence of patients with ME/CFS did not appear to differ significantly from the one of healthy controls and patients with multiple sclerosis alike.

Table 3.3: Seroprevalence (%) by cutoff method for seropositivity and by study group. c_- and c_+ are on the linear scale (U/ml). Seroprevalence was calculated based on c_+. The method denoted by ”M” refers to the cutoff suggested by the protocol of the commercial kit. The confidence intervals (CI) refer to the Clopper-Pearson exact confidence interval for a proportion.

				Seroprevalence (95% CI)				
Virus	Method	c_-	c_+	**Global**	**HC**	**ME-M**	**ME-S**	**MS**
CMV	M	8.0	12.0	33.5 (28.9-38.4)	37.4 (28.2-47.3)	28.6 (22.4-35.4)	33.3 (21.1-47.5)	36.7 (23.4-51.7)
	1	-	12.6	33.5 (28.9-38.4)	37.4 (28.2-47.3)	28.6 (22.4-35.4)	33.3 (21.1-47.5)	36.7 (23.4-51.7)
	2	-	13.5	33.2 (28.6-38.1)	37.4 (28.2-47.3)	28.6 (22.4-35.4)	31.5 (19.5-45.6)	36.7 (23.4-51.7)
	3	9.4	14.1	32.9 (28.4-37.9)	37.4 (28.2-47.3)	28.1 (21.9-34.9)	31.5 (19.5-45.6)	36.7 (23.4-51.7)
EBV	M	8.0	12.0	87.3 (83.6-90.4)	87.9 (80.1-93.4)	86.2 (80.6-90.7)	81.5 (68.6-90.7)	75.5 (61.1-86.7)
	1	-	249.5	2.0 (0.09-3.9)	1.9 (0.02-6.6)	1.5 (0.03-4.4)	0.0 (0.0-6.6)	6.1 (1.3-16.9)
	2	-	11.5	87.3 (83.6-90.4)	87.9 (80.1-93.4)	86.2 (80.6-90.7)	81.5 (68.6-90.7)	75.5 (61.1-86.7)
	3	5.6	20.4	85.5 (81.7-88.9)	87.9 (80.1-93.4)	82.7 (76.6-87.7)	81.5 (68.6-90.7)	75.5 (61.1-75.5)
HSV-1	M	8.0	12.0	45.2 (40.2-50.2)	42.1 (32.6-51.9)	41.8 (34.8-49.1)	51.9 (37.8-65.6)	46.9 (32.5-61.7)
	1	-	271.0	0.0 (0.0-0.1)	0.0 (0.0-3.4)	0.0 (0.0-1.2)	0.0 (0.0-6.6)	0.0 (0.0-7.3)
	2	-	46.9	34.5 (29.8-39.4)	28.0 (19.8-37.5)	34.7 (28.1-41.8)	38.9 (25.9-53.1)	34.7 (21.7-49.6)
	3	42.7	83.2	30.7 (26.2-35.5)	24.3 (16.5-33.5)	32.1 (25.7-39.2)	33.3 (21.1-47.5)	28.6 (16.6-43.3)
HSV-2	M	8.0	12.0	38.1 (33.3-43.1)	33.6 (24.8-43.4)	38.8 (31.9-45.9)	40.7 (27.6-54.9)	32.7 (19.9-47.5)
	1	-	12.0	38.1 (33.3-43.1)	33.6 (24.8-43.4)	38.8 (31.9-45.9)	40.7 (27.6-54.9)	32.7 (19.9-47.5)
	2	-	10.7	38.8 (33.9-43.8)	33.6 (24.8-43.4)	39.3 (32.4-46.5)	40.7 (27.6-54.9)	36.7 (23.4-51.7)
	3	7.1	12.6	37.8 (33.0-42.8)	33.6 (24.8-43.4)	38.8 (31.9-45.9)	40.7 (27.6-54.9)	30.6 (18.3-45.4)

3.5 Conclusions

This study aimed to review the finite mixture models based on SMSN and to recommend their routine use in serological data analysis. Such recommendation sets its foundation in the high flexibility of these models in describing different patterns of randomness, as illustrated with the analysis of antibodies against 6 different herpesviruses. In particular, a high modelling flexibility is necessary given that right and left asymmetry could emerge from hypothetical seronegative and seropositive populations, respectively. In this regard, most popular distributions used in statistics are not able to exhibit either left or right asymmetry depending on the parameters specified. A less-known family of distributions that shows such remarkable stochastic property is the so-called the Generalized Tukey's λ distribution (Ramberg et al., 1974; Freimer et al., 1988). This distribution offers a great variety of shapes owing to four parameters controlling the location, the scale, the skewness, and the flatness of the resulting distribution. However, the Generalized Tukey's λ distribution is only defined in terms of its quantile function and, hence, its estimation is cumbersome. This distribution has already been proposed for mixture modelling, but there are only theoretical and computational developments available for the two-component case (Su, 2007; Su, 2011). This limits the applicability of these alternative models, namely, in data where there is evidence for more than two serological populations, such as the case of the antibodies against HSV-1 here analyzed or against the influenza virus reported elsewhere (Nhat et al., 2017). Therefore, finite mixture models based on SMSN would appear the most general and flexible approach so far for analysing serological data.

For data analysis, we recommend the use of the package `mixsmsn` for estimating the finite mixture models (Prates et al., 2013). Notwithstanding this recommendation, the package only allows to estimate finite mixture models where all mixing distributions belong to the same class of SMSN probability distributions. Hence, it is only possible to fit 4 different models per number of components. In theory, there are $4^2 = 16$ possible two-component mixture models resulting from the combination of Normal, Skew-Normal, Generalized Student's t, and Skew-t distributions as mixing distributions. Note that these possible models result from imposing parametric restrictions to the most general mixture model based on the Skew-t distribution. For three-component mixture models, the number of possible models increases to $4^3 = 64$. Therefore, the package `mixsmsn` excludes a vast number of possible models, which ultimately affects the detection of the true best model for the data. This computational limitation might be the reason for some inconsistencies that can be found in the example of application. For instance, a single Skew-Normal distributions was considered the best model for the antibodies against HHV-6. However, the hypothesis of a single Skew-Normal distribution against a mixture of two Skew-Normal distributions could be rejected by bootstrap at the 5% significance level. A possible explanation for this contradicting evidence is that the best model for these data could be a mixture of a Normal distribution for the seronegative population and a Skew-Normal distribution for the seropositive population.

Another limitation of using `mixsmsn` package is that, for mathematical tractability, the mixtures of generalized Student t and Skew-t distributions were assumed to have the same degrees of freedom in all the mixing distributions. In theory, this assumption could be relaxed so this parameter could vary from one component of the mixture to another. This modelling option was available in the package `EMMIXuskew` for the mixture of Skew-t distributions (G. McLachlan and S. Lee, 2013). However, this package is currently discontinued. In practice, we expect some degree of numerical instability when trying to estimate different degrees of freedom for mixtures in which different components overlap with each other substantially. In this regard, future research could be conducted in order to determine under which conditions different degrees of freedom could infer for the different components.

The problem of determining the optimal cutoff value for seropositivity has been intensively investigated, discussed, and revisited over the years (Ridge et al., 1993; Kafatos et al., 2016; Migchelsen et al., 2017; Saraswati et al., 2019). In this regard, the most popular cutoffs for seropositivity are simply defined by the mean plus a given number of times the standard deviation of the hypothetical seronegative population without checking the Normality assumption of the hypothetical seronegative population. The resulting cutoffs are associated with high-order quantiles of the Normal distribution, such as 97.7% or 99.9% for the 2σ and 3σ rules, respectively. In practice, these cutoffs imply a high specificity but show an arbitrary sensitivity for the respective serological classification. When the hypothetical seronegative population shows a right skew distribution, similar cutoffs can be obtained by calculating same high quantiles of the estimated SMSN, as done here. The reverse argument can be made when analysing antibodies where seropositivity is expected to be the default serological state of an individual, such as the case of antibodies against HHV-6 and VZV here analyzed or vaccine-related antibodies in populations where vaccination is mandatory. For these antibodies, similar cutoffs can be determined by the mean minus a given number of times the standard deviation of the hypothetical seropositive population assumed to be normally distributed. For a hypothetical seropositive population with a left skew distribution, the cutoff values for seropositive are now calculated using the low order quantiles (e.g., 2.3% and 0.1%-quantiles for the 2σ and 3σ rules, respectively). Inversely, these cutoffs generate a high sensitivity but an arbitrary specificity for the respective serological classification. It is worth noting that, as expected, several authors advocate a free-cutoff approach for serological analysis

(Chis Ster, 2012; Bouman et al., 2020). However, a detailed discussion about the advantages and disadvantages of free-cutoff approaches was considered to be out of the scope of this study.

In terms of the results concerning the example of application, there is no evidence for a different level of exposure of the patients with ME/CFS to these herpesviruses when compared to healthy controls and patients with multiple sclerosis. This finding seems independent of the method for determining the seropositivity and it is in line with the findings reported in the original study (Cliff et al., 2019) and with another serological investigation of these herpesviruses when comparing patients with healthy controls only (Blomberg, Rizwan, et al., 2019). A possible explanation for this ”negative” finding might be explained by the choice of antibodies against highly immunogenic antigens used in this serological study. It is then possible that there is a specific set of viral-derived antigens associated with ME/CFS, as suggested by a comprehensive study about the role of antibodies against EBV in this disease (Loebel et al., 2017). Finally, a more detailed analysis of these data is currently carried out in order to understand whether the lack of association between ME/CFS and these antibodies could be explained by putative confounding effects of age and gender on the underlying antibody distributions. This detailed analysis will be reported elsewhere.

In summary, the finite mixture models based on SMSN show a good potential to become a routine tool for serological data analysis. They have the advantage of including the popular Gaussian mixture models as special cases. However, given the statistical complexity of these models, we recommend a closer collaboration between biomedical researchers who generate the serological data and biostatisticians who have in principle the knowledge and skills to fit and compared them properly.

Chapter 4

Association analysis between herpesviruses serology and ME/CFS using a varying cutoff approach for seropositivity

The evidence of an association between Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and chronic herpesviruses infections remains inconclusive. Two possible reasons for this lack of consistent evidence are the large heterogeneity of the patients' population with different disease triggers and the use of arbitrary cutoffs for defining seropositivity. In this work we re-analyzed previously published serological data related to 7 herpesvirus antigens. These data were collected as part of the United Kingdom ME/CFS Biobank (UKMEB). In our re-analysis, patients with ME/CFS were subdivided into four major subgroups related to the disease triggers: S_0 - 42 patients who did not know their disease trigger; S_1 - 43 patients who reported a non-infection trigger; S_2 - 93 patients who reported an infection trigger, but that infection was not confirmed by a lab test; and S_3 - 48 patients who reported an infection trigger and that infection was confirmed by a lab test. In accordance with a sensitivity analysis, the data were compared to those from 99 healthy controls allowing the seropositivity cutoffs to vary within a wide range of possible values. We found a negative association between S_1 and seropositivity to Epstein-Barr virus (VCA and EBNA1 antigens) and Varicella-Zoster virus. However, the significance of this finding was affected by the seropositivity cutoff used. We also found that S_3 had a lower seroprevalence to the human cytomegalovirus when compared to healthy controls for all cutoffs used for seropositivity. In summary, herpesviruses serology could distinguish subgroups of ME/CFS patients according to their disease trigger.

4.1 Introduction

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a complex disease whose affected patients experience persistent fatigue that cannot be alleviated by rest and suffer from post-exertional malaise upon minimal physical and/or mental activity (Rivera et al., 2019). Disease prevalence has been estimated around 0.4% after pooling data from different epidemiological studies (Lim et al., 2020). However, this estimate might be conservative (Valdez et al., 2019, Hanson et al., 2020) due to poor societal recognition of the disease including amongst health professionals (Raine et al., 2004), the inexistence of an objective disease-specific biomarker for the corresponding diagnosis (Scheibenbogen et al., 2017), a small number of well-designed epidemiological studies (Estévez-López et al., 2020), and limited funding opportunities for more comprehensive and integrative research (Pheby et al., 2020).

The pathogenesis of ME/CFS remains a topic under intense debate with the proposal of many competing hypotheses (Underhill, 2015; Sotzny et al., 2018; Blomberg, Gottfries, et al., 2018; Hatziagelaki et al., 2018; Sepúlveda, Carneiro, et al., 2019; Morris et al., 2019; Wirth et al., 2020; Stanculescu et al., 2021). However, there is a general consensus that the disease could be initiated by a combination of genetic predisposing factors and environmental triggers (e.g., exposure to toxins, chronic emotional and physical stress) (Blomberg, Gottfries, et al., 2018; L. Nacul, O'Boyle, et al., 2020). In this regard, a large proportion of patients report an acute infection at the onset of their symptoms (Johnston et al., 2016; Chu et al., 2019). With the objective of finding the master infectious agent of the disease cause and progression, many serological investigations were conducted with inconclusive or even contradicting findings (Rasa et al., 2018). Possible reasons for this contrasting evidence could be related to disease misclassification and selection bias (L. Nacul, E. Lacerda, et al., 2019; Malato et al., 2021), the necessity of dividing patients into different subtypes (Jason et al., 2005), the low number of patients recruited (Hatziagelaki et al., 2018), or differences in the antigen and experimental assays used (Ariza, 2020). An additional but often ignored reason is that serological studies are typically based on arbitrary cutoff values for identifying seropositive individuals or high antibody responders, as illustrated in two serological studies on herpesviruses (Loebel et al., 2017; Blomberg, Rizwan, et al., 2019).

Recently, the analysis of serological data from the United Kingdom ME/CFS Biobank (UKMEB) did not find any association between ME/CFS and the presence of antibodies against chronic infections by different herpesviruses (Cliff et al., 2019). In this work, we re-analyzed these data by dividing the ME/CFS patients into four subgroups related to non-infection versus infection disease triggers. We also performed a sensitivity analysis of the association between ME/CFS and each herpesvirus as a function of the cutoff defining seropositivity.

4.2 Materials and Methods

4.2.1 Study participants

All study participants are part of the UKMEB as described before (Tengvall et al., 2019). In summary, the data refer to a cohort of 226 patients with ME/CFS and 99 healthy controls (HC). At biobank enrollment, patients had to fill in a symptom's assessment questionnaire in which they were asked a specific question about whether they had an infection at the disease onset. This question had four categories of response, which we used to divide the patients into the following subgroups (Table 4.1): subgroup S_0 - she/he did not know whether she/he had an infection at the disease onset ($n = 42, 18.5\%$); subgroup S_1 - she/he did not have an infection at the disease onset ($n = 43, 18.9\%$); subgroup S_2 - she/he had an infection at the disease onset, but this infection was not confirmed with a lab test ($n = 93, 41.0\%$); subgroup S_3 - she/he had an infection at the disease onset and this infection was confirmed with a lab test ($n = 48, 21.1\%$). In the participant questionnaire, patients were also asked to narrate the factors that could have triggered or contributed to the disease. Given that this was an open question, we only performed a brief description of the respective responses (Table 4.2).

All individuals had age between 18 and 60 years old. Patients with ME/CFS were referred for a possible inclusion in the UKMEB by general practitioners working in the United Kingdom National Health System (NHS). The respective diagnosis was confirmed using the 1994 Centers for Disease Control and Prevention (CDC) (Fukuda et al., 1994) or the 2003 Canadian Consensus Criteria (Carruthers et al., 2003) by the UKMEB dedicated clinical research team, according to their designed clinical protocol (E. M. Lacerda, Bowman, et al., 2017). Healthy controls were either family member or friends of the recruited patients with ME/CFS, or they were volunteers recruited by advertisement within Higher Education Institutions. The exclusion and inclusion criteria of the UKMEB and additional information about recruitment and sample processing can be found elsewhere (E. M. Lacerda, Bowman, et al., 2017; E. M. Lacerda, Mudie, et al., 2018).

4.2.2 Herpesviruses serology

Serological data and the respective laboratory procedures were previously described in the original study (Cliff et al., 2019). However, given that the main focus of this early study was cellular immunology, the description of herpesviruses serology was kept to a minimum. We have now provided some additional details. The following commercial ELISA assays from Demeditec Diagnostics (Kiel, Germany) were used to quantify the plasma concentrations of IgG antibodies against the following viruses: the human cytomegalovirus (CMV; Prod. Ref. DECMV01), EBV - VCA antigen (Prod. Ref. DEEBVG0150), EBV - EBNA1 antigen (Prod. Ref. DE4246), herpes simplex virus-1 (HSV1) (Prod. Ref. DEHSV1G0500), herpes simplex virus-2 (HSV2; Prod. Ref. DEHSV2G0540), Varicella-Zoster virus (VZV; Prof. Ref. DE-VZVG0490). The commercial ELISA-VIDITEST from VIDIA (Vestec, Czech Republic) was used for IgG quantification against the human herpesvirus 6 (HHV6; Prod. Ref. ODZ-235). Antibody quantification was expressed in arbitrary units per milliliter (U/ml). According to manufacturer's instructions, seropositivity was considered for all samples with concentration $\geq$ 12 U/ml for HSV1, HSV2, VZV, CMV and EBV antigens. Likewise, individuals with antibody concentrations against HHV6 $\geq$ 12.5 U/ml were considered seropositive.

4.2.3 Statistical analysis

To compare the age and gender distributions of different study groups and/or subgroups of ME/CFS patients, we used the non-parametric Kruskal-Wallis test and the Pearson's χ^2 test for two-way contingency tables, respectively. For simplicity of the analysis, we only reported frequencies and the respective percentages of different disease triggers in the subgroups of ME/CFS that mentioned the occurrence of such triggers.

We previously performed thorough analyses of different cutoff values for seropositivity to each viral antigen (Dias Domingues et al., 2020; Dias Domingues et al., 2021). These earlier analyses were based on the comparison and selection of different scale mixture of skew-normal distributions and four different criteria to define seropositivity. In accordance with a sensitivity analysis, instead of selecting a fixed cutoff, we here allowed this cutoff to vary between 10 U/ml and 100 U/ml with a lag of 1 U/ml. For each cutoff of a given antibody, we first estimated the unadjusted seropositivity odds ratio (OR) between each ME/CFS subgroup and the healthy controls using a logistic regression model in which seropositivity status of the individuals and a group indicator were the outcome and the covariate, respectively. We then adjusted this OR using a similar logistic regression model but including age, gender, and a group indicator variable as covariates. In both unadjusted and adjusted analyses, the effect of healthy controls was set as the reference of the group indicator variable. We used the Wald's score test to assess the significance of different log-ORs in relation to healthy controls.

Finally, we estimated the statistical power of the hypothetical associations using a parametric Bootstrap (Efron et al., 1993). For each antibody, we used the following algorithm: (i) determine the optimal seropositive cutoff by maximizing the likelihood ratio statistic as a function of the seropositivity cutoff when comparing the above logistic models with and without the group indicator covariate; (ii) generate the seropositivity data resulting from the optimal cutoff; (iii) estimate a logistic model including the group indicator only (unadjusted analysis) or a logistic model including age, gender and group indicator variables as covariates (adjusted analysis) using the seropositivity data obtained in (ii); simulate 1,000 data sets using the seropositive probability estimates obtained from models fitted in (iii); (iv) calculate the power of the association between seropositivity and each study group by the proportion of simulated data sets in which the association was deemed significant at the 5% significance level using the Wald's score test as described above.

The statistical analysis was conducted in the R software version 4.0.3. In particular, the estimation of the logistic regression models was done using the "glm" command. The corresponding scripts are available from the first or the corresponding author upon request. The significance level of each executed test was set at 5%.

4.2.4 Ethical approval

All participants provided written informed consent for data collection (questionnaire, clinical measurement and laboratory tests), and for allowing their samples to be available to any research receiving ethical approval (E. M. Lacerda, Bowman, et al., 2017). Participants received an extensive information sheet and consent form in which there was an option for participation withdrawal from the study at any time. Ethical approval was granted by the London School of Hygiene & Tropical Medicine (LSHTM) Ethics Committee (Ref. 6123) and the National Research Ethics Service (NRES) London-Bloomsbury Research Ethics Committee (REC ref. 11/10/1760, IRAS ID: 77765).

4.3 Results

4.3.1 Basic characterization of study participants

The four subgroups of ME/CFS had the same age distribution approximately (Kruskal-Wallis test, $p = 0.30$) with means of 44.6, 40.7, 43.3, and 40.9 years old for S_0, S_1, S_2, and S_3, respectively. Similarly, the percentages of female patients ranged from 70.8% to 80.6%, but they were not statistically different (Pearson's χ^2 test, $p = 0.62$). Overall, the percentage of severely affected patients significantly differed among the subgroups (Pearson's χ^2 test, $p = 0.003$). In particular, the percentage of these patients in S_0 and S_1 was approximately 9%. This value was in clear contrast with the 30% of severely affected patients belonging to S_2 and S_3, both groups related to infection triggers. In terms of the number of narrated disease factors or triggers reported in the participant's questionnaire, the subgroup S_1 had the lowest percentage of patients reporting a single factor or trigger for their disease (44%) when compared to infection-related subgroups S_2 and S_3 (56% and 67%, respectively; Table 4.1). The same subgroup was the one with the highest percentage of missing data to this question (33% for S_1 versus 6% and 10% for S_2 and S_3, respectively). Overall, the distribution of the number of reported disease factors or triggers was significantly different among subgroups S_1, S_2, and S_3 (Pearson's χ^2 test, $p < 0.001$) mostly due to differences in the amount of missing data.

These subgroups of ME/CFS patients were well matched for gender and age with respect to the healthy control group (Pearson's χ^2 and Kruskal-Wallis tests, $p = 0.69$ and 0.44, respectively).

4.3.2 Disease factors or triggers reported by different subgroups of ME/CFS

When the 184 patients belonging to the subgroups S_1, S_2, and S_3 were asked to narrate the factors or triggers of their disease in the participant questionnaire, 103 (56%) and 56 (30%) of them reported single and multiple factors (or triggers), respectively. However, 25 patients (14%) did not mention any specific trigger or factor contributing to their disease. The following non-infection factors or triggers were mentioned by patients mostly belonging to the subgroup S_1: stress subdivided into general anxiety (9%, n=20), personal (8%, n=18) or professional-related stress (5%, n=11); accidents/injuries/surgeries (5%, n=11); pregnancy, childbirth and other problems related to women's reproduction system (3%, n=6), and other non-infection triggers (Table 4.2). The remaining factors are related to microbial infections and/or infectious diseases: upper respiratory tract infections - glandular fever (GF), tonsillitis, EBV infections, or throat infection (21%, n=48); lower respiratory tract infections - chest infection or pneumonia (4%, n=10); flu- or cold-like illness (11%, n=26); gastrointestinal problems and related infections (4%; n=9); and tropical infectious diseases – Dengue fever and schistosomiasis (1%, n=3); and other viral or bacterial infection, and unspecified infections (22%, n=51). Note that 6 patients from subgroup S_1 mentioned an infection in the narrative question about the factors or triggers of their disease. However, the same patients also reported other possible non-infection triggers, such as trauma, bereavement, and stress. We speculate that these patients attributed a higher likelihood to these non-infection disease triggers when answering the related question in the symptoms' assessment questionnaire. Interestingly, patients belonging to the subgroup S_3 reported the highest percentage of disease factors or triggers consistent with an EBV infection (46%, n=22). Patients from subgroup S_2 also self-reported a high frequency of EBV-related factors or triggers (27%, n=25), but closely matched by a flu-like infection or illness (22%, n= 20).

4.3.3 Serological data analysis by subgroup of ME/CFS

We then compared serology data of these ME/CFS subgroups of patients with healthy controls (Figure 4.1). In this analysis, we intended to assess the impact of cutoff on the resulting seropositivity odds ratio between each study group and healthy controls.

With the respect to unadjusted analysis, we could not find any significant association of herpesviruses serology with subgroups S_0 and S_2 (Figures 4.2A and C). The only exception was a putative association for subgroup S_0 using a cutoff of 37 U/ml for the antibodies against EBV-VCA (Figure 4.2A). Interestingly, we found significant negative associations between the subgroup S_1 and antibodies against EBV-VCA, EBV-EBNA1, and VZV depending on the cutoff used (Figure 4.2B). These negative associations suggested decreased seroprevalences to these herpesviruses in this subgroup when compared to healthy controls. We also found a strong negative association between subgroup S_3 and CMV seropositivity (Figure 4.2D). This association was consistent across the range of cutoffs specified for the analysis and it suggested decreased antibody levels in this subgroup of patients in relation to healthy controls. All the above findings remained significant after adjusting for age and gender (Figures

A.2A, B, C, and D). This finding was consistent with a good matching between the different ME/CFS subgroups of patients and healthy controls in terms of age and gender. However, the significance of adjusted ORs was slightly reduced due to these putative confounding factors.

Finally, we estimated the statistical power related to the identified associations using the optimal seropositivity cutoff for each herpesvirus antibody. For the unadjusted analysis, these optimal cutoffs varied from 11 to 90 (Figure A.1 and Table A.1). Similar optimal cutoffs were obtained for the analysis adjusting for age and gender (Figure A.3 and Table A.1) with the exception of EBV-EBNA1 for which the optimal cutoffs were 72 and 88 for the unadjusted and adjusted analyses, respectively. The maximum power ($\approx$ 90%) was obtained for the association between CMV seropositivity and ME/CFS subgroup S_3 in either unadjusted or adjusted analyses (Figure A.4). A high power ($\approx$75%) was also obtained for the associations between VZV seropositivity and ME/CFS subgroup S_1. The remaining associations between each study group and herpesvirus seropositivity had a power that did not exceed 60%. In conclusion, the manufacturer's seropositivity cutoffs were not the most adequate to maximize the chance of finding an association of ME/CFS subgroups with the herpesvirus serology and only three associations between the study groups and herpesviruses seropositivity had a high statistical power.

4.4 Discussion

In contrast with the original study where we could not find differences related to herpesviruses serology between healthy controls and ME/CFS patients divided according to their disease severity (Cliff et al., 2019), our re-analysis of the same data identified two subgroups of ME/CFS patients (S_1 and S_3) in which such differences are now statistically significant. This new finding was only possible due to the stratification of patients according to a question related to the occurrence of an infection at disease onset together with a sensitivity analysis of the seropositivity cutoff used. Patients' stratification or subtyping was performed in line with past recommendations for ME/CFS research (Jason et al., 2005). Following this recommendation, we previously performed an immunological investigation based on a stratification of ME/CFS patients according to the severity of their symptoms (Cliff et al., 2019). By using this stratification, we showed perturbations in the T-cell compartment, namely, in effector CD8+ T cells and in the mucosal-associated invariant T cells. In another study using similar stratification of the samples from the UKMEB, the levels of the cellular stress biomarker GDF15 were found to be increased in severely affected patients at different time points (Melvin et al., 2019). We speculate that immunological and other perturbations could be detected if our alternative stratification would have been used. This investigation will be carried out in the near future.

In line with our findings, evidence has been emerging that the occurrence of an acute infection at the onset of disease symptoms is indeed a key stratifying factor to detect genetic and immunological differences between subgroups of ME/CFS patients when compared to healthy controls (Steiner et al., 2020; Szklarski et al., 2021). However, the simplistic approach of dividing patients according to non-infection and infection triggers might not be sufficient to obtain relatively homogeneous subgroups of ME/CFS patients, which affects the statistical power to detect any disease-specific effects. Besides the limited choice of antibodies against different herpesvirus-related antigens, the large heterogeneity in infectious triggers seems a possible explanation for the lack of association between the subgroup S2 and herpesviruses seropositivity. Notwithstanding not having their infection trigger confirmed in the lab, patients from this subgroup reported the highest proportion of flu-like illnesses, which could have been caused by the influenza virus, the rhinovirus, or the respiratory syncytial virus (Monto, 2002). It is then conceivable that these patients exhibit different pathological mechanisms of ME/CFS according to the causative virus, some of which without any direct impact on antibody responses against herpesviruses. To overcome these problems, we recommend the collection of infection-trigger data as detailed and accurate as possible.

Our most consistent association was obtained between CMV seropositivity and patients from the subgroup S_3. These patients tended to be less seropositive to this herpesvirus when compared to healthy controls, irrespective of the seropositivity cutoff value used. Previously, different serological investigations did not provide conclusive evidence for the role of CMV on ME/CFS pathogenesis, as reviewed in Rasa et al., 2018. The lack of or the use of an inadequate stratification could also explain these past findings. In this regard, the unveiled association was obtained in a subgroup in which the accuracy of the reporting might be the highest, because the disease-triggering infections were supposedly confirmed in the lab. However, we cannot ignore the fact that this subgroup has a large fraction of patients whose disease trigger was related to an EBV infection, one of the most reported causative agents of ME/CFS. Therefore, it is possible that our finding resulted from a coincidence between a low-resolution patient's stratification and a random enrichment of a specific infection trigger in one of the subgroups.

A supposedly decreased seropositivity (or antibody levels) to CMV in an EBV-infection trigger could be explained by the hyperregulation hypothesis (Sepúlveda, Carneiro, et al., 2019). According to this hypothesis, a possible pathological mechanism of ME/CFS is related to an expansion of regulatory T cells (Tregs) driven by an autoimmune response against a viral antigen that mimics a self-antigen. This expansion of Tregs upon herpesvirus infection or reactivation locks the (adaptive) immune system in an active state of hyperregulation where different infections are more difficult to be cleared from the body. Frequent infections are in fact reported by patients with ME/CFS (E. Lacerda et al., 2019). The question is then how the expansion of Tregs could affect antibody responses against CMV. The so-called follicular Tregs might hold the answer to this question. These specialized Tregs are derived from Treg precursors with the ability to migrate to germinal center reactions (GCRs) to inhibit the respective antibody production and antibody maturation (Maceiras et al., 2017). In particular, experiments with animal models demonstrated that the amount of IgG antibodies against different foreign antigens is increased in immunized mice depleted of follicular Tregs (Chung et al., 2011; Wollenberg et al., 2011). In this line of thought, it is reasonable to assume that an increased proportion of Tregs in ME/CFS could be translated into an increased proportion of follicular Tregs. This increase could in turn decrease the antibody production derived from GCRs. We can then hypothesize that an EBV infection triggered an autoimmune response that disrupted the normal balance between Tregs and effector T cells; a peptide of the viral EBNA6 was found to share a high sequence homology with the human lactoperoxidase and thyroid peroxidase (Loebel et al., 2017). The disruption of this balance could lead to an increase of both Tregs and follicular Tregs. A possible consequence of this increase is a diminished antibody production against a posterior CMV infection or reactivation. Note that several peptides from CMV were also found as putative candidate for molecular mimicry with human proteins (Lunardi et al., 2005). Similar to the situation of immunosuppression, a reduction in the humoral immunity against CMV would render ME/CFS patients more susceptible to a possible reactivation of this virus (Krmpotić et al., 2019). It is worth noting that the role of follicular Tregs was never investigated on ME/CFS.

Another interesting finding is the possible association between the subgroup S_1 and EBV and VZV seropositivity. This subgroup refers to patients who reported non-infectious triggers, mostly related to stressful or stress-related events. It is also a group where ME/CFS was triggered in many women who had problems during and after pregnancy, had difficult childbirth or had disorders related to women's reproduction system. In line with this finding, stressful conditions and events such as the ones experienced by astronauts increase the chance of herpesvirus reactivation, specifically, EBV, VZV and CMV (Rooney et al., 2019). Reactivation of latent herpesvirus infections could be explained by an increase in production of stress-related hormones together with an inflammatory cytokine signature that debilitates the immune system. This subgroup is then expected to have a higher prevalence of active herpesvirus infections than the remaining subgroups of ME/CFS patients and healthy controls. Given that this subgroup could represent less than 50% of the patients (Johnston et al., 2016; Chu et al., 2019), it is likely to have insufficient statistical power to detect any differences in herpesvirus reactivation rates between ME/CFS and healthy controls even in the case of a proper stratification of the patients' populations. This limitation is then likely to explain the inconsistent findings on herpesvirus reactivation across many studies on ME/CFS.

We did not find any association between the subgroup S_0 and herpesvirus seropositivity. This subgroup represented 18.5% of the patients' cohort, a value compatible with the percentages of patients that did not report any disease triggering event from past epidemiological studies (10%, ref. (Chu et al., 2019); 24%, ref. (L.C. Nacul et al., 2011)). The sample size of this subgroup was not very large and, therefore, we cannot rule out that our lack of associations could be simply due to insufficient statistical power to detect putative associations between this subgroup and herpesvirus seropositivity. Finally, in our association analysis, we allowed the seropositivity cutoff to vary within a given range of possible values, similarly done in a recent study of molecular mimicry between Anoctamin 2 and EBNA1 in multiple sclerosis (Tengvall et al., 2019). This analytical approach seems reasonable given the difficulty to choose the best seropositivity cutoff among the different criteria and methods available, as illustrated in earlier analyses of the same data (Dias Domingues et al., 2020; Dias Domingues et al., 2021). This approach is also in line with several discussions about seropositivity estimation and the sensibility to use a fixed cutoff (Ridge et al., 1993; Kafatos et al., 2016; Migchelsen et al., 2017; Bouman et al., 2020). However, we should note that a high cutoff for the data might not define seropositivity per se, but rather a high antibody response whose detection could be the primary objective of the analysis (Loebel et al., 2017; Blomberg, Rizwan, et al., 2019). The use of a high cutoff is also in accordance with a modelling approach where seropositivity might be subdivided into different levels (Pothin et al., 2016; Nhat et al., 2017; Moreira da Silva et al., 2020). Therefore, our sensitivity-like approach seems to have the capacity to detect further serological associations beyond the ones based on the classical seroprevalence. Such a capacity could increase the chance of reaching scientific reproducibility. We then recommend the routine use of this approach in future serological investigations of ME/CFS.

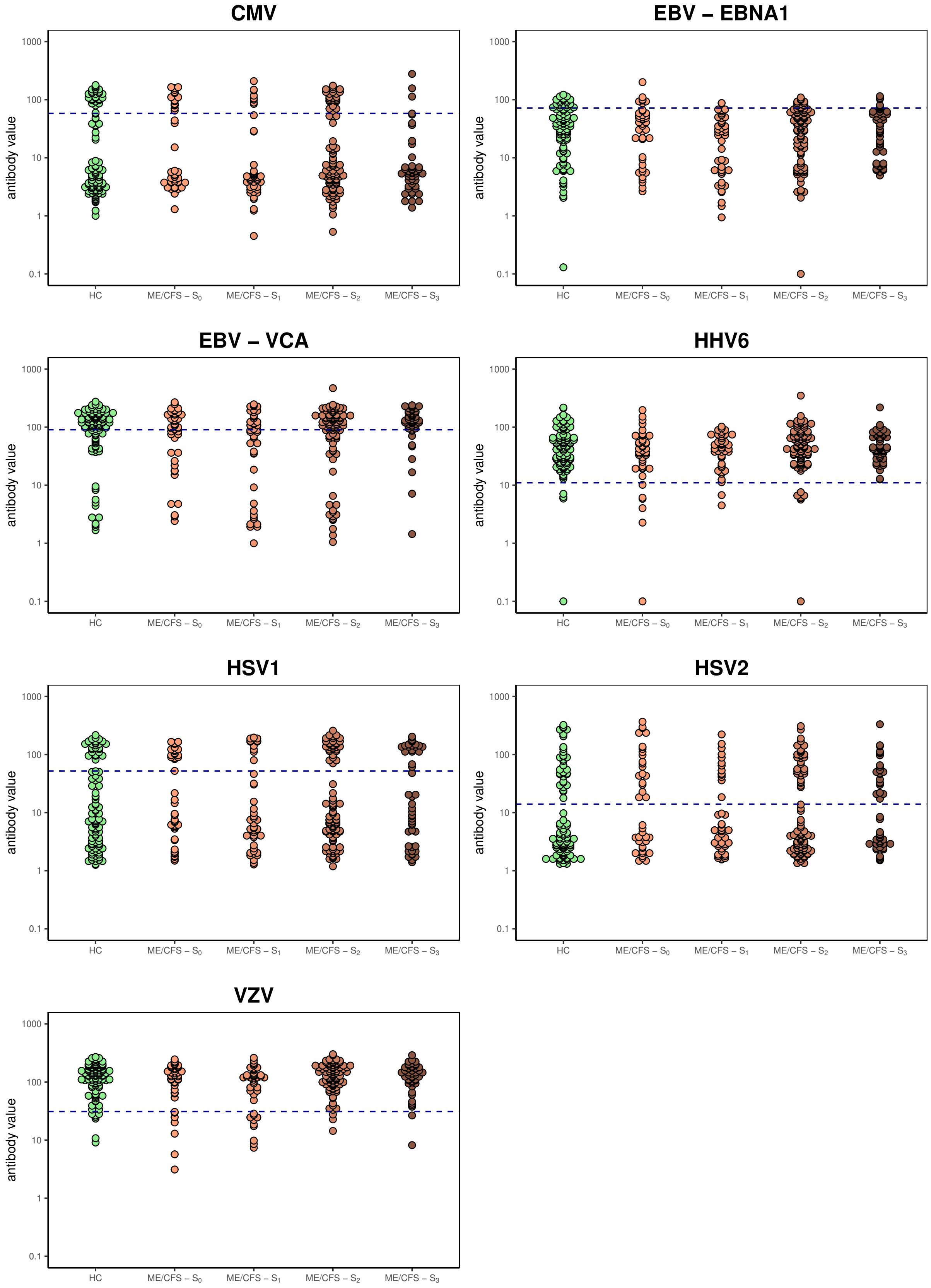


Figure 4.1: Herpesvirus serology data per study group including the four ME/CFS subgroups. Horizontal dashed lines represent the optimal seropositivity cutoff for the unadjusted analysis according to the maximization of likelihood ratio statistic for testing the significance of the group indicator covariate in the logistic models. Antibody concentration in y axis is given in U/ml.

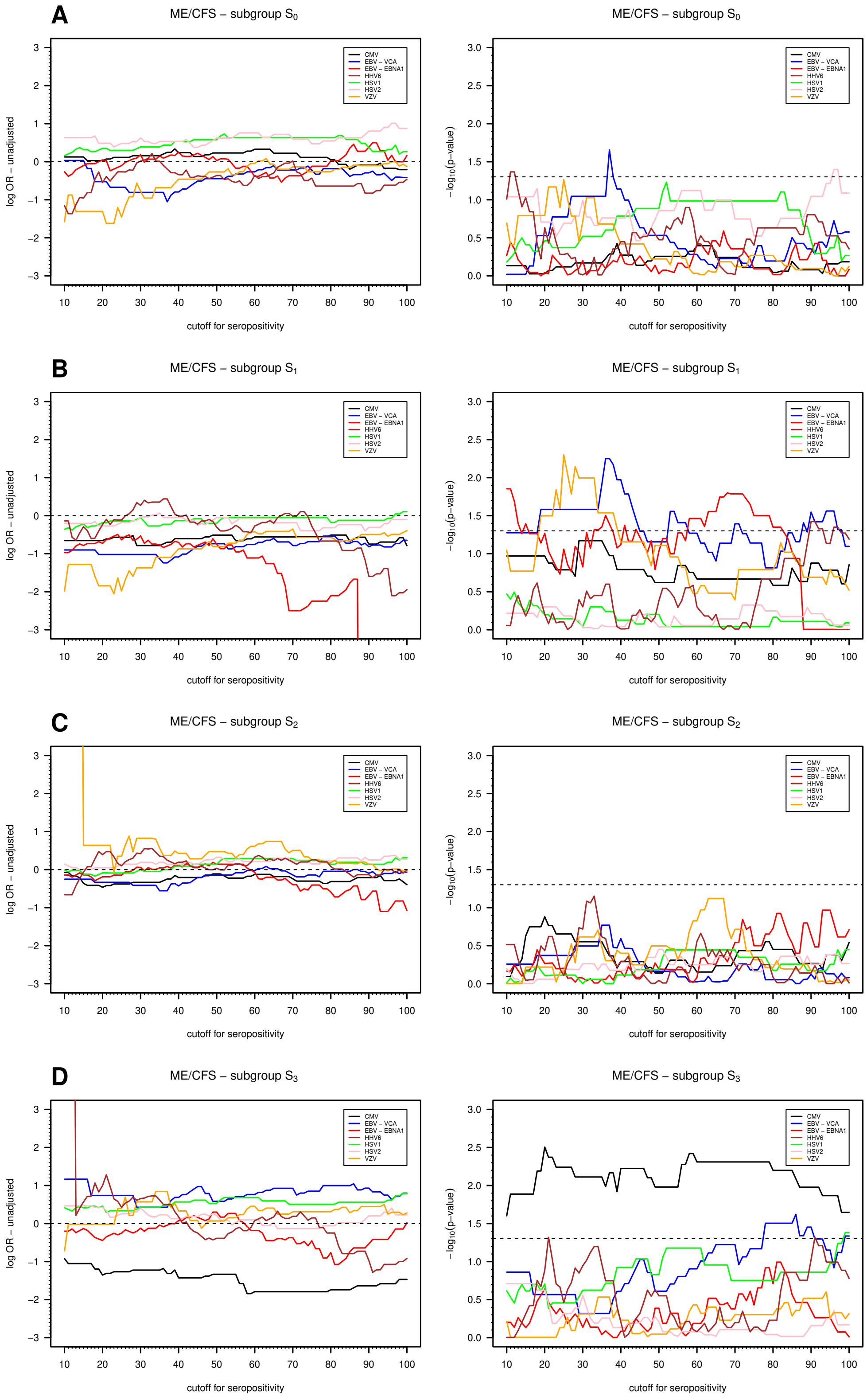


Figure 4.2: Unadjusted association analysis of seropositivity to different herpesvirus antigens based on log-OR of four subgroups of patients with ME/CFS in relation to healthy controls. For convenience, statistical significance was calculated as -log10(p-value). The dashed lines in the statistical significance plots represent the threshold associated with the 5% significance level (i.e., -log10(0.05)). Cutoff values in which -log10(p-values) are above these thresholds provided evidence for significant associations.

Table 4.1: Basic characteristics of study participants where patients with ME/CFS were split into four subgroups according to the responses about their disease triggers in the symptoms' assessment questionnaire: S_0 – Do not know; S_1 – Non-infection trigger; S_2 – An infection trigger but not confirmed with a lab test; and S_3 – An infection trigger confirmed with a lab test. IQR denotes the interquartile range.

		Subgroups of ME/CFS patients				**Comparison (p-values)**	
	Healthy controls (n=99)	$S_0(n=42)$	$S_1(n=43)$	$S_2(n=93)$	$S_3(n=48)$	**ME/CFS subgroups**	**ME/CFS subgroups + Healthy controls**
Female (%)	73 (73.7)	33 (78.6)	33 (76.7)	75 (80.6)	34 (70.8)	0.62	0.69
Mean age (IQR)	41.9 (32.0-51.5)	44.6 (35.0-53.8)	40.7 (28.0-52.0)	43.3 (35.0-53.0)	40.9 (32.0-50.3)	0.30	0.44
Disease severity at recruitment							
Mild/moderate (%)	-	38 (90.5)	39 (90.7)	64 (68.8)	34 (70.8)	0.003	-
Severely affected (%)	-	4 (9.5)	4 (9.3)	29 (31.2)	14 (29.2)	-	-
Number of self-reported disease triggers							
Single	-	-	19 (44)	52 (56)	32 (67)	<0.001[a]	-
Multiple	-	-	10 (23)	35 (38)	11 (23)	-	-
Missing	-	-	14 (33)	6 (6)	5 (10)	-	-

[a] Pearson's χ^2 test including the missing as a category for the number of disease factors/triggers.

Table 4.2: Frequency and the respective percentage within brackets of specific disease factors or triggers narrated by patients from the subgroups S_1 ($n = 43$), S_2 ($n = 93$), and S_3, ($n = 48$) in the participant's questionnaire.

	Subgroups of ME/CFS patients			
Reported disease trigger	S_1(%)	S_2(%)	S_3(%)	Total(%)
Glandular Fever; tonsilitis; EBV infection	1 (2)	25 (27)	22 (46)	48 (21)
Respiratory infection; pneumonia	1 (2)	6 (6)	3 (6)	10 (4)
Flu-like infection or illness	2 (5)	20 (22)	4 (8)	26 (11)
Gastrointestinal infection	0 (0)	6 (6)	3 (6)	9 (4)
Tropical infections	0 (0)	1 (1)	2 (4)	3 (1)
Other infections including unspecified viral infections	2 (5)	33 (35)	13 (27)	51 (22)
General Stress or Anxiety	6 (14)	11 (12)	3 (6)	20 (9)
Stress due to personal events	9 (21)	6 (6)	3 (6)	18 (8)
Stress at work or school	4 (9)	5 (5)	2 (4)	11 (5)
Vaccinations	0 (0)	4 (4)	6 (12)	10 (4)
Chemical exposure	1 (2)	6 (6)	0 (0)	7 (3)
Accidents/Injuries/Surgeries	7 (16)	2 (2)	2 (4)	11 (5)
Pregnancy/Childbirth/Postnatal/Hysterectomy/Endometriosis	6 (14)	0 (0)	0 (0)	6 (3)
Other	4 (9)	6 (6)	0 (0)	7 (3)

Chapter 5

Analysis of cutoff point estimation for determining seropositivity in the context of SARS-CoV-2 infections

This chapter will apply mixture models based on distributions from the SMSN family to antibody data against four SARS-CoV-2 virus antigens. Furthermore, since the true infection status of individuals is known *a priori*, performance measures will be calculated for the methods proposed in chapters 3 and 4 for cutoff point estimation such as sensitivity, specificity and accuracy. The results of a simulation study will also be presented.

5.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection that causes the devastating and often lethal COVID-19 disease was first detected in China, province of Wuhan in December 2019 (Rosado et al., 2020). Rapidly, SARS-CoV-2 infection spread over the entire world and the COVID-19 disease was declared as a pandemic by the World Health Organization.

The detection of the virus is so far done by the so-called reverse transcription quantitative PCR (RT-qPCR) on samples from nasopharyngeal or throat swabs (Rosado et al., 2020). In general, only symptomatic individuals or people who were in close contact with detected cases are tested, which might lead to an underestimation of the proportion of individuals infected with SARS-CoV-2 (Stringhini et al., 2020). Alternatively, serological testing allows to detect asymptomatic individuals exposed to the infection. In addition, serological testing is able to quantify the degree of exposure to the infection in the population. In this context, it is important to estimate seroprevalence at the population level, i.e., the proportion of seropositive individuals that show antibodies against any SARS-CoV-2 antigen (Larremore et al., 2020).

The presence of antibodies in a serum sample can be regarded as an indicator of immunity against a given infectious agent or as an indicator of past infection in the absence of vaccination (Gay, 1996). The detection of antibodies in the serum samples is classically done via enzyme linked immunosorbent assays (ELISA), where the resulting data are light intensities, also called optical density, which reflects the underlying antibody concentration in the samples (Dias Domingues et al., 2021). For statistical convenience, the analysis of serological data proceeds by dichotomizing the amount of antibodies present in the serum of an individual using an arbitrary cutoff point in the antibody distribution to achieve a certain sensitivity and specificity. This allows the classification of individuals into seronegative (with antibody levels below the cutoff point) and seropositive (with antibody levels above the cutoff point) (Rosado et al., 2020).

Given the possible impact of the cutoff chosen, different criteria for seropositivity determination have a direct impact on the sensitivity and specificity of the respective serological classification (Parker et al., 1990). In addition, it might also impact the estimation of the seroprevalence (Kafatos et al., 2016) and the following (epidemiological) decision that can be taken when facing a given estimate of this epidemiological parameter. This means that when determining the cutoff point for a serological test, one should take into account the benefit of the test, the economic and social consequences of serological misclassification and the prevalence of the disease in the population. It turns out that these aspects are often ignored in practice (Ridge et al., 1993).

One of the traditional methods to establish the cutoff point in serological assays is to consider the logarithmic transformation of the antibody concentration of a known seronegative population and proceed to calculate the mean plus 2 or 3 standard deviations (Ridge et al., 1993; Maple et al., 2006; Baughman et al., 2006; Tong et al., 2007). This method is more adequate when the antibody distribution of the seronegative population is normally distributed (Baughman et al., 2006). However, our previous studies of different serological data (Moreira da Silva et al., 2020; Dias Domingues et al., 2021) showed evidence against a normality assumption for the antibody levels associated with a putative seronegative population. In the case where the true infection (or disease) status is known, ROC curve-based methods are most commonly used to determine the cutoff point for defining seropositivity. These methods are widely discussed in the literature (Perkins et al., 2006; Hasibi et al., 2013; M. Rota et al., 2014; Habibzadeh et al., 2016; Blacksell et al., 2016; Unal, 2017; Migchelsen et al., 2017).

Alternatively, finite mixture models can be used to determine the seropositivite cutoff directly from the data (Baughman et al., 2006; Sepúlveda, Stresman, et al., 2015; Kafatos et al., 2016; Migchelsen et al., 2017; Dias Domingues et al., 2021). In our previous work, three methods for determining seropositivity cutoff were explored using the so-called scale mixtures of Skew-normal distributions in the case where the true infection status is unknown (Dias Domingues et al., 2021). In this paper we applied the same methods and models in order to evaluate their performance in freely available serological data concerning SARS-CoV-2 virus (Rosado et al., 2020). We also used simulation to understand the performance of the cutoff estimators associated with different criteria for seropositivity determination.

5.2 Serological data concerning SARS-CoV-2 virus

In this study we analyzed IgG antibody responses against four SARS-CoV-2 spike or nucleoprotein antigens: RBD – glycoprotein receptor-binding domain; S^{tri} — S trimeric spike protein; S1 — spike glycoprotein S1 domain; S2 – SARS-CoV-2 spike glycoprotein S2 domain. Antibodies were measured in serum samples collected up to 39 days after symptom onset from 215 adults in four French hospitals (53 patients and 162 health-care workers) with quantitative RT-PCR-confirmed SARS-CoV-2 infection. A total of 335 negative control serum samples were collected from France, Thailand, and Peru before the start of the COVID-19 pandemic (Rosado et al., 2020). A detailed description of lab procedures can be found in the original study (Rosado et al., 2020). The data is freely available at https://github.com/MWhite-InstitutPasteur/SARSCoV2SeroDXphase2.

5.3 Statistical methods

Serological data can be viewed as arising from two or more latent populations; each population is assumed to represent different levels of exposure to a given antigen. For simplicity, individuals that were never exposed or exposed a long time ago to an infectious agent are considered as seronegative. In contrast, individuals exposed to the same infectious agent are considered seropositive. In this scenario, the antibody distribution can be described by a mixture of two or more probability distributions (Dias Domingues et al., 2020). However, the true serological state of the individuals is unknown and therefore it needs to be estimated.

In the particular case of the SARS-CoV-2 data, we know which individuals were exposed to the virus and, therefore, we can assume to know which individuals are true seronegative and seropositives.

In many serological studies, it is common to assume a normal distribution for the basis of the mixture models. However, the behaviour of antibody distribution is not constant over time and their concentration decreases after infection (Rosado et al., 2020). This fact makes the distribution of the seropositive population skewed to the left (Gay, 1996). In order to accommodate the possible skewness in the seropositive population we use the scale mixture of Skew-Normal (SMSN) class of distributions that include the Skew-Normal and the Skew-t distributions, which will be the focus of our study. A brief description of these alternative distributions can be found below.

5.3.1 Skew-Normal and Skew-t distributions

Let $W \frown SN(\mu, \sigma^2, \alpha)$ a random variable with a Skew-Normal distribution. In this distribution, the parameters μ, σ^2, and α can be seen as the location, scale, and shape parameters, respectively. Then the probability density function (pdf) is given by

$$\begin{aligned} f_W(w) &= 2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{w-\mu}{\sigma})^2} \times \int_0^{\alpha(\frac{w-\mu}{\sigma})} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = \\ &= \frac{2}{\sigma}\phi\left(\frac{w-\mu}{\sigma}\right)\Phi\left(\alpha(\frac{w-\mu}{\sigma})\right), w \in \mathbb{R}, \end{aligned} \tag{5.1}$$

where $\phi(.)$ and $\Phi(.)$ is the pdf and the cumulative distribution function of the standard Normal distribution, respectively (Basso et al., 2010; Azzalini, 2014; Dias Domingues et al., 2021).

The Skew-Normal distribution is part of a family of distributions called the Scale Mixtures of Skew-Normal distributions (SMSN), of which the Skew-t distribution is also a particular case (Dias Domingues et al., 2021).

A random variable W is said to have a Skew-t distribution, $W \frown ST(\mu, \sigma^2, \alpha, v)$, if the pdf is given by

$$f_W(w) = 2f_T(w; \mu, \sigma^2, v+1)F_T\left(A(w)\sqrt{\frac{v+1}{d(w)+v}}; v+1\right), w \in \mathbb{R}, \tag{5.2}$$

where $f_T(.; \mu, \sigma^2, v+1)$ and $F_T(.; \mu, \sigma^2, v+1)$ represents the pdf and the cumulative distribution function of the generalized Student's t distribution with $v+1$ degress of freedom, $A(w) = \alpha\frac{(w-\mu)}{\sigma}$ and $d(w) = \left(\frac{w-\mu}{\sigma}\right)^2$ (Basso et al., 2010; Azzalini, 2014; Dias Domingues et al., 2021).

5.3.2 Finite mixture models

Let G_1 and G_2 be the seronegative and seropositive subpopulations from a population G, respectively. Let π_1 and π_2 the probabilities of sampling a seronegative and a seropositive individual, respectively (with the usual restriction of $\sum_{k=1}^{2} \pi_k = 1$ and $0 \leq \pi_k \leq 1$) and considering Z the random variable that represents the antibody level. The probability density function (pdf) of Z is given by

$$f(z; \Theta) = \sum_{k=1}^{2} \pi_k f_k(z; \boldsymbol{\theta}_k), \tag{5.3}$$

where $f_k(z; \boldsymbol{\theta}_k)$ is the mixing probability density function of Z associated with the $k-th$ latent population and parameterized by the vector $\boldsymbol{\theta}_k$. Θ is the vector of all unknown parameters of the mixture model, i.e., $\Theta = (\pi_1, \pi_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. In our application, $f_k(z; \boldsymbol{\theta}_k)$, is given by the Skew-normal or the Skew-t distributions.

In general, the estimation of a finite mixture model can be done by the classical EM algorithm (S.X. Lee et al., 2016). The EM algorithm is an iterative method widely used in incomplete data problems where the maximum likelihood estimators (MLE) have no closed expression (Dempster et al., 1977). Considering $(z_1, z_2, ..., z_n)$ the observed sample of size n and $Y_i \equiv Y_{ik}, (i = 1, .., n; k = 1, 2)$, the binary vector representing the component from which the data comes from. Thus, $Y_i \frown Bernoulli(\pi_2)$ and the pdf of Y_i is given by

$$f(y_i; \Theta) = \pi_2^{yi2}(1-\pi_2)^{1-y_{i2}}. \tag{5.4}$$

From what was explained in Chapter 2, section 2.1.4, we have that the complete data is the pair (z_n, y_n) and the joint pdf is given by

$$f((z_i, y_i); \Theta) = [(1-\pi_2)(f_1(z_i; \boldsymbol{\theta}_1))]^{1-yi2}[\pi_2 f_2(z_i; \boldsymbol{\theta}_2)]^{yi2}. \tag{5.5}$$

Then, the log-likelihood function is given by

$$\log L(\Theta) = \sum_{i=1}^{n}(1-y_{i2})\log\{(1-\pi_2)(f_1(z_i; \boldsymbol{\theta}_1))\} + y_{i2}\log\{\pi_2 f_2(z_i; \boldsymbol{\theta}_2)\}. \tag{5.6}$$

The step E of the EM algorithm consists in obtaining

$$Q(\Theta, \Theta^{(p)}) = E_{\Theta^{(p)}}\{\log L(\Theta)|z_i\} = \sum_{i=1}^{n} w_{i1}^{(p+1)}\log\{(1-\pi_2 f_1(z_i; \boldsymbol{\theta}_1)\} + w_{i2}^{(p+1)}\log\{(\pi_2 f_2(z_i; \boldsymbol{\theta}_2)\}, \tag{5.7}$$

where $w_{ik}^{(p+1)} = E_{\Theta^{(p)}}\{Y_{ik}|z_i\} = P_{\Theta^{(p)}}\{Y_{ik} = 1|z_i\}, k = 1, 2.$

The step M consists in maximizing $Q(\Theta, \Theta^{(p)})$ as function of the unknown parameters. However, if the model has many parameters that need to be estimated, then step M may incur in computational problems such as excessive time consuming or estimate instability. In this sense, it is possible to break the step M into several sub-steps $(S > 1)$ that allow to get around these computational constraints by performing some restrictions on the parameters. This method is called expectation-conditional-maximization (ECM) algorithm (Meng et al., 1993; Liu et al., 1994; G.J. McLachlan et al., 2008). Considering that $\Theta^{(p+s)}$ represents the value of Θ in the sth CM step of the iteration $p+1$ in order to maximize $Q(\Theta, \Theta^{(p)})$ and the constraint function $g_s(\Theta) = g_s(\Theta^{(p+(s-1))})$, the ECM algorithm is performed as follow (Liu et al., 1994):

1. calculate the expected complete-data log-likelihood given the current estimates of the parameters, $\Theta^{(p)}$. The calculations are the same as for the EM algorithm;

2. fix $\Theta^{(p)}$ and calculate $\Theta^{(p+s)}$ to maximise the expected complete-data log-likelihood;

3. fix $\Theta^{(p+s)}$ and calculate $\Theta^{(p+(s+1))}$ to maximise the expected complete-data log-likelihood on the $s+1$ sub-step iteration and continuing until you have gone through all the S sub-steps.

In this way, it can be seen that $Q(\Theta, \Theta^{(p+s)} \geq Q(\Theta, \Theta^{(p)})$ for all $\Theta \in \Omega_s(\Theta^{(p+s)})$, where $\Omega_s(\Theta^{(p+s)}) = \{\Theta \in \Omega : g_s(\Theta) = g_s(\Theta^{(p+(s-1))})\}$ (Meng et al., 1993; Liu et al., 1994; G.J. McLachlan et al., 2008).

Considering the SMSN family of distributions, namely the Skew-Normal and the Skew-t distributions, the application of the ECM algorithm in the context of mixtures can be found in (Lin et al., 2007b; Basso et al., 2010).

In order to decide which model is the best one among all the models fitted to the same data, we used the Bayesian Information Criterion (BIC) (Dias Domingues et al., 2021).

5.3.3 Definition of seropositivity

Seroprevalence is an epidemiological measure defined by the proportion of seropositive individuals in the sample. For its estimation, it is then necessary to define the serological status of the i-th individual by dychotomization the variable, Z_i, which represents the antibody concentration of the individual. This dychotomization is done by determining a value c such that for antibody values equal to or greater than c, the individual is classified as seropositive and seronegative, otherwise. Thus, let Y be the random variable representing the number of seropositive individuals in a sample of size n, we have to

$$Y = \sum_{i=1}^{n} I_{\{Z_i \geq c\}} \frown Binomial(n, \pi_2),$$

where π_2 represents the seroprevalence, i.e, $\pi_2 = P[Z_i \geq c]$ and $I_{\{.\}}$ is the indicator variable. Considering that the random variable representing the antibody levels Z_i is modelled by a finite mixture of distributions, the way to estimate the cutoff c from the observed data is not standard. To facilitate the determination of this cutoff value, we below present three estimation methods or criteria.

- **Method 1 (M1):** It is based on the 99.9%-quantile associated with the estimated seronegative population. This method is the most popular in sero-epidemiology (Sepúlveda, Stresman, et al., 2015; Saraswati et al., 2019). It is often called as the 3σ rule, because the 99.9%-quantile is given by the mean plus 3 times the standard deviation of a normally distributed seronegative population;

- **Method 2 (M2):** It relies on the minimum of the density mixture functions. In the case of two latent populations, the cutoff corresponds to the absolute minimum, and in the case of three or more latent populations the cutoff corresponds to the lowest relative minimum. This point can be calculated using the Dekker's algorithm (Brent, 1973). It should be noted that the minimum of the mixing function is not expected to coincide with the point of intersection of the probability densities of each individual subpopulation;

- **Method 3 (M3):** It imposes a threshold in the the so-called conditional classification curves (Sepúlveda, Stresman, et al., 2015). Under the assumption that all components but the first one refer to seropositive individuals, the conditional classification curve for the i-th individual given the antibody level $Z_i = x$ is defined as

$$p_{+|Z_i=x} = \frac{\pi_2 f_2(Z_i = x; \boldsymbol{\theta}_2)}{\sum_{k=1}^{2} \pi_k f_k(Z_i = x; \boldsymbol{\theta}_k)}. \tag{5.8}$$

In turn, the classification curve of seronegative individuals is simply given by

$$p_{-|Z_i=x} = 1 - p_{+|Z_i=x}. \tag{5.9}$$

After calculating these curves, one can impose a minimum value for the classification of each individual. In this case, two cutoff values arise in the antibody distribution, one for the seronegative individuals and another for seropositive individuals. Mathematically, the classification rule is given as follows

$$C_i = \begin{cases} \text{seronegative} & \text{, if } x_i \leq c_- \\ \text{equivocal} & \text{, if } c_- < x_i < c_+ \\ \text{seropositive} & \text{, if } x_i \geq c_+ \end{cases} \tag{5.10}$$

where c_- and c_+ are the cutoff values in the antibody distribution that ensure a minimum classification probability, say 90%. To calculate these cutoff values in practice, one can use the bisection method providing an initial interval where they might be located (Sepúlveda, Stresman, et al., 2015).

5.3.4 Performance of the proposed methods for cutoff point estimation

In order to evaluate the performance of each of the cutoff points, we estimated the respective sensitivity and specificity. Let D and D^* be the true and estimated serological classification (or infection status), respectively. Sensitivity is defined as the conditional probability

$$sens = P(D^* = +|D = +), \tag{5.11}$$

In turn, the specificity is defined as

$$spec = P(D^* = -|D = -). \tag{5.12}$$

The overall performance of each method is given by the accuracy (ACC) of the proposed method which corresponds to the proportion of correct results, that is,

$$ACC = sens \times P(D = +) + spec \times P(D = -). \tag{5.13}$$

5.3.5 Simulation study

We performed a small simulation study to assess the performance of cutoff points proposed by each method. With this purpose, we assume two simulation scenarios regarding the mixture model assumed for the data: (i) a mixture model based on the Skew-Normal distributions and (ii) a mixture model based on the Skew-t distribution.

For each scenario, we simulated 1000 samples with dimensions 100, 500 and 1000. In addition, for each simulation cycle, the weight of the mixture model was varied to check the ability of the model to identify the seropositive component even when the weight assigned to that component is very low. The implications of varying the weight of the seronegative and seropositive population are as follows: in the case where the proportion of seronegative individuals is very high relative to seropositive individuals, more effective decisions can be made to control the number of infections in the population. The opposite scenario is important in the case of effectiveness of vaccination in the population, particularly for individuals who may have lost immunity.

To this end, it was considered that the proportion of seronegative individuals could take the values 90%, 60% and 30%, being the respective proportion of seropositive individuals 10%, 40% and 70%, respectively. For each simulated sample, the parameters of the mixture model were estimated by maximum likelihood (via the EM algorithm) according to the distributional scenarios described above, as well as the respective cutoff points according to the methods M1, M2 and M3. Considering θ^* the estimated parameter, θ the true value of the parameter, than we calculate the relative error that is $\frac{1}{1000}\sum_{i=1}^{1000}[(\theta^* - \theta)/\theta] \times 100\%$ and the mean squared error (MSE), i.e, $\frac{1}{1000}\sum_{i=1}^{1000}[(\theta^* - \theta)^2]$.

5.3.6 R packages

We used the package `mixsmsn` to fit different mixture models based on SMSN (Prates et al., 2013). In particular, we used the function `smsn.mix` to estimate the model parameter via the EM algorithm For fitting the Student's t-distribution, we considered the R package `extraDistr` (Wolodzko, 2020), namely, the function `dlst` to calculate their density, the function `plst` to define the cumulative distribution function and the function `rlst` to generate random samples in the simulation study. The fitting of the Skew-Normal distributions was performed with the package `sn` (Azzalini, 2020). The functions `dsn`, `psn` and `rsn` were used to calculate the probability density function, the cumulative distribution function and generate random samples of the Skew-Normal distribution, respectively. In the case of the Skew-t distribution, the functions `dst`, `pst` and `rst` were used to calculate the probability density function, the cumulative distribution function and generate random samples, respectively.

5.4 Results

5.4.1 Patients characteristic's

For this study, data relating to 549 individuals was analysed. Serum samples were collected from individuals with confirmed SARS-CoV-2 infection by PCR test in four hospital units from Paris, namely: 4 (0.7%) from the Hôpital Bichat, 49 (9.0%) from the Hôpital Cochin and 161 (29.3%) from the Nouvel Hôpital (Strasbourg). Regarding the negative controls, 68 (12.4%) are from the Thai Red Cross (TRC), 90 (16.4%) from the Peruvian donors (NHP) and 177 (32.2%) from the France blood donors (Établissement Français du Sang). For each antigen under analysis, the logarithmic transformation of base 10 was considered for the concentration of antibodies against that antigen.

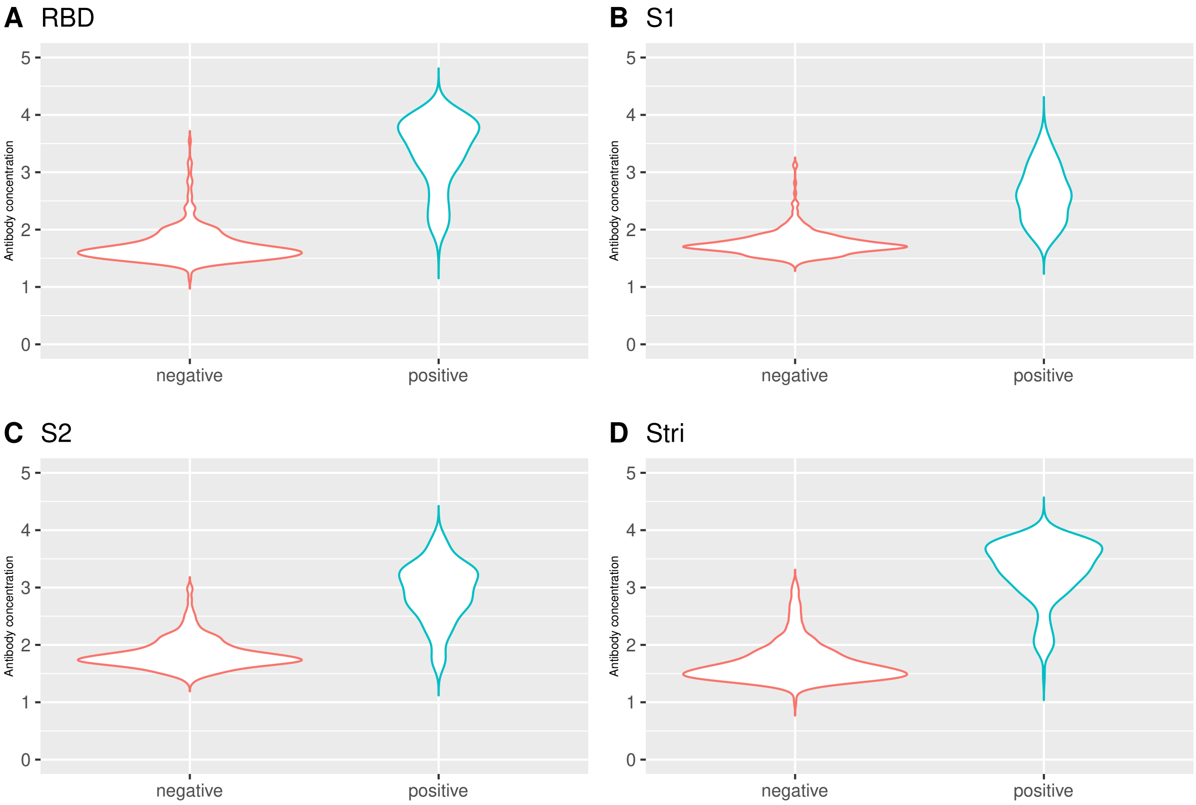


Figure 5.1: Antibody distribution by infection status. **A.** Antibody distribution for RBD antigen. **B.** Antibody distribution for S1 antigen. **C.** Antibody distribution for S2 antigen. **D.** Antibody distribution for S^{tri} antigen. Number of negative individuals: 335; number of positive individuals: 214. Antibody concentration in y axis is given in log10 units.

Regarding the analysis of antibodies by the individuals who performed PCR test, there were statistically significant differences between individuals who tested negative and positive for SARS-CoV-2 by Mann-Whitney test (RBD: 1.64 vs. 3.48, $p < 0.001$; S1: 1.72 vs. 2.59, $p < 0.001$; S2: 1.79 vs. 2.99, $p < 0.001$; Stri: 1.59 vs. 3.43, $p < 0.001$) (Figure 5.1). Such differences were expected given the general knowledge about the infection status, i.e., individuals who have already been exposed to the virus will have a higher concentration of antibodies than those who are still susceptible.

5.4.2 Mixture Model approach

We performed the fitting of the different mixture models considering two subpopulations, i.e., a seronegative population and a seropositive population. According to the BIC values, the model based on the Skew-Normal distribution was considered for the following antigens: RBD (BIC=852.25), S1 (BIC=561.63), S2 (BIC=775.29). For the case of the Stri antigen, the best model was found to be the Skew-t distribution (BIC=915.82) (Figure 5.2 and table 5.2). As has been observed in previous studies, there is a marked skew to the right of the data for the seronegative population and a skewed for the left in the seropositive population, although not very marked for the S1 ($\alpha_{S1} = 1.062$) and S2 ($\alpha_{S2} = 0.450$) antigens (Table 5.1).

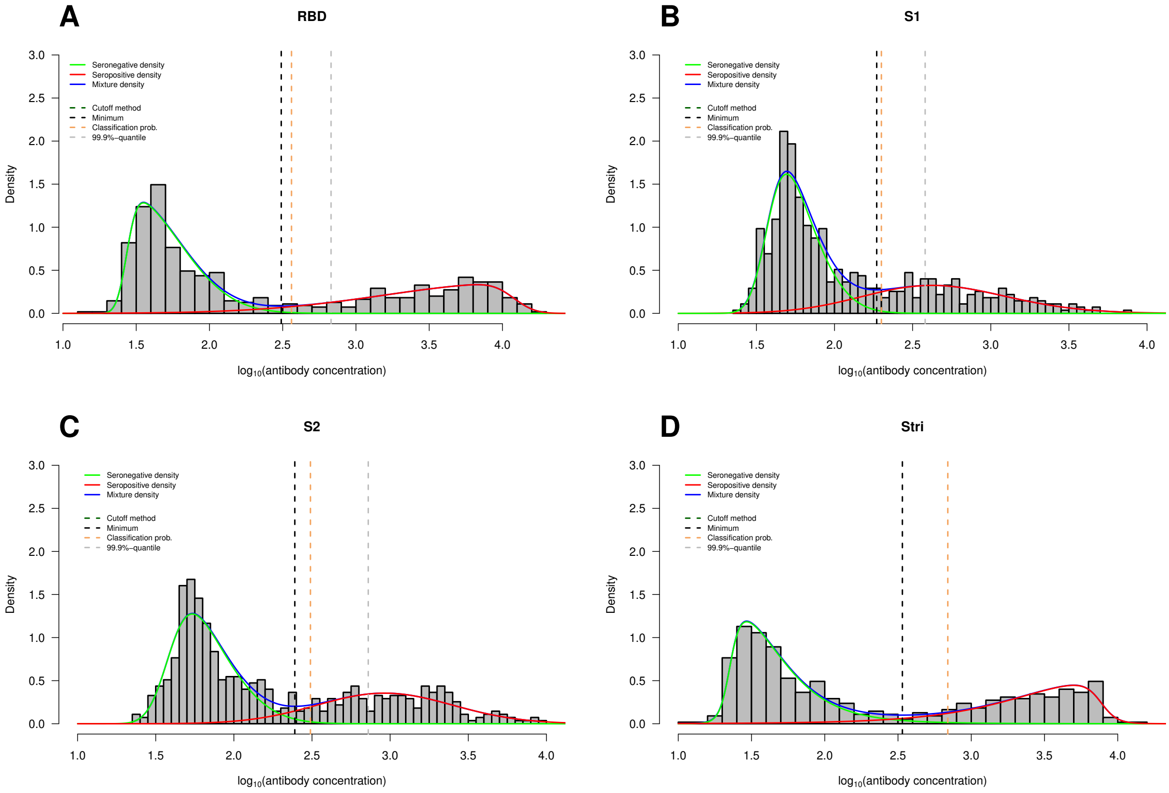


Figure 5.2: Best models with two components for the data under analysis. **A.** Antibody distribution for RBD antigen. **B.** Antibody distribution for S1 antigen. **C.** Antibody distribution for S2 antigen. **D.** Antibody distribution for S^{tri} antigen. Antibody concentration in x axis is given in $\log_{10}$ units.

Table 5.1: Parameter estimates for the best model

		Seronegative population				**Seropositive population**			
Antigen	**Distribution**	μ	σ^2	α	v	μ	σ^2	α	v
RBD	Skew-Normal	1.435	0.125	6.318	NA	4.077	0.767	-7.634	NA
S1	Skew-Normal	1.569	0.062	2.687	NA	2.339	0.321	1.062	NA
S2	Skew-Normal	1.583	0.096	2.804	NA	2.817	0.212	0.450	NA
S^{tri}	Skew-t	1.352	0.121	5.751	4.873	3.885	0.367	-6.482	4.873

5.4.3 Seropositivity estimation

After defining the model that best fits the data, we proceeded to categorize the amount of antibodies for each antigen by estimating the cutoff point. For this, we used the methods M1, M2 and M3 already described and whose results are shown in Figure 5.1 and table 5.2.

Estimation of the cutoff point based on the minimum densities of the mixture model (M2) proved to be the method with the highest sensitivity for classifying seropositive individuals, as well as the one that produces the highest proportion of correct results (accuracy) for the RBD antigen ($cutoff = 2.49, sens = 86.45\%, ACC = 92.89\%$), S1 ($cutoff = 2.27, sens = 71.03\%, ACC = 86.89\%$) and S2 ($cutoff = 2.39, sens = 83.64\%, ACC = 90.89\%$). In the case of the S^{tri} antigen, it was not possible to calculate the sensitivity and accuracy of the method based on the 99.9%-quantile (M1), given the high values that the quantile assumes leading to the seropositive population being fully absorbed by it. Thus, for comparison purposes, the application of each methods to the Skew-Normal distribution was considered, again verifying that the method based on the minimum densities of the mixture model produces the highest sensitivity ($cutoff = 2.46, sens = 90.19\%$). However, for this antigen, the method with the highest accuracy is based on the conditional probability (set at 90%) of classifying an individual as being seropositive (ACC=93.44%) (Figure 5.3 and table 5.2).

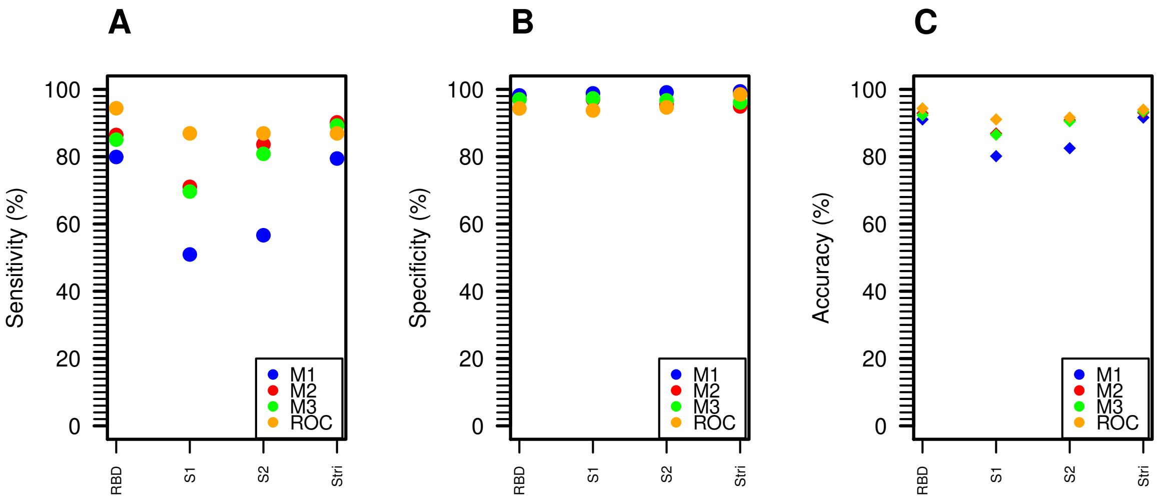


Figure 5.3: Performance of each method to estimate the cutoff value. **A.** Sensitivity values for each method. **B.** Specificity values for each method. **C.** Accuracy values for each method.

Table 5.2: Bayesian Information Criterion (BIC) values, cutoff value estimates, sensitivity, specificity and accuracy for each method by antigen according to the best model. *C* denotes the cutoff point estimate.

			M1				**M2**				**M3**			
Antigen	**Distribution**	**BIC**	**C**	**sens (%)**	**spec (%)**	**ACC (%)**	**C**	**sens (%)**	**spec (%)**	**ACC (%)**	**C**	**sens (%)**	**spec (%)**	**ACC (%)**
RBD	Normal	953.00	2.65	84.11	97.61	92.35	2.33	90.18	95.52	93.44	2.37	88.79	95.82	93.08
	Skew-Normal	**852.25**	2.83	79.91	98.21	91.07	2.49	86.45	97.01	92.89	2.56	85.05	97.01	92.35
	Student t	959.60	4.16	0.09	100	61.38	2.34	90.18	95.52	93.44	2.38	88.79	96.42	93.44
	Skew-t	854.78	4.80	NA	100.00	NA	2.60	84.58	97.61	92.53	2.89	78.97	98.51	90.89
S1	Normal	561.81	2.43	63.08	97.91	84.34	2.13	81.31	95.52	89.98	2.12	82.71	95.52	90.53
	Skew-Normal	**561.63**	2.58	50.93	98.81	80.15	2.27	71.03	97.01	86.89	2.30	69.63	97.31	86.52
	Student t	568.98	3.15	15.42	100.00	67.03	2.14	80.37	95.52	89.62	2.12	82.71	95.52	90.53
	Skew-t	568.27	3.27	10.28	100.00	65.03	2.27	71.03	97.01	86.89	2.31	69.16	97.31	86.34
S2	Normal	778.76	2.66	72.89	98.51	88.52	2.23	89.72	92.23	91.26	2.24	88.32	92.84	91.07
	Skew-Normal	**775.29**	2.86	56.54	99.10	82.51	2.39	83.64	95.52	90.89	2.49	80.84	96.72	90.53
	Student t	785.73	3.51	9.35	100.00	64.66	2.24	88.32	92.84	91.07	2.25	87.38	93.13	90.89
	Skew-t	781.75	3.72	4.21	100.00	62.66	2.39	83.64	95.52	90.89	2.50	80.37	97.01	90.53
$S^t ri$	Normal	1010.18	2.75	87.85	97.91	93.98	2.37	91.12	94.63	93.26	2.47	90.17	94.93	93.08
	Skew-Normal	916.15	2.98	79.44	99.40	91.62	2.46	90.19	94.93	93.08	2.58	89.25	96.12	93.44
	Student t	1016.84	4.34	NA	100.00	NA	2.39	90.65	94.63	93.08	2.48	89.72	95.22	93.08
	Skew-t	**915.82**	5.49	NA	100.00	NA	2.53	89.25	96.12	93.44	2.84	85.51	98.51	93.44

In order to evaluate the quality of methods M1, M2 and M3, the optimal cutoff point was estimated using the ROC curve. This is possible since the true infection status of the individuals is known. It is interesting to see that in terms of specificity and accuracy the results are similar to the method that is traditionally used (ROC curve). However, it is possible to observe a poor performance of the M1 method with regard to its sensitivity. (Figure 5.3, table 5.2 and table 5.3).

Table 5.3: Cutoff point estimates, sensitivity, specificity, accuracy and area under the curve (AUC) for the empirical ROC curve method

Antigen	**Cutoff**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**AUC (CI 95%)**
RBD	2.15	94.39	94.33	94.35	98.50 (97.80, 99.30)
S1	2.07	86.92	93.73	91.07	96.10 (94.60, 97.60)
S2	2.33	86.92	94.63	91.62	94.90 (92.80, 97.00)
S^{tri}	2.81	86.92	98.51	93.98	98.30 (97.40, 99.20)

5.4.4 Simulation results

To conduct the simulation study, two scenarios were considered: the first consists of the scenario where the model that best fits the data is a Skew-Normal distribution, and the second where the model that best fits the data is a Skew-t distribution. For this purpose, the results for the RBD antigen (Skew-Normal distribution) and the S^{tri} antigen (Skew-t distribution) were selected. For each scenario the sample size was varied, as well as the proportion of seronegative individuals in the population. The results are shown in table 5.4 and table 5.5.

In general it is found that as the sample size increases, both the relative error and the root mean square error tend to decrease. It is also found that for small samples and extreme π_1 values ($\pi_1 = 0.3$ or $\pi_1 = 0.9$), the models tend to have some difficulty in identifying a seronegative and seropositive population. This is a result that alerts to the existence of possible false positives and false negatives in the case of small samples.

In situations where there is an ongoing vaccination plan and therefore the majority of the population is seropositive (e.g. $\pi_1 = 0.9$) it is important to know if it is possible to identify seronegative individuals in this population given the time of immunization. If the timing of immunization is short, it is important to identify these individuals early in order to take action and prevent a further increase in infections.

Table 5.4: Relative bias and Mean Square Error (MSE) of the 99.9%-quantile method (M1); minimum of mixture densities method (M2) and conditional probability method (M3) for the RBD antigen. opt_{M1} denotes the theoretical cutoff point for the 99.9%-quantile; opt_{M2} denotes the theoretical cutoff point for the minimum of the density mixture method; opt_{M3} denotes the theoretical cutoff point for conditional probability method. π_1 denotes the weight of the seronegative population; c_{M1} denotes the cutoff estimated by M1 method after N=1000 simulations; c_{M2} denotes the cutoff estimated by M2 method after N=1000 simulations; c_{M3} denotes the cutoff estimated by M3 method after N=1000 simulations.

Normal distribution; $opt_{M1} = 2.65$**;** $opt_{M2} = 2.33$**;** $opt_{M3} = 2.37$

Sample size	π_1	c_{M1}	c_{M2}	c_{M3}	**Relative bias** c_{M1} **(%)**	**MSE** (c_{M1})	**Relative bias** c_{M2} **(%)**	**MSE** (c_{M2})	**Relative bias** c_{M3} **(%)**	**MSE** (c_{M3})	**% Two comp. retained**
100	0.3	5.67	2.34	2.46	113.9314	0.0914	0.6185	0.0001	3.8938	0.0004	93.1
	0.6	5.17	2.51	2.58	95.2155	0.0641	7.6017	0.0004	8.8272	0.0006	100.0
	0.9	3.68	2.72	2.75	39.0584	0.0114	16.6388	0.0017	15.9472	0.0016	99.9
500	0.3	5.68	2.35	2.48	114.3252	0.0183	0.9958	0.000007	4.5243	0.00003	100.0
	0.6	5.19	2.51	2.59	95.9969	0.0129	7.7858	0.00007	9.2046	0.0001	100.0
	0.9	3.72	2.70	2.73	40.3432	0.0023	16.1593	0.0002	15.3683	0.0002	100.0
1000	0.3	5.69	2.35	2.47	114.6181	0.0092	0.8462	0.000002	4.3660	0.00001	100.0
	0.6	5.19	2.51	2.59	95.9990	0.0064	7.7956	0.00003	9.1638	0.00005	100.0
	0.9	3.73	2.70	2.73	40.6998	0.0011	16.0422	0.0001	15.3499	0.0001	100.0
Skew-Normal distribution; $opt_{M1} = 2.83$**;** $opt_{M2} = 2.49$**;** $opt_{M3} = 2.56$											
100	0.3	4.63	2.50	2.74	63.3181	0.0345	0.5088	0.0001	7.0728	0.0010	96.9
	0.6	5.73	2.75	2.74	102.2463	0.0846	10.3808	0.0010	6.9046	0.0006	99.5
	0.9	3.94	3.04	2.89	39.2453	0.0131	22.0631	0.0043	12.7332	0.0016	94.7
500	0.3	4.44	2.48	2.68	56.7727	0.0053	-0.5071	0.000009	4.6945	0.00007	100.0
	0.6	5.76	2.74	2.73	103.2662	0.0171	10.2602	0.0001	6.3299	0.00006	100.0
	0.9	3.94	3.14	2.89	39.2537	0.0025	26.1894	0.0011	13.0415	0.0002	100.0
1000	0.3	4.39	2.48	2.68	55.3506	0.0024	-0.5008	0.000003	4.6036	0.00003	100.0
	0.6	5.76	2.75	2.72	103.3116	0.0085	10.4370	0.00009	6.0958	0.00003	100.0
	0.9	3.94	3.16	2.89	38.9617	0.0012	27.1796	0.0005	12.9545	0.0001	100.0
Student t distribution; $opt_{M1} = 4.16$**;** $opt_{M2} = 2.34$**;** $opt_{M3} = 2.38$											
100	0.3	5.85	2.15	2.23	40.6111	0.0296	-7.8239	0.0004	-6.3481	0.0004	99.9
	0.6	15.22	2.31	2.45	265.6374	57.9703	-1.2775	0.0001	2.7550	0.0003	100.0
	0.9	33.74	2.60	2.86	710.4484	13.8984	11.5284	0.0013	20.084	0.0035	84.3
500	0.3	5.85	2.16	2.25	40.4626	0.0057	-7.3517	0.00006	-5.4927	0.00004	100.0
	0.6	5.39	2.31	2.47	29.3408	0.0030	-0.9376	0.00004	3.7417	0.00006	100.0
	0.9	25.38	2.59	2.92	509.6060	1.0029	11.2388	0.0001	22.5757	0.0006	100.0
1000	0.3	5.85	2.16	2.25	40.5011	0.0028	-7.4162	0.00003	-5.5194	0.00002	100.0
	0.6	5.37	2.31	2.47	28.8965	0.0014	-1.0348	0.000001	3.7648	0.00001	100.0
	0.9	24.52	2.59	2.93	489.0068	0.4401	11.2467	0.00007	23.0901	0.0003	100.0
Skew-t distribution; $opt_{M1} = 4.80$**;** $opt_{M2} = 2.60$**;** $opt_{M3} = 2.89$											
100	0.3	4.59	2.35	2.53	-4.5079	0.0031	-9.9120	0.0009	-12.7692	0.0021	99.3
	0.6	8.14	2.47	2.68	69.4317	0.4065	-5.1031	0.0004	-7.5949	0.0012	100.0
	0.9	NA	2.78	3.06	NA	NA	6.6832	0.0008	5.5875	0.0014	40.6
500	0.3	4.43	2.31	2.49	-7.8118	0.0004	-11.0747	0.0001	-14.2211	0.0003	100.0
	0.6	6.81	2.48	2.71	41.8089	0.0114	-4.6537	0.00004	-6.5774	0.0001	100.0
	0.9	22.93	2.83	3.22	377.5419	0.7006	8.8679	0.0001	11.1838	0.0005	98.7
1000	0.3	4.42	2.32	2.50	-7.9676	0.0002	-10.8456	0.00008	-13.6653	0.0001	100.0
	0.6	6.82	2.49	2.73	42.1175	0.0048	-4.2254	0.00001	-5.8102	0.00004	100.0
	0.9	22.81	2.85	3.28	374.9435	0.3352	9.6048	0.00007	13.0902	0.0002	99.5

Table 5.5: Relative bias and Mean Square Error (MSE) of the 99.9%-quantile method (M1); minimum of mixture densities method (M2) and conditional probability method (M3) for the S^{tri} antigen. opt_{M1} denotes the theoretical cutoff point for the 99.9%-quantile; opt_{M2} denotes the theoretical cutoff point for the minimum of the density mixture method; opt_{M3} denotes the theoretical cutoff point for conditional probability method. π_1 denotes the weight of the seronegative population; c_{M1} denotes the cutoff estimated by M1 method after N=1000 simulations; c_{M2} denotes the cutoff estimated by M2 method after N=1000 simulations; c_{M3} denotes the cutoff estimated by M3 method after N=1000 simulations.

Normal distribution; $opt_{M1} = 2.75$; $opt_{M2} = 2.37$; $opt_{M3} = 2.47$											
Sample size	π_1	c_{M1}	c_{M2}	c_{M3}	**Relative bias** c_{M1} **(%)**	**MSE** (c_{M1})	**Relative bias** c_{M2} **(%)**	**MSE** (c_{M2})	**Relative bias** c_{M3} **(%)**	**MSE** (c_{M3})	**% Two comp. retained**
100	0.3	5.52	2.33	2.51	100.7315	0.0769	-2.4407	0.0002	1.6126	0.0004	95.2
	0.6	5.09	2.51	2.64	84.9707	0.0549	5.3154	0.0002	6.8551	0.0004	100.0
	0.9	3.69	2.75	2.81	34.1160	0.0094	15.1483	0.0015	13.6833	0.0014	99.7
500	0.3	5.53	2.33	2.53	101.0684	0.0154	-2.1549	0.00001	2.3597	0.00002	100.0
	0.6	5.09	2.51	2.64	85.4509	0.0110	5.2410	0.00003	6.9538	0.00006	100.0
	0.9	3.72	2.75	2.81	35.1543	0.0018	15.0652	0.0002	13.6792	0.0002	100.0
1000	0.3	5.53	2.33	2.52	101.0995	0.0077	-2.2169	0.000004	2.2269	0.000006	100.0
	0.6	5.10	2.51	2.64	85.6346	0.0055	5.3327	0.00002	7.0909	0.00003	100.0
	0.9	3.72	2.74	2.81	35.4638	0.0009	14.9128	0.0001	13.6695	0.0001	100.0
Skew-Normal distribution; $opt_{M1} = 2.98$; $opt_{M2} = 2.46$; $opt_{M3} = 2.58$											
100	0.3	4.19	2.38	2.69	40.7854	0.0155	-3.3167	0.0002	4.3221	0.0007	95.8
	0.6	5.69	2.69	2.78	91.0185	0.0741	9.2839	0.0007	7.6265	0.0007	98.5
	0.9	4.03	3.03	2.91	35.1676	0.0116	23.2536	0.0046	12.7065	0.0016	77.8
500	0.3	4.13	2.37	2.73	38.4556	0.0026	-3.7077	0.00003	5.4574	0.00008	100.0
	0.6	5.73	2.69	2.78	92.1103	0.0151	9.2010	0.0001	7.6095	0.00009	100.0
	0.9	3.99	3.14	2.96	34.1244	0.0020	27.5261	0.0011	14.4625	0.0002	100.0
1000	0.3	4.12	2.37	2.72	38.2737	0.0013	-3.7890	0.00001	5.2385	0.00003	100.0
	0.6	5.73	2.69	2.78	92.1627	0.0075	9.1482	0.00006	7.6591	0.00004	100.0
	0.9	3.99	3.19	2.96	33.8530	0.0010	29.4491	0.0006	14.6881	0.0001	100.0
Student t distribution; $opt_{M1} = 4.34$; $opt_{M2} = 2.39$; $opt_{M3} = 2.48$											
100	0.3	6.12	2.27	2.47	40.9502	0.0499	-4.9286	0.0002	-0.1643	0.0002	100.0
	0.6	7.47	2.42	2.66	71.9891	5.5903	1.2376	0.00009	7.5574	0.0006	100.0
	0.9	25.93	2.68	2.97	496.9340	7.0674	12.0814	0.0011	19.9530	0.0031	87.9
500	0.3	5.96	2.28	2.49	37.1156	0.0052	-4.6806	0.00003	0.8053	0.00008	100.0
	0.6	5.49	2.43	2.69	26.4182	0.0027	1.5251	0.000006	8.8431	0.0001	100.0
	0.9	21.13	2.67	3.04	386.3460	0.6201m	11.8413	0.0001	22.7049	0.0006	100.0
1000	0.3	5.95	2.28	2.49	37.0179	0.0026	-4.7472	0.00001	0.7876	0.000002	100.0
	0.6	5.49	2.43	2.70	26.3405	0.0013	1.6033	0.000002	9.1507	0.000005	100.0
	0.9	20.91	2.67	3.05	381.2481	0.2865	11.9272	0.000008	23.1211	0.0003	100.0
Skew-t distribution; $opt_{M1} = 5.49$; $opt_{M2} = 2.53$; $opt_{M3} = 2.84$											
100	0.3	4.20	2.33	2.59	-23.4201	0.0186	-7.7389	0.0005	-8.9259	0.0012	99.9
	0.6	6.97	2.49	2.79	26.9224	0.1472	-1.2113	0.0001	-2.0286	0.0006	100.0
	0.9	NA	2.76	3.05	NA	NA	9.2127	0.0010	7.1952	0.0013	43.7
500	0.3	4.19	2.31	2.59	-23.6233	0.0035	-8.7076	0.0001	-8.8653	0.0001	100.0
	0.6	7.09	2.51	2.86	29.2639	0.0094	-0.5573	0.000009	0.6079	0.00003	100.0
	0.9	19.83	2.83	NA	261.1660	0.4367	12.0420	0.0002	NA	NA	97.2
1000	0.3	4.19	2.31	2.59	-23.6742	0.0017	-8.8327	0.000005	-8.9670	0.000008	100.0
	0.6	7.27	2.52	2.88	32.4641	0.0043	-0.3427	0.0000003	1.2370	0.000001	100.0
	0.9	19.8	2.85	3.28	260.0015	0.2104	12.5592	0.0001	15.3388	0.0003	97.8

5.5 Conclusions

The purpose of this study was to use a flexible class of mixture models to antibody data against the SARS-CoV-2 virus. In particular, we used a class of models that allows capturing the skewness present in this type of data, namely the Skew-Normal and Skew-t distributions.

It has become clear that diagnostic tests play a key role in the early identification of infected individuals, allowing us to act to control a pandemic by isolating and tracing the contacts of an infected person. Diagnostic tests can classify an individual as seronegative or seropositive by defining a cutoff point that can take on different values depending on the technique used by the manufacturer to develop the test. Most of the time, this cutoff point is relaxed and is calculated using the 3σ-rule, which assumes that the underlying distribution of the data is Normal. However, as we have seen in our application, this assumption cannot always be made, making this method unfeasible.

Note that this study has the advantage that the true cases and controls of the infection are known, allowing us to compare different methods for obtaining the cutoff point that allows classifying an individual as seropositive.

In Dias Domingues et al., 2021, three methods for obtaining the cutoff point had been presented that could not yet be validated because the true infection status of the individuals was not known. In this sense, we proceeded to use these methods in this study, and it was verified that the three methods under analysis present high accuracy, compared to methods used in literature, namely through the empirical ROC curve. However, the proposed methods proved to be more specific than sensitive. Note that the performance of the method based on the 99.9% probability quantile may be overestimated, especially when the fitted distribution corresponds to a heavy-tailed distribution (such as the Skew-t distribution). This is because the calculation of this quantile involves only and exclusively the population of seropositive individuals, so that if the distribution is too skewed to the right, then the seropositive population is totally absorbed by this quantile.

When a new virus is present in the population, there is a natural tendency for the proportion of susceptible individuals to be much higher than the seropositive individuals. This is the phase in which early identification of the infected people is essential for pandemic control, although total control of the spread of the virus only occurs when there is vaccination or eradication of the virus. In this sense, with the simulation study developed in this work, we intend to analyze the pandemic evolution scenarios and understand the behavior of different methods for determining the cutoff point. It was found that as the sample size increases, there is a tendency for the relative error and the mean square error of the cutoff point estimates in skewed distributions to decrease, while this tendency is not linear in the case of the usual symmetric distributions (Normal and Student t). This fact may be due to the fact that symmetrical distributions are not the most appropriate for these types of data, or even that the proposed methods should not be used when considering the usual distributions.

As we expected, for small sample sizes and for large imbalances in the serological populations, the proposed models were found to have problems in identifying two components. Note that in the case of skewed distributions, it will be natural that if the weight of the seronegative population is very high, then observations relating to the seropositive population are considered false negatives and false positives otherwise.

A limitation of this study is the fact that the adjustment of the different mixture models was performed using the same distribution for the two components (through the package mixsmsn). If the components of the mixture model were distinct, this would have a direct implication on the estimated cutoff points. However, the package that would allow this analysis is now discontinued.

In conclusion, we recommend the use of mixture models based on distributions of the SMSN family for the analysis of serological data given the flexibility of these models, as well as the use of the proposed methods for determining cutoff points as an alternative to the method based on the 3σ rule.

Chapter 6

General conclusions

This work was intended to provide new tools for the analysis of serological data. More specifically, it was shown that serological data can be well described by mixture of distributions based on the SMSN family, namely through the Skew-Normal and Skew-t distributions.

In addition to using these distributions to model serological data, we also wanted to work and deepen an issue that was hitherto little discussed in this type of data and that is related to the cutoff point to be used to define the serological status of an individual. In the case where the true infection status is known, several statistical methods are available to estimate the optimal cutoff point, namely through ROC curve methods. However, when the true infection status of an individual is unknown the majority of seropeidemiological studies uses a standard rule based on the normality of the population which consists in defining a control population (seronegative) and calculating the mean plus 3 standard deviations. In this sense, we presented two applications where this theme is discussed: one in which the serological status of the individual is unknown and another in which the serological status is known. The second scenario was used to validate the proposed methods with those already used in literature, namely the ROC curve.

The methods proposed in this study showed the same capacity as the reference method already used in the literature, thus proving to be viable alternatives for estimating the cutoff point when the true serological status is unknown.

In addition to evaluating the impact that the choice of cutoff point has on the seroprevalence of the population, it was also possible through a simulation study to understand what are the desirable conditions for the proposed models to have the ability to identify two serological populations (seronegative and seropositive). It was found that the two serological populations were well detected for large samples and in balanced proportions.

Several points of this project may still be subject to analysis in future work. More specifically, for application in chronic fatigue data, it is possible to perform a multivariate analysis considering several antibodies that are related to each other and understanding which factors influence the seropositivity of an individual.

With regard to the package used for adjusting the mixture models, it was found that it was not possible to perform the adjustment for the case where the mixture model components have distinct distributions (for example, a mixture model based on Skew Normal and Skew-t distributions). Thus, in the future, an update to the package `mixsmsn` could be performed to extend its applicability. Furthermore, in this work we only consider the Skew-Normal and Skew-t distributions of the SMSN family. However, this can be generalized to the use of the Skew-slash and Skew-contaminated distributions that are also part of this family of distributions.

Regarding the application to the study of SARS-CoV-2 infections, more specifically to the simulation study carried out, it is still possible to complement the analysis by calculating confidence intervals via non-parametric bootstrap and calculating the coverage of each method to estimate the cutoff point.

Bibliography

Ariza, M.E. (2020). "Commentary: Antibodies to Human Herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Patients". In: *Frontiers in Immunology* 11, p. 1945. DOI: https://doi.org/10.3389/fimmu.2020.01400.

Azadeh, Mitra et al. (2017). "Calibration Curves in Quantitative Ligand Binding Assays: Recommendations and Best Practices for Preparation, Design, and Editing of Calibration Curves". In: *The AAPS Journal* 20, pp. 1–16. DOI: https://doi.org/10.1208/s12248-017-0159-4.

Azzalini, A. (1985). "A Class of distributions which includes the normal Ones". In: *Scandinavian Journal of Statistics* 12, pp. 171–178.

— (2014). "The skew-normal and related families". In: *Cambridge University Press*.

— (2020). "The Skew-Normal and Related Distributions Such as the Skew-t". In: *R CRAN*. URL: https://cran.r-project.org/web/packages/sn/sn.pdf.

Azzalini, A. and A. Capitanio (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution". In: *Journal of the Royal Statistical Society: Series B* 65.2, pp. 367–389. DOI: https://doi.org/10.1111/1467-9868.00391.

Basso, R.M. et al. (2010). "Robust mixture modelling based on scale mixtures of skew-normal distributions". In: *Computational Statistics and Data Analysis* 54.12, pp. 2926–2941. DOI: https://doi.org/10.1016/j.csda.2009.09.031.

Baughman, A. L. et al. (2006). "Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels". In: *Statistics in Medicine* 25, pp. 2994–3010. DOI: https://doi.org/10.1002/sim.2442.

Blacksell, S. et al. (2016). "Optimal cutoff and accuracy of an IgM enzyme-linked immunosorbent assay for diagnosis of acute scrub typhus in northern Thailand: an alternative reference method to the IgM immunofluorescence assay”. In: *Journal of clinical microbiology* 54.6, pp. 1472–1478. DOI: https://doi.org/10.1128/JCM.02744-15.

Blomberg, J., C.G. Gottfries, et al. (2018). “Infection elicited autoimmunity and Myalgic encephalomyelitis/chronic fatigue syndrome: An explanatory model”. In: *Frontiers in Immunology* 9. DOI: https://doi.org/10.3389/fimmu.2018.00229.

Blomberg, J., M. Rizwan, et al. (2019). “Antibodies to Human Herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Patients”. In: *Frontiers in Immunology* 10, p. 1946. DOI: https://doi.org/10.3389/fimmu.2019.01946.

Bouman, J. A., S. Bonhoeffer, and R. R. Regoes (2020). “Estimating seroprevalence with imperfect serological tests: exploiting cutoff-free approaches”. In: *bioRxiv*. DOI: https://doi.org/10.1101/2020.04.29.068999.

Braun, D. K., G. Dominguez, and P. E. Pellett (1997). “Human herpesvirus 6”. In: *Clinical Microbiology Reviews* 10.3, pp. 521–567. DOI: https://doi.org/10.1128/CMR.10.3.521.

Brent, R.P. (1973). “Algorithms for Minimization Without Derivatives”. In: *Prentice-Hall, Englewood Cliffs, New Jersey*, pp. 73–76.

Carruthers, B. M. et al. (2003). “Myalgic Encephalomyelitis/Chronic Fatigue Syndrome”. In: *Journal of Chronic Fatigue Syndrome* 11.1, pp. 7–115. DOI: https://doi.org/10.1300/J092v11n01_02.

Chis Ster, I. (2012). “Inference for serological surveys investigating past exposures to infections resulting in long-lasting immunity – an approach using finite mixture models with concomitant information”. In: *Journal of Applied Statistics* 39.11, pp. 2523–2542. DOI: https://doi.org/10.1080/02664763.2012.722608.

Chu, L. et al. (2019). “Onset patterns and course of myalgic encephalomyelitis/chronic fatigue syndrome”. In: *Frontiers in Pediatrics* 7.12. DOI: https://doi.org/10.3389/fped.2019.00012.

Chung, Y. et al. (2011). “Follicular regulatory T cells expressing Foxp3 and Bcl-6 suppress germinal center reactions”. In: *Nature Medicine* 17.8, pp. 983–988. DOI: https://doi.org/10.1038/nm.2426.

Cliff, J.M. et al. (2019). “Cellular Immune Function in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)”. In: *Frontiers in immunology* 10, p. 796. DOI: https://doi.org/10.3389/fimmu.2019.00796.

Cook, J. et al. (2011). “Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, Equatorial Guinea”. In: *Plos One* 6.9, e25137. DOI: `https://doi.org/10.1371/journal.pone.0025137`.

Dempster, A.P. and D.B. Rubin (1977). “Maximum likelihood estimation from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society* 39.1, pp. 1–38.

Dias Domingues, T., H. Mouriño, and N. Sepúlveda (2020). “A statistical analysis of serological data from the UK myalgic encephalomyelitis/chronic fatigue syndrome biobank”. In: *AIP Conference Proceedings* 2293.1. DOI: `https://doi.org/10.1063/5.0026633`.

— (2021). “Analysis of antibody data using Finite Mixture Models based on Scale Mixtures of Skew-Normal distributions”. In: *medRxiv*. DOI: `https://doi.org/10.1101/2021.03.08.21252807`.

Efron, B. and R.J. Tibshirani (1993). “An Introduction to the Bootstrap”. In: *1st ed. Dordrecht: Springer Science+Business Media*. DOI: `https://doi.org/10.3389/fimmu.2020.01400`.

Estévez-López, F. et al. (2020). “Systematic Review of the Epidemiological Burden of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Across Europe: Current Evidence and EUROMENE Research Recommendations for Epidemiology”. In: *Journal of Clinical Medicine* 9.1557. DOI: `https://doi.org/10.3390/jcm9051557`.

Feng, Z.D and C.E. McCullogh (1996). “Using Bootstrap Likelihood Ratios in Finite Mixture Models”. In: *Journal of the Royal Statistical Society* 58.3, pp. 609–617. DOI: `https://doi.org/10.1111/j.2517-6161.1996.tb02104.x`.

Ferreira, C.S., Bolfarine H., and Lachos V.H. (2011). “Skew scale mixtures of normal distributions: Properties and estimation”. In: *Statistical Methodology* 8.2, pp. 154–171. DOI: `https://doi.org/10.1016/j.stamet.2010.09.001`.

Figueiredo, F. and M.I. Gomes (2013). “The skew-normal distribution in SPC.” In: *REVSTAT–Statistical Journal* 11.1, pp. 83–104.

Freimer, M. et al. (1988). “A Study of the Generalised Tukey Lambda Family”. In: *Communications in Statistics – Theory and Methods* 17, pp. 3547–3567. DOI: `https://doi.org/10.1080/03610928808829820`.

Fukuda, K. et al. (1994). “The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group”. In: *Annals of*

Internal Medicine 121.12, pp. 953–959. DOI: https://doi.org/10.7326/0003-4819-121-12-199412150-00009.

Gay, N.J. (1996). "Analysis of serological surveys using mixture models: application to a survey of parvovirus B19". In: *Statistics in Medicine* 15, pp. 1567–1573. DOI: https://doi.org/10.1002/(SICI)1097-0258(19960730)15:14<1567::AID-SIM289>3.0.CO;2-G.

Habibzadeh, F., P. Habibzadeh, and M. Yadollahie (2016). "On determining the most appropriate test cut-off value: the case of tests with continuous results". In: *Biochemia medica* 26.3, pp. 297–307. DOI: https://doi.org/10.11613/BM.2016.034.

Hanson, M.R. and A. Germain (2020). "Letter to the editor of metabolites". In: *Metabolites*. DOI: https://doi.org/10.3390/metabo10050216.

Hasibi, M. et al. (2013). "Determination of the accuracy and optimal cut-off point for ELISA test in diagnosis of human brucellosis in Iran". In: *Acta Medica Iranica*, pp. 687–692.

Hatziagelaki, E. et al. (2018). "Myalgic encephalomyelitis/chronic fatigue syndrome—metabolic disease or disturbed homeostasis due to focal inflammation in the hypothalamus?" In: *Journal of Pharmacology and Experimental Therapeutics* 367, pp. 155–167. DOI: https://doi.org/10.1124/jpet.118.250845.

Helb, D. A. et al. (2015). "Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.32, E4438–E4447. DOI: https://doi.org/10.1073/pnas.1501705112.

Hipólito, Ana (2017). "Deteção e quantificação de leite pelo método ELISA no Laboratório SGS". In: *Dissertação para obtenção do grau de mestre em Tecnologia e Segurança Alimentar, Universidade Nova de Lisboa.*

Hoogen, L. L. van den et al. (2020). "Quality control of multiplex antibody detection in samples from large-scale surveys: the example of malaria in Haiti". In: *Scientific Reports* 10.1, p. 1135. DOI: https://doi.org/10.1038/s41598-020-57876-0.

Hsiang, M. S. et al. (2012). "Surveillance for malaria elimination in Swaziland: a national cross-sectional study using pooled PCR and serology". In: *Plos One* 7.1, e29550. DOI: https://doi.org/10.1371/journal.pone.0029550.

Jackman, S. (2009). "Bayesian Analysis for the Social Sciences". In: *Wiley*, p. 507.

Jason, L.A. et al. (2005). "Chronic fatigue syndrome: The need for subtypes". In: *Neuropsychology Review* 15, pp. 29–58. DOI: https://doi.org/10.1007/s11065-005-3588-2.

Johnston, S.C., D.R. Staines, and S.M. Marshall-Gradisnik (2016). “Epidemiological characteristics of chronic fatigue syndrome/myalgic encephalomyelitis in Australian patients”. In: *Clinical epidemiology* 8, pp. 97–107. DOI: https://doi.org/10.2147/CLEP.S96797.

Kafatos, G. et al. (2016). “Is it appropriate to use fixed assay cut-offs for estimating seroprevalence?” In: *Epidemiology and infection* 144.4, pp. 887–895. DOI: https://doi.org/10.1017/S0950268815001958.

Krmpotić, A. et al. (2019). “Role of antibodies in confining cytomegalovirus after reactivation from latency: three decades’ résumé”. In: *Medical Microbiology and Immunology* 208, pp. 415–429. DOI: https://doi.org/10.1007/s00430-019-00600-1.

Lacerda, E. M., E. W. Bowman, et al. (2017). “The UK ME/CFS Biobank for biomedical research on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and Multiple Sclerosis”. In: *Open Journal of Bioresources* 4, p. 4. DOI: https://doi.org/10.5334/ojb.28.

Lacerda, E. M., K. Mudie, et al. (2018). “The UK ME/CFS Biobank: A Disease-Specific Biobank for Advancing Clinical Research Into Myalgic Encephalomyelitis/Chronic Fatigue Syndrome”. In: *Frontiers in Neurology* 9, p. 1026. DOI: https://doi.org/10.3389/fneur.2018.01026.

Lacerda, E.M. et al. (2019). “A logistic regression analysis of risk factors in ME/CFS pathogenesis”. In: *BMC Neurology* 19. DOI: https://doi.org/10.1186/s12883-019-1468-2.

Lachos Dávila, V. H., C. B. Zeller, and C. R. B. Cabral (2018). “Finite mixture of skewed distributions”. In: *Springer Briefs in Satistics - ABE*.

Lammie, P. J. et al. (2012). “Development of a new platform for neglected tropical disease surveillance”. In: *International Journal for Parasitology* 42.9, pp. 797–800. DOI: https://doi.org/10.1016/j.ijpara.2012.07.002.

Larremore, D. et al. (2020). “Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys”. In: *medRxiv*. DOI: https://doi.org/10.1101/2020.04.15.20067066.

Lee, S.X., K.L. Lee, and G.J. McLachlan (2016). “A simple multithreaded implementation of the EM algorithm for mixture models”. In: *arXiv preprint arXiv:1606.02054*.

Lim, E.J. et al. (2020). “Systematic review and meta-analysis of the prevalence of chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME)”. In: *Journal of Translational Medicine* 18.100. DOI: https://doi.org/10.1186/s12967-020-02269-0.

Lin, T.I., J.C. Lee, and S.Y. Yen (2007a). “Finite mixture modelling using the Skew-Normal distribution”. In: *Statistica Sinica* 17, pp. 909–927.

— (2007b). “Finite mixture modelling using the Skew-Normal distribution”. In: *Statistica Sinica* 17.3, pp. 909–927.

Liu, C. and D.B. Rubin (1994). “The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence”. In: *Biometrika* 81.4, pp. 633–648. DOI: `https://doi.org/10.1093/biomet/81.4.633`.

Loebel, M. et al. (2017). “Serological profiling of the EBV immune response in Chronic Fatigue Syndrome using a peptide microarray”. In: *PloS One* 12.6, e0179124. DOI: `https://doi.org/10.1371/journal.pone.0179124`.

Lukočienė, O. and J. K. Vermunt (2009). “Determining the number of components in mixture models for hierarchical data”. In: *Advances in data analysis, data handling and business intelligence. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg*, pp. 241–249. DOI: `https://doi.org/10.1007/978-3-642-01044-6_22`.

Lunardi, C. et al. (2005). “Induction of endothelial cell damage by hCMV molecular mimicry”. In: *Trends in Immunology* 26, pp. 19–24. DOI: `https://doi.org/10.1016/j.it.2004.10.009`.

Maceiras, A.R. et al. (2017). “T follicular regulatory cells in mice and men”. In: *Immunology* 152, pp. 25–35. DOI: `https://doi.org/10.1111/imm.12774`.

Malato, J. et al. (2021). “Statistical challenges of investigating a disease with a complex diagnosis”. In: *medRxiv*. DOI: `https://doi.org/10.1101/2021.03.19.21253905`.

Maple, P.A.C. et al. (2006). “Application of a noninvasive oral fluid test for detection of treponemal IgG in a predominantly HIV-infected population”. In: *European Journal of Clinical Microbiology and Infectious Diseases* 25.12, pp. 743–749. DOI: `https://doi.org/10.1007/s10096-006-0216-x`.

McLachlan, G. and S. Lee (2013). “EMMIXuskew: An R Package for Fitting Mixtures of Multivariate Skew t Distributions via the EM Algorithm”. In: *Journal of Statistical Software* 55.12, pp. 1–22. DOI: `https://doi.org/10.18637/jss.v055.i12`.

McLachlan, G. and D. Peel (2000). “Finite Mixture Models”. In: *John Wiley & Sons, New York*.

McLachlan, G.J. and T. Krishnan (2008). “The EM algorithm and extensions”. In: *John Wiley & Sons*.

Mehrjou, A., R. Hosseini, and B. N. Araabi (2016). “Improved Bayesian information criterion for mixture model selection”. In: *Pattern Recognition Letters* 69, pp. 22–27. DOI: `https://doi.org/10.1016/j.patrec.2015.10.004`.

Melvin, A. et al. (2019). “Circulating levels of GDF15 in patients with myalgic encephalomyelitis/chronic fatigue syndrome”. In: *Journal of Translational Medicine* 17.409. DOI: `https://doi.org/10.1186/s12967-019-02153-6`.

Meng, X.L. and D.B. Rubin (1993). “Maximum likelihood estimation via the ECM algorithm: A general framework”. In: *Biometrika* 80.2, pp. 267–278. DOI: `https://doi.org/10.1093/biomet/80.2.267`.

Migchelsen, S. J. et al. (2017). “Defining Seropositivity Thresholds for Use in Trachoma Elimination Studies”. In: *PLoS Neglected Tropical Diseases* 11.1, e0005230. DOI: `https://doi.org/10.1371/journal.pntd.0005230`.

Monto, A.S (2002). “Epidemiology of viral respiratory infections”. In: *American Journal of Medicine (Elsevier Inc.)*, pp. 4–12. DOI: `https://doi.org/10.1016/s0002-9343(01)01058-0`.

Moreira da Silva, J. et al. (2020). “Detection and modeling of anti-Leptospira IgG prevalence in cats from Lisbon area and its correlation to retroviral infections, lifestyle, clinical and hematologic changes”. In: *Veterinary and Animal Science* 10, p. 100144. DOI: `https://doi.org/10.1016/j.vas.2020.100144`.

Morris, G. et al. (2019). “Myalgic encephalomyelitis or chronic fatigue syndrome: how could the illness develop?” In: *Metabolic Brain Disease* 34, pp. 385–415. DOI: `https://doi.org/10.1007/s11011-019-0388-6`.

Nacul, L., E.M. Lacerda, et al. (2019). “How have selection bias and disease misclassification undermined the validity of myalgic encephalomyelitis/chronic fatigue syndrome studies?” In: *Journal of Health Psychology* 24, pp. 1765–1769. DOI: `https://doi.org/10.1177/1359105317695803`.

Nacul, L., S. O’Boyle, et al. (2020). “How Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) Progresses: The Natural History of ME/CFS”. In: *Frontiers in Neurology* 11. DOI: `https://doi.org/10.3389/fneur.2020.00826`.

Nacul, L.C. et al. (2011). “Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) in three regions of England: A repeated cross-sectional study in primary care”. In: *BMC Medicine* 9.91. DOI: `https://doi.org/10.1186/1741-7015-9-91`.

Nhat, N. et al. (2017). “Structure of general-population antibody titer distributions to influenza A virus”. In: *Scientific Reports* 7.1, p. 6060. DOI: https://doi.org/10.1038/s41598-017-06177-0.

O’Connell, M.A., B.A. Belanger, and P.D. Haaland (1993). “Calibration and assay development using the four-parameter logistic model”. In: *Chemometrics and Intelligent Laboratory Systems* 20.2, pp. 97–114. DOI: https://doi.org/10.1016/0169-7439(93)80008-6.

Oliveira-Brochado, A. and F. V. Martins (2005). “Assessing the number of components in mixture models: a review”. In: *Universidade do Porto, Faculdade de Economia do Porto* 194.

Parker, R. A., D. D. Erdman, and L. J. Anderson (1990). “Use of mixture models in determining laboratory criterion for identification of seropositive individuals: application to parvovirus B19 serology”. In: *Journal of Virological Methods* 27.2, pp. 135–144. DOI: https://doi.org/10.1016/0166-0934(90)90130-8.

Perkins, N.J. and E.F. Schisterman (2006). “The inconsistency of “optimal” cut-points using two ROC based criteria”. In: *American Journal of Epidemiology* 163.7, pp. 670–675. DOI: https://doi.org/10.1093/aje/kwj063.

Pheby, D. F.H. et al. (2020). “The Development of a Consistent Europe-Wide Approach to Investigating the Economic Impact of Myalgic Encephalomyelitis (ME/CFS): A Report from the European Network on ME/CFS (EUROMENE)”. In: *Healthcare* 8.88. DOI: https://doi.org/10.3390/healthcare8020088.

Pothin, E. et al. (2016). “Estimating malaria transmission intensity from Plasmodium falciparum serological data using antibody density models”. In: *Malaria Journal* 15.79. DOI: https://doi.org/10.1186/s12936-016-1121-0.

Prates, M.O., V.H. Lachos, and C. Cabral (2013). “Fitting finite mixture of scale mixture of skew-normal distributions”. In: *Journal of Statistical Software* 54, pp. 1–20.

Raine, R. et al. (2004). “General practitioners’ perceptions of chronic fatigue syndrome and beliefs about its management, compared with irritable bowel syndrome: Qualitative study”. In: *BMJ* 328, pp. 1354–1356. DOI: https://doi.org/10.1136/bmj.38078.503819.ee.

Ramberg, J. and B. Schmeiser (1974). “An Approximate Method for Generating Asymmetric Random Variables”. In: *Communications of the Association for Computing Machinery* 17, pp. 78–82. DOI: https://doi.org/10.1145/360827.360840.

Rasa, S. et al. (2018). “Chronic viral infections in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)”. In: *Journal of Translational Medicine* 16.1, p. 268. DOI: https://doi.org/10.1186/s12967-018-1644-y.

Ridge, S. E. and A. L. Vizard (1993). “Determination of the optimal cutoff value for a serological assay: an example using the Johne’s Absorbed EIA”. In: *Journal of Clinical Microbiology* 31.5, pp. 1256–1261. DOI: https://doi.org/10.1128/jcm.31.5.1256-1261.1993.

Rivera, M.C. et al. (2019). “Myalgic encephalomyelitis/chronic fatigue syndrome: A comprehensive review”. In: *Diagnostics* 9. DOI: https://doi.org/10.3390/diagnostics9030091.

Rogier, E. et al. (2015). “Multiple comparisons analysis of serological data from an area of low Plasmodium falciparum transmission”. In: *Malaria Journal* 14, p. 436. DOI: https://doi.org/10.1186/s12936-015-0955-1.

Rooney, B. V. et al. (2019). “Herpes Virus Reactivation in Astronauts During Spaceflight and Its Application on Earth”. In: *Frontiers in Microbiology* 10.16. DOI: https://doi.org/10.3389/fmicb.2019.00016.

Rosado, J. et al. (2020). “Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study”. In: *The Lancet Microbe*. DOI: https://doi.org/10.1016/S2666-5247(20)30197-X.

Rota, M. and L. Antolini (2014). “Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers”. In: *Computational Statistics and Data Analysis* 69, pp. 1–14. DOI: https://doi.org/10.1016/j.csda.2013.07.015.

Rota, M. C. et al. (2008). “Measles serological survey in the Italian population: interpretation of results using mixture model”. In: *Vaccine* 26.34, pp. 4403–4409. DOI: https://doi.org/10.1016/j.vaccine.2008.05.094.

Salazar, Alberto et al. (2017). “Allergen-Based Diagnostic: Novel and Old Methodologies with New Approaches”. In: *IntechOpen*. DOI: http://dx.doi.org/10.5772/intechopen.69276.

Saraswati, K. et al. (2019). “The validity of diagnostic cut-offs for commercial and in-house scrub typhus IgM and IgG ELISAs: A review of the evidence”. In: *PLoS Neglected Tropical Diseases* 13.2, e0007158. DOI: https://doi.org/10.1371/journal.pntd.0007158.

Scheibenbogen, C. et al. (2017). “The European ME/CFS Biomarker Landscape project: An initiative of the European network EUROMENE”. In: *Journal of Translational Medicine* 15.165. DOI: `https://doi.org/10.1186/s12967-017-1263-z`.

Sepúlveda, N., J. Carneiro, et al. (2019). “Myalgic Encephalomyelitis/Chronic Fatigue Syndrome as a Hyper-Regulated Immune System Driven by an Interplay Between Regulatory T Cells and Chronic Human Herpesvirus Infections”. In: *Frontiers in Immunology* 10. DOI: `https://doi.org/10.3389/fimmu.2019.02684`.

Sepúlveda, N., G. Stresman, et al. (2015). “Current Mathematical Models for Analyzing Anti-Malarial Antibody Data with an Eye to Malaria Elimination and Eradication”. In: *Journal of Immunology Research* 10, p. 738030. DOI: `https://doi.org/10.1155/2015/738030`.

Shikova, E. et al. (2020). “Cytomegalovirus, Epstein-Barr virus, and human herpesvirus-6 infections in patients with myalgic encephalomyelitis/chronic fatigue syndrome”. In: *Journal of Medical Virology* 92.12, pp. 3682–3688. DOI: `https://doi.org/10.1002/jmv.25744`.

Sotzny, F. et al. (2018). “Myalgic Encephalomyelitis/Chronic Fatigue Syndrome – Evidence for an autoimmune disease”. In: *Autoimmunity Reviews* 17, pp. 601–609. DOI: `https://doi.org/10.1016/j.autrev.2018.01.009`.

Sowa, M. et al. (2017). “Next-Generation Autoantibody Testing by Combination of Screening and Confirmation-the CytoBead® Technology”. In: *Clinical Reviews in Allergy & Immunology* 53.1, pp. 87–104. DOI: `https://doi.org/10.1007/s12016-016-8574-3`.

Stanculescu, D., L. Larsson, and J. Bergquist (2021). “Hypothesis: Mechanisms That Prevent Recovery in Prolonged ICU Patients Also Underlie Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)”. In: *Frontiers in Medicine* 8, p. 628029. DOI: `https://doi.org/10.3389/fmed.2021.628029`.

Steiner, S. et al. (2020). “Autoimmunity-Related Risk Variants in PTPN22 and CTLA4 Are Associated With ME/CFS With Infectious Onset”. In: *Frontiers in Immunology* 11. DOI: `https://doi.org/10.3389/fimmu.2020.00578`.

Stringhini, S. et al. (2020). “Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study”. In: *The Lancet* 396.10247, pp. 313–319. DOI: `https://doi.org/10.1016/S0140-6736(20)31304-0`.

Su, S. (2007). “Fitting Single and Mixture of Generalized Lambda Distributions to Data via Discretized and Maximum Likelihood Methods: GLDEX in R”. In: *Journal of Statistical Software* 21.9, pp. 1–17. DOI: `https://doi.org/10.18637/jss.v021.i09`.

— (2011). “Maximum Log Likelihood Estimation using EM Algorithm and Partition Maximum Log Likelihood Estimation for Mixtures of Generalized Lambda Distributions”. In: *Journal of Modern Applied Statistical Methods* 10.2, p. 17. DOI: https://doi.org/10.22237/jmasm/1320120960.

Szklarski, M. et al. (2021). “Delineating the association between soluble CD26 and autoantibodies against G-protein coupled receptors, immunological and cardiovascular parameters identifies distinct patterns in post-infectious vs. non-infection-triggered Myalgic Encephalomyelitis/ Chronic Fatigue Syndrome”. In: *Frontiers in Immunology* 12, p. 1077. DOI: https://doi.org/10.3389/fimmu.2021.644548.

Tengvall, K. et al. (2019). “Molecular mimicry between Anoctamin 2 and Epstein-Barr virus nuclear antigen 1 associates with multiple sclerosis risk”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.34, pp. 16955–16960. DOI: https://doi.org/10.1073/pnas.1902623116.

Tong, D.D., S. Buxser, and T.J. Vidmar (2007). “Application of a mixture model for determining the cutoff threshold for activity in high-throughput screening”. In: *Computational Statistics and Data Analysis* 51.8, pp. 4002–4012. DOI: https://doi.org/10.1016/j.csda.2006.12.014.

Unal, I. (2017). “Defining an optimal cut-point value in ROC analysis: an alternative approach”. In: *Computational and mathematical methods in medicine*. DOI: https://doi.org/10.1155/2017/3762651.

Underhill, R.A. (2015). “Myalgic encephalomyelitis, chronic fatigue syndrome: An infectious disease”. In: *Medical Hypotheses* 85, pp. 765–773. DOI: https://doi.org/10.1016/j.mehy.2015.10.011.

Valdez, A. R. et al. (2019). “Estimating prevalence, demographics, and costs of ME/CFS using large scale medical claims data and machine learning”. In: *Frontiers in Pediatrics* 6, p. 412. DOI: https://doi.org/10.3389/fped.2018.00412.

Vila Nova, B. et al. (2018). “Evaluation of the humoral immune response induced by vaccination for canine distemper and parvovirus: a pilot study”. In: *BMC Veterinary Research* 16.14, p. 348. DOI: https://doi.org/10.1186/s12917-018-1673-z.

Wang, S. S. et al. (2003). “Seroprevalence of human papillomavirus-16, -18, -31, and -45 in a population-based cohort of 10000 women in Costa Rica”. In: *British Journal of Cancer* 89.7, pp. 1248–1254. DOI: https://doi.org/10.1038/sj.bjc.6601272.

Wirth, K. and C. Scheibenbogen (2020). “A Unifying Hypothesis of the Pathophysiology of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS): Recognitions from the finding of autoantibodies against β2-adrenergic receptors”. In: *Autoimmunity Reviews* 19. DOI: https://doi.org/10.1016/j.autrev.2020.102527.

Wollenberg, I. et al. (2011). “Regulation of the Germinal Center Reaction by Foxp3 + Follicular Regulatory T Cells”. In: *Journal of Immunology* 187, pp. 4553–4560. DOI: https://doi.org/10.4049/jimmunol.1101328.

Wolodzko, T. (2020). “Additional Univariate and Multivariate Distributions”. In: *R CRAN*. URL: https://cran.r-project.org/web/packages/extraDistr/index.html.

Wu, L. et al. (2020). “Optimisation and standardisation of a multiplex immunoassay of diverse Plasmodium falciparum antigens to assess changes in malaria transmission using seroepidemiology”. In: *Wellcome Open Research* 4, p. 26. DOI: https://doi.org/10.12688/wellcomeopenres.14950.2.

Xie, C. H., J. Y. Chang, and Y. J. Liu (2013). “Estimating the number of components in Gaussian mixture models adaptively for medical image”. In: *Optik* 124.23, pp. 6216–6221. DOI: https://doi.org/10.1016/j.ijleo.2013.05.028.

Yalçınkaya, Abdullah, Birdal Şenoğlu, and Ufuk Yolcu (2018). “Maximum likelihood estimation for the parameters of skew normal distribution using genetic algorithm”. In: *Swarm and Evolutionary Computation* 38, pp. 127–138. DOI: https://doi.org/10.1016/j.swevo.2017.07.007.

Yu, Y. and J.L. Harvill (2019). “Bootstrap likelihood ratio test for Weibull mixture models fitted to grouped data”. In: *Communications in Statistics - Theory and Methods* 48.18, pp. 4550–4568. DOI: https://doi.org/10.1080/03610926.2018.1494838.

Appendix A

Supplementary material of Chapter 4

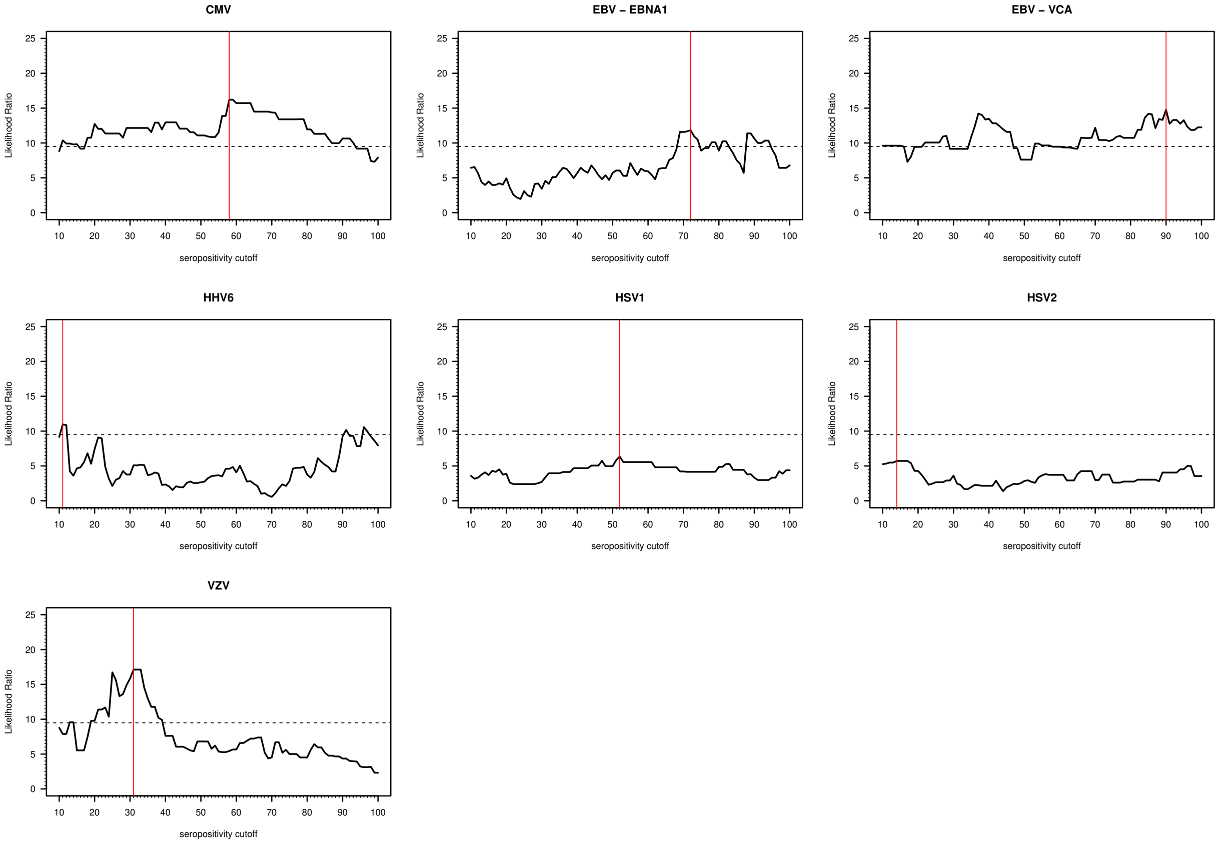


Figure A.1: Seropositivity cutoff versus the likelihood ratio statistic for testing the significance of a group indicator covariate in the logistic models used in unadjusted analysis. The vertical red lines represent the optimal cutoff in which the maximization of the likelihood ratio statistic is achieved. The dashed horizontal lines represent the critical point of the likelihood ratio test. This critical point is defined by the 95% quantile of the χ^2 distribution with 5 degrees of freedom for the likelihood ratio statistic under the null hypothesis.

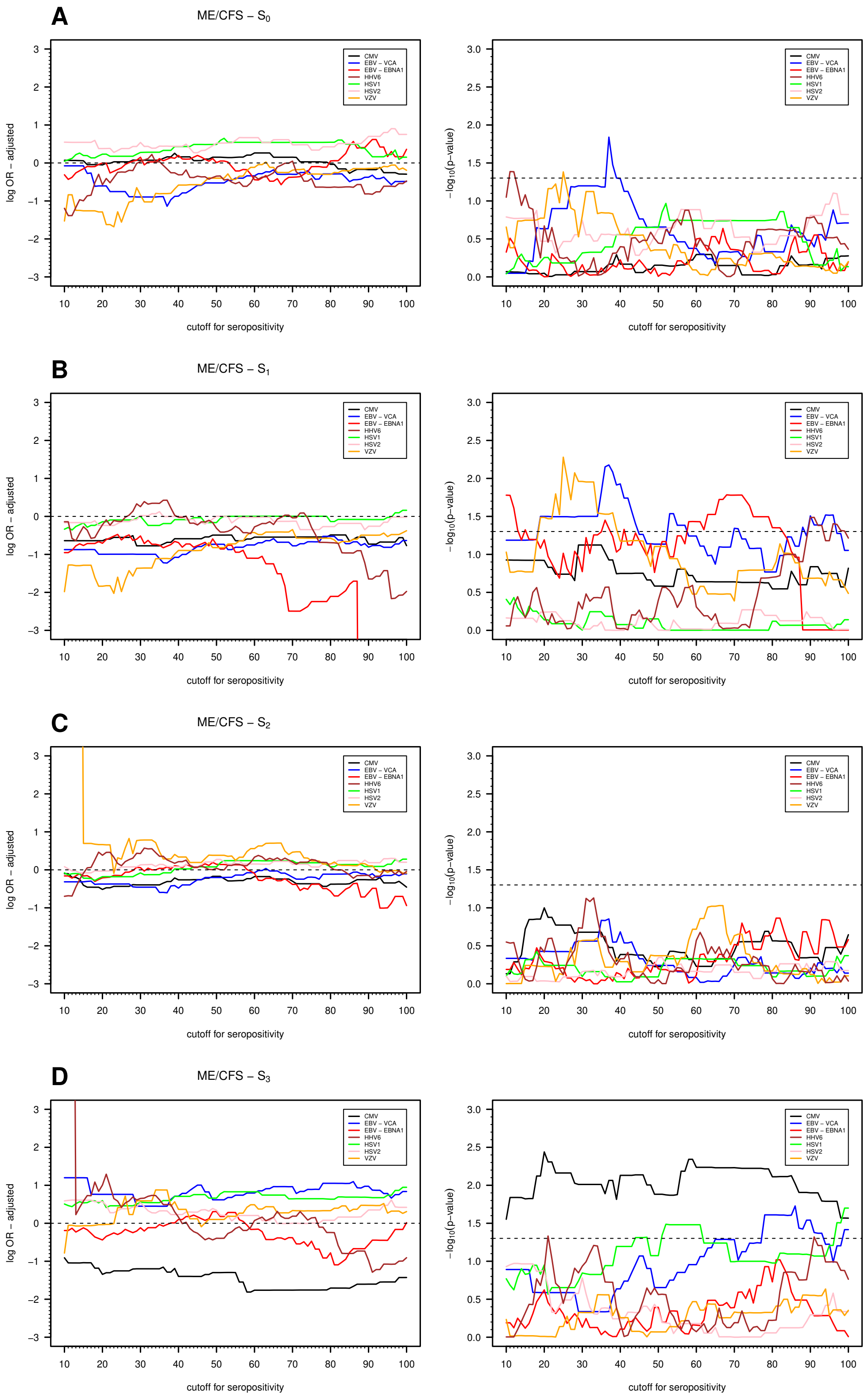


Figure A.2: Age-gender adjusted association analysis of seropositivity to different herpesvirus antigens based on log-OR of the 4 subgroups of patients with ME/CFS in relation to healthy controls. See Figure 4.2 for further information.

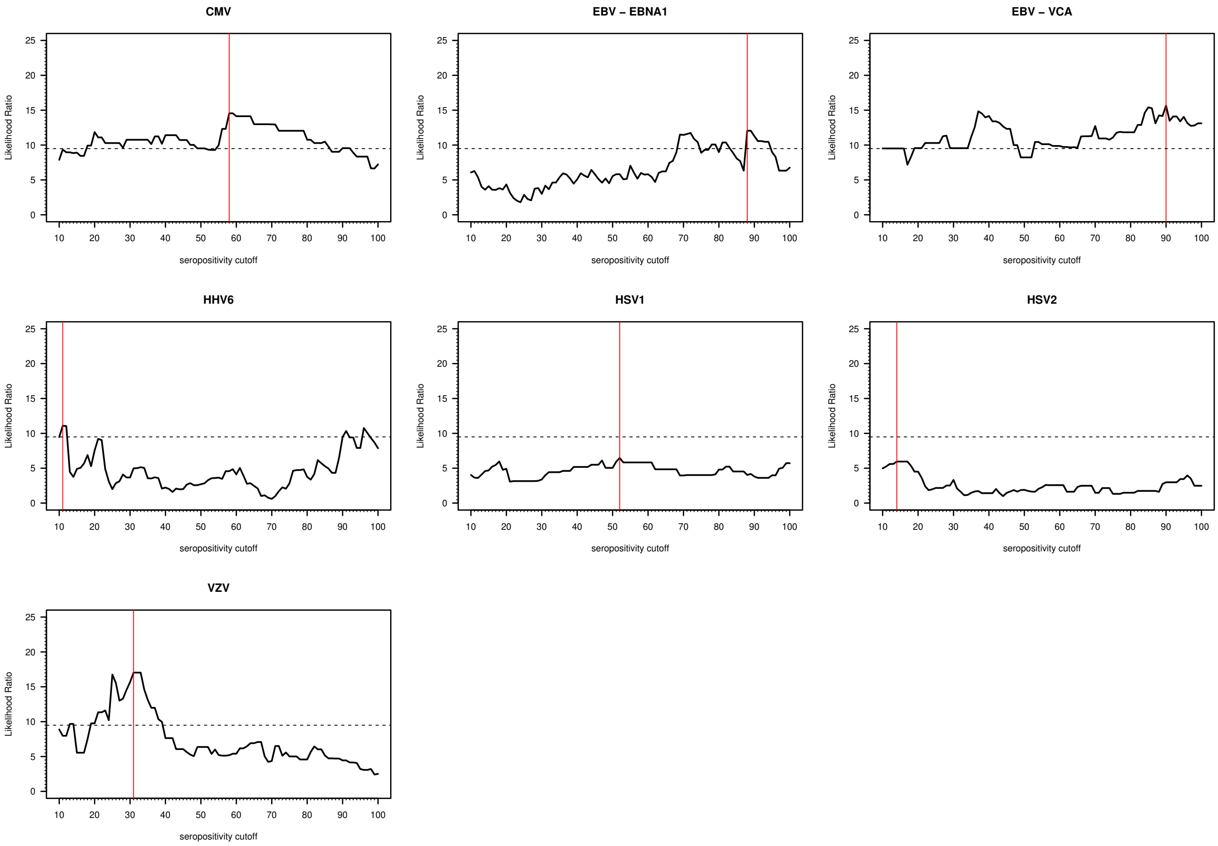


Figure A.3: Seropositivity cutoff versus the likelihood ratio statistic for testing the significance of a group indicator covariate in the logistic models used in the analysis controlling for age and gender. See Figure A.1 for further information.

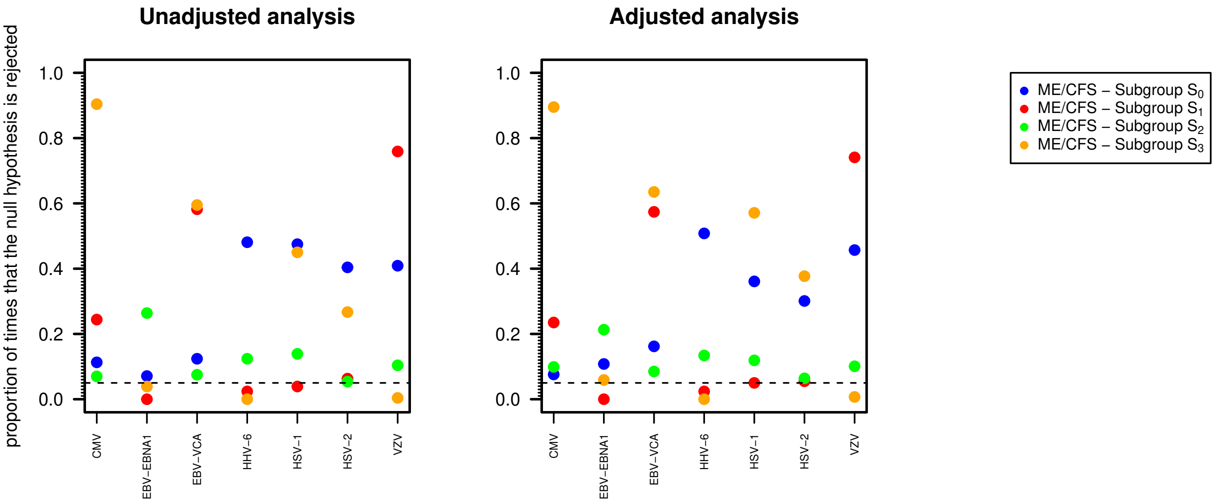


Figure A.4: Statistical power to detect an association between each study group and the seropositivity to a given herpesvirus. Statistical power was estimated by the proportion of times when the log odds ratio between a given study group and healthy controls was deemed statistically significant at 5% significance level in 1000 simulated data sets from logistic models using the optimal cutoffs shown in Figures A.1 and A.3. Horizontal dashed lines represent the 5% significance level.

Table A.1: Optimal seropositivity cutoff for each herpesvirus antibody as shown in Figures A.1 (unadjusted analysis) and A.3 (age and gender adjusted analysis).

Herpesvirus serology	**Unadjusted analysis**	**Adjusted analysis**
CMV	58	58
EBV-EBNA1	72	88
EBV-VCA	90	90
HHV6	11	11
HSV1	52	52
HSV2	14	14
VZV	31	31

Appendix B

Proceedings of the International Conference of Numerical Analysis and Applied Mathematics, ICNAAM 2019

A statistical analysis of serological data from the UK myalgic encephalomyelitis/chronic fatigue syndrome biobank

Cite as: AIP Conference Proceedings **2293**, 420099 (2020); https://doi.org/10.1063/5.0026633
Published Online: 25 November 2020

Tiago Dias Domingues, Helena Mouriño, and Nuno Sepúlveda

View Online Export Citation

A Statistical Analysis of Serological Data from the UK Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Biobank

Tiago Dias Domingues,[1, a)] Helena Mouriño[2] and Nuno Sepúlveda[1, 3]

[1]*CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2]*CMAFcIO, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3]*London School of Hygiene and Tropical Medicine, UK*

[a)]Corresponding author: tmdomingues@fc.ul.pt

Abstract. Myalgic encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a debilitating disease of a unknown cause without a specific biomarker for the respective diagnostic. Here we analyzed anti-viral antibody data from healthy controls, patients with multiple sclerosis, and patients with ME/CFS. Finite mixture models based on flexible skew Normal and skew distributions were first fitted to the data in order to determine the seropositivity to each virus. A logistic regression analysis was then carried out to distinguish between patients with ME/CFS and the remaining study groups.

INTRODUCTION

ME/CFS is a disease with unknown cause characterized by prolonged tiredness and persistence of many non-specific symptoms that limit patient's quality-of-life. The most frequent symptoms are post-exertion malaise, chronic pain, sleep disturbances and frequent viral infections. The prevalence of this condition has been estimated between 0.2% and 0.3%. This disease is believed to be triggered by common viral infections, which elicit in turn a chronic activation of the immune system and its possible exhaustion. Similar disease triggers, immunological deregulation and clinical symptoms are often reported for autoimmune diseases such as multiple sclerosis or rheumatoid arthritis. The similarity between these diseases suggested an autoimmune origin for ME/CFS [1].

Until now there is no disease-specific biomarker that could clearly identify putative patients in the population. Instead patients are identified by symptoms' assessment questionnaires, which may vary from country to country. To help searching a putative biomarker, a ME/CFS biobank was recently created in the United Kingdom [2]. This biobank comprises data from healthy controls (HC), patients with multiple sclerosis (MS), patients with ME/CFS. Here we explore the potential of using antibody data from this biobank in discriminating between patients with ME/CFS and remaining study groups.

DATA & STATISTICAL METHODOLOGY

Antibody data

Antibody data refer to quantitative measurements of optimal density determined by ELISA assays, as described elsewhere [3]. For each individual, there are antibody data related to four common herpes viruses: human Cytomegalovirus (CMV), Herpes Simplex Virus types 1 and 2 (HSV-1 and HSV-2), and Epstein-Barr virus (antibodies against EBNA and VCA proteins).

Estimating seropositivity

When analyzing data of the antibody responses against a specific virus, it is usually assumed the existence of two or more latent, unobservable populations representing different serological states (*e.g.*, seronegative and seropositive). In this scenario, data is typically described by a mixture of two or more probability distributions. In general, let $Z_1,...,Z_n$ be the identical and independent random variables representing the antibody levels for a sample of n individuals, $G_1,...,G_g$ be the partition from a superpopulation G (sample space) and $\pi_1,...,\pi_g$ the probabilities of sampling an individual from each serological population (with the usual restriction of $\sum_{k=1}^{g}\pi_k = 1$ and $0 < \pi_k \leq 1$). The probability

International Conference of Numerical Analysis and Applied Mathematics ICNAAM 2019
AIP Conf. Proc. 2293, 420099-1–420099-4; https://doi.org/10.1063/5.0026633
Published by AIP Publishing. 978-0-7354-4025-8/$30.00

density function (pdf) of the antibody level of *i*-th individual is then given by

$$f(z_i) = \sum_{k=1}^{g} \pi_k f_k(z_i;\theta_k), \quad i = 1,...,n, \tag{1}$$

where $f_k(z_i;\theta_k)$ is the so-called mixing pdf associated with *k*-th serological group and parameterized by a vector θ_k.

The first statistical analysis was then to fit different finite mixture models to the data and select the best one. With this purpose, we first log-transformed the data and then fitted distinct mixture models based the normal, skew normal, student's *t* and skew-*t* distributions. The pdf distribution of skew normal and skew-*t* are respectively described by the following formulas

$$f(z_i;\theta) = \sum_{k=1}^{g} 2\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\big(-\frac{(z_i-\mu_k)^2}{2\sigma_k^2}\big) \times \int_{-\infty}^{\lambda_k \frac{(z_i-\mu_k)}{\sigma_k}} \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{x^2}{2}\big)dx, \quad i = 1,...,n, \tag{2}$$

and

$$f(z_i;\theta) = \sum_{k=1}^{g} \pi_k \frac{\Gamma(\frac{\nu_k+1}{2})}{\Gamma(\frac{\nu_k}{2})\sqrt{\pi\nu_k}\sigma_k} \big(1+\frac{d}{\nu_k}\big)^{-\frac{\nu_k+1}{2}} T\big(\sqrt{\frac{\nu_k+1}{d+\nu_k}}A;\nu_k+1\big), \quad i = 1,...,n. \tag{3}$$

where $T(\cdot;\nu_k+1)$ denotes the distribution function of the standard *t* distribution with ν_k+1 degrees of freedom, location equal to 0 and scale equal to 1; $d = (z_i-\mu_k)^2/\sigma_k^2$ and $A = \lambda_k(z_i-\mu_k)/\sigma_k$. Note that, if $\lambda_k = 0$, the above formulas correspond to the Normal and non-standard *t* distributions, respectively [4].

Model estimation was performed using maximum likelihood method as available in the R package `mixsmsn` [5]. Model selection was performed using the Akaike's Information Criteria (AIC) where the goodness-of-fit was penalised by the number of parameters of an given model. The best model for the data is then the one with the lowest AIC value.

Seropositivity to a given virus was determined by calculating a cutoff τ in the respective antibody distribution above which individuals would be considered seropositive. For this calculation, the seronegative population was interpreted as the component of the mixture distribution with the lowest average value while the remaining components were interpreted as different levels of seropositivity upon recurrent infections. In this case, a cutoff for a seropositivity was calculated by the estimated 99.9%-quantile associated with the distribution of the putative seronegative population.

In the end, the seropositivity of *i*-th individual can be seen as resulting from a Bernoulli random variable $Y_i \frown Ber(p)$ where p is the so-called seroprevalence, which is the probability of an individual having an antibody level greater than τ. Seroprevalence can be easily estimated as the proportion of putative seropositive individuals in the sample.

Once the cut-off point was estimated for each antibody under analysis, the next step was to assess the sensitivity and specificity of the respective serological classification. With this purpose, it was assumed that the seronegative population is associated with the first component of the mixture distribution while the remaining components are associated with seropositivity. In this case, the probability of classifying an individual as seronegative S^- given an antibody level z_i is defined as

$$P(S^-|z_i) = \frac{\pi_1 f_1(z_i;\theta_1)}{\sum_{k=1}^{2} \pi_k f_k(z_i;\theta_k)}. \tag{4}$$

In turn, the probability of classifying an individual as seropositive S^+ given an antibody level z_i is simply defined as

$$P(S^+|z_i) = 1 - P(S^-|z_i). \tag{5}$$

Logistic regression

A logistic regression model was constructed with the objective of understanding whether serological data is able to distinguish patients with ME/CFS from healthy controls and patients with MS. In this model, the response variable was the study group allowing a pairwise comparison between ME/CFS and HC and between ME/CFS and MS. The respective covariates are the seropositivity for the different virus. The predictive performance of the logistic regression model was assessed using the area under receiver operating characteristic curve (AUC).

RESULTS

The analysis was carried out in a total of 335 individuals with complete information for all the variables under analysis. Eighty-one (24.2%) were male and 254 (75.8%) were female with ages at data collection of 43.47 ± 11.24 years. The distribution of different study groups was the following: healthy controls (HC, n=92; 27.5%), patients with multiple sclerosis (MS, n=36; 10.8%), and patients with ME/CFS (n=207; 61.8%).

Estimating seroprevalence associated with each virus

TABLE 1. Results of the best model for the antibody data of 4 common herpesviruses.

Virus	Distribution	No of components	Population	Mean (SD)	Cut-off	AIC	Seroprevalence (95% CI)
CMV	skew-*t*	2	S^-	1.61 (0.28	2.503	957.62	31.94
			S^+	5.09 (0.54)			(22.41-41.47)
EBV (EBNA)	skew-*t*	2	S^-	2.07 (0.34)	2.321	916.97	76.12
			S^+	4.45 (0.74)			(67.41-84.83)
EBV (VCA)	skew-*t*	2	S^-	0.67 (0.39)	3.521	779.22	83.88
			S^+	5.37 (0.49)			(76.37-91.39)
HSV1	skew-*t*	3	S^-	0.51 (0.09)	1.058	1020.38	76.42
			S^+ 1	1.43 (0.73)			(67.74-85.09)
			S^+ 2	5.22 (0.24)			
HSV2	skew normal	2	S^-	0.50 (0.65)	2.622	1029.82	37.61
			S^+	3.84 (0.79)			(27.71-47.51)

Most antibody distributions could be described by two latent populations with the exception of HSV1 antibody data where the best model supported three latent populations (I). In this case, the latent population with the lowest mean was interpreted as the seronegative population while the remaining latent populations were considered to be the putative seropositive populations with distinct degrees of exposure to the virus. Seroprevalence was estimated to be the highest for the Epstein-Barr virus (EBNA or VCA) and HSV1. These results were in line with the known fact that these viruses are quite common in the human population. As an example, Figure 1 shows the classification probability of the putative seropositive and seronegative population as function of the antibody level associated with CMV.

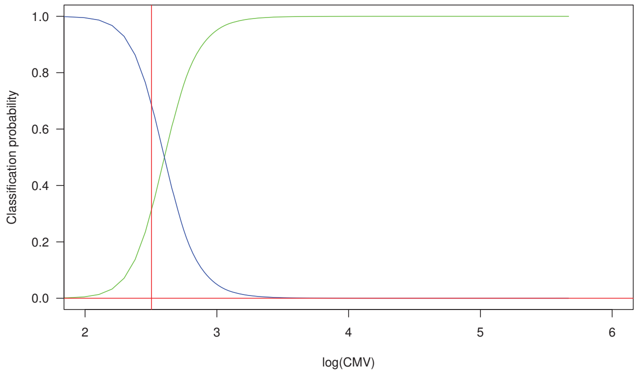


FIGURE 1. Conditional classification probabilities for CMV antibody data as defined by equations (4) and (5) (blue line - putative seronegative population, green line - putative seropositive populatio, red vertical line - cutoff for seropositivity as reported in Table I.

Comparing ME/CFS with Multiple Sclerosis (Model 1) and Healthy controls (Model 2)

After defining which individuals were seropositive and seronegative to each virus, we then determined the potential of using serology as way to distinguish ME/CFS patients from healthy controls and patients with MS. Most of serological predictors were not statistically significant at 5% significance level with some exceptions shown in Table II. In the case of ME/CFS patients versus healthy controls, the fitted regression model had limited predictive power because the estimate of AUC was close to a random guess situation (0.60; 95% CI=(0.53-0.67)). A slight better prediction was obtained for the comparison between patients with ME/CFS and MS (AUC=0.75; 95% CI=(0.68-0.82)). However, this prediction was still far from being useful for diagnostic purposes.

TABLE 2. Statistically significant results from the multiple logistic regression comparing ME/CFS group against healthy controls and patients with MS.

Comparison	Predictors	OR (95% CI)	p-value	AUC CI (95%)
ME/CFS vs. HC	CMV	0.60 (0.36-1.02)	0.06	0.60 (0.53-0.67)
	EBNA (VCA)	0.53 (0.24-1.07)	0.09	
ME/CFS vs. MS	EBV (EBNA)	0.07 (0.00-0.33)	0.01	0.75 (0.68-0.82)
	HSV-1	0.06 (0.00-0.32)	0.01	
	HSV-2	2.22 (1.03-4.92)	0.04	

CONCLUSIONS

In conclusion, the results suggested that serological data to these common herpesviruses have limited power to be used as a diagnostic tool for ME/CFS. A limitation of this study is the small number of antibodies used for predicting ME/CFS cases. It is then expected that an increased number of antibodies under analysis might improve model prediction. In this scenario, high-throughput serology data might be a solution as available elsewhere [6]. Another limitation is related to a possible misclassification of seropositivity for some individuals. This influences the results and might introduce bias in the analysis specially for individuals who tend to show antibody levels close to cutoff for seropositive. Finally, these herpes viruses might elicit similar antibody responses and, therefore, one should expected some correlation between the respective data. This correlation suggest the use of multivariate methods, which will be reported elsewhere.

ACKNOWLEDGMENTS

The authors would like to thank Jackie-Cliff, Ji-Sook Lee, Eliana Lacerda and Luis Nacul for sharing the data used in the work. This work is partially funded by Fundação para a Ciência e a Tecnologia, Portugal (ref: UID/MAT/00006/2019) and resulted from a short-term scientific mission of TDS to the London School of Hygiene & Tropical Medicine under the umbrella of the EUROMENE Cost Action from the European Union.

REFERENCES

1. F. Sotzny, J. Blanco, E. Capelli, J. Castro-Marrero, S. Steiner, M. Murovska, and C. Scheibenbogen, "Myalgic encephalomyelitis/chronic fatigue syndrome – evidence for an autoimmune disease," Autoimmunity Reviews **17**, 601 – 609 (2018).
2. E. M. Lacerda, E. W. Bowman, J. M. Cliff, C. C. Kingdon, E. C. King, J.-S. Lee, T. G. Clark, H. M. Dockrell, E. M. Riley, H. Curran, *et al.*, "The uk me/cfs biobank for biomedical research on myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) and multiple sclerosis," Open journal of bioresources **4** (2017).
3. J. M. Cliff, E. C. King, J.-S. Lee, N. Sepúlveda, A.-S. Wolf, C. Kingdon, E. Bowman, H. M. Dockrell, L. C. Nacul, E. Lacerda, *et al.*, "Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs)," Frontiers in immunology **10**, 796 (2019).
4. T. I. Lin, J. C. Lee, and S. Y. Yen, "Finite mixture modelling using the skew normal distribution," Statistica Sinica **17**, 909–927 (2007).
5. M. O. Prates, V. H. Lachos, and C. Cabral, "mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions," Journal of Statistical Software **54**, 1–20 (2013).
6. M. Loebel, M. Eckey, F. Sotzny, E. Hahn, S. Bauer, P. Grabowski, J. Zerweck, P. Holenya, L. G. Hanitsch, K. Wittke, *et al.*, "Serological profiling of the ebv immune response in chronic fatigue syndrome using a peptide microarray," PloS one **12**, e0179124 (2017).

Appendix C

Additional work

Veterinary and Animal Science 10 (2020) 100144

Contents lists available at ScienceDirect

Veterinary and Animal Science

journal homepage: www.elsevier.com/locate/vas

Detection and modeling of anti-*Leptospira* IgG prevalence in cats from Lisbon area and its correlation to retroviral infections, lifestyle, clinical and hematologic changes

Joana Moreira da Silva[a], Sara Prata[a], Tiago Dias Domingues[b,c], Rodolfo Oliveira Leal[a], Telmo Nunes[a], Luís Tavares[a], Virgílio Almeida[a], Nuno Sepúlveda[b,d], Solange Gil[a,*]

[a] *CIISA – Centre for Interdisciplinary Research in Animal Health, Faculty of Veterinary Medicine, University of Lisbon, 1300-477 Lisbon, Portugal*
[b] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*
[c] *Department of Statistics and Operational Research (DEIO), Faculty of Sciences, University of Lisbon, Portugal*
[d] *Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, United Kingdom*

ARTICLE INFO

Keywords:
Leptospirosis
IgG
FIV positive
Free-roaming cats
One Health

ABSTRACT

Leptospirosis is a zoonosis of global importance caused by *Leptospira* species. Rodents are the main reservoirs, known to shed the bacteria in urine, thus contaminating water and soil and infecting other animals and people. Leptospirosis has been re-emerging in both developing and developed countries including Europe. It has been hypothesized that cats could be asymptomatic carriers of *Leptospira*. This study aims to evaluate cats' exposure to *Leptospira* in Lisbon, Portugal, by measuring IgG titres and correlating them with possible factors that may increase the risk of exposure in urban cats. Two hundred and forty-three samples were collected from the biobank. An ELISA test followed by a seroprevalence analysis using a finite mixture model was performed to detect and measure anti-*Leptospira* IgG antibodies titres. In parallel, a survey was conducted to identify possible risk factors for seropositivity.

According to the ELISA test protocol, only twenty-three cats (9.5%; 95% CI =(6.1%;13.9%)) could be considered as seropositive to *Leptospira* antigens. However, when the same data were analysed by the best different mixture models, one hundred and forty-four cats (59.3%; 95%CI = (52.8%; 65.5%)) could be classified as intermediate and high antibody responders to *Leptospira* antigens. Seropositivity to Feline Immunodeficiency Virus infection (FIV) was found to be the only significant risk factor associated with anti-*Leptospira* IgG antibodies. In conclusion, the present studies raises the possibility of a higher exposure of cats to *Leptospira* than previously thought due to the identification of a subpopulation of cats with intermediate antibody levels.

Introduction

Leptospirosis is one of the leading zoonotic diseases in terms of morbidity and mortality worldwide, often in regions where the burden of leptospirosis is underestimated. Globally, the total number of leptospirosis cases has been estimated at 1.03 million and 58.900 deaths every year (Costa et al., 2015). It is also considered by many the most widespread bacterial zoonotic disease (Costa et al., 2015) and a silent epidemic disease by the World Health organization (WHO), Pan American Health organization (PAHO) and Health and Climate Foundation (Schneider et al., 2013).

Leptospirosis is caused by an infection with spirochete bacterium of the genus *Leptospira* and affects humans as well as a broad spectrum of animal hosts. There are currently 27 *Leptospira* species as delineated by DNA–DNA hybridization (Masuzawa et al., 2019). Phylogenetic analysis of these species using the 16S rRNA gene has resulted in the broad

Abbreviations: IgG, Immunoglobulin G; ELISA, enzyme-linked immunosorbent assay; FIV, Feline Immunodeficiency Virus infection; FeLV, Feline Leukemia Virus; WHO, World Health organization; PAHO, Pan American Health organization; MAT, microscopic agglutination tests; IDIU, Infectious Diseases Isolation Unit; VTH, Veterinary Teaching Hospital; CBC, complete blood count; CKD, Chronic Kidney Disease; USG, Urine Specific Gravity; ALP, serum alkaline phosphatase; AST, aspartate aminotransferase; ALT, alanine aminotransferase

[*] Corresponding author.

E-mail addresses: joanacmoreirasilva3@gmail.com (J. Moreira da Silva), saraprata@fmv.ulisboa.pt (S. Prata), tmdomingues@fc.ul.pt (T.D. Domingues), rleal@fmv.ulisboa.pt (R.O. Leal), tnunes@fmv.ulisboa.pt (T. Nunes), ltavares@fmv.ulisboa.pt (L. Tavares), vsa@fmv.ulisboa.pt (V. Almeida), Nuno.Sepulveda@lshtm.ac.uk (N. Sepúlveda), solange@fmv.ulisboa.pt (S. Gil).

https://doi.org/10.1016/j.vas.2020.100144
Received 20 May 2020; Received in revised form 25 September 2020; Accepted 25 September 2020
Available online 01 October 2020

classification of the species into pathogenic, saprophytic and intermediate (Perolat et al., 1998). Within the species *Leptospira interrogans* over 500 serovars are recognized (Caimi & Ruybal, 2020).

Although the respective molecular classification is not problematic for clinicians and individual treatment, it poses a problem regarding public health and epidemiology as it does not have enough discriminatory power to determine the infecting serovar (Levett, Paul, 2001). However, methods are being developed to improve this situation (Bezerra da Silva, Carvalho, Hartskeerl & Ho, 2011).

Most mammalian species are natural carriers of pathogenic *Leptospira* (Hartskeerl, Collares-Pereira & Ellis, 2011). These include feral, semi-domestic, farm and pet animals. Leptospirosis is commonly diagnosed in livestock species such as cattle, sheep, goats, horses, pigs and dogs (Pal, Mahendra, 1996).

Leptospira infection is also commonly described and investigated in dogs, while in cats it is less well described. Recently, the role of cats as concurrent carriers for illness has been questioned. Cats are the main predators for rodents and can act as reservoir hosts, with some studies proving transmission of pathogenic *Leptospira* between cats and other animals (Ojeda, Salgado, Encina, Santamaria & Monti, 2018)· There is also the premise that feral cats or cats living in shelters are more likely to have been infected with these bacteria. Considering the predator-prey relationship cats have with rodents and their close proximity to Humans, their role as a potential source for this agent needs to be better evaluated. These last options are more likely to happen with stray cats, cats who have outdoor access or even cats that live in rural environments. The direct contact with other cats, dogs or cattle is also considered to be a risk factor for this infection (Arbour, Blais, Carioto & Sylvestre, 2012; Hartmann et al., 2013).

Laboratory diagnosis of leptospirosis is not straightforward and may involve tests to detect *Leptospira* (Musso & La Scola, 2013), leptospiral antigens, or leptospiral nucleic acid in animal tissues or body fluids and/or to detect anti-leptospiral antibodies. Serological testing includes microscopic agglutination tests (MAT), enzyme-linked immunosorbent assay (ELISA) and rapid immunomigration tests (Kodjo, Calleja, Loenser, Lin & Lizer, 2016; Lizer, Velineni, Weber, Krecic & Meeus, 2018).

Diagnosis of infection by antibody detection in cats is appealing since they are not currently vaccinated, and therefore, the chance of finding false positives is much smaller. Testing is not too expensive and it can be performed in veterinary hospitals with supporting diagnostic laboratories. However, the international market supply of *Leptospira* IgG ELISA kits applicable to cat samples is limited. For example, in Portugal, there is only one commercial kit available at the time of the study, developed by the Bioassay Technology Laboratory (BT Lab). Production of other Cat *Leptospira* IgG ELISA test kit had been discontinued, possibly due to lack of sales. The rationale for using antibody data is that the antibody concentrations in the serum could be an indicator of bacteria exposure, thus providing epidemiological information about cats which are currently or have been infected. Antibody quantification is usually done by means of traditional enzyme linked immunosorbent assays. Optical densities or titres in arbitrary units are then used for the subsequent data analysis. In this epidemiological scenario of extremely low frequency of disease, scarcity of ELISA tests to measure cat *Leptospira* IgG and weakness of these tests validation methods, it is timely to apply statistical approaches to determine in antibody data analysis for diseases like malaria to cat leptospirosis in an attempt to optimize an ELISA test result interpretation (Sepúlveda, Stresman, White & Drakeley, 2015).

The present study falls under the "One Health" scope, as cats are exposed to environmental risks and share a great proximity to their owners, consequently placing them at risk of contracting leptospirosis.

The objectives of this study were: (1) to determine the seroprevalence of *Leptospira* spp. antibodies in cats presented to the Veterinary Teaching Hospital (VTH) of the Veterinary Faculty (FMV) of the University of Lisbon (ULisboa) by analysing the data directly with a statistical modeling approach; (2) to investigate associated risk factors, namely indoor/outdoor lifestyle and presence of retroviral infection.

Materials and methods

Sample collection

Previously collected blood samples from 243 cats was used to assess the performance of an ELISA kit for the presence of anti-*Leptospira* IgG. Blood samples were collected from a biobank which was developed using cat's serum samples obtained from a well characterized population of cats. Biobank stored blood samples were previously collected by venipuncture of the jugular vein. To allow a better evaluation and simpler blood sample collections, cats were subjected to mild sedation with 0•2 to 0•5 mg/kg butorphanol solution sc (Dolorex, Intervet Portugal). Serum samples were collected after clotting of the sample had occurred by centrifugation (5000 *g*, 10 min), and were subsequently frozen at −80 °C until analysed. This population is composed of cats from three different locations: two different animal shelters in the Lisbon area; cats which went to a consultation at the Veterinary Teaching Hospital – University of Lisbon (*VTH*) and cats which were hospitalized at the Infectious Disease Unit at the hospital (*IDIU*). All samples were stored at −80 °C.

Data collection

Data was collected using clinical database software from 2014 to December 2018. Collected data included age, lifestyle (indoor or outdoor) and contact with other animals. Cat's lifestyle was considered unknown when this information was missing or it was stated that cats were indoor but had contact with other animals with unknown lifestyle. Cats were considered to have an indoor lifestyle if they had not been outdoors for more than 10 years.

Plasma samples and FIV/FeLV infection

A total of 243 samples were analysed. All blood samples were tested to confirm their viral infection status by means of commercially available ELISA kits (ViraCHEK/FIV and ViraCHEK/FeLV, Synbiotics). Therefore, two groups were set: one group of retroviral negative cats (Status FIV/Feline Leukemia Virus (FeLV) negative) and one group of positive-retroviral cats (which were positive for FIV, FeLV or both).

Leptospiral IGG screening

All samples were screened for the presence of anti- *Leptospira* IgG antibodies by using IgG ELISA test kit by Bioassay Technology Laboratory. The manufacturer's guidelines were followed for the making of this test and the OD values were read at 450 nm. The cut-off value to consider a sample positive was the sum of the value obtained for the negative control plus 0.15. For quality control purposes, both the OD value for a blank well (no solutions at all) and the OD for negative control had to be ≤ 1. For statistical purposes, data under analysis referred to the average antibody values from two independent replicates of ELISA performed in the same biological samples.

Estimation of IGG seroposivity to Leptospira antigens

Finite mixture models based on flexible Skew-Normal and Skew-t distributions were fitted to data. In theory, the basic assumption of these models is that the antibody distribution could represent different serological populations (e.g., seronegative population and different seropositive populations possibly describing different levels of exposure to *Leptospira*). Note that these mixture models were chosen, because they extend the classical mixture models based on normal and t distributions by introducing an additional parameter that controls the degree of asymmetry of the mixing distributions. The above models were estimated assuming one serological population (e.g., seronegative) to 5 different serological populations. In the case of models with more than one serological population, it was assumed that the serological population with the lowest average referred to hypothetical seronegative individuals, while the remaining serological populations referred to putative seropositive individuals with different degree of exposure to leptospira. Model estimation was performed based on the maximum likelihood method using the Expectation-Maximization algorithm. Akaike's Information Criterion (AIC), which is defined by the deviance minus twice the number of model parameters, was used to determine the best fitted model for the data. According to this criterion, the best model was the one showing with the lowest AIC estimate among all models tested (Supplementary Table 1).

After determining the best model for the data, the putative seronegative population was identified as the hypothetical serological population with the lowest antibody average. A cut-off for seropositivity was calculated using the estimated 99.9% quantile of the hypothetical seronegative population. Cats whose antibody values were above this cut-off, were considered seropositive and seronegative otherwise. After classifying each cat as either seropositive or seronegative, the seroprevalence of the sample was estimated by the proportion of the seropositive cats among all cats tested. Note that the choice of a cut-off value defining seropositivity was somehow arbitrary. Therefore, a more stringent cut-off value could have been used for estimating seroprevalence. Finally, univariate and multivariate logistic regression models were used to determine which factors were associated with seropositivity to leptospiral antigen.

All statistical analysis was conducted in the software R version 3.4.3. using the package mixsmsn to estimate finite mixture models (Prates, Cabral & Lachos, 2013) and glm function to fit regression to the corresponding seropositivity data. The significance level for statistical testing was specified at 5%.

Results

Characterization of the studied population: Retroviral status, lifestyle, other pets and concomitant diseases

One hundred and twenty two out of the 243 cats tested were seropositive for either FIV or FeLV (74 FIV positive, 37 FeLV positive and 11 FIV and FeLV co-infected) (Fig. 1A). Twelve and 159 cats (65.43%) had indoor and outdoor lifestyles, respectively, according to criteria defined in the section of data collection from Materials and Methods. The remaining 72 cats (29.63%) had an unknown lifestyle (Fig. 1B).

With respect to home contact with other animals, 115 cats had other cats in the same house, 64 cats were considered to have unknown lifestyles, 40 cats did not cohabit with other pets, 16 cats lived with other cats and dogs, 6 cats cohabited with dogs and 2 cats cohabited with other pets (birds, reptile, fish or others) (Fig. 1C).

Estimation of Leptospiral seroprevalence

All samples were screened for the presence of anti- *Leptospira* IgG antibodies by using IgG ELISA test kit by Bioassay Technology Laboratory. Twenty-three out of the 243 cats (9.5%; 95% CI=(6.1%;13.9%)) analysed were tested positive using the seropositive cut-off value suggested by the manufacturer of the commercial ELISA used. An alternative seropositive cut-off value based on flexible finite mixture models was calculated for the same data. In brief, among all the ten mixture models fitted to the data (Supplementary Table 1), the lowest AIC estimate was obtained for the mixture model based on Skew-t distribution with three serological populations (Fig. 2). These serological populations were assumed to represent a seronegative population (with the lowest average) and two seropositive populations. Under this assumption, the cut-off for seropositivity was re-estimated at 0.40, which contrasts with the cut-off value of 1 as instructed by the manufacturer of the serological kits. Therefore, the intermediate serological population could be interpreted as putative seropositive cats on their way to sero-reversion. According to this new cut-off value, the seroprevalence to leptospirosis was re-estimated at 59.3% (n = 144/243; 95%CI = (52.8%; 65.5%)).

Analysis of potential factors contributing to Leptospiral seroprevalence

Regarding the analysis comparing retroviral infection and Leptospiral IgG seropositivity, 52.78% animals negative for retroviral infections tested positive for IgG anti-*Leptospira* (76/144) and 46.53% were positive for retroviral infections and IgG anti-*Leptospira* (67/144 - 38/67 for FIV$^+$; 24/67 for FELV$^+$ and 5/67 FIV$^+$/ FELV$^+$) (Fig. 3).

From the risk assessment analysis, FIV is the only significant factor (p-value = 0.02 - (Table 1)). Four cats lived in animal refugees, with the other 140 positive cats being domestic. However, most of positive cats (n = 96) had an outdoor lifestyle (66.67%), with a small percentage being indoor cats (n = 5; 3.47%). 18.75% (27/144) of animals had some form of renal and/or hepatic laboratory parameter alteration and of these, 51.85% (14/27) were immunocompetent (data not shown).

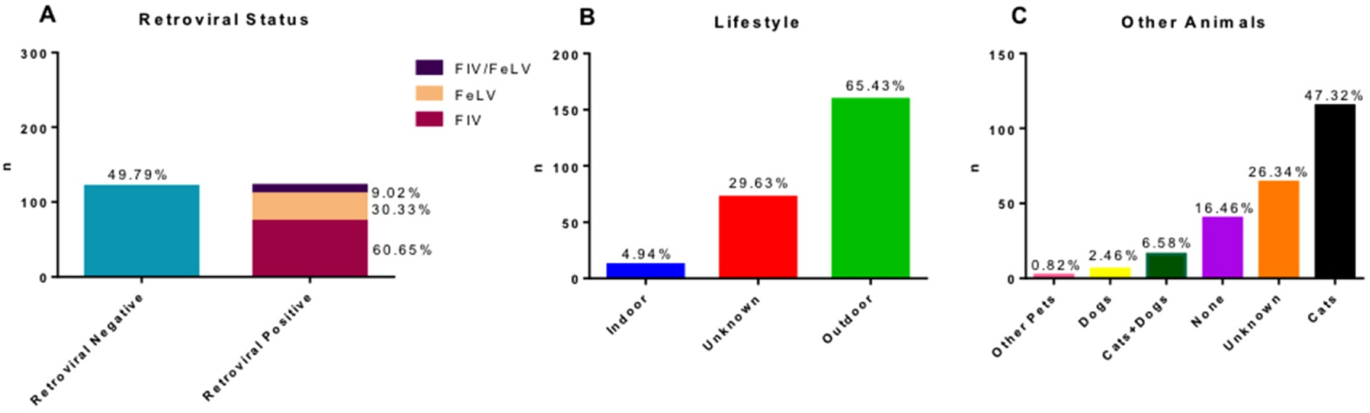


Fig. 1. Sample distribution between different characteristics analysed. **A** – 50.21% (n = 122) were retroviral positive, of which 60.65% (n = 74) were FIV$^+$, followed by 30.33% (n = 37) for FeLV$^+$ and 9.02% (n = 11) had both retroviral infections; 49.79% (n = 121) were retroviral negative. **B** - 4.94% had an indoor lifestyle (n = 12), followed by 29.60% (n = 72) whose lifestyle could not be characterized, and 65.43% (n = 159) had access to the outdoor. **C** – 0.82% (n = 2) had close contact with other pets, followed by contact just dogs (2.42%, n = 6); 6.58% (n = 16) lived with other cats and dogs, followed by cats which lived alone (16.46%, n = 40); for 26.34% (n = 64), co-habitation with other animals could not be determined and 47.32% (n = 115) co-habited with other cats.

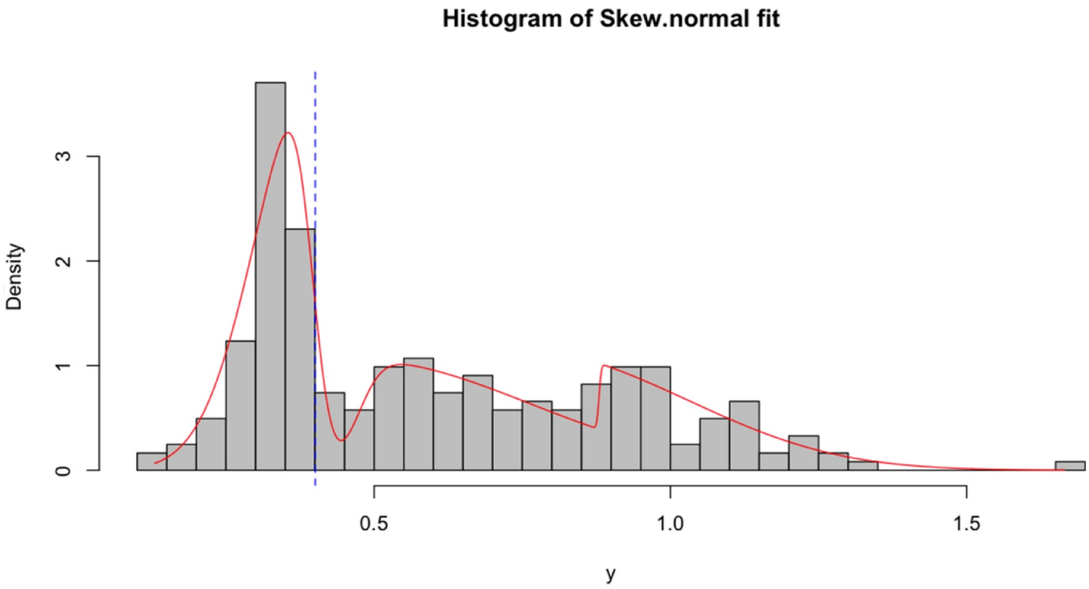
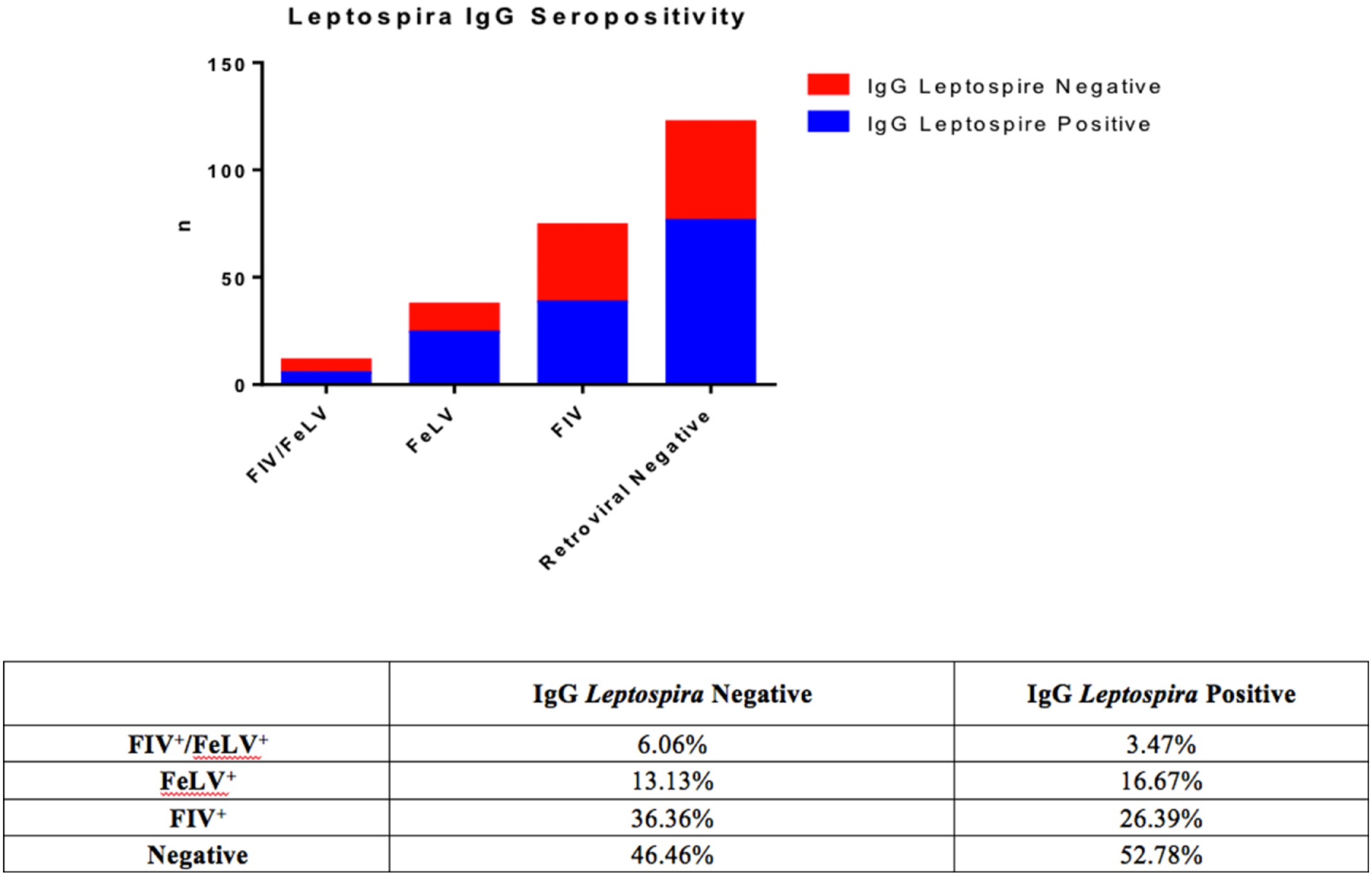


Fig. 2. Positive samples distribution across population. The first peak represents samples positive for the presence of IgG anti-*Leptospira*. The dotted line represents the estimated cut-off value for the ELISA assay. The second and third peak represent samples with a low positivity result or false positives.

Discussion

This study comes under the perspective of One Health (where human and animal health are inexorably linked) with leptospirosis being an emerging infectious disease and zoonosis.

The frequency of clinical illness is low in cats, despite the presence of leptospiral antibodies in the feline free-roaming population indicating a high probability of leptospiral exposure. Serological surveys carried out in Europe found a serological prevalence of 10% in Glasgow (Agunloye & Nash, 1996), 18% in Munich (Weis et al., 2017) and 48%



	IgG *Leptospira* Negative	IgG *Leptospira* Positive
FIV$^+$/FeLV$^+$	6.06%	3.47%
FeLV$^+$	13.13%	16.67%
FIV$^+$	36.36%	26.39%
Negative	46.46%	52.78%

Fig. 3. IgG *Leptospira* positive samples distribution within the retroviral status previously determined. Within positive IgG results, 3.47% (n = 5) had both retroviral infections, followed by FeLV$^+$ individuals (16.67%, n = 24) and FIV$^+$ (26.39%, n = 38). 52.78% (n = 76) were negative for retroviral infections.

Table 1
Different variables analysed and respective significance values and intervals. Through multivariate logistic regression model, FIV is the only characteristic that can influence positiveness for IgG.

Variable	Estimate (SE)	*P*	OR (CI 95%)
Intercept	−1.22 (1.87)	0.52	–
Gender (M)	0.16 (0.43)	0.72	1.17 (0.50–2.73)
FIV	***−1.05 (0.45)***	***0.02***	***0.35 (0.14–0.85)***
FeLV	−0.15 (0.53)	0.77	0.86 (0.30–2.42)
Lifestyle (Outdoor)	0.73 (0.75)	0.33	2.08 (0.48–9.07)
Dog	1.01 (0.67)	0.13	2.74 (0.73–10.22)
Cat	0.98 (1.64)	0.55	2.67 (0.11–66.29)
Other	1.14 (1.99)	0.57	3.13 (0.06–155.27)

****CI** – confidence interval; **OR** – *odds-ratio*; **P** – p-value; **SE** – Standard Error.

in France (Andre-Fontaine, Geneviève, 2006). Serovars Canicola, Grippotyphosa and Pomona have been isolated from cats (Adler & de la Peña Moctezuma, 2010). Clinical signs in cats are usually mild or not apparent, despite the presence of leptospiraemia and leptospiruria and histological evidence of renal and hepatic inflammation.

Several ELISAs have been developed and are primarily used for the detection of recent infections in dogs and livestock species. Problems with validation are a major constraint (OIE, 2018). A small number of ELISA tests for dogs (Hartman and Houten, 1984; Hartman, van den Ingh & Rothuizen, 1986) and cattle (Cousins, Robertson & Hustas, 1985) have been validated, using sequential serum samples from experimental animals but not beyond 6 months post-challenge. Laboratory variation and differences in host-specific humoral immune responses sometimes make correct assignment of antibody tests even more difficult. Many different serogroup antigens are tested in the assay, but false-negative results occur when the infecting serogroup is not included (Hartmann et al., 2013).

Evidence for a high seroprevalence in cats suggests that exposure in these species and its role for transmitting the bacteria to humans could be of more clinical importance in this species than previously recognized (Azócar-Aedo, Monti & Jara, 2014), posing a non-negligible public health risk to cat owners. However, the role of cats in the epidemiology of this zoonosis has not received much attention and clinical reports of leptospirosis in cats are rare (Arbour, Blais, Carioto and Sylvestre, 2012; Hartmann et al., 2013). Cats are predators of rodents (Loss & Marra, 2017; Parsons, Banks, Deutsch & Munshi-South, 2018), so prey-predator transmission between cats and rodents is likely to occur, and adding the free access to the outside or leash-walking, the concern should be reinforced. Scarce number of available ELISA tests and their limited validation for cats are a problem when analysing the frequency of exposure to *Leptospira* in the cat population.

Antibody quantification regarding diseases with extremely low prevalence, for which few ELISA tests are available plus the weakness of these tests' validation methods, advocate for a deeper analysis of the raw data in order to improve the accuracy of the serological classification of individuals/cats. This method was applied for other diseases with low rates of seropositive infected individuals such as malaria, where alternative measures based on antibodies have gained recent interest due to the possibility of estimating past disease exposure in absence of infected individuals. (Sepúlveda et al., 2015). This can be done using flexible mathematical models that could distinguish seronegative individuals from seropositive ones, thus, allowing the estimation of the seroprevalence, the proportion of seropositive individuals in the sample. The main advantage of using finite mixture models is allowing the raw data to determine the cut-off value for seroprevalence, rather than using the pre-established cut-off value. Such models also help determine specificity and sensitivity from the serologic classification obtained, which cannot be done when using a universal cut-off value.

Out of the 243 cats tested, (50% positive for retroviral infections and the rest immunocompetent), 23 samples tested positive for *Leptospira* IgG following the manufacturer's instruction and the indicated cut-off value for seropositivity. Applying the best model for quantitative antibody data used to determine the seropositivity of each individual to *Leptospira* antigen, 59.25% ($n = 144$) tested positive for the presence of IgG against *Leptospira*. This value is higher when compared to other studies, with positive results ranging from 18% to 43% (Dybing, Jacobson, Irwin, Algar & Adams, 2017; Lapointe, Plamondon & Dunn, 2013; Weis et al., 2017).

In these previous studies, the presence of *Leptospira* species was determined by IgM or PCR. However, it is valid to compare such results with the one obtained in this study when it comes to contact with the bacteria, because IgG remains after IgM and the bacteria have been cleared from the host. Nevertheless, the cut-off value used in these studies was pre-established by the manufacturer, which poses the question of whether or not one can trust such seroprevalence results as it is not known how the cut-off used was calculated. Seroprevalence comparison would be easier if cut-off was determined through the use of finite mixture models. Comparisons would also be more reliable since it would be known how the cut-off had been calculated.

The first risk factor analysed was retroviral infection status. This study showed that FIV infected cats have significantly lower anti-*Leptospira* IgG titres when compared to immunocompetent or FeLV infected cats. As immunocompromised cats often present a decreased capability on creating a memory immune response (Machado et al., 2019), a possible explanation relies on the fact that FIV infected cats have an impaired memory response towards Leptospira antigens used in this ELISA. Contrary to expectations, the same is not observed in FeLV cats. This can be due to a lower number of FeLV-infected cats comparing to FIV-infected ones as well as a non-standardization of lifestyle status between groups, which can reflect a different antigen-stimulus and, consequently, differences on memory immune response. Further studies are needed to clarify this finding.

This study analysed other risk factors that may eventually lead to cats' contact with *Leptospira* species, despite no significant association was found. Some individuals had missing data, which reduced the statistical power to detect an association with IgG *Leptospira* seropositivity. Although only one of the proposed risk factors was significantly associated with the anti-*Leptospira* IgG titres observed (FIV^{+}), careful examination of outdoor lifestyle and sharing the household with other animals should be further analysed using more indoor confined cats. Furthermore, the rough definition of "outdoor lifestyle" can also justify that no association was found – better parameters should have been determined in order to categorize into outdoor or indoor lifestyle.

The imbalanced frequency between indoor and outdoor distribution within the studied population, as well as the number of other animals sharing the household, may difficult the statistical analysis regarding these variables and their association to seropositivity. This aspect is a limitation of the study.

Nineteen of the cats presented azotaemia (renal or post-renal). After careful record review, most of the azotaemic patients exhibit some form of post-renal azotaemia which, by itself, justifies renal tubular damage and impaired urine concentrating ability. However, there is evidence that *Leptospira* can lodge itself in the renal tissue of carrier species like the cat (Parsons et al., 2018). Anti-*Leptospira* IgG seropositivity show that these cats have had contact with the bacteria and, therefore, azotaemia due to renal colonization by the bacteria cannot be ruled out without doing further testing. Some studies confirm that seropositivity is significantly greater in Chronic Kidney Disease (CKD) diagnosed cats (Rodriguez et al., 2014). Although not conclusive, several clinical studies in human subjects also assert that there may be a strong association between *Leptospira* infection and the development of CKD, suggesting infection as a risk factor for CKD (Carrillo-Larco, Altez-Fernandez, Acevedo-Rodriguez, Ortiz-Acha & Ugarte-Gil, 2019; Yang, Chang & Yang, 2019).

Some of the sampled cats revealed changes in the complete blood count (CBC) (mainly leucocytosis, leukopenia and anemia). These changes can be explained by them presenting a suspicion of or diagnosed illness (urinary tract inflammation or infection, colangiohepatitis, among others). Many of these animals were FIV and/or FeLV infected, which can likely account for these CBC abnormalities in some part (Tvedten & Raskin, 2013). Overall, no consistent clinical signs pattern was associated with the IgG Leptospiral positive population.

The present work using a statistical analysis of the raw antibody data contributes to a better knowledge of feline leptospirosis exposure in the context of One Health. The more evidence is gathered by studies like this one, the easier it will be to raise public awareness for cats and their possible role in transmission of this zoonotic disease. In a future study, it would be interesting to evaluate the owner's *Leptospira* IgM/IgG seroprevalence, alongside their cats.

Conclusions

In conclusion, the direct analysis of antibody data provided evidence for an additional subpopulation of cats with intermediate antibody levels to *Leptospira* antigen. Since this putative subpopulation was considered as seronegative according to the commercial ELISA test, we hypothesize that the real exposure of cats to *Leptospira* bacteria might be in fact higher than previously reported and therefore, it might pose a health hazard for both animals and humans in the context of One Health. To assess the validity of this hypothesis, similar serological assessment should be conducted in the other cohorts of cats.

Funding

This work was supported by CIISA - Centro de Investigação Interdisciplinar em Sanidade Animal, Faculdade de Medicina Veterinária, Universidade de Lisboa, Lisboa, Portugal, Project UIDP/CVT/00,276/2020 (funded by FCT). The authors received funding from Boehringer Ingelheim Animal Health – Portugal for sponsoring the purchase of the ELISA kits.

Ethical approval

This work involved the use of non-experimental animals only (including owned or unowned animals and data from prospective or retrospective studies). Established internationally recognised high standards ('best practice') of individual veterinary clinical patient care were followed. Samples regarding shelters in the Lisbon area had previously been collected as part of a PhD project – this study was approved by the FMV's Ethics Committee CEBEA (CEBEA - CIISA502010). The remaining samples were collected as part of routine consultation at VTH or hospitalization at UIDI. In these situations, owners also signed a written consent, which allowed the use for their pet's clinical history and surplus of biological samples for research purposes.

Authors' contributions

JMS, SP, ROL and SG performed the experiments and analyzed the data. NS, TD and TN performed the statistical analysis and helped drafting and revising the manuscript. LT and VA contributed to the analysis, interpretation of data and revised the manuscript. SG and NS conceived the study, analyzed the data and participated in its coordination, helped to draft the manuscript and supervised throughout. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to acknowledge Dr. Pedro Fabrica, from Boehringer Ingelheim Animal Health – Portugal, for sponsoring the purchase of the ELISA kits.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.vas.2020.100144.

References

Adler, B., & de la Peña Moctezuma, A. (2010). Leptospira and leptospirosis. *Veterinary Microbiology, 140*, 287–296. https://doi.org/10.1016/j.vetmic.2009.03.012.

Agunloye, C. A.&, & Nash, A. S. (1996). Investigation of possible leptospiral infection in cats in Scotland. *The Journal of Small Animal Practice, 37*, 126–129.

André-Fontaine, G. (2006). Canine leptospirosis-Do we have a problem? *Veterinary Microbiology, 117*, 19–24. https://doi.org/10.1016/j.vetmic.2006.04.005.

Arbour, J., Blais, M.-. C., Carioto, L., & Sylvestre, D. (2012). Clinical Leptospirosis in Three Cats (2001–2009). *Journal of the American Animal Hospital Association, 48*, 256–260. https://doi.org/10.5326/JAAHA-MS-5748.

Azócar-Aedo, L., Monti, G., & Jara, R. (2014). Leptospira spp. in domestic cats from different environments: Prevalence of antibodies and risk factors associated with the seropositivity. *Animals, 4*, 612–626.

Bezerra da Silva, J., Carvalho, E., Hartskeerl, R.& ., & Ho, P. (2011). Evaluation of the Use of Selective PCR Amplification of LPS Biosynthesis Genes for Molecular Typing of *Leptospira* at the Serovar Level. *Current Microbiology, 62*, 518–524.

Caimi, K., & Ruybal, P. (2020). Leptospira spp., a genus in the stage of diversity and genomic data expansion. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases, 81*, Article 104241. https://doi.org/10.1016/j.meegid.2020.104241.

Carrillo-Larco, R. M., Altez-Fernandez, C., Acevedo-Rodriguez, J. G., Ortiz-Acha, K., & Ugarte-Gil, C. (2019). Leptospirosis as a risk factor for chronic kidney disease: A systematic review of observational studies. *PLOS Neglected Tropical Diseases, 13*, 1–10. https://doi.org/10.1371/journal.pntd.0007458.

Costa, F., Hagan, J. E., Calcagno, J., Kane, M., Torgerson, P., & Martinez-Silveira, M. S. (2015). Global Morbidity and Mortality of Leptospirosis: A Systematic Review. *PLOS Neglected Tropical Diseases, 9*, 0–1. https://doi.org/10.1371/journal.pntd.0003898.

Cousins, D. V., Robertson, G. M., & Hustas, L. (1985). The use of the enzyme-linked immunosorbent assay (ELISA) to detect the IgM and IgG antibody response to Leptospira interrogans serovars hardjo, pomona and tarassovi in cattle. *Veterinary Microbiology, 10*, 439–450.

Dybing, N. A., Jacobson, C., Irwin, P., Algar, D., & Adams, P. J. (2017). *Leptospira* species in feral cats and black rats from Western Australia and Christmas Island. *Vector-Borne Zoonotic Dis. 17*, 319–324. https://doi.org/10.1089/vbz.2016.1992.

Hartman, E. G., & Houten, M. V. A. N. (1984). 1984 *Veterinary Immunology and Immunopathology, 7*, 245–254 7, 245–254.

Hartman, E. G., van den Ingh, T. S. G. A. M., & Rothuizen, J. (1986). Clinical, pathological and serological features of spontaneous canine leptospirosis. An evaluation of the IgM- and IgG-specific ELISA. *Veterinary Immunology and Immunopathology, 13*, 261–271. https://doi.org/10.1016/0165-2427(86)90078-4.

Hartmann, K., Egberink, H., Pennisi, M. G., Lloret, A., Addie, D., & Belák, S., Boucraut-Baralon, C., Frymus, T., Gruffydd-Jones, T., Hosie, M.J., Lutz, H., Marsilio, F., Möstl, K., Radford, A.D., Thiry, E., Truyen, U., & (2013). Leptospira Species Infection in Cats: ABCD guidelines on prevention and management. *Journal of Feline Medicine and Surgery, 15*, 576–581. https://doi.org/10.1177/1098612X13489217.

Hartskeerl, R. A., Collares-Pereira, M., & Ellis, W. A. (2011). Emergence, control and re-emerging leptospirosis: Dynamics of infection in the changing world. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases, 17*, 494–501. https://doi.org/10.1111/j.1469-0691.2011.03474.x.

Kodjo, A., Calleja, C., Loenser, M., Lin, D., & Lizer, J. (2016). A rapid in-clinic test detects acute leptospirosis in dogs with high sensitivity and specificity. *BioMed Research International, 2016*, 1–3. https://doi.org/10.1155/2016/3760191.

Lapointe, C., Plamondon, I., & Dunn, M. (2013). Feline leptospirosis serosurvey from a Quebec referral hospital. *The Canadian Veterinary Journal. La Revue Veterinaire Canadienne, 54*, 497–499.

Levett, P. N. (2001). Leptospirosis. *Clinical Microbiology, 14*, 296–326. https://doi.org/10.1128/CMR.14.2.296.

Lizer, J., Velineni, S., Weber, A., Krecic, M., & Meeus, P. (2018). Evaluation of 3 serological tests for early detection of leptospira-specific antibodies in experimentally infected Dogs. *Journal of veterinary internal medicine / American College of Veterinary Internal Medicine, 32*, 201–207. https://doi.org/10.1111/jvim.14865.

Loss, S. R., & Marra, P. P. (2017). Population impacts of free-ranging domestic cats on mainland vertebrates. *Frontiers in Ecology and the Environment, 15*, 1633.

Machado, I., Carvalho, A., Gomes, J., Cunha, E., Tavares, L., & Almeida, V.& .; Gil, S. (2019). 2019. Feline retrovirus-infected hospitalised cats – aetiology, co-infections and survival rates. *Clinical / research abstracts accepted for presentation at ISFM Congress 2019* (pp. 843–852). 26-30 June 2019.

Masuzawa, T., Saito, M., Nakao, R., Nikaido, Y., Matsumoto, M., Ogawa, M., et al. (2019). Molecular and phenotypic characterization of *Leptospira johnsonii* sp. nov., *Leptospira* *ellinghausenii* sp. nov. and *Leptospira ryugenii* sp. nov. isolated from soil and water in Japan. *Micro and Immuno, 63*(3–4), 89–99.

Musso, D., & La Scola, B. (2013). Laboratory diagnosis of leptospirosis: A challenge. *Journal of Microbiology, Immunology and Infection, 46*, 245–252. https://doi.org/10.1016/j.jmii.2013.03.001.

O.I.E. (2018). *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2019.* Paris, France: World Organization for Animal Health503–551 Chapter 3.1.12.

Ojeda, J., Salgado, M., Encina, C., Santamaria, C., & Monti, G. (2018). Evidence of interspecies transmission of pathogenic leptospira between livestock and a domestic cat dwelling in a dairy cattle farm. *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science, 80*, 1305–1308. https://doi.org/10.1292/jvms.16-0361.

Pal, M. (1996). Leptospirosis: A contemporary zoonosis. *The Veterinarian, 20*, 11–12.

Parsons, M. H., Banks, P. B., Deutsch, M. A., & Munshi-South, J. (2018). Temporal and space-use changes by rats in response to predation by feral cats in an urban ecosystem. *Frontiers in Ecology and Evolution, 6*, 1–8. https://doi.org/10.3389/fevo.2018.00146.

Perolat, P., Chappel, R. J., Adler, B., Baranton, G., Bulach, D. M., & Billinghurst, M. L. (1998). Leptospira fainei sp. nov., isolated from pigs in Australia. *International Journal of Systematic Bacteriology, 48*, 851–858. https://doi.org/10.1099/00207713-48-3-851.

Prates, M. O., Cabral, C. R. B., & Lachos, V. H. (2013). Mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software, 54*, 1–20. https://doi.org/10.18637/jss.v054.i12.

Rodriguez, J., Blais, M.-. C., Lapointe, C., Arsenault, J., Carioto, L., & Harel, J. (2014). Serologic and urinary PCR survey of leptospirosis in healthy cats and in cats with kidney disease. *Journal of Veterinary Internal Medicine / American College of Veterinary Internal Medicine, 28*, 284–293. https://doi.org/10.1111/jvim.12287.

Schneider, M. C., Jancloes, M., Buss, D. F., Aldighieri, S., Bertherat, E., & Najera, P., Galan, D. I., Durski, K., & (2013). Leptospirosis: A silent epidemic disease. *International Journal of Environmental Research and Public Health, 10*, 7229–7234. https://doi.org/10.3390/ijerph10127229.

Sepúlveda, N., Stresman, G., White, M. T., & Drakeley, C. J. (2015). Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication. *Journal of Immunology Research*. 2015 https://doi.org/10.1155/2015/738030.

Tvedten, H., & Raskin, R. E. (2013). Leukocyte Disorders. eds In M. Willard, & H. Tvedten (Eds.). *Small Animal Clinical Diagnosis by Laboratory Methods* (pp. 126–155). (5th ed.). St. Louis, Missouri: Elsevier Saunders.

Weis, S., Rettinger, A., Bergmann, M., Llewellyn, J. R., Pantchev, N., Straubinger, R. K., et al. (2017). Detection of Leptospira DNA in urine and presence of specific antibodies in outdoor cats in Germany. *Journal of Feline Medicine and Surgery, 19*, 470–476. https://doi.org/10.1177/1098612X16634389.

Yang, H. Y., Chang, C. H., & Yang, C. W. (2019). Leptospirosis and chronic kidney disease. *Translational Research in Biomedicine, 7*, 27–36. https://doi.org/10.1159/000500380.