UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Métodos de análisis de factores y predictores de la evolución en la supervivencia de la enfermedad cardiovascular

D. Álvaro Hernández Vicente

2023

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Métodos de análisis de factores y predictores de la evolución en la supervivencia de la enfermedad cardiovascular

Autor: D. Álvaro Hernández Vicente

Director/es: D. Domingo Andrés Pascual Figal

D. Manuel Franco Nicolás

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR
Aprobado por la Comisión General de Doctorado el 19-10-2022

D./Dña. Álvaro Hernández Vicente

doctorando del Programa de Doctorado en

Matemáticas

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Métodos de análisis de factores y predictores de la evolución en la supervivencia de la enfermedad cardiovascular

y dirigida por,

D./Dña. Domingo Andrés Pascual Figal

D./Dña. Manuel Franco Nicolás

D./Dña.

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Si la tesis hubiera sido autorizada como tesis por compendio de publicaciones o incluyese 1 o 2 publicaciones (como prevé el artículo 29.8 del reglamento), declarar que cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 6 de febrero de 2023

Firmado por HERNANDEZ VICENTE ALVARO - ***0295** el día 06/02/2023 con un certificado emitido por AC FNMT Usuarios

Fdo.:

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Información básica sobre protección de sus datos personales aportados	
Responsable:	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación:	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad:	Gestionar su declaración de autoría y originalidad
Destinatarios:	No se prevén comunicaciones de datos
Derechos:	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

A mi familia y amigos por toda su comprensión,
su apoyo y su paciencia infinita conmigo.

A Bernardo Cascales, por creer en mí.

A María, por aguantar tantas noches en soledad
y tanto tiempo robado.

A Diego.

Agradecimientos

Este documento de tesis cierra una etapa de más de cinco años de trabajo. En el mundo académico quizá no es mucho pero, echando la vista atrás, ha dado tiempo a ver pandemias, crisis y guerras. Me he despedido de familiares y amigos, he visto a amigos graduarse, aprobar oposiciones, ser padres o casarse, como yo. En todo este tiempo lo único que se ha mantenido inmutable ha sido esta tesis.

Si he llegado hasta aquí ha sido gracias a muchas personas que me han acompañado, me han animado y han confiado en mí más que yo mismo.

A Bernardo Cascales, que creyó en mí en un momento importante y fue un ejemplo de profesor y de persona para todos los que tuvimos la suerte de conocerle.

A Antonio y Elvira, el equipo SAE, con los que empecé este camino de la estadística y el análisis de datos y con los que senté las bases de todo lo que sé.

A todos mis compañeros del grupo de investigación en cardiología y de la plataforma de bioinformática del IMIB. Todos ellos (tanto los pasados como los presentes) me recibieron con los brazos abiertos, y su enorme trabajo y sacrificio, muy poco reconocido a veces, es uno de los pilares fundamentales de la investigación.

A Alejandro y Laura por su infinita hospitalidad al abrirme las puertas de su casa en tantas ocasiones y haberme aguantando en todas ellas. Si mis estancias en Madrid han sido imprescindibles para la consecución de esta tesis, vosotros también lo habéis sido al hacerlas posibles.

A Fátima y Carlos, del CNIC, por todo lo que me han enseñado sobre estadística bayesiana. Sin vuestro grandísimo apoyo esta tesis no habría salido adelante.

A mis directores y mi tutora de tesis. A Manolo y Juana Mari, por darme tanta libertad para avanzar por mi cuenta. Y a Domingo, por confiar en mí para formar parte de este grupo y forzarme a no tirar la toalla; su enorme dedicación y pasión por la ciencia, la cardiología y la salud de los pacientes es un ejemplo para todo el mundo de la investigación.

A la Universidad de Murcia que, en su apuesta por la ciencia de la Región mediante su Plan Propio de Fomento de la Investigación R-455/2019, ha financiado todo el desarrollo y trabajo de esta tesis.

A mi nueva familia alemana, por la gran acogida que me dieron, por esforzarse tanto conmigo con el idioma y por su gran disposición siempre a ayudar, como queda patente en el resumen de este documento. *Vielen Dank*.

A mis padres, hermanos, cuñados, familia cercana y amigos íntimos, que desde bien pequeño me siguen aguantando (que no es poco) y siempre están ahí. Aunque sepan que es un tema complejo y que no soy muy bueno explicándome, hacen un esfuerzo tremendo por preguntarme, interesarse y comprender lo que hago y de qué trata esta tesis.

Y por último, a María, la que más ha sufrido con diferencia mis dificultades con esta tesis. Ha sido mi apoyo en la desmotivación, el desánimo y el agotamiento habituales en mí que han marcado estos meses. Espero poder compensar todas las noches que la he dejado dormir sola, las vacaciones que no hemos tenido, los viajes que hemos pospuesto y todo el tiempo que le he robado al escribir esta tesis. *Ich liebe dich*.

—Álvaro Hernández Vicente
Enero, 2023

Prefacio

Este trabajo se enmarca dentro de la bioestadística, también conocida como biometría o estadística médica (términos menos habituales en España pero con mucha más historia en otros países como Reino Unido). Esta rama de la estadística trata de desarrollar métodos y formas de análisis para aplicarlos a la medicina, la biología u otras ramas de las ciencias de la salud y ciencias de la vida. La bioestadística como disciplina, podemos decir, vive con un pie en cada uno de esos mundos: el mundo de la salud (o de lo «bio» donde se encuentran las ciencias de la salud y la medicina), que es un mundo principalmente práctico, y el mundo de las matemáticas y la estadística, que es sobre todo teórico y abstracto.

Igual que otras disciplinas aplicadas (como la bioinformática) el hecho de situarse en el límite entre dos disciplinas como la estadística y la medicina tiene tanto ventajas como inconvenientes. Desde el punto de vista de alguien que viene de la estadística la ventaja principal es poder ver la utilidad a los métodos aprendidos y buscar soluciones a problemas reales, no quedándose solo en la teoría. Mientras, aquel que viene de la medicina puede analizar los datos de los que dispone y obtener conclusiones útiles en la práctica clínica. La principal desventaja para ambos es que deben acercarse en la medida de lo posible al mundo del otro para poder escoger los modelos adecuados o poder interpretar los resultados obtenidos.

Este trabajo se ha intentado redactar con el objetivo de hacerlo lo más accesible posible a investigadores de ambas ramas o incluso a otros que quieran acercarse al análisis de datos en cardiología. Con este objetivo en mente, y pensando en los más ajenos a la rama de las matemáticas, se ha añadido además un apéndice con un repaso de la notación matemática básica utilizada en este documento.

Resumen

La insuficiencia cardíaca (IC) aparece cuando el corazón pierde su capacidad para ejercer su función de bomba circulatoria y da lugar a la aparición de síntomas y signos a consecuencia de su disfunción. Se define como un síndrome porque sus manifestaciones clínicas, síntomas y signos son diversos, al igual que las enfermedades que terminan desembocando en ella, lo que provoca que su diagnóstico sea complejo. Se sitúa entre las primeras causas de hospitalización y muerte en el mundo y, debido a su alto coste asociado, destaca como uno de los asuntos sociosanitarios de mayor relevancia en la actualidad.

Dada su alta mortalidad, la mayor parte de la investigación de la IC se ha centrado en el desarrollo de fármacos para reducirla y evitar la progresión de la enfermedad. Algunos de los aspectos que no han sido tan estudiados son cómo evolucionan estos pacientes y qué factores afectan a su mejoría durante el transcurso de la enfermedad. El aspecto más relevante a la hora de definir la IC es la presencia de disfunción contráctil, que vendría reflejada en la fracción de eyección de ventrículo izquierdo (FEVI o FE). De esta forma, la enfermedad se clasifica en IC con FE reducida (menor o igual al 40 %) o FE preservada. Para la IC con FE reducida se dispone de fármacos que previenen la progresión de la enfermedad y reducen la enfermedad. Sin embargo, no todos los pacientes se benefician de estos fármacos, la respuesta es variable y, por tanto, se necesitan mejores modelos predictivos para estudiar el comportamiento de la IC (y de la FE en particular), considerando su interacción con otras características de los pacientes.

El objetivo principal de este trabajo de tesis es el desarrollo e implementación de un modelo estadístico con el que estudiar la evolución de pacientes con IC, mediante el uso de la fracción de eyección del ventrículo izquierdo (FEVI), y determinar qué factores tiene asociados.

Este estudio se plantea como un estudio longitudinal, retrospectivo, observacional y descriptivo. En él analizamos los datos clínicos de 251 pacientes que acuden a la unidad de IC del Hospital Virgen de la Arrixaca de Murcia durante los años 2005 a 2017 con FEVI reducida (menor o igual del 40 %). Los datos de FEVI los obtenemos de las 1504 ecocardiografías (media de 6.0 ± 3.7 por paciente) recogidas en el hospital. Para todos ellos se ha revisado la mortalidad (que se produce en un 45.4 % de los pacientes) y tiempos de seguimiento (media 8.9 ± 4.7 años).

El modelo que proponemos y desarrollamos para estudiar los datos de estos pacientes está basado en los conocidos como *joint models* para datos longitudinales y de supervivencia. Para describir la evolución de la FEVI usamos una función no lineal determinada por un valor

inicial, una determinada velocidad a la que aumentan los valores y un punto máximo al que se aproximan, que llamamos de saturación.

Para la estimación y definición de este modelo usamos la estadística bayesiana y el lenguaje probabilístico Stan, que nos ofrecen una mayor libertad a la hora de establecer las distintas características del modelo.

Los resultados de nuestro análisis indican que los pacientes parten con unos valores iniciales de FEVI en torno a un 25 % (intervalo de credibilidad al 95 % de [22.6, 26.8]). El aumento de estos valores se produce en los primeros 6 meses, siendo más corto en presencia de cardiopatía isquémica. El aumento medio se estima en un 15.5 % (intervalo de credibilidad al 95 % de [12.7, 18.3]).

Las variables que se asocian con los valores iniciales de FEVI o bien con el aumento posterior son: el infarto aguo de miocardio (IAM), la hipertensión arterial (HTA), la etiología isquémica, el tiempo desde el diagnóstico o el volumen telediastólico (VTD, que es una medida del remodelado cardíaco).

Nuestros resultados muestran que la mejora de los valores de la FEVI se asocia con un descenso en el riesgo de muerte en torno a un 18 % (*hazard ratio* de 0.82 con intervalo de credibilidad al 95 % de [0.71, 0.94]).

Este documento de tesis está estructurado en cinco grandes partes, siguiendo el estilo habitual de un artículo científico.

La primera parte corresponde a la introducción y comprende los dos primeros capítulos. En el primero de ellos se presenta y define la insuficiencia cardíaca, se explica su complejidad, su diagnóstico y la gran problemática que lleva asociada. En el segundo capítulo se hace un repaso sobre los métodos estadísticos habituales para el análisis de datos médicos, se introducen los modelos centrales de este trabajo como son los *joint models* para datos longitudinales y de supervivencia y se presentan los dos grandes paradigmas de la estadística: frecuentista y bayesiano.

La segunda parte, con un único capítulo, comprende la justificación y los objetivos principales de este trabajo, que se puede resumir en el desarrollo de un modelo estadístico para el estudio de la evolución de pacientes con insuficiencia cardíaca.

El diseño del estudio, el desarrollo completo de este modelo estadístico para el análisis de datos, así como toda la metodología utilizada en el trabajo se describen en la tercera parte del documento. Cada uno de estos puntos se corresponden con los capítulos que componen esta parte: diseño, modelización y metodología.

Los resultados de la aplicación del modelo propuesto a los datos clínicos, junto con el análisis descriptivo y exploratorio de estos últimos, se recogen en la cuarta parte.

La última parte de este documento comprende los capítulos de discusión de los resultados y de conclusiones que se han alcanzado a partir de los resultados obtenidos en este trabajo.

Abstract

Heart failure (HF) occurs when the heart loses its ability to perform its function as a circulatory pump resulting in the appearance of symptoms and signs consequences of its dysfunction. It is defined as a syndrome because its clinical manifestations, symptoms and signs are diverse, as well as the diseases that lead to it, which makes its diagnosis complex. Heart failure is one of the leading causes of death and hospitalization in the world and, due to its high associated cost, it stands out as one of the most important social and health care issues nowadays.

Due to its high mortality, most HF research has focused on the development of drugs to reduce it and prevent disease progression. Some less studied aspects are how these patients evolve and what factors affect their improvement during the course of the disease.

The most relevant aspect in defining HF is the presence of contractile dysfunction, manifested in the left ventricular ejection fraction (LVEF or EF). Thus, the disease is classified into HF with reduced EF (less than or equal to 40 %) or preserved EF. For HF patients with reduced EF drugs that prevent disease progression and reduce disease are available. However, not all patients benefit from these drugs, the response is variable and, therefore, better predictive models are needed to study HF behavior (and EF in particular), considering its interaction with other patient characteristics.

The main objective of this work is the development and implementation of a statistical model to study the evolution of patients with HF, by using left ventricular ejection fraction (LVEF), and to determine factors associated with it.

This is a longitudinal, retrospective, observational and descriptive study. In it, we analyzed the clinical data of 251 patients attending the HF unit of the Hospital Virgen de la Arrixaca of Murcia during the years 2005 to 2017 with reduced LVEF (less than or equal to 40 %). LVEF data were obtained from 1504 echocardiographies (average of 6.0 ± 3.7 per patient) collected in the hospital. For all of them, mortality (occurring in 45.4 % patients) and follow-up times (average of 8.9 years) were reviewed.

To study these data we propose a model based on joint models for longitudinal and time-to-event data. To describe the evolution of LVEF values, we use a nonlinear function determined by an initial value, a growth rate of the values and a maximum point called saturation where they approach.

To estimate and define this model we use bayesian statistics and the probabilistic language Stan. This language provides more freedom to establish different characteristics of the model.

The results of our analysis indicate that patients start with initial LVEF values around 25 % (with 95 % credible interval of [22.6, 26.8]). The increase in these values is produced in the first 6 months and it is shorter in the presence of ischemic heart disease. The average increase is estimated at 15.5 % (with 95 % confidence interval of [12.7, 18.3]).

The variables associated with LVEF initial values or with posterior increment are: acute myocardial infarction, hypertension, ischemic etiology, time since diagnosis or end-diastolic volume (a cardiac remodeling measure).

Our results show that LVEF improvement reduces the risk of death by 18 % (hazard ratio of 0.82 with a 95 % credible interval of [0.71, 0.94]).

This manuscript is structured in five main parts, following the usual style of a scientific article.

The first part corresponds to the introduction and comprises the first two chapters. In the first chapter we present and define heart failure, explain its complexity, diagnosis and the major problems associated with it. In chapter 2 we review usual statistical methods for the analysis of clinical data, introduce the central models of this work, such as the joint models for longitudinal and time-to-event data, and present the principal statistical paradigms: frequentist and Bayesian.

Second part, with a single chapter, includes justification and main objectives of this work. It can be summarized as the development of a statistical model to study the evolution of patients with heart failure.

The study design, complete statistical model development for data analysis, as well as the entire methodology used are described in the third part of this document. Each point corresponds to a chapter comprising this part: design, modeling and methodology.

The fourth part includes the results of the application of the proposed model to clinical data, with the descriptive and exploratory analysis of the latter.

The last part of this document comprises the final chapters with the discussion of the results and conclusions we draw from them.

Zusammenfassung

Eine Herzinsuffizienz (HF) liegt vor, wenn das Herz seine Funktion als Kreislaufpumpe nicht mehr erfüllen kann und als Folge der Funktionsstörung Symptome und Anzeichen auftreten. Es wird als Syndrom definiert, weil ihre klinischen Erscheinungsformen, Symptome und Anzeichen vielfältig sind, ebenso wie die Krankheiten, die zu ihr führen, was ihre Diagnose komplex macht. Die HF ist eine der häufigsten Ursachen für Krankenhausaufenthalte und Todesfälle in der Welt, und ist aufgrund der damit verbundenen hohen Kosten, heute eines der wichtigsten sozialen und gesundheitlichen Probleme.

Angesichts der hohen Sterblichkeitsrate konzentrierte sich die HF-Forschung bisher vor allem auf die Entwicklung von Medikamenten, die die Sterblichkeitsrate senken und das Fortschreiten der Krankheit verhindern sollen. Wenig erforscht ist unter anderem, wie sich diese Patienten entwickeln und welche Faktoren ihre Verbesserung im Verlauf der Krankheit beeinflussen. Der wichtigste Aspekt bei der Bestimmung von HF ist das Vorhandensein einer kontraktilen Dysfunktion, die sich in der linksventrikulären Ejektionsfraktion (LVEF oder EF) widerspiegelt; so wird die Krankheit in HF mit reduzierter EF (kleiner oder gleich 40 %) oder erhaltener EF eingeteilt. Für HF mit reduzierter EF stehen Medikamente zur Verfügung, die das Fortschreiten der Krankheit verhindern und die Krankheit Symptome reduzieren. Allerdings profitieren nicht alle Patienten von diesen Medikamenten, da das Ansprechen variabel ist; daher werden bessere Vorhersagemodelle benötigt, um das Verhalten von HF (und insbesondere von EF) unter Berücksichtigung der Wechselwirkung mit anderen Patientenmerkmalen zu untersuchen.

Das Hauptziel dieser Arbeit ist die Entwicklung und Implementierung eines statistischen Modells, um die Entwicklung von Patienten mit HF anhand der linksventrikulären Ejektionsfraktion (LVEF) zu untersuchen und zu ermitteln, welche Faktoren damit verbunden sind.

Bei dieser Arbeit handelt es sich um eine retrospektive, beobachtende und beschreibende Dauerstudie. Wir analysierten die klinischen Daten von 251 Patienten, die zwischen 2005 und 2017 in der HF-Abteilung des Hospitals Virgen de la Arrixaca in Murcia behandelt wurden und eine reduzierte LVEF (kleiner oder gleich 40 %) aufwiesen. Die LVEF-Daten wurden aus 1504 Echokardiografien (Mittelwert von 6,0±3,7 pro Patient) gewonnen, die in diesem Krankenhaus durchgeführt wurden. Die Sterblichkeit (die bei 45,4 % der Patienten eintritt) und die Nachbeobachtungszeit (durchschnittlich 8,9±4,7 Jahre) wurden für alle Patienten überprüft.

Das Modell, das wir zur Untersuchung der Daten dieser Patienten vorschlagen und entwickeln haben, basiert auf den so genannten „joint models zur Modellierung von Längsschnitt-

und Überlebenszeitdaten“. Um die Entwicklung der LVEF zu beschreiben, verwenden wir eine nichtlineare Funktion, die durch einen Anfangswert, eine bestimmte Rate, mit der die Werte ansteigen, und einen Maximalpunkt, dem sie sich nähern, bestimmt wird.

Für die Schätzung und Definition dieses Modells verwenden wir die Bayesianische Statistik und die probabilistische Sprache Stan, die uns mehr Freiheit bei der Festlegung der verschiedenen Merkmale des Modells geben.

Die Ergebnisse unserer Analyse deuten darauf hin, dass die Patienten mit anfänglichen LVEF-Werten um die 25 % beginnen (95 % Glaubwürdigkeitsintervall von [22,6, 26,8]). Der Anstieg dieser Werte erfolgt in den ersten 6 Monaten und ist bei Vorliegen einer ischämische Herzkrankheit kürzer. Der mittlere Anstieg wird auf 15,5 % geschätzt (95 % Glaubwürdigkeitsintervall [12,7, 18,3]).

Zu den Variablen, die mit den anfänglichen LVEF-Werten oder einem anschließenden Anstieg in Verbindung stehen, gehören: akuter Myokardinfarkt (AMI), arterielle Hypertonie (AHT), ischämische Ätiologie, Zeit seit der Diagnose oder enddiastolisches Volumen (EDV, ein Maß für den Kardiales Remodeling).

Unsere Ergebnisse zeigen, dass eine Verbesserung der LVEF-Werte mit einer Verringerung des Sterberisikos um etwa 18 % verbunden ist (Hazard Ratio von 0,82 mit einem 95-prozentigen Glaubwürdigkeitsintervall von [0,71, 0,94]).

Die vorliegende Arbeit ist in fünf Hauptteile gegliedert, die dem üblichen Stil eines wissenschaftlichen Artikels entsprechen.

Der erste Teil ist die der Einleitung und umfasst die ersten beiden Kapitel. Im ersten von ihnen wird die Herzinsuffizienz dargestellt und definiert sowie ihre Komplexität, ihre Diagnose und die damit verbundene große Problematik der HF erläutert. Das zweite Kapitel gibt einen Überblick über die gängigen statistischen Methoden zur Analyse medizinischer Daten, führt in die zentralen Modelle dieser Arbeit ein, wie die „joint models zur Modellierung von Längsschnitt- und Überlebenszeitdaten“ und stellt die beiden wichtigsten Paradigmen der Statistik vor: frequentistisch und Bayesianisch.

Der zweite Teil mit einem einzige Kapitel umfasst die Begründung und die Hauptziele dieser Arbeit, die sich als Entwicklung eines statistischen Modells für die Untersuchung der Entwicklung von Patienten mit Herzinsuffizienz zusammenfassen lassen.

Das Design der Studie, die vollständige Entwicklung dieses statistischen Modells für die Datenanalyse sowie die gesamte Methodik, die in der Arbeit verwendet wird, werden im dritten Teil des Dokuments beschrieben. Jeder dieser Punkte entspricht den Kapiteln, aus denen sich dieser Teil zusammensetzt: Entwurf, Modellierung und Methodik.

Die Ergebnisse der Anwendung des vorgeschlagenen Modells auf die klinischen Daten werden zusammen mit der deskriptiven und explorativen Analyse der klinischen Daten in Teil 4 vorgestellt.

Der letzte Teil dieses Dokuments umfasst die abschließenden Kapitel mit der Diskussion der Ergebnisse und den Schlussfolgerungen aus den in dieser Arbeit erzielten Ergebnissen.

Índice de contenidos

Acrónimos

ARA-II	Antagonistas del receptor de la angiotensina II (en inglés ARBs, de *angiotensin II receptor blockers*)
ARM	Antagonistas del receptor de mineralocorticoides, también conocidos como inhibidores de la aldosterona o antialdosterónicos (en inglés MRA, de *mineralocorticoid receptor antagonist*)
CIPA	Código de identificación personal autonómico
BB	Bloqueadores beta o betabloqueantes (en inglés *betablockers*)
BNP	Péptido natriurético cerebral (en inglés *Brain Natriuretic Peptide*)
DAG	Grafo acíclico dirigido (del inglés *directed acyclic graph*)
DAI	Desfibrilador automático implantable
DM	Diabetes *mellitus*
DTD	Diámetro telediastólico
E-BFMI	Estimación bayesiana de la fracción de información faltante (del inglés *estimated bayesian fraction missing information*)
ECG	Electrocardiograma
SEC	Sociedad española de cardiología
ES	Error estándar de la media
ESC	Sociedad europea de cardiología (por su siglas en inglés *European Society of Cardiology*)
FE	Fracción de eyección (en inglés EF, de *ejection fraction*)
FEVI	Fracción de eyección del ventrículo izquierdo (en inglés *Left Ventricle Ejection Fraction* o *LVEF*)
HCUVA	Hospital Clínico Universitario Virgen de la Arrixaca de Murcia
HR	Razón de riesgos (del inglés *Hazard ratio*)
HTA	Hipertensión arterial
IAM	Infarto agudo de miocardio
INE	Instituto Nacional de Estadística
INRA	Inhibidor de la neprilisina el receptor de la angiotensina (en inglés ARNi, de *angiotensin receptor-neprilysin inhibitor*)
IC	Insuficiencia cardíaca (en inglés *Heart Failure* o *HF*)
IC95	Intervalo de credibilidad al 95 %, usado en las tablas y gráficos de resultados
ICA	Insuficiencia cardíaca aguda
ICC	Insuficiencia cardíaca crónica

IC-FEr Insuficiencia cardíaca con fracción de eyección del ventrículo izquierdo reducida (en inglés *HFrEF*, de *reduced*)

IC-FElr Insuficiencia cardíaca con fracción de eyección del ventrículo ligeramente reducida (en inglés *HFmrEF*, de *midly reduced*)

IC-FEc Insuficiencia cardíaca con fracción de eyección del ventrículo izquierdo conservada (en inglés *HFpEF*, de *preserved*)

IECA Inhibidor de la enzima de conversión de la angiotensina (en inglés ACEI, *angiotensin-converting-enzyme inhibitors*)

IR Insuficiencia renal

iSGLT2 Inhibidores del cotransportador de sodio-glucosa tipo 2 (del inglés *sodium-glucose transport protein 2 inhibitor*)

MCMC Métodos de Montecarlo basados en cadenas de Markov (del inglés *Markov chain Montecarlo*)

NA Valor perdido (del inglés *not available*)

NHC Número de historia clínica

NHST Contrastes de hipótesis (del inglés *null hypothesis significance testing*)

NT-proBNP Fracción N-terminal del propéptido natriurético cerebral

NUTS Método hamiltoniano de Montecarlo utilizado por Stan para el muestreo (del inglés *No-U-Turn sampler*)

NYHA Clase de la asociación del corazón de Nueva York (del inglés, *New York Heart Association*)

PP Probabilidad de un parámetro de ser positivo, *i. e.* $P(\cdot > 0)$

RAE Real Academia Española

RIQ Rango intercuartílico, *i. e.* $[Q_1, Q_3]$

TRC Terapia de resincronización cardíaca

VI Ventrículo izquierdo (en inglés LV, de *left ventricle*)

VTD Volumen telediastólico (o volumen diastólico final, VDF)

VTS Volumen telesistólico (o volumen sistólico final, VSF)

Parte I

Introducción

Capítulo 1

Nociones de insuficiencia cardíaca

1.1 | Definición y diagnóstico

La insuficiencia cardíaca o IC (*heart failure* o *HF* en inglés) es un síndrome complejo que se caracteriza por la incapacidad del corazón de bombear la cantidad suficiente de sangre al organismo. Debido a su complejidad, la definición de insuficiencia cardíaca ha ido cambiando durante años, desde definiciones más simples que se centraban en la parte hemodinámica a otras más complejas que añadían factores como la respuesta hormonal (Feldman y Mohacsi 2019; McDonagh 2011).

En la actualidad, el consenso es definirlo como un síndrome clínico caracterizado por unos síntomas típicos (como disnea —dificultad respiratoria o falta de aire—, fatiga o inflamación de tobillos) junto a otros signos clínicos (como crepitantes pulmonares —sonido anormal escuchado en la auscultación—, edema periférico —inflamación de las extremidades— o presión venosa yugular alta) y producido por una anomalía estructural o funcional del corazón que resulta en un aumento de la presión dentro del corazón y una reducción del gasto cardíaco —volumen de sangre bombeada por minuto— (Feldman y Mohacsi 2019; Jurgens *et al.* 2022; McDonagh *et al.* 2022).

El hecho de hablar de «síndrome», que se define como un conjunto de síntomas que se dan juntos y que tiene una cierta identidad, se debe a que es causada por otras enfermedades como la cardiopatía isquémica —que es un desequilibrio del flujo sanguíneo en el corazón—, el infarto agudo de miocardio (IAM), la hipertensión arterial (HTA), la diabetes (DM), arritmias como la fibrilación auricular, valvulopatías como la estenosis aórtica, o la amiloidosis cardíaca —acumulación de proteína amiloide en el corazón—. Esto hace que muchos pacientes «distintos» (que llegan por distintas vías al hospital y deben ser tratados de forma diferente) desemboquen en insuficiencia cardíaca, haciendo el diagnóstico complicado y haciendo necesaria la determinación de esa causa subyacente —lo que se conoce como etiología— (ver figura 1.1).

Aunque son muchos los problemas que pueden causar o contribuir a una insuficiencia cardíaca (fallos en las válvulas, las aurículas, los ventrículos, el ritmo cardíaco, etc.) lo más común para valorar la mejoría de un paciente es centrarse en el funcionamiento del ventrículo izquierdo (VI). Después de oxigenarse en los pulmones, la sangre llega a través de las venas

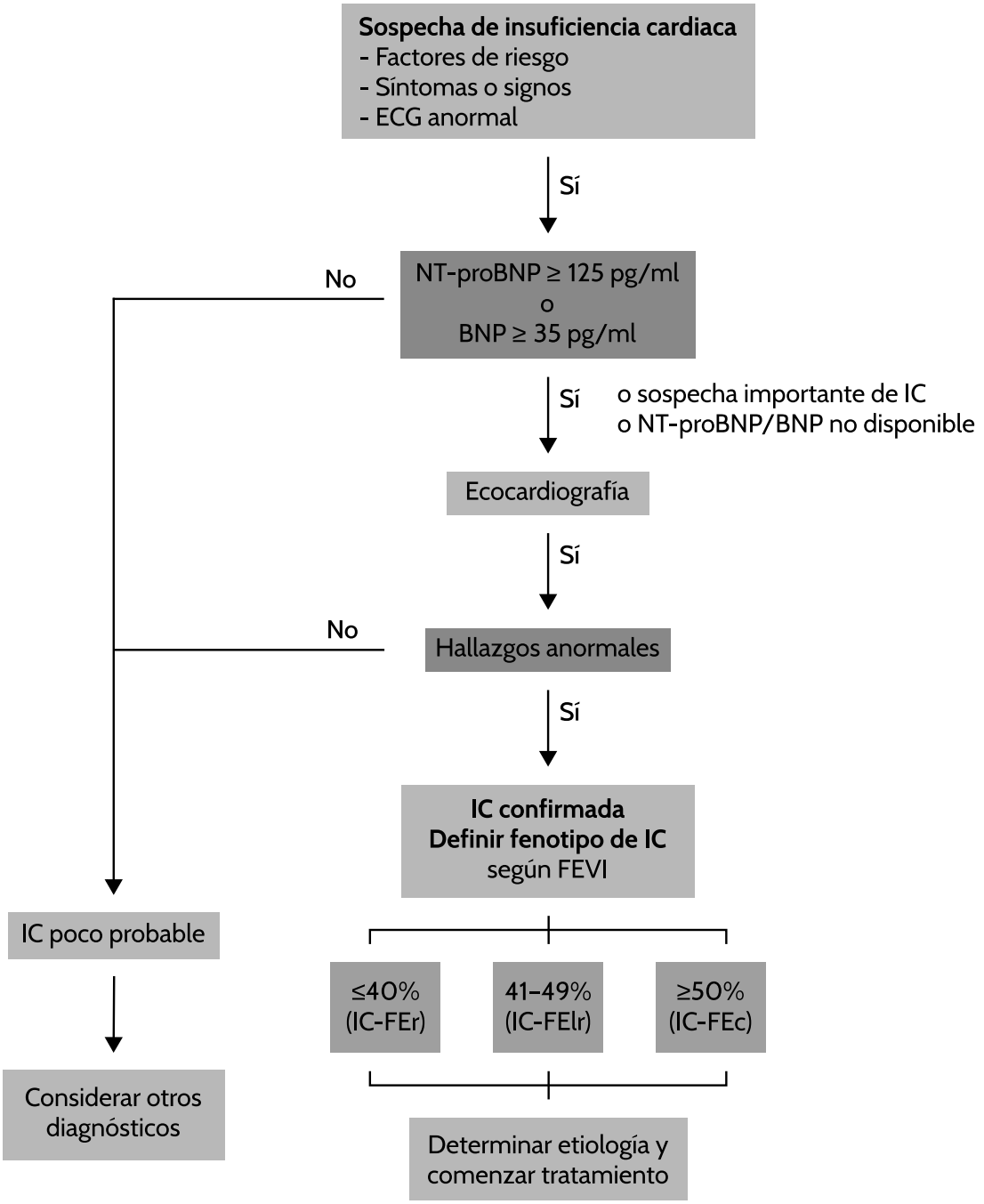

Figura 1.1. *Algoritmo de diagnóstico de insuficiencia cardíaca. Figura basada en la figura 1 de McDonagh et al. 2022. BNP (por las siglas en inglés de Brain Natriuretic Peptide): péptido natriurético cerebral; NT-proBNP: fracción N-terminal del propéptido natriurético cerebral; ECG: electrocardiograma; IC: insuficiencia cardíaca; FEVI: fracción de eyección del ventrículo izquierdo; IC-FEc: IC con FEVI conservada; IC-FElr: IC con FEVI ligeramente reducida; IC-FEr: IC con FEVI reducida.*

pulmonares al ventrículo izquierdo y este es el responsable de bombearla por la arteria aorta para transportarla al resto del cuerpo. Para conseguirlo, este ventrículo necesita impulsar la sangre con suficiente presión, lo que hace que sea más grueso y más muscular que el ventrículo derecho. Cuando se produce un deterioro de la contractilidad ventricular se produce una congestión intracardíaca, que se traduce en un aumento de la presión que retrógradamente llevan al edema pulmonar y del resto del sistema venoso, a la vez que a una perdida de flujo anterógrado, con hipoperfusión de tejidos y órganos, como músculos, riñones y cerebro. Así, la sangre se acumula en pulmones y venas pulmonares y el aporte de oxígeno al organismo (músculos, cerebro, etc.) se ve mermado. Todo ello provoca falta de aire, cansancio muscular, fatiga, mareos o confusión, sobre todo a la hora de realizar alguna actividad física[1] (Camm *et al.* 2018).

Específicamente, para medir el correcto funcionamiento del ventrículo izquierdo una de las medidas más extendidas y utilizadas es la fracción de eyección (FEVI), que se obtiene mediante ecocardiografía. Esta medida es un indicador del volumen de sangre que eyecta el corazón en cada latido hacia la aorta. La FEVI se calcula como un ratio (volumen de sangre eyectado respecto al «total posible», entendiendo este como el volumen de llenado máximo del ventrículo) para así tener una medida porcentual (multiplicando por cien). La fórmula para calcular la FEVI es

$$\text{FEVI} = \frac{\text{VTD} - \text{VTS}}{\text{VTD}} \cdot 100,$$

donde VTD representa el volumen telediastólico (también conocido como volumen diastólico final o VDF), que es el volumen del ventrículo cuando está en diástole o relajado, y donde VTS es el volumen telesistólico (también conocido como volumen sistólico final o VSF), que es el volumen una vez el corazón se contrae o se encuentra en sístole.

Debido principalmente a que en los primeros ensayos clínicos sobre tratamientos para la IC se observaba un beneficio en pacientes con FEVI $\leq 40\,\%$ esta medida es la más utilizada hoy en día para su clasificación. Aunque las clasificaciones variaban ligeramente entre países y sociedades científicas (Bozkurt *et al.* 2021, pág. 360) según las más recientes guías clínicas europeas (McDonagh *et al.* 2022) y las últimas iniciativas sobre una estandarización de los términos entre sociedades y grupos (Bozkurt *et al.* 2021) se pueden diferenciar tres grandes grupos de pacientes:

- aquellos con FEVI $\leq 40\,\%$, denominada «FEVI reducida» y designada como IC-FEr,
- aquellos con FEVI en el rango 40–49 % (hasta 50 %), denominada como «ligeramente reducida» y designada como IC-FElr,
- y aquellos con FEVI $\geq 50\,\%$, conocida como «conservada» y que se designa como IC-FEc.

Cabe destacar que cuando los pacientes son diagnosticados con IC-FEr y posteriormente presentan una mejoría superando su FEVI el 50 % se les pasa a considerar como pacientes con

[1] En la web se puede encontrar mucha información al respecto sobre la detección y síntomas de IC. El lector interesado puede visitar las páginas web de: la Clínica Universidad Navarra, la Fundación del Corazón, o el programa Mimocardio.

«IC con FEVI mejorada» en lugar de IC-FEc. Las recomendaciones generales, en estos casos, es continuar tratándolos como IC-FEr (McDonagh *et al.* 2022, sección 8.3).

1.2 | Fisiopatología y remodelado cardíaco

Debido a las múltiples causas que pueden producir una insuficiencia cardíaca, la fisiopatología de la enfermedad —esto es, los mecanismos y procesos que se producen a nivel molecular, celular, etc. y que resultan en la enfermedad— se hace también compleja.

Las distintas condiciones que producen la enfermedad, como un infarto agudo de miocardio (IAM) —donde por una obstrucción en las arterias coronarias una parte del músculo cardíaco se necrosa— o una miocardiopatía —donde por una anomalía de alguna proteína contráctil se produce una contracción deficiente—, provocan que el corazón se intente adaptar y compensar el daño para continuar supliendo las necesidades de oxígeno del organismo, lo que puede conseguirse, al inicio, mediante una dilatación de las cavidades cardíacas o una hipertrofia de las células del miocárdicas. Estos mecanismos del corazón, que inicialmente ayudan a paliar los efectos negativos que está sufriendo, contribuyen de forma crónica a la progresión de la enfermedad y afectan a su propia estructura y a la pérdida de su función, proceso que se conoce como remodelado cardíaco (Cohn *et al.* 2000; McDonagh 2011, pág. 133).

El remodelado cardíaco, de esta manera, se sabe que influye en distintos parámetros que se utilizan como métrica de evaluación del desempeño del corazón, como la FEVI y los volúmenes cardíacos —al final de la diástole o llenado (volumen telediastólico o VTD) y al final de la sístole o contracción (telesistólico o VTS)— (Konstam *et al.* 2011).

1.3 | Epidemiología

La insuficiencia cardíaca es un síndrome fuertemente relacionado con la edad, por lo que, debido al envejecimiento general de la población tanto la incidencia —número de casos nuevos en una población— como la prevalencia —proporción de casos en la población— están en aumento (Benjamin *et al.* 2018; McDonagh *et al.* 2022).

En el estudio PATHWAYS-HF, realizado haciendo uso de la base de datos BIG-PAC —donde se monitorizan a más de un millón de pacientes—, se estima que la prevalencia de insuficiencia cardíaca en España se sitúa en un 1.89 % con una tasa de incidencia de 2.78 casos por cada 1000 personas-años[2] (Sicras-Mainar *et al.* 2022). Además, varios estudios españoles y europeos sostienen que esta prevalencia aumenta con la edad, pudiendo superar el 17 % en personas mayores de 70 años (Sayago-Silva *et al.* 2013).

[2]Esta medida de «personas-años» refleja tanto el número de pacientes como el de años de seguimiento en conjunto. Dos estudios, donde uno monitoriza a 100 pacientes durante un año y otro a 10 pacientes durante 10 años, tendrán ambos 100 personas-años de información.

Por grupos de FEVI (definidos en el apartado 1.1) el registro de IC de la Sociedad Europea de Cardiología *ESC HF Long-Term* calcula que la proporción de pacientes que acuden a los departamentos de cardiología o a unidades especializadas de IC se sitúa en un 60 % para la IC-FEr, un 24 % para la IC-FElr y un 16 % para la IC-FEc (Chioncel *et al.* 2017). En la población general, sin embargo, muchos pacientes con IC-FEc están infradiagnosticados y no llegan a los departamentos de cardiología (Riet *et al.* 2014). Las proporciones reales se estima que son similares entre pacientes con IC-FEr y pacientes con IC-FEc o IC-FElr (McDonagh *et al.* 2022).

1.4 | Historia natural

El curso clínico de la insuficiencia cardíaca es progresivo y se caracteriza por un empeoramiento de la calidad de vida del paciente y un aumento de la necesidad de asistencia sanitaria, fundamentalmente debido a hospitalizaciones muy frecuentes (Chaudhry y Stewart 2016). Una vez establecido el diagnóstico en las primeras etapas y superada la fase aguda (IC aguda o ICA) en la que el paciente sufre la enfermedad que precede al síndrome —siendo la hospitalización la forma más frecuente de inicio (Fernández-Gassó *et al.* 2019a)— comienza la fase crónica de este (IC crónica o ICC).

En esta fase inicial tras el diagnóstico de la insuficiencia cardíaca, además de las recomendaciones sobre los hábitos de vida (ejercicio, ingesta hídrica y de sal, tóxicos, etc.), la ingesta donde se limita el agua, la sal, las grasas o se prohíbe el tabaco, se prescriben los diferentes tratamientos farmacológicos (que veremos más en detalle en el próximo apartado 1.6) u otros no farmacológicos como la implantación de un desfibrilador automático implantable (DAI), una terapia de resincronización cardíaca (TRC, como un marcapasos) o, en última instancia, un trasplante cardíaco (TC) en casos necesarios (McDonagh *et al.* 2022).

Pese a todo ello, el pronóstico de los pacientes sigue siendo malo, con una calidad de vida baja, y solo se consigue mejorar de forma significativa en aquellos diagnosticados con IC-FEr, en los que se dispone de fármacos que evitan la progresión de la enfermedad y reducen la mortalidad (McDonagh *et al.* 2022).

Según los últimos datos publicados por el INE (2022) sobre defunciones, correspondientes al año 2020, la insuficiencia cardíaca se sigue situando como una de las causas de muerte más frecuentes tanto a nivel nacional como regional (ver figura 1.2), posicionándose como la quinta y sexta causa más frecuente[3]. Con estos mismos datos, además, podemos ver su carácter progresivo, donde la mayoría de defunciones se concentran a partir de los 80 años de edad[4].

Respecto a hospitalizaciones, tras el diagnóstico principal de IC, se estima que los pacientes ingresan una vez al año de media, aunque con el envejecimiento de la población también

[3]Conviene destacar, además, que esto es teniendo en cuenta la excepcionalidad del año 2020, donde las defunciones por la COVID-19 han hecho que esta se sitúe en primer lugar a nivel nacional con más de 60000 muertes, sin contar aquellos casos sospechosos en los que no se ha identificado el virus (ver figura 1.2).

[4]El número de muertes por IC en las franjas de más edad según el INE en el año 2020 son: 388 (60 a 64 años), 390 (65 a 69 años), 678 (70 a 74 años), 1031 (75 a 79 años), 2519 (80 a 84 años), 5189 (85 a 89 años), 5274 (90 a 94 años) y 3013 (de 95 años y más).

Figura 1.2. *Gráficos de causas de muerte más frecuentes en España (A, arriba) y en la Región de Murcia (B, abajo) en el año 2020. Las causas están ordenadas verticalmente según la frecuencia, donde se destaca en color rojo la defunción por insuficiencia cardíaca. Gráficos descargados de las infografías que publica el INE (disponibles en la web* `https://www.ine.es/explica/explica_infografias.htm`*).*

se espera observar un aumento en los próximos años, quizás llegando al 50 % (McDonagh *et al.* 2022). Según datos de la Región de Murcia cada año se producen en torno a dos hospitalizaciones por IC por cada 1000 habitantes, entre los que un 57 % corresponderían a primeras hospitalizaciones —esto es, de pacientes de nuevo diagnóstico— y el otro 43 % a reingresos —hospitalizaciones de pacientes que ya han sido diagnosticados previamente— (Fernández-Gassó *et al.* 2019a).

1.5 | Coste sociosanitario

Debido al elevado número de hospitalizaciones causadas por la IC y a la atención sanitaria adicional que requieren los pacientes que la padecen, el coste tanto social como sanitario que consume es alto. La mayor parte de este coste recae en las hospitalizaciones, que suponen dos tercios del gasto y se estima que consumen del 1.5 % al 2 % del gasto sanitario del Sistema Nacional de Salud (Oliva *et al.* 2010).

A nivel global el coste sanitario varía mucho entre países. Según datos de una revisión sistemática llevada a cabo desde 2004 a 2016 el coste sanitario anual de un paciente con IC se sitúa entre los 26 mil dólares de Alemania hasta los casi 900 dólares de Corea del Sur, siendo el coste total en todo el transcurso de la enfermedad de más de 125 mil dólares por paciente (Lesyuk *et al.* 2018).

En el caso de España, un estudio de 2013 (revisado también en Lesyuk *et al.* 2018) sitúa los costes totales sanitarios en los 4860 euros (Delgado *et al.* 2014). Considerando otros costes adicionales como la medicación, los centros de día, las residencias o los cuidados esta cifra puede llegar a superar los 12 mil euros (Delgado *et al.* 2014).

1.6 | Tratamiento farmacológico

Los diferentes tratamientos farmacológicos para la insuficiencia cardíaca se centran en mejorar los síntomas, enlentecer la progresión de la enfermedad y reducir el riesgo de muerte y hospitalizaciones directamente relacionadas con la patología.

En las últimas versiones de las guías clínicas europeas sobre el diagnóstico y tratamiento de la IC (McDonagh *et al.* 2022) los distintos fármacos disponibles se categorizan según su nivel de evidencia y su grado de recomendación; diferenciando a los pacientes según el grupo de FEVI al que pertenecen (McDonagh *et al.* 2022, sección 5.3).

Por su complejidad, cabe destacar que, en la actualidad, todavía no se ha demostrado la efectividad de ningún fármaco para el tratamiento de los pacientes con IC-FEc (o IC-FElr) en la práctica clínica real. Así, la mayoría de recomendaciones actuales se centran en el tratamiento del grupo de IC-FEr, mucho más estudiado en la práctica. Para este grupo, el tratamiento

principal comprende tres familias de fármacos, que conforman la piedra angular del tratamiento en estos pacientes (McDonagh *et al.* 2022, secciones 5.2 y 5.3):

- los inhibidores de la enzima de conversión de la angiotensina (IECA), como el Ramipril o el Enalapril, —donde los antagonistas del receptor de la angiotensina II (ARA-II) son un alternativa en pacientes con alguna intolerancia a los IECA—
- los antagonistas del receptor de mineralocorticoides (ARM), también conocidos como inhibidores de la aldosterona o antialdosterónicos, como la Espironolactona, y
- los betabloqueantes, bloqueadores β o bloqueadores de los receptores beta adrenérgicos (BB), como el Bisoprolol o el Carvedilol.

Estos tres grupos de medicamentos han demostrado su relación con una menor mortalidad en múltiples ensayos clínicos y estudios (McDonagh *et al.* 2022). En el estudio de Fonarow *et al.* (2011), que evaluaba a más de dos millones y medio de estadounidenses, se estimaba una reducción del riesgo de mortalidad de un 17 %, un 34 % y un 30 % con los tratamientos con IECA/ARA-II, betabloqueantes y antialdosterónicos, respectivamente (Fonarow *et al.* 2011, tabla I).

La búsqueda de nuevos fármacos para el tratamiento de la IC es una de las líneas de investigación con mayor interés. En los últimos años, han probado su eficacia algunas familias adicionales de fármacos entre las que podemos destacar los de la familia de inhibidores de la neprilisina y del receptor de la angiotensina (INRA), como el fármaco combinado Sacubitrilo/Valsartán (McMurray, Packer *et al.* 2014; Velazquez *et al.* 2019), o los inhibidores del cotransportador de sodio-glucosa tipo 2 (iSGLT2), como la Dapaglifozina y la Empaglifozina (McMurray, Solomon *et al.* 2019; Packer *et al.* 2020).

Todos estos fármacos han demostrado su utilidad principalmente en ensayos clínicos y centrados en la reducción de la mortalidad y la mejora del pronóstico. Los estudios en un contexto de práctica clínica real son escasos, desconociéndose su efecto, en particular, sobre la progresión a largo plazo de la enfermedad y sobre la FEVI, como marcador más relevante en estos pacientes.

Capítulo 2

Análisis de datos y estadística médica

2.1 | Análisis de datos, estadística y *machine learning*

Cada vez es más común escuchar o leer términos como «ciencia de datos» (en inglés *data science*) o *machine learning* («aprendizaje automático» en español). Estos conceptos, junto con el de «estadística», se utilizan siempre en un contexto similar y con el objetivo principal de analizar datos para obtener conclusiones sobre el mundo real (ya sean datos médicos para mejorar el tratamiento de pacientes, biológicos para el desarrollo de nuevos fármacos o climatológicos para hacer predicciones sobre el calentamiento global, entre otros). La definición y los límites de cada uno de ellos, sin embargo, no está del todo clara (Murphy 2012; Schutt y O'Neil 2013) y, en algunas ocasiones, las técnicas que utilizan son muy similares —cuando no las mismas— (Murphy 2012).

Algunos autores consideran, por ejemplo, a la estadística y a la ciencia de datos básicamente como la misma disciplina; otros, en cambio, consideran tanto a la estadística como al *machine learning* como partes (y no las más importantes) de la ciencia de datos; y otros consideran que puede hacerse ciencia de datos perfectamente sin hacer uso de la estadística ni del *machine learning* (Donoho 2017).

En la práctica, quizá el *machine learning* se pueda relacionar más con la programación, la computación y la informática; la estadística, con los modelos probabilísticos y la inferencia; y la ciencia de datos, con todo lo que engloba a los datos en su conjunto como su tratamiento, depuración o visualización.

Si aceptamos la diferenciación anterior (aunque sea algo difusa), las partes principales que se suelen considerar para realizar un análisis de datos son[5]:

1. Depuración[6] y tratamiento de datos
2. Exploración y visualización de datos

[5] Cabe señalar que estos pasos no están estandarizados y algunos autores, en un intento de generalizar este proceso, hacen planteamientos distintos como Peng y Matsui 2016, pág. 6 o Schutt y O'Neil 2013

[6] Con depuración de datos (también conocida en otros ámbitos como preprocesamiento) se entiende la parte de revisión y detección de posibles valores erróneos que haya que corregir o completar, como valores perdidos (no disponibles) o valores que se salen de lo que cabría esperar (denominados *outliers* o valores atípicos).

3. Modelización e inferencia estadística
4. Interpretación y comunicación de resultados

Todos estos pasos, aunque se planteen ordenados, conviene señalar que nunca siguen un proceso lineal y están todos relacionados entre sí (Peng y Matsui 2016). La exploración de los datos (paso 2) puede detectar «fallos» y advertir de la necesidad de una depuración adicional de los datos (paso 1) o los resultados obtenidos de un modelo (paso 4) pueden llevar a replantear cambios en el mismo modelo (paso 3) —que a su vez puede necesitar algún cambio adicional en el formato de los datos (paso 1)—.

2.2 | Datos clínicos y medidas repetidas

De estos pasos descritos en el apartado 2.1 para el análisis de datos, por lo general, el que suele conllevar más tiempo es el primero de ellos: la depuración y el tratamiento de los datos. Esta dedicación y tiempo, además, puede ampliarse significativamente cuando se trata de datos reales de práctica clínica donde no se hace un control de calidad ni son revisados en busca de posibles errores como, por ejemplo, en un ensayo clínico.

En un ensayo clínico el investigador controla qué pacientes cumplen las condiciones necesarias para participar, qué variables se van a medir y durante cuánto tiempo se va a hacer el seguimiento. Los datos que se recogen en estos casos están muy controlados y, por lo general, hay pocos valores que queden sin completar. Además, los datos pasan por un proceso de revisión exhaustivo (denominado monitorización) para controlar que toda la información recogida se corresponde con la que aparece en los informes médicos (Gallin *et al.* 2017).

Los ensayos clínicos están considerados como las mejores pruebas científicas para apoyar un tratamiento y los que mejor reducen los posibles sesgos (Frieden 2017) y, aunque su coste (tanto en tiempo para seleccionar los pacientes como económico) es alto, permiten dar respuestas definitivas sobre el efecto de una intervención (Grimes y Schulz 2002). Su peso en el desarrollo de guías clínicas a nivel internacional, sin embargo, es pequeño, siendo mayor el aportado por estudios observacionales o las opiniones de los expertos (Fanaroff *et al.* 2019).

Los estudios observacionales, por el contrario, aunque ofrezcan menor nivel de prueba que los ensayos clínicos pueden ser de mucha utilidad para comprobar la efectividad y la seguridad de los tratamientos en la práctica clínica real (Frieden 2017). Este hecho ha provocado que en los últimos años los estudios observacionales hayan aumentado de forma exponencial (sobre todo aquellos que hacen uso de grandes bases de datos) y algunas sociedades y asociaciones internacionales recomienden su uso (Torp-Pedersen *et al.* 2019).

Uno de los problemas de los estudios observacionales donde se analizan datos provenientes de la práctica clínica real es que estos no pasan los controles y revisiones que pasarían en un ensayo clínico. Si estos datos, además, comprenden varios años con el objetivo de estudiar la evolución de una enfermedad, su complejidad y sus problemas se multiplican.

Para poder analizar una base de datos de práctica clínica con varios años de seguimiento de una determinada enfermedad las características y particularidades que se han de tener en cuenta son múltiples:

- valores perdidos tanto en el momento inicial como en los distintos seguimientos,
- medidas repetidas a lo largo del tiempo de cada paciente (datos longitudinales o también llamados de panel),
- medidas espaciadas irregularmente en el tiempo, dependiendo de cada paciente y de su mejoría o empeoramiento (ni son medidas equidistantes en el tiempo ni son el mismo número de medidas para cada paciente),
- medidas censuradas en el tiempo por fin de seguimiento o defunción (denominados datos de supervivencia y tiempos hasta un evento —*time-to-event* o *survival data*—),
- distintos tratamientos que se pueden ver modificados (tanto en medicación como en dosis) en el transcurso de la enfermedad o
- posibles eventos recurrentes relacionados con la progresión de la enfermedad (como rehospitalizaciones/reingresos).

2.3 | Métodos habituales en medicina y zona de confort

En investigación (y sobre todo en medicina) la estadística forma un papel destacado dentro de la metodología. Hoy en día, se nos hace casi impensable encontrar un artículo científico sin una parte de métodos estadísticos donde se expliquen todos los análisis que se han realizado. En cambio, los métodos que se utilizan más habitualmente siguen siendo métodos muy conocidos y estudiados como las pruebas *t* de Student, la pruebas χ^2 de Pearson o los modelos de regresión cuando se necesita ajustar por alguna variable como la edad o el sexo. En medicina, donde juega un papel importante el estudio de la supervivencia de los pacientes, destacan además métodos como las estimaciones de Kaplan-Meier o los modelos de riesgos proporcionales de Cox (dos métodos que se encuentran entre los más utilizados y citados de la historia de acuerdo con Noorden *et al.* 2014). Por otro lado, todos estos métodos se suelen estimar utilizando la estadística clásica (conocida como frecuentista), realizando contrastes de hipótesis e intervalos de confianza.

Para analizar datos complejos como nuestro caso y no está claro el método o el modelo más adecuado a utilizar, una práctica común es desglosar el problema o simplificarlo, de tal manera que se pueda analizar por partes o utilizar métodos conocidos como los comentados anteriormente. Así, por ejemplo, cuando se dispone de datos a lo largo de muchos años una opción es reducir todo el seguimiento a momentos concretos (considerar la fecha de ingreso, la de alta, pasados 6 meses o 1 año desde el reclutamiento, la fecha de muerte, etc.) o cuando se tienen distintos eventos (como reingresos y muerte) considerar el evento combinado de «primer evento».

Aunque estas «simplificaciones» de los datos son muy útiles y nos permiten avanzar en su comprensión, las estimaciones obtenidas de los análisis en algunas ocasiones pueden estar sobrestimadas (Pintilie 2011) o darnos una imagen sesgada de los resultados. Debido a esto, cada vez son más los investigadores que recomiendan salir de esta «zona de confort» y utilizar métodos avanzados que tengan en cuenta toda la información de los datos y su complejidad (Claggett *et al.* 2013).

2.4 | Modelos estadísticos

Un modelo estadístico es una representación matemática que hace uso de la probabilidad para establecer relaciones entre los datos e, idealmente, describir el proceso de generación de los mismos (Cox 2006). El objetivo principal de un modelo estadístico es representar la realidad para poder entenderla mejor y, así, predecir, decidir una acción o tomar una decisión sobre el mundo real (Hand 2014).

2.4.1. Modelos de regresión lineal

En la práctica del análisis de datos habitualmente no se dispone de una sola variable, si no que se dispone de multitud de información relacionada entre sí. Esto hace que las relaciones que hay que establecer (modelos) entre los datos se vuelvan complejas para poder representar mejor la realidad.

Regresión lineal simple

Uno de los modelos estadísticos más conocidos y utilizados es el modelo de regresión lineal simple para establecer la relación entre dos variables (una denominada variable «dependiente», «resultado» o «respuesta», y la otra, variable «independiente», «predictora» o «covariable»). Por ejemplo —y adelantando un poco nuestro caso concreto— si se quisiera explorar la relación entre un marcador de insuficiencia cardíaca y (como puede ser la FEVI) con la edad del paciente en el momento del diagnóstico podríamos establecer un modelo como

$$y_p = \alpha + \beta_{\mathsf{edad}}\mathsf{edad}_p + \varepsilon_p, \qquad \text{donde } \varepsilon_p \sim \text{Normal}(0, \sigma) \tag{2.1}$$

donde $p = 1, \ldots, N_p$ representa cada paciente (siendo N_p el número total de pacientes) e y_u y edad_p son los datos disponibles del marcador y la edad para cada uno de ellos.

Al escribir este modelo, al igual que con cualquier otro, estamos estableciendo una relación entre los datos (observados) y los parámetros (no observados). Y, al igual que todos los modelos, al establecer esta relación estamos asumiendo ciertas condiciones que hay que contemplar antes de utilizarlo en un caso concreto. Con este modelo (ver ecuación (2.1)), por ejemplo, estamos suponiendo, entre otras cosas:

1. que la relación entre los valores del marcador y y la edad es una relación lineal que parte de un valor aproximado de α (denominado «intercepto») para y y que aumenta un valor aproximado de β_{edad} por cada año (sea el valor de este mayor o menor),
2. que las observaciones son independientes entre sí y
3. que la desviación típica σ de todas ellas es común (lo que se conoce como homocedasticidad).

Los modelos estadísticos, dado que se utilizan distribuciones de probabilidad, se pueden reescribir de diversas formas haciendo uso de las reglas generales de la teoría de la probabilidad (Blitzstein y Hwang 2019). El modelo anterior (ecuación (2.1)), por ejemplo, puede escribirse como

$$y_p - \alpha - \beta_{\mathsf{edad}}\mathsf{edad}_p = \varepsilon_p \sim \text{Normal}(0, \sigma)$$

y de esta, teniendo en cuenta las propiedades de la distribución normal, se puede obtener la formulación equivalente

$$y_p \sim \text{Normal}(\alpha + \beta_{\mathsf{edad}}\mathsf{edad}_p, \sigma), \qquad \text{donde } p = 1, \ldots, N_p. \tag{2.2}$$

Regresión múltiple

En el caso de que se tengan otras variables que se necesite tener en cuenta en el modelo (por ejemplo, el sexo, que debe ser una variable de ceros y unos para poder introducirla en la fórmula) la regresión lineal simple se puede generalizar con cierta facilidad añadiendo términos a la suma, lo que da lugar a la denominada regresión múltiple[7]

$$y_p = \alpha + \beta_{\mathsf{edad}}\mathsf{edad}_p + \beta_{\mathsf{sexo}}\mathsf{sexo}_p + \varepsilon_p, \qquad \text{donde } p = 1, \ldots, N_p. \tag{2.3}$$

Al igual que la regresión lineal simple del apartado anterior hay múltiples maneras equivalentes para escribir este modelo. Cuando hay más covariables (no solo edad y sexo, como el ejemplo anterior) una forma útil es utilizar notación vectorial (o matricial)

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N_p} \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \end{pmatrix} + \begin{pmatrix} \beta_{\mathsf{edad}}\mathsf{edad}_1 + \beta_{\mathsf{sexo}}\mathsf{sexo}_1 \\ \beta_{\mathsf{edad}}\mathsf{edad}_2 + \beta_{\mathsf{sexo}}\mathsf{sexo}_2 \\ \vdots \\ \beta_{\mathsf{edad}}\mathsf{edad}_{N_p} + \beta_{\mathsf{sexo}}\mathsf{sexo}_{N_p} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{N_p} \end{pmatrix} =$$

$$= \alpha \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \mathsf{edad}_1 & \mathsf{sexo}_1 \\ \mathsf{edad}_2 & \mathsf{sexo}_2 \\ \vdots & \vdots \\ \mathsf{edad}_{N_p} & \mathsf{sexo}_{N_p} \end{pmatrix} \begin{pmatrix} \beta_{\mathsf{edad}} \\ \beta_{\mathsf{sexo}} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{N_p} \end{pmatrix}.$$

A partir de esta ecuación, si denotamos al vector de observaciones como $\mathbf{y} = (y_1, \ldots, y_{N_p})^\top$,

al vector de unos como $\mathbf{J_{N_p,1}}$, a la matriz con las covariables (edad y sexo) por columnas como $\mathbf{C}$, al vector de efectos de las covariables como $\boldsymbol{\beta} = (\beta_{\mathsf{edad}}, \beta_{\mathsf{sexo}})^\top$ y al vector de errores como $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{N_p})^\top$, entonces el modelo lineal se puede escribir como

$$\mathbf{y} = \alpha \mathbf{J_{N_p,1}} + \mathbf{C}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.4}$$

2.4.2. Modelos para datos longitudinales

Cuando se dispone de varias medidas para un mismo paciente —*i. e.*, varios valores del marcador para una misma persona pero medidos en distintos tiempos— se habla de medidas repetidas, datos longitudinales o datos de panel. En este caso, lo normal es que no se pueda asumir que las observaciones son independientes (los valores en un momento dado pueden estar afectados por el valor del mismo paciente en un momento anterior) y no sería adecuado aplicar un modelo de regresión lineal.

La opción más sencilla y directa sería, como hemos comentado anteriormente, simplificar los datos para poder usar algún método conocido (ver apartado 2.3). Por ejemplo, si se seleccionaran los valores iniciales y finales para cada paciente se podrían calcular las diferencias. Con esto tendríamos el cambio en el marcador para cada paciente y, en este caso sí, se podría estimar un modelo de regresión lineal.

Sumado al hecho de que con estas simplificaciones se pierde mucha información que puede ser relevante, el objetivo que se suele buscar cuando se dispone de datos longitudinales consiste en describir la trayectoria o la evolución del marcador principal, para lo que es necesario utilizar todos los datos disponibles.

Una de las estrategias más comunes es, partiendo de un modelo más simple como el modelo de regresión lineal, generalizarlo de tal modo que se tengan en cuenta las medidas repetidas.

Modelos de regresión multinivel

Una primera generalización sería estimar un modelo de regresión lineal independiente para cada uno de los pacientes. Así, por ejemplo, para un paciente cualquiera p que tenga n medidas del marcador y en distintos tiempos t podríamos definir un modelo como

$$y_j = \alpha + \beta t_j + \varepsilon_j, \qquad j = 1, \ldots, n$$

donde j representa cada observación. Así, el valor del marcador y para este paciente concreto empezaría (esto es, $t_1 = 0$) en un nivel de α y aumentaría una cantidad β por cada unidad de tiempo.

[7] Cabe mencionar que esta terminología es la común en matemáticas y estadística; sin embargo, en la estadística aplicada a medicina es habitual encontrar el término «multivariante» para referirse a estos modelos, lo que crea cierta confusión con aquellos con múltiples variables dependientes $(y, z) = \alpha + \beta x + \varepsilon$. Aquí seguiremos la terminología estándar en matemáticas y utilizaremos «multivariante» para referirnos a estos últimos modelos.

El punto clave en este modelo es cómo generalizarlo a todos los pacientes. Las variables y y t se podrían escribir en el modelo como $y_{p,j}$ y $t_{p,j}$ para indicar que corresponden a las observaciones j-ésimas de cada paciente p. Pero cada uno de los parámetros α y β se podría generalizar de varias formas. Centrándonos solo en α podríamos:

1. considerar para cada paciente p un valor independiente α_p

$$y_{p,j} = \alpha_p + \beta t_{p,j} + \varepsilon_{p,j},$$

2. considerarlo común a todos los pacientes, que se podrían dejar como α,

$$y_{p,j} = \alpha + \beta t_{p,j} + \varepsilon_{p,j},$$

3. o bien considerar que viene determinado por una distribución de probabilidad; así, por ejemplo, todos los pacientes podrían partir de un valor medio (α_p) pero alejándose cada uno más o menos de él,

$$\begin{aligned} y_{p,j} &= \alpha_p + \beta t_{p,j} + \varepsilon_{p,j}, & \varepsilon_{p,j} &\sim \text{Normal}(0, \sigma) \\ \alpha_p &= \alpha + \widetilde{\varepsilon}_{p,j}, & \widetilde{\varepsilon}_{p,j} &\sim \text{Normal}(0, \widetilde{\sigma}). \end{aligned}$$

Estos últimos modelos —aunque muy simplificados— son lo que se conocen como modelos multinivel[8] y, además de permitir tener en cuenta el carácter longitudinal de los datos, permiten definir modelos más complejos que puedan representar mejor el proceso de generación de los datos: los parámetros α y β pueden depender de otras covariables como sexo y edad, estar relacionadas entre sí, etc. Además, no necesitan que los pacientes tengan el mismo número de medidas ni que estas estén espaciadas de manera regular en el tiempo (Rizopoulos 2012, pág. 16).

Una forma sencilla de generalizar este modelo es escribiéndolo como

$$y_{p,j} = \mu_{p,j} + \varepsilon_{p,j}, \qquad \varepsilon_{p,j} \sim \text{Normal}(0, \sigma), \tag{2.5}$$

donde a $\mu_{p,j}$ se le suele denominar función trayectoria del modelo (Ibrahim, Chu *et al.* 2010).

Esta cantidad $\mu_{p,j}$ se puede especificar utilizando la variable de tiempo t de múltiples maneras, pudiendo variar mucho según los casos contemplados por cada autor —donde se pueden distinguir los vectores de efectos aleatorios de los de efectos fijos o se pueden tener en cuenta los distintos tipos de censura posibles— (Rizopoulos 2012; Brilleman, Crowther *et al.* 2018; Papageorgiou *et al.* 2019).

[8] Estos modelos también se conocen como modelos jerárquicos, modelos mixtos, modelos de efectos aleatorios, modelos anidados o modelos de coeficientes aleatorios. También es habitual encontrarlos como modelos de efectos aleatorios y fijos, debido a que se puede dejar algún parámetro «fijo» —como el parámetro β en nuestro ejemplo—.

Referencias y *software*

Para profundizar más en los modelos multinivel destacamos los libros de Molenberghs y Verbeke (2000), Gelman y Hill (2006, parte 2A) y Goldstein (2010, capítulo 5). En R las librerías más utilizadas para estimar modelos de este tipo son lme4 (D. Bates *et al.* 2015) y nlme (Pinheiro y D. M. Bates 2000).

2.4.3. Modelos para datos de supervivencia

En medicina, como ya se ha comentado en el apartado 2.3, uno de los aspectos que se estudian con mayor interés es la mortalidad. A la hora de analizar un evento como la muerte hay varias particularidades a tener en cuenta:

- si el seguimiento que se hace de los pacientes para determinar si han fallecido o no es muy corto puede ser que ninguno de ellos lo haya hecho y
- si, en cambio, el seguimiento es suficientemente largo todos los pacientes terminarán por sufrirlo.

En la práctica, casi siempre se va a estar entre esos dos extremos, con pacientes que han sufrido el evento y otros que todavía no lo han sufrido. Este tipo de datos se denominan datos censurados[9] (Leung *et al.* 1997) y para analizarlos se debe considerar tanto si el evento se ha producido como el tiempo que ha tardado en producirse. En aquellos pacientes en los que no se haya producido el evento, al menos, se debe conocer el tiempo de seguimiento —si un paciente lleva dos años vivo el tiempo hasta la muerte será mayor de dos años—.

El estudio y la modelización de este tipo de datos donde se tiene si ha ocurrido un evento y en qué tiempo lo ha hecho (si lo ha hecho), dadas las particularidades asociadas, han dado lugar a toda una rama dentro de la estadística denominada «análisis de supervivencia» (Clark *et al.* 2003; Kleinbaum 2012).

El análisis de supervivencia se modeliza a través de dos funciones: la función de supervivencia, que se denota como $S(t)$, y la función de riesgo, que se denota como $h(t)$, donde t representa la variable tiempo. La función de supervivencia $S(t)$ para cada tiempo concreto t indica la probabilidad de que una persona sobreviva al menos hasta el tiempo t. De forma opuesta —ya que, se centra más en el hecho de que se produzca el evento—, la función de riesgo indica la probabilidad instantánea que tiene el evento de producirse en ese momento concreto t, dado que el paciente ya haya sobrevivido hasta ese tiempo t.

Denominando T a la variable aleatoria que representa los tiempos de supervivencia de los individuos estas funciones vienen determinadas por las ecuaciones

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F_T(t),$$

[9] Este tipo de censura se denomina censura por la derecha — ya que, se produce solamente en el seguimiento— y es en la que nos centraremos en este documento. Existen otros tipos de censura como la censura por la izquierda (producida en el tiempo inicial) o en un intervalo (cuando se producen ambas a la vez, por la derecha y por la izquierda).

$$h(t) = \lim_{\varepsilon \to 0} \left(\frac{P(t < T \leq t + \varepsilon \mid T > t)}{\varepsilon} \right)$$

donde $F_T(t)$ es la función de distribución acumulada de T para cada tiempo t.

Así definidas, estas funciones cumplen lo esperable para lo que representan, como que la supervivencia al inicio es total, $S(0) = 1$, que la supervivencia es decreciente (a mayor tiempo menor supervivencia) o que, si pasa el tiempo suficiente, la supervivencia será nula, $S(+\infty) = 0$ (Kleinbaum 2012).

Partiendo de la función de riesgo $h(t)$ y teniendo en cuenta distintas reglas de probabilidad y de análisis matemático se puede obtener la relación $h(t) = p(t)/S(t)$. A partir de esta, simplemente reescribiendo la ecuación, se obtiene la función de densidad $p(t) = h(t)S(t)$ (ver Collett 2015, pág. 12).

En este punto se podría pensar que para modelizar la supervivencia basta con modelizar las funciones $h(t)$ y $S(t)$. Sin embargo, es suficiente con modelizar una de ellas, ya que, una se puede obtener a partir de la otra.

Esto se puede probar partiendo de la definición de $S(t)$, con la que podemos llegar, tomando derivadas a ambos lados, a

$$S(t) = 1 - F_T(t) \quad \Rightarrow \quad S'(t) = -F'_T(t) = -p(t)$$

donde F' denota a la derivada. Usando esta igualdad junto con la la relación $p(t) = h(t)S(t)$, se obtienen

$$h(t) = \frac{p(t)}{S(t)} = \frac{-S'(t)}{S(t)}$$

y

$$H(t) = \int_0^t h(u)\, du = -\int_0^t \frac{S'(t)}{S(u)}\, du = -\log(S(t)), \tag{2.6}$$

estableciendo la relación entre ambas funciones $S(t)$ y $h(t)$. $H(t)$ denota a la función de riesgo acumulada y puede interpretarse como el riesgo del evento hasta el tiempo t, teniendo en cuenta que todavía no ha ocurrido hasta ese tiempo.

A partir de la ecuación (2.6) se puede escribir la función de supervivencia como

$$S(t) = \exp\left(-\int_0^t h(u)\, du \right) = \exp(-H(t)). \tag{2.7}$$

Función de verosimilitud para modelos de supervivencia

Cabe destacar que la censura juega un papel fundamental a la hora de estimar, a partir de unos datos concretos, los parámetros que mejor se ajustan a ellos. Este proceso se lleva a cabo utilizando la conocida como «función de verosimilitud» (*likelihood function*, en inglés). Esta función representa lo verosímiles que son los datos dependiendo de los parámetros y se verá más en detalle en el apartado 2.5.

Así, la función de verosimilitud para la supervivencia informa sobre la probabilidad de observar los datos que, en este caso, son los tiempos de supervivencia t_p para cada paciente p (tiempo hasta el evento o hasta la censura),

$$t_p = \text{mín}(t_p^{(\text{evento})}, t_p^{(\text{censura})}),$$

y la variable, e_p, que indica si ha ocurrido el evento o no

$$e_p = \begin{cases} 1 & \text{si el evento se ha producido para el paciente } p \\ 0 & \text{en caso de censura.} \end{cases}$$

Cuando se observa el evento para un paciente concreto su contribución a la verosimilitud viene determinada por su distribución $p(t_p) = h(t_p)S(t_p)$. Pero cuando el evento está censurado, la única información que se tiene es que el paciente ha sobrevivido hasta ese tiempo, por lo que su contribución solo es de $S(t_p)$ (Rizopoulos 2012). De esta manera, la función de verosimilitud para los datos (t_p, e_p) de un paciente p se puede unificar y escribir en una sola función como

$$p(t_p, e_p|h, S) = \left.\begin{cases} h(t_p)S(t_p) & \text{en caso de evento, } e_p = 1 \\ S(t_p) & \text{en caso de censura, } e_p = 0 \end{cases}\right\} = h(t_p)^{e_p} S(t_p). \tag{2.8}$$

Modelos de riesgos proporcionales

Cuando solo se dispone de los datos de supervivencia y no hay más variables que se necesite relacionar o estudiar cómo le afectan —o, como mucho, solo hay dos grupos a comparar— existen métodos para aproximar las funciones de supervivencia y riesgo. Entre estos métodos el más conocido es el estimador de Kaplan-Meier (Collett 2015, pág. 32). Sin embargo, cuando se quieren tener en cuenta otras variables como el sexo o la edad, hay que recurrir a modelos estadísticos para describir la función $h(t)$, ya que, como hemos visto en el capítulo anterior, $S(t)$ vendrá determinada a partir de esta.

Aunque hay muchos modelos para la función de riesgo uno de los más utilizados es el modelo de riesgos proporcionales de Cox (Collett 2015, pág. 57). En este modelo la función de riesgo $h(t)$ se separa en dos factores: una función de riesgo basal —que denotaremos como $\lambda(t)$— y el efecto de las covariables que hacen que esta función aumente o disminuya. Así, utilizando directamente la notación vectorial (como en la ecuación (2.4)) un modelo con vector de covariables $\mathbf{C}$ y con un vector de efectos $\boldsymbol{\gamma}$ se puede escribir como

$$h(t) = \lambda(t) \exp(\mathbf{C}\boldsymbol{\gamma}). \tag{2.9}$$

El uso de la exponencial en la parte de las covariables se debe al hecho que el riesgo debe ser positivo, por lo que en lugar de interpretar los coeficientes γ (como en la regresión lineal) se suelen interpretan los términos $\exp(\gamma)$, que se conocen como *hazard ratios* (o HR).

El modelo de Cox solo ofrece estimaciones de los parámetros γ y deja sin definir la función de riesgo basal $\lambda(t)$. Por esta razón, se enmarca dentro de los conocidos como métodos semiparámetricos. En los casos en que se requiera una estimación de la función de riesgo se debe estimar esta función de riesgo basal. Los métodos que la determinan explícitamente se conocen como paramétricos.

Referencias y *software*

Una introducción al análisis y los modelos de supervivencia la podemos encontrar en los artículos de Bradburn *et al.* (2003) y Clark *et al.* (2003). Para profundizar más recomendamos los libros de Collett (2015) y Kleinbaum (2012) y, desde una perspectiva bayesiana (que veremos en el apartado 2.5.2), destacamos el libro de Ibrahim, Chen *et al.* (2001). Para la estimación de estos modelos en R hay numerosas librerías. Las más destacables se recogen en el *task view* sobre supervivencia de R (Allignol y Latouche 2022), donde la más usada es survival (Therneau 2022; Therneau y Grambsch 2000).

2.4.4. Modelos para datos longitudinales y de supervivencia

En los estudios longitudinales ocurre habitualmente que el seguimiento del paciente finaliza debido a un evento terminal como la muerte. En estos casos —y, sobre todo, cuando el marcador es un indicador de la progresión de la enfermedad— cabría preguntarse en qué medida los cambios y la evolución de esos marcadores observados a lo largo del tiempo se asocian con un riesgo mayor de muerte.

Con ese objetivo, se hace necesario utilizar algún modelo estadístico que utilice ambas fuentes de información a la vez: los datos longitudinales y los datos de supervivencia.

Joint models para datos longitudinales y de supervivencia

Para poder analizar estos datos conjuntamente, en los últimos años han ganado mucha importancia los conocidos como *joint models* para datos longitudinales y de supervivencia[10] (Papageorgiou *et al.* 2019; Rizopoulos 2012).

La idea principal detrás de los *joint models* es la unión y estimación conjunta de dos modelos: un modelo para la parte longitudinal de los datos, como el modelo multinivel (ecuación (2.5)), y un modelo para los datos de supervivencia, como el modelo de riesgos proporcionales (ecuación (2.9)). La forma básica de unir ambos modelos es mediante el uso de un parámetro común como la trayectoria del modelo multinivel, μ.

[10] Al mencionar estos modelos es conveniente remarcar qué tipo de datos se analizan en conjunto, ya que, con el término «joint model» se pueden encontrar otros modelos para analizar datos de un tipo diferente. Un ejemplo lo podemos encontrar en Lu (2017).

Poniendo ambos submodelos en forma vectorial una manera simplificada de escribir un *joint model* es

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu} + \varepsilon \\ h(t) &= \lambda(t)\exp(\mathbf{C}\boldsymbol{\gamma} + \gamma_\mu \boldsymbol{\mu}). \end{aligned} \tag{2.10}$$

donde se sigue la notación utilizada en ambos submodelos (longitudinal y de supervivencia), se deja $\boldsymbol{\mu}$ sin especificar y solo cambian los parámetros $\boldsymbol{\gamma}$, que representa los efectos de las covariables, y γ_μ, que denota el efecto de la trayectoria sobre el riesgo de evento.

Las ventajas de estos modelos conjuntos en lugar de estimar los dos submodelos por separado está en explorar la relación entre los niveles del marcador y (mediante su evolución media μ) y la supervivencia, lo que permite obtener mejores estimaciones (Ibrahim, Chu *et al.* 2010; Hickey *et al.* 2018).

Referencias y *software*

La gran atención que han recibido estos modelos en los últimos años ha hecho que cada vez sean más los artículos y libros donde se utilizan y mayor la cantidad de librerías y paquetes disponibles para analizarlos.

Por la visión general que ofrece, para profundizar más en estos modelos, recomendamos las revisiones realizadas por Hickey *et al.* (2018) y Papageorgiou *et al.* (2019). Otras referencias recomendables son Rizopoulos (2012) y Brilleman, Crowther *et al.* (2018), donde se utiliza una perspectiva bayesiana.

Para la estimación de *joint models* en la misma revisión de Papageorgiou *et al.* (2019) se puede encontrar un breve resumen de distintos paquetes y librerías. Para el *software* estadístico R podemos destacar JM (Rizopoulos 2010), JMBayes (Rizopoulos 2016) o joineR (Williamson *et al.* 2008; Philipson *et al.* 2018). Recientemente, además, el paquete rstanarm (Goodrich *et al.* 2022) ha añadido funcionalidades que permiten estimarlos de forma bayesiana haciendo uso del software Stan (Stan Development Team 2021).

2.5 | Inferencia estadística

Un modelo estadístico, como hemos visto en la sección anterior, establece una relación entre unos datos observados y unos parámetros que determinan las funciones de distribución. La relación funciona en ambos sentidos, en el sentido de que si se conociera exactamente los valores de los parámetros la generación de datos con esa distribución se denomina simulación de datos (ver figura 2.1, izquierda). Recíprocamente, disponiendo de los datos, el cálculo de los parámetros que definen la distribución se denomina inferencia estadística (ver figura 2.1, derecha).

Dado un conjunto de observaciones o datos, el problema principal que se plantea es que son muchos los posibles valores de los parámetros de un modelo estadístico, por lo que ¿cómo

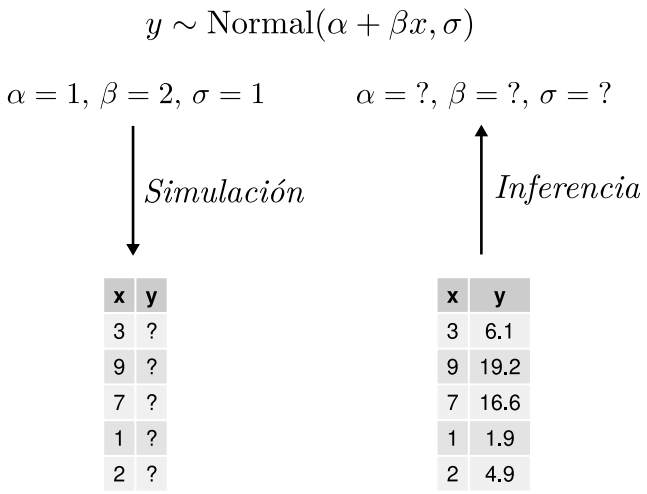

Figura 2.1. *Diagrama de simulación de datos e inferencia a partir de un modelo estadístico. Como ejemplo se ha establecido un modelo de regresión lineal simple (ver apartado 2.4.1). En la simulación de datos (izquierda), dados los parámetros α, β y σ, se calculan (simulan) posibles valores de la variable y para los valores observados de x. En la inferencia estadística (derecha), dados los pares de valores observados x e y, se estiman los valores de los parámetros α, β y σ.*

decidir entre unos y otros? La inferencia estadística trata de resolver esta cuestión marcándose objetivos como: calcular intervalos o conjuntos de valores donde sea más probable que se sitúen los parámetros y evaluar la consistencia de los datos para valores concretos de los parámetros (Cox 2006, pág. 7).

En la inferencia estadística existen dos grandes enfoques o escuelas enfrentadas: la frecuentista (o clásica) y la bayesiana. Sobre los problemas y ventajas de cada una y sobre cuál es mejor de las dos se ha escrito y discutido ampliamente (Wagenmakers *et al.* 2008; Vallverdú 2016; Zyl 2018).

En general, las críticas al enfoque frecuentista se suelen centrar en el mal uso y malinterpretación del p-valor, en la significancia estadística y en los contrastes de hipótesis (ver apartado 2.5.1). Al enfoque bayesiano, por otro lado, se le critica una mayor subjetividad y una mayor exigencia computacional (ver apartado 2.5.2).

Ambos paradigmas parten de un modelo estadístico y utilizan la función de verosimilitud que este determina. Esta función, como hemos mencionado en el apartado 2.4.3, representa lo creíbles que son los datos para unos parámetros fijos y se define mediante la función de probabilidad (Held y Sabanés Bové 2014, pág. 14). Si se denota al conjunto de parámetros con $\boldsymbol{\theta}$ y con L a la función de verosimilitud esta se puede escribir como

$$L(\boldsymbol{\theta}) = p(datos|\boldsymbol{\theta}).$$

El objetivo de la inferencia es el inverso, habiendo obtenido los datos poder obtener información sobre los parámetros.

2.5.1. Inferencia frecuentista

La inferencia frecuentista considera que los parámetros son unos valores fijos pero desconocidos y lo único que se tiene son unos datos medidos con mayor o menor error. Si se repitiera el proceso de tomar datos (repetir el experimento) entonces a largo plazo la frecuencia que se observaría en los datos se aproximaría a la «real» estipulada por los parámetros, esto es, a su probabilidad (Lambert 2018, pág. 17).

Este concepto de probabilidad, ya que los parámetros son vistos como valores fijos, hace que no tenga sentido hablar de la probabilidad de un parámetro y no sea posible decir cómo de probable es que se encuentre entre ciertos valores (intervalo). En este contexto se plantea bajo qué parámetros es más probable obtener nuestros datos o calcular la probabilidad de que un intervalo contenga el valor del parámetro. Así, algunas de las técnicas centrales de la estadística frecuentista son:

- la estimación puntual de parámetros por máxima verosimilitud, que calcula bajo qué parámetros puntuales se consigue la mayor probabilidad de nuestros datos,
- los contrastes de hipótesis (conocidos en la literatura por sus siglas NHST, del inglés *null hypothesis significance testing*), que se basan en enfrentar dos hipótesis sobre los parámetros —denominadas hipótesis nula (H_0) e hipótesis alternativa (H_1)— y calculan la probabilidad de error al tomar una decisión a partir de los datos observados, y
- los intervalos de confianza, que calculan un rango de valores admisibles donde el parámetro esté incluido si se repitiera el experimento múltiples veces y que lo contendrán en al menos un determinado porcentaje.

Los más utilizados son los contrastes de hipótesis, donde se calcula el conocido como p-valor. El p-valor se define como la probabilidad de obtener unos datos como los nuestros (o «más extremos», en el sentido de la hipótesis alternativa H_1) partiendo o suponiendo como cierta la hipótesis nula H_0.

Así, cuando este p-valor es pequeño (habitualmente menor a 0.05), se concluye que, puesto que la probabilidad de obtener nuestros datos bajo la suposición de que se cumple la hipótesis nula es muy baja, entonces la hipótesis nula debe ser falsa.

El p-valor se ha visto que genera gran confusión en los investigadores, que tienden a resumir los resultados en base solo a este valor, a su significancia y a hablar sobre la existencia o no de un efecto (Nuzzo 2014; Gelman y J. Carlin 2017; O'Connor 2019). La mala interpretación y el mal uso del p-valor ha provocado multitud de críticas durante más de veinte años (Cohen 1994; Hauer 2004; Zyl 2018) y a generar una crisis de reproducibilidad en la ciencia (Ioannidis 2005; Baker 2016), llegando la Sociedad Americana de Estadística a publicar un comunicado oficial en 2016 sobre los peligros asociados al p-valor (Wasserstein *et al.* 2019).

Las recomendaciones que vienen planteando muchos investigadores pasan por abandonar la significancia (Leek *et al.* 2017; Amrhein *et al.* 2019), directamente no utilizar los p-valores

(Wasserstein *et al.* 2019) o reemplazarlos por otros métodos como los bayesianos (Nuzzo 2014; Wasserstein *et al.* 2019). Estos cambios, sin embargo, aunque cada vez es más habitual ver la utilización de otros métodos distintos de los contrastes de hipótesis, se producen lentamente (Matthews 2021).

2.5.2. Inferencia bayesiana

La inferencia bayesiana, al contrario que la frecuentista, considera la probabilidad como una medida de certeza en general que nos ayuda a expresar la incertidumbre sobre el valor de los parámetros. Así, lo único que la estadística bayesiana considera como seguro y conocido son los datos que pueden venir con mayor o menor probabilidad de una población con unos parámetros u otros (Lambert 2018).

La estadística bayesiana hace uso del conocido teorema de Bayes para, dadas las observaciones o los datos, estimar las probabilidades de los parámetros (Held y Sabanés Bové 2014, pág. 168). Este teorema se obtiene de la definición de probabilidad condicionada y se puede formular (siguiendo la misma notación de este apartado) como

$$p(\boldsymbol{\theta}|datos) = \frac{p(\boldsymbol{\theta}, datos)}{p(datos)} = \frac{p(\boldsymbol{\theta})p(datos|\boldsymbol{\theta})}{p(datos)} \propto p(\boldsymbol{\theta})p(datos|\boldsymbol{\theta}) \tag{2.11}$$

donde el símbolo «$\propto$» significa «proporcional a», y se puede obtener debido a que el término $p(datos)$ del denominador, conocido como constante de normalización, no depende de los parámetros.

El teorema de Bayes establece así la relación entre tres cantidades principales —junto con la constante de normalización— (Schoot *et al.* 2021):

- $p(\boldsymbol{\theta})$, que se denomina probabilidad a *priori* y representa la probabilidad de los valores de los parámetros para un determinado modelo en la población,
- $p(datos|\boldsymbol{\theta})$, que corresponde a la función de verosimilitud vista anteriormente y que representa la probabilidad condicional de los datos observados dados los parámetros del modelo, y
- $p(\boldsymbol{\theta}|datos)$, denominada probabilidad a *posteriori* y que representa la probabilidad de los parámetros conocidos los datos.

Al establecer un modelo estadístico lo que estamos definiendo es su función de verosimilitud —esto es, la forma en que se relacionan los datos con los parámetros— por lo que definiendo además las distribuciones para todos los parámetros del modelo quedaría establecido el modelo de probabilidad total, que es la distribución de probabilidad conjunta para todas las cantidades del modelo. Con esta función, por el teorema de Bayes, se puede calcular la probabilidad a *posteriori*, $p(\boldsymbol{\theta}|datos)$. Esta es la distribución que juega el papel principal en la inferencia bayesiana, ya que, a partir de ella se pueden calcular estadísticos de interés como la media o intervalos de credibilidad para los parámetros (Gelman, J. B. Carlin *et al.* 2013).

Para muchos investigadores, en contraste con la inferencia frecuentista, la inferencia bayesiana es una aproximación lógica y coherente que permite obtener conclusiones más cercanas a la intuición (Vallverdú 2016; Zyl 2018). Uno de los ejemplos habituales son los intervalos de credibilidad, cuya interpretación (valores entre los que se situará el parámetro con cierta probabilidad) se acerca más a la interpretación que erróneamente se le suelen dar a los intervalos de confianza.

Entre las ventajas de la estadística bayesiana están su mayor flexibilidad para la definición de los modelos y sus mejores resultados a la hora de estimar modelos multinivel más complejos (Gelman, J. B. Carlin *et al.* 2013). Una crítica importante que recibe es su subjetividad respecto a las distribuciones a *priori*, aunque otros autores ven este aspecto como una ventaja puesto que permite introducir en el modelo el conocimiento previo que podamos tener y no consideran que ninguna esté libre de toda subjetividad (Hoijtink *et al.* 2008, sección 9.3).

El problema principal de la inferencia bayesiana reside en la normalización para realizar el cálculo de la distribución a *posteriori*. Esta constante de normalización sirve para que la probabilidad esté bien definida. Por el teorema de la probabilidad total esta se puede escribir como[11]

$$p(datos) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(datos|\boldsymbol{\theta})$$

En algunos casos, la elección de una determinada distribución a *priori* y de una función de verosimilitud resultan en una distribución a *posteriori* que se puede calcular analíticamente[12]. Sin embargo, en general, esto no es posible y esta integral se convierte en intratable, lo que impide evaluar la distribución a *posteriori* (Held y Sabanés Bové 2014; Schoot *et al.* 2021).

La intratabilidad de esta integral ha sido la principal razón de que la estadística bayesiana haya sido descartada históricamente por muchos investigadores en favor de la frecuentista (Schoot *et al.* 2021). En la actualidad, sin embargo, los avances en técnicas de muestreo de la probabilidad como los métodos de Montecarlo basados en cadenas de Markov (MCMC, del inglés *Markov chain Montecarlo*)[13], permiten indirectamente la inferencia de la distribución a *posteriori* mediante la realización de simulaciones (Schoot *et al.* 2021).

En la práctica son varios los lenguajes de programación probabilísticos que implementan algoritmos concretos de muestreo basados en los métodos de MCMC. Algunos ejemplos son el algoritmo de Metrópolis-Hastings, en el que se basa el método de muestreo de Gibbs, o los métodos hamiltonianos de Montecarlo. La principal ventaja de estos lenguajes es su flexibilidad para la definición de modelos estadísticos complejos, haciendo relativamente sencilla su especificación e implementación. La mayor desventaja, por otro lado, es que dejan muchas opciones computacionales para construir y evaluar los modelos, lo que ha llevado a que una línea de investigación activa sea la propuesta de *workflows* de trabajo con pasos

[11] En el caso discreto se escribiría como $p(datos) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(datos|\boldsymbol{\theta})$.

[12] Ver distribuciones a *priori* conjugadas en Held y Sabanés Bové (2014, pág. 179) para más información.

[13] Una visualización interactiva de los diferentes métodos de muestreo MCMC se puede encontrar en la web https://chi-feng.github.io/mcmc-demo/.

recomendados para mejorar la construcción, la evaluación, la validación y el refinamiento de los modelos (Betancourt 2020; Gabry, Simpson *et al.* 2019; Gelman, Vehtari *et al.* 2020).

Referencias y *software*

Para una introducción práctica a la estadística bayesiana destacamos los libros de Gelman, J. B. Carlin *et al.* (2013), Lambert (2018) y McElreath (2020) y el artículo de Schoot *et al.* (2021). Otros libros que profundizan más en las bases teóricas de la inferencia son Held y Sabanés Bové (2014) y Cox (2006).

Un buen resumen del software enfocado en la estadística bayesiana se puede encontrar en la tabla 2 de Schoot *et al.* (2021) entre el que podemos destacar JAGS (Plummer 2003, 2022), OpenBUGS (Lunn *et al.* 2012; Spiegelhalter *et al.* 2014), R-INLA (Lindgren y Rue 2015) o Stan (Carpenter *et al.* 2017; Stan Development Team 2021).

Dado que en este trabajo el lenguaje probabilístico que utilizaremos para especificar y evaluar nuestros modelos estadísticos es Stan dedicamos el siguiente apartado a hacer un breve repaso sobre su estructura y utilización.

2.5.3. Stan

Stan (Stan Development Team 2021) es un lenguaje de programación probabilístico que permite estimar modelos estadísticos mediante inferencia bayesiana. Esto lo consigue al tener implementados métodos hamiltonianos de Montecarlo, como el *No-U-Turn sampler* (NUTS) (Betancourt 2017), que pertenecen a la familia de métodos de MCMC para el muestreo de distribuciones de probabilidad. Estos métodos son más eficientes y robustos que el muestreo de Gibbs o el algoritmo Metrópolis-Hastings para modelos complejos (Carpenter *et al.* 2017).

Cuando Stan ejecuta un modelo estadístico evalúa el logaritmo de la función de probabilidad para un conjunto dado de parámetros (Stan Development Team 2021, sección 7.3).

Al trabajar con múltiples observaciones y_i (con $i = 1, \ldots, n$) de una variable y la función de verosimilitud que nos define el modelo es la multiplicación de todas las probabilidades de las observaciones (Held y Sabanés Bové 2014)

$$L(\boldsymbol{\theta}) = p(\textit{datos}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}) = p(y_1, \ldots, y_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}).$$

Si esta ecuación además se multiplica por las funciones de distribución de los parámetros tendríamos establecida la función de probabilidad completa de nuestro modelo.

En Stan, al trabajar con logaritmos, todas estas multiplicaciones se convierten en sumas, y cada sentencia de probabilidad incrementa el logaritmo de la probabilidad total, que en Stan se representa por `target`.

De esta manera, en Stan, para definir un modelo estadístico donde una variable y sigue una distribución normal se puede utilizar la notación con distribuciones (código 2.1) o, de manera equivalente, se puede incrementar el logaritmo de la función de probabilidad (código 2.2).

```
y ~ normal(mu, sigma);
```

Código 2.1. **Stan**. Sentencia de Stan con distribución de probabilidad.

```
target += normal_lpdf(y | mu, sigma);
```

Código 2.2. **Stan**. Sentencia de Stan con incremento del logaritmo de la probabilidad.

Estructura

Un programa en Stan se compone de ocho bloques diferentes (ninguno obligatorio) donde se puede especificar (separando en partes) un modelo estadístico a través de su función de probabilidad total (Stan Development Team 2021, apartado 8).

En cada uno de estos bloques se especifican los datos y observaciones que necesita el modelo (bloque `data`), los parámetros que utiliza (bloque `parameters`) o las sentencias de probabilidad, donde se especifican las distribuciones a *priori* y la función de verosimilitud (bloque `model`). Todos estos bloques, además de los que no hemos comentado (como el de `functions` que se utiliza para definir funciones adicionales), se pueden ver en el código 2.3.

```
functions {
  // definición de funciones
}
data {
  // datos de entrada del modelo
}
transformed data {
   // definición de constantes y transformación de datos
}
parameters {
   // parámetros del modelo
}
transformed parameters {
   // transformación de parámetros
}
model {
   // declaraciones del modelo y función log de probabilidad
}
generated quantities {
   // cálculos basados en datos y parámetros
}
```

Código 2.3. **Stan**. Estructura y esqueleto de un modelo Stan.

Modelos

Como se ha visto en el apartado anterior, para aplicar los métodos bayesianos es necesario establecer la función de distribución total, esto es, la función de verosimilitud y las distribuciones a *priori* de todos los parámetros.

Si recuperamos el ejemplo del modelo lineal visto en la ecuación (2.3) para relacionar los valores de un marcador y en una muestra de pacientes con su edad en el diagnóstico y su sexo, esta podría reescribirse como

$$\begin{aligned} \mathsf{y}_p &\sim \text{Normal}(\alpha + \beta_{\mathsf{edad}}\mathsf{edad}_p + \beta_{\mathsf{sexo}}\mathsf{sexo}_p, \sigma) \\ \alpha &\sim \text{Normal}(0, 1) \\ \beta_{\mathsf{edad}} &\sim \text{Normal}(0, 1) \\ \beta_{\mathsf{sexo}} &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{Cauchy}(0, 1)\big|_{x>0} \end{aligned} \tag{2.12}$$

donde, por simplicidad para el ejemplo, hemos utilizado distribuciones normales de media 0 y varianza 1 para los parámetros β y una Cauchy truncada a valores positivos para la varianza.

En este ejemplo se definen explícitamente las distribuciones de probabilidad de cada variable y cada parámetro (distribuciones a *priori*), se puede observar cómo pasar la ecuación (2.12) directamente a Stan, donde el único paso adicional necesario es especificar los datos que debe utilizar (ver código 2.4).

```
data {
  int<lower=0> N_p;
  vector[N_p] y;
  vector[N_p] edad;
  vector[N_p] sexo;
}
parameters {
  real alpha;
  real beta_edad;
  real beta_sexo;
  real<lower=0> sigma;
}
model {
  alpha ~ normal(0, 1);
  beta_edad  ~ normal(0, 1);
  beta_sexo ~ normal(0, 1);
  sigma ~ cauchy(0, 1);
  y ~ normal(alpha + beta_edad * edad + beta_sexo * sexo, sigma);
}
```

Código 2.4. **Stan**. Modelo lineal para un marcador y con edad y sexo como covariables.

CmdStanR

Para analizar los resultados que se obtienen al estimar un modelo estadístico con Stan lo más habitual es usar un lenguaje de programación como Python, R, Stata o Julia, entre otros. Por esta razón, una opción es ejecutar los modelos de Stan y obtener los resultados sin salir del lenguaje utilizado a través de paquetes conocidos como interfaces. Para Stan existen múltiples interfaces: CmdStan para la línea de comandos, RStan y CmdStanR para R, PyStan y CmdStanPy para Python, MatlabStan para MATLAB, Stan.jl para Julia o StataStan para Stata[14].

En este trabajo, puesto que los análisis de datos los llevaremos a cabo en R, en lugar de RStan (Stan Development Team 2022), utilizaremos la interfaz CmdStanR (Gabry y Češnovar 2022). Esta interfaz es más reciente que RStan y tiene como principal ventaja respecto a esta que utiliza la versión más reciente de Stan y permite —haciendo uso de la clase R6 (Chang 2022) para definir objetos en R— una experiencia de uso similar y unas implementaciones comunes entre las distintas interfaces (como CmdStanPy, de Python).

Por ejemplo, si se supone que el modelo especificado en el código 2.4 se tiene guardado en un archivo de texto con nombre `modelo_lineal_fevi.stan` este se puede estimar desde R con CmdStanR ejecutando

```
library("cmdstanr")
datos_fevi <- list(N_p = 6,
                   y = c(40, 35, 37, 29, 38, 30),
                   edad = c(57, 63, 60, 72, 74, 55),
                   sexo = c(0, 0, 1, 0, 1, 1))
mod <- cmdstan_model("modelo_lineal_fevi.stan")
fit <- mod$sample(data = datos_fevi, seed = 123,
                  chains = 4, parallel_chains = 4,
                  iter_warmup = 1000, iter_sampling = 2000)
)
```

Código 2.5. **R**. Modelo lineal en R con la librería CmdStanR.

Antes de terminar este apartado cabe mencionar que para modelos sencillos como el modelo lineal, si comparamos el código que hemos necesitado con Stan (códigos 2.4 y 2.5) frente al código que se suele utilizar para estimar el mismo modelo con R (código 2.6) se puede ver que hay una clara diferencia en la extensión del código.

```
datos_fevi <- data.frame(y = c(40, 35, 37, 29, 38, 30),
                         edad = c(57, 63, 60, 72, 74, 55),
                         sexo = c(0, 0, 1, 0, 1, 1))
fit <- lm(y ~ 1 + edad + sexo, data = datos_fevi)
```

Código 2.6. **R**. Modelo lineal en R con la función `lm`.

Esto ha provocado que, con el objetivo de acercar la estadística bayesiana y hacerla accesible a un público más general, vayan apareciendo algunos paquetes —entre los que destacan brms

[14]La lista completa de interfaces de Stan se puede consultar en `https://mc-stan.org/users/interfaces/`.

(Bürkner 2017) y rstanarm (Goodrich *et al.* 2022) para R— que simplifican muchos de estos modelos y nos permiten estimarlos de una forma más abreviada similar a la función `lm` de R (ver código 2.7 y comparar con el código 2.6).

```
library("rstanarm")
datos_fevi <- data.frame(y = c(40, 35, 37, 29, 38, 30),
                         edad = c(57, 63, 60, 72, 74, 55),
                         sexo = c(0, 0, 1, 0, 1, 1))
fit <- stan_lm(y ~ 1 + edad + sexo, data = datos_fevi)
```

Código 2.7. **R**. Modelo lineal en R con la librería rstanarm.

En todo caso, para modelos más complejos o en aquellos en los que se necesite modificar ciertos detalles se podría necesitar escribir el modelo directamente en Stan.

Parte II

Justificación y objetivos

Capítulo 3

Justificación y objetivos

3.1 | Justificación

La insuficiencia cardíaca constituye uno de los mayores problemas de salud pública que existen en la actualidad. La gran complejidad de este síndrome, con las diversas enfermedades que desembocan en ella, hace que su prevalencia sea muy alta e, incluso, siga en aumento debido a su relación con la edad y al envejecimiento de la población. Su elevada mortalidad, la gran cantidad de hospitalizaciones que lleva asociadas —y, con ello, el gran coste sanitario que implica— y la destacada pérdida de calidad de vida que se le relaciona la sitúan en uno de los asuntos sanitarios de mayor relevancia.

La alta mortalidad y morbilidad de la IC ha hecho que el grueso de la atención haya recaído sobre todo en la investigación de fármacos centrados en reducir la mortalidad. Así, los efectos pronósticos de muchos de estos fármacos han quedado probados en numerosos ensayos clínicos y estudios. Lo que no está todavía suficientemente claro es su efecto sobre la mortalidad en la práctica clínica real, ni su relación con la mejoría de la FEVI, cómo evoluciona este parámetro a lo largo de todo el seguimiento ni qué papel desempeña el remodelado cardíaco y el VTD.

En esta «era de la información», del *machine learning*, de la ciencia de datos y del *big data* ha ganado mucha importancia la posibilidad de descargar y analizar grandes cantidades de datos, siendo cada vez más habitual encontrar estudios médicos observacionales con decenas de miles de pacientes y varios años de seguimiento.

Aunque todavía hay mucho trabajo por hacer, en España también se va avanzando en este sentido y, en la actualidad, ya se pueden empezar a descargar datos clínicos de forma automática y masiva a nivel hospitalario y, en algunos casos, a nivel regional.

Este trabajo surge precisamente de la posibilidad de recopilar los datos recogidos durante años de los pacientes ingresados por IC en el Hospital Clínico Universitario Virgen de la Arrixaca de Murcia (HCUVA): datos de ingresos (y reingresos), seguimientos, tratamientos, ecocardiografías, mortalidad, etc. Toda esta información en conjunto ofrece una visión general de la IC, permitiendo estudiar y analizar la evolución completa de los pacientes.

Las características de los datos que provienen de la práctica clínica y que comprenden varios años de seguimiento —como su carácter longitudinal con medidas irregulares para

cada paciente, la censura debido a un evento terminal como la muerte o los cambios en los tratamientos— hacen difícil su análisis estadístico.

En el ámbito del *machine learning* hay, en general, pocos métodos desarrollados para datos longitudinales no equidistantes en el tiempo y, menos aún, para datos con algún tipo de censura. Por extensión, no hemos encontrado ningún método de *machine learning* que permita analizar ambas cosas simultáneamente.

Para poder analizar conjuntamente la evolución de la FEVI y la mortalidad en los últimos años están ganando mucho protagonismo los conocidos como *joint models* para datos longitudinales y de supervivencia. Una ventaja de estos modelos es que no requieren que los pacientes tengan el mismo número de medidas ni que estas sean regulares en el tiempo.

A la hora de estimar un modelo estadístico como los *joint model* la inferencia bayesiana es una opción que plantea muchas ventajas. Entre ellas se encuentran dar mejores estimaciones en modelos más complejos y ser una de las opciones recomendadas como alternativa a los los contrastes de hipótesis clásicos. Sin embargo, desde un punto de vista práctico, su mayor ventaja es que concede flexibilidad para definir el modelo que represente mejor el proceso de generación de los datos.

3.2 | Objetivos

La finalidad principal de este trabajo consiste en aplicar los métodos y modelos estadísticos adecuados que nos permitan analizar la mayor cantidad de información conjuntamente de la base de datos disponible de pacientes con IC (ver apartado 4.2).

Los objetivos que nos planteamos los podemos enumerar en:

1. Desarrollo e implementación de un modelo estadístico para analizar la evolución de la FEVI en pacientes con IC-FEr, teniendo en cuenta que esta evolución se ve afectada por un evento terminal como la muerte.
2. Estimación de los efectos de los antecedentes e historia clínica al inicio en la evolución de la FEVI.
3. Estimación del efecto del remodelado cardíaco, mediante los valores de VTD, en la evolución de la FEVI.
4. Estimación en el modelo de los efectos de los tratamientos sobre la evolución de la FEVI.
5. Estimación del efecto de los niveles de FEVI sobre la mortalidad.

Parte III

Material y métodos

Capítulo 4

Diseño del estudio y bases de datos

4.1 | Diseño del estudio

Este estudio se puede definir como un estudio longitudinal, retrospectivo, observacional y descriptivo. Es un estudio longitudinal porque nos centraremos en la evolución temporal de los pacientes; retrospectivo, debido a que los pacientes ingresaron en el pasado; observacional, dado que no hay intervención por nuestra parte; y descriptivo, por el hecho de no haber un grupo control (Grimes y Schulz 2002; Gallin *et al.* 2017, pág. 231).

Para el desarrollo de este estudio se han seguido la normativa y las regulaciones vigentes, siendo aprobado por el comité ético local del Hospital Clínico Universitario Virgen de la Arrixaca de Murcia.

4.2 | Bases de datos disponibles

En este estudio analizamos una base de datos de uso clínico de pacientes con insuficiencia cardíaca tratados en el HCUVA. Esta base consta de la información personal de cada paciente, la historia clínica, los antecedentes familiares, los diferentes tratamientos farmacológicos que ha llevado en el transcurso de la enfermedad y los resultados de las ecocardiografías realizadas en el seguimiento.

Para completar más la información —mediante los identificadores personales como los números de historia clínica hospitalaria (NHC) y de tarjeta sanitaria (código de identificación personal autonómico o CIPA)— recopilamos y unificamos todas las ecocardiografías disponibles de cada paciente, las fechas de mortalidad y seguimiento, las fechas de reingresos hospitalarios y comprobamos con la unidad de trasplantes cardíacos del hospital a qué pacientes se les realizó este procedimiento.

4.3 | Procesamiento de los datos

Después de la recopilación de todos los datos disponibles de la base inicial de pacientes, con el objetivo de darles un formato adecuado para poder analizarlos, es necesario llevar a

cabo un procesamiento de los datos (ver apartado 2.1). En este trabajo dividimos este proceso en dos grandes partes:

1. procesamiento estándar de los datos y
2. procesamiento especial para definir nuestro modelo en Stan.

4.3.1. Procesamiento estándar

Este primer paso consiste en la fusión de todas las bases de datos que tenemos disponibles, en la depuración y validación de la información (revisión y búsqueda de incongruencias, estudio de valores perdidos, valores atípicos, etc.), y en la estructuración de datos a la manera estándar de análisis.

Las distintas fuentes de información o bases de datos, como es habitual, están en múltiples formatos —*mdb* de Microsoft Access, *xlsx* de Microsoft Excel e, incluso, *docx* de Microsoft Word—. Para fusionar todas estas tablas de datos, en primer lugar, las transformamos a un formato común y estándar como es *csv* —ficheros en texto plano, sin estilos, con los valores separados por comas (*comma separated values* por sus siglas en inglés)—. Una vez que están en este formato las importamos y fusionamos con el software estadístico R, quedando toda la información estructurada en tres tablas:

1. una primera tabla con toda la información inicial de los pacientes junto a los datos de mortalidad y eventos (en esta tabla cada fila corresponde a un paciente),
2. una tabla con las variables de ecocardiografías realizadas durante todo el seguimiento a los pacientes (cada fila tiene la información de una ecocardiografía de un paciente en un momento concreto) y
3. una tercera tabla con los tratamientos farmacológicos prescritos a cada paciente durante todo su seguimiento (cada fila tiene la información del tratamiento de cada paciente en cada fecha).

Estas tres tablas las exportamos en formato largo o *tidy*, por ser este el formato más adecuado para tabular datos longitudinales. En este formato de tabla cada fila corresponde a una observación (paciente, ecocardiografía o tratamiento, respectivamente) y cada columna a una variable (Wickham 2014).

En la parte de depuración de datos revisamos todas las variables útiles para nuestro análisis en busca de valores perdidos (o NA, del inglés *not available*) y valores atípicos o *outliers*. El análisis de valores perdidos mostró que habían pocos datos faltantes en la parte de antecedentes e historia clínica. Considerando que los perdidos se han producido de forma aleatoria optamos por excluirlos como norma general, aproximación que se conoce como análisis de casos completos (Gelman y Hill 2006, capítulo 25).

La única variable con una cantidad importante de valores perdidos es el VTD proveniente de ecocardiografía; en este caso, utilizamos el diámetro del ventrículo (DTD) y aplicamos la

fórmula de Teicholz (Wandt *et al.* 1999) para aproximarlo:

$$\text{VTD} = \frac{7 \cdot \text{DTD}}{2.4 + \text{DTD}^3}.$$

Una vez revisados los datos, los estructuramos y transformamos para poder utilizarlos en los análisis. Respecto a eventos y tratamientos, al disponer de fechas, calculamos tiempos desde el inicio (fecha de la primera ecocardiografía). El trasplante cardíaco, al realizarse en casos en los que el corazón no puede recuperarse, lo consideramos como muerte (aunque sea del corazón, no del paciente).

Las variables numéricas las centramos y escalamos para tener una mejor interpretación de los resultados. Las variables categóricas, como es habitual, las codificamos con ceros y unos para construir las distintas variables de diseño —también conocidas como ficticias o *dummy variables* en inglés (Law *et al.* 2020)— para poder añadirlas a los modelos como covariables (ver tabla 4.1). La única imposición necesaria para usarlas es tomar una categoría como referencia —por ejemplo, para la variable «sexo» se puede establecer «hombre» como referencia y utilizar «mujer» en el modelo— para así evitar el problema de colinealidad perfecta —las covariables estarían perfectamente correlacionadas y no se podría estimar correctamente el modelo— (Suits 1957).

Tabla 4.1. *Proceso de conversión de una tabla con variables categóricas (izquierda) en una tabla con variables de diseño (derecho) para poder utilizarlas en un modelo estadístico.*

Sexo	**Etiología**		**Mujer**	**Hombre**	**Et. Idiopática**	**Et. Isquémica**
Hombre	Isquémica		0	1	0	1
Mujer	Isquémica		1	0	0	1
Hombre	Idiopática	$\longrightarrow$	0	1	1	0
Hombre	Idiopática		0	1	1	0
⋮	⋮		⋮	⋮	⋮	⋮

4.3.2. Procesamiento para el modelo en Stan

Con todas las variables procesadas como se ha explicado en la sección anterior (variables numéricas y variables de diseño de las categóricas) se pueden realizar la gran mayoría de modelos estadísticos que necesitemos. De hecho, muchos programas estadísticos como R, por defecto, ya realizan algunas de estas tareas automáticamente como omitir los valores perdidos o calcular la matriz de variables de diseño.

Sin embargo, a la hora de definir un modelo utilizando métodos bayesianos y un lenguaje probabilístico como Stan, en ocasiones, en conveniente transformar los datos para mejorar la computación o para simplificar la notación.

En general, en informática lo que más tiempo de computación suele necesitar son los bucles (donde una sentencia se ejecuta repetidas veces) y los condicionales (donde una sentencia

solo se ejecuta si cumple cierta condición). Aunque en Stan estas sentencias no son demasiado lentas es más eficiente trabajar con vectores y operaciones de matrices (Stan Development Team 2021, sección 14.8).

Como veremos en el siguiente capítulo el modelo que desarrollaremos en este trabajo será un *joint model* por intervalos. Este modelo lo escribiremos, en la medida de lo posible, de forma matricial (como ecuación (2.4)) para mejorar los tiempos de computación.

Las transformaciones necesarias en nuestros datos para describir nuestro modelo en forma vectorial son tres:

1. unificación de los tiempos de las ecocardiografías en los intervalos de tiempo —de esta manera, si tenemos dos ecocardiografías muy próximas en el tiempo que pertenecen al mismo intervalo, las consideramos juntas—,
2. conversión de las variables de eventos y tiempos de supervivencia en dos matrices (en ambos casos cada fila corresponderá a un paciente y cada columna a uno de los intervalos de tiempo):
   - una matriz de eventos por intervalos, donde se identificará en cuál de los intervalos se produce el evento o la censura (ver tabla 4.2), y

Tabla 4.2. *Proceso de conversión de las variables de tiempos y eventos (izquierda) en una matriz con los intervalos de tiempo y los eventos (derecha).*

Tiempo	**Evento**		[0, 1)	[1, 2)	[2, 3)	…
1.5	0		0	1	0	…
0.5	1		1	0	0	…
2.3	1	$\longrightarrow$	0	0	1	…
1.1	0		0	1	0	…
⋮	⋮		⋮	⋮	⋮	⋱

   - una matriz de tiempos de supervivencia, donde se indicará el tiempo de supervivencia dentro de cada intervalo (ver tabla 4.3);

Tabla 4.3. *Proceso de conversión de las variables de tiempos y eventos (izquierda) en una matriz con los tiempos de supervivencia en cada intervalo (derecha)*

Tiempo	**Evento**		[0, 1)	[1, 2)	[2, 3)	…
1.5	0		1	0.5	0	…
0.5	1		0.5	0	0	…
2.3	1	$\longrightarrow$	1	1	0.3	…
1.1	0		1	0.1	0	…
⋮	⋮		⋮	⋮	⋮	⋱

3. conversión de las variables de tiempos y tratamientos en una matriz para cada tratamiento, donde se calcularán las proporciones de tiempo que cada paciente ha tomado el tratamiento en cada intervalo, así, si un paciente ha tomado un fármaco durante todo el intervalo el valor será de 1 y si, en la mitad del intervalo ha dejado de tomarlo, valdrá 0.5.

Capítulo 5

Modelización

En este trabajo desarrollaremos y estimaremos un *joint model* para datos longitudinales y de supervivencia con el objetivo de estudiar la evolución del marcador FEVI de IC y su relación con la mortalidad. Debido a que los tratamientos pueden cambiar durante el seguimiento —con independencia de la disponibilidad de ecocardiografía— este modelo lo ajustaremos por intervalos.

Dados los patrones de evolución de la FEVI que se pueden observar en los datos (ver apartado 7.2) para el modelo longitudinal implementaremos un modelo de crecimiento no lineal que desarrollaremos en el apartado 5.1.

Para el modelo de supervivencia utilizaremos un modelo de riesgos proporcionales con riesgo basal constante por intervalos. Este modelo es similar al descrito por Y. Wang y Taylor (2001) o Ibrahim, Chen *et al.* (2001, sección 3.2) y lo desarrollaremos en el apartado 5.2.

La estimación de los parámetros que describirán nuestro *joint model* completo la haremos con inferencia bayesiana utilizando el lenguaje estadístico Stan (Stan Development Team 2021). Por esta razón, además de matemáticamente, definiremos el modelo en este lenguaje.

Para poder ejecutar nuestro modelo estadístico bayesiano, como hemos visto en el apartado 2.5.2, necesitamos establecer la función de verosimilitud, $p(datos|\boldsymbol{\theta})$, y las distribuciones a *priori* de los parámetros $p(\boldsymbol{\theta})$. En este capítulo nos centraremos en definir las relaciones entre las observaciones y los parámetros, esto es, en la función de verosimilitud.

5.1 | Modelo longitudinal

5.1.1. Evolución de la FEVI

Para describir la evolución temporal de nuestro marcador principal, FEVI, en lugar de utilizar la clásica forma lineal, utilizamos una función que pueda describir mejor el patrón no lineal que se puede observar tanto en nuestros datos (ver capítulo 7 y figura 7.2) como en otros trabajos publicados (Lupón *et al.* 2018).

De esta manera, la curva que utilizaremos para describir la evolución de la FEVI es una curva con crecimiento exponencial y nivel de saturación —esto es, un nivel máximo al que se aproxima la curva al aumentar el tiempo—.

Esta curva de evolución la podemos representar con la función

$$y(t) = \chi + \delta - \delta e^{-\nu t}. \tag{5.1}$$

Como podemos ver, esta función[15] establece la relación entre los valores de la variable dependiente y (que en nuestro caso será la FEVI media) según aumenta el tiempo t utilizando para ello tres parámetros principales:

- el parámetro χ, que describe el valor inicial de la variable y, esto es, cuando el tiempo es cero,1 $t = 0$,
- el parámetro δ, que representa el valor de saturación al que se aproxima la variable y al aumentar el tiempo, $t \to +\infty$, y
- el parámetro ν, que se puede interpretar como la velocidad con la que se produce ese aumento.

En esta función el valor $\chi + \delta$ representa el nivel de saturación de la curva, esto es, el nivel máximo hasta el que aumenta. Adicionalmente, un índice que nos ayuda en la interpretación de la velocidad ν a la que crece la curva es la vida media $\log(2)/\nu$, que se corresponde con el tiempo que tarda la curva en aumentar la mitad del aumento total, esto es, $\delta/2$. La figura 5.1 ilustra todos los parámetros que describen la curva. Además, en la figura 5.2 se muestran los cambios que se producen en la curva al variar cada uno de los parámetros, manteniendo los demás constantes.

5.1.2. Modelo para FEVI

Una vez establecida la curva de evolución en la ecuación (5.1) podemos plantear nuestro modelo suponiendo que nuestros valores de FEVI y para cada paciente $p = 1, \ldots, N_p$ y cada intervalo $i = 1, \ldots, N_i$ seguirán de media esta curva para cada tiempo $\bar{t}_i$, lo que podemos escribir como

$$\begin{cases} y_o \sim \text{Normal}\left(\mu_{p_o,i_o}, \sigma\right) \\ \mu_{p,i} = \chi_p + \delta_p - \delta_p e^{-\nu_p \bar{t}_i} \end{cases} \tag{5.2}$$

utilizando la distribución normal dada en la ecuación (2.2) y donde, debido a que nuestra tabla de ecocardiografías está en formato largo (ver apartado 4.3.1), hemos necesitado especificar los vectores p_o e i_o para saber a qué paciente p y a qué intervalo i corresponde cada observación $o = 1, \ldots, N_o$.

[15] Esta función coincide con la media del proceso estocástico denominado proceso de Ornstein-Uhlenbeck (Maller *et al.* 2009; Uhlenbeck y Ornstein 1930). En estos casos, una forma habitual de encontrarla definida en la literatura es $y(t) = x_0 e^{-\alpha t} + \mu(1 - e^{-\lambda t})$, donde x_0 se corresponde con χ, $\chi + \delta$ se corresponde con μ y λ con ν, coincidiendo δ en ambos casos.

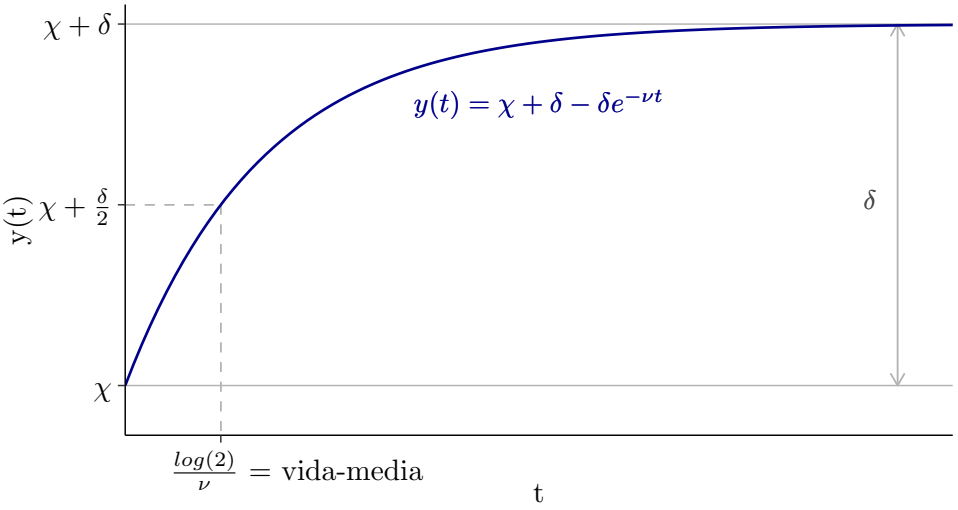

Figura 5.1. *Curva de crecimiento exponencial con nivel de saturación. Se representa la curva (azul) junto con todos los parámetros que la describen.*

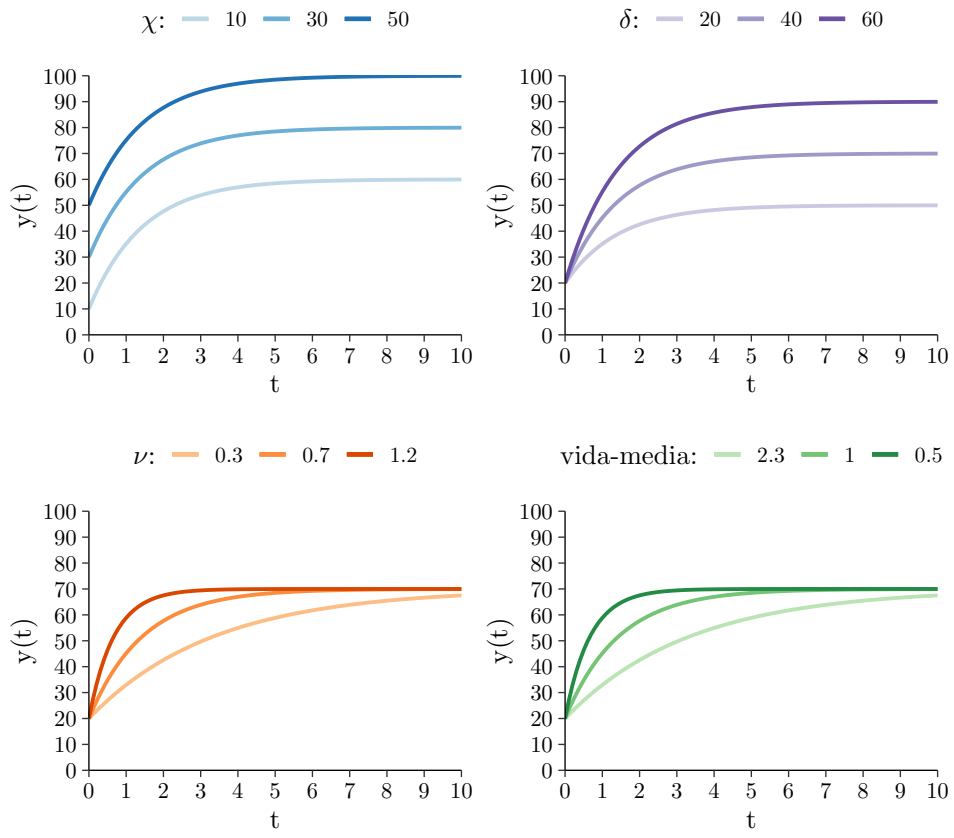

Figura 5.2. *Curvas de crecimiento exponencial con nivel de saturación variando los distintos parámetros: χ, figura arriba izquierda, color azul; δ, arriba derecha, color morado; ν, abajo izquierda, color naranja; y vida media, abajo derecha, color verde. Los valores constantes de cada gráfico (a excepción del que varía en cada uno) son: $\chi = 20$, $\delta = 50$ y $\nu = 0.7$ (vida-media $= 1$).*

Adicionalmente, en la ecuación (5.2) podemos considerar que los fármacos que toma cada paciente en cada intervalo de tiempo también tienen un efecto sobre los valores de FEVI, haciendo que estos sean superiores o inferiores a la media $\mu_{p,i}$. Añadiendo esta parte, la ecuación se puede escribir como

$$\begin{cases} y_o \sim \text{Normal}\left(\mu_{p_o,i_o}, \sigma\right) \\ \mu_{p,i} = \chi_p + \delta_p - \delta_p e^{-\nu_p \bar{t}_i} + \displaystyle\sum_{f=1}^{N_f} \phi_f \{F_{p,i}\}_f \end{cases} \tag{5.3}$$

donde $\{F_{p,i}\}_f$ es la matriz que nos indica si el paciente p en el intervalo i tomaba el fármaco $f = 1, \ldots, N_f$ (ver apartado 4.3.2) y ϕ_f denota el efecto que tiene cada uno sobre la media.

5.1.3. Parámetros de la curva: χ, δ y ν

Con el modelo establecido para cada paciente p tendremos unos valores χ_p, δ_p y ν_p que describen el valor de la curva en el tiempo $\bar{t}_i$ de cada intervalo i. El siguiente paso, por tanto, es modelizar estos parámetros (ver apartado 2.4.2).

Los modelos que hemos establecido para estos parámetros son modelos múltiples (ver ecuación (2.3)) donde, en cada caso, habrá un valor medio α_* y distintas covariables como sexo y edad lo aumentarán o disminuirán en una cantidad β_*.

Llamamos α_χ, α_δ y α_ν a los valores iniciales de cada uno de estos parámetros; $\mathbf{C_\chi}$, $\mathbf{C_\delta}$ y $\mathbf{C_\nu}$ a sus matrices de covariables; $\boldsymbol{\beta_\chi}$, $\boldsymbol{\beta_\delta}$ y $\boldsymbol{\beta_\nu}$ a los efectos de las covariables; y σ_χ, σ_δ y σ_ν a sus varianzas. Con esta notación los modelos se pueden escribir como

$$\begin{cases} \chi_p \sim \text{Normal}\left(\alpha_\chi + \displaystyle\sum_{c=1}^{N_{c_\chi}} \beta_{\chi_c} C_{\chi_{p,c}},\, \sigma_\chi\right) \\ \delta_p \sim \text{Normal}\left(\alpha_\delta + \displaystyle\sum_{c=1}^{N_{c\delta}} \beta_{\delta_c} C_{\delta_{p,c}},\, \sigma_\delta\right) \\ \nu_p \sim \text{Normal}\left(\alpha_\nu + \displaystyle\sum_{c=1}^{N_{c_\nu}} \beta_{\nu_c} C_{\nu_{p,c}},\, \sigma_\nu\right) \end{cases} \tag{5.4}$$

En muchas ocasiones, al ejecutar un modelo como el anterior en Stan aparecen problemas de divergencias que se pueden solucionar reparametrizando el modelo, esto es, expresándolo de una forma alternativa pero equivalente. Una de las reparametrizaciones más habituales es la parametrización no centrada (McElreath 2020, pág. 421). Esta parametrización consiste en utilizar la propiedad de la distribución normal para reescribir el modelo estandarizado, *i. e.*, de media cero y varianza uno.

En nuestro caso, si reparametrizamos $\boldsymbol{\chi}$ y $\boldsymbol{\delta}$ podemos reescribir la ecuación (5.4) como

$$\begin{cases} \chi_p = \alpha_\chi + \sum_{c=1}^{N_{c\chi}} \beta_{\chi c} C_{\chi p,c} + \sigma_\chi \widehat{\chi}_p \\ \delta_p = \alpha_\delta + \sum_{c=1}^{N_{c\delta}} \beta_{\delta c} C_{\delta p,c} + \sigma_\delta \widehat{\delta}_p \end{cases} \tag{5.5}$$

donde los nuevos vectores de parámetros $\widehat{\boldsymbol{\chi}}$ y $\widehat{\boldsymbol{\delta}}$ siguen distribuciones normales centradas en 0 y con varianza 1.

El parámetro $\boldsymbol{\nu}$ también se podría reparametrizar, pero en este caso, conlleva posteriormente la aparición de problemas de convergencia en la ejecución.

5.1.4. Modelo longitudinal final

A partir de las ecuaciones (5.3) a (5.5) tenemos la función de verosimilitud para el modelo longitudinal, por lo que solo faltaría establecer las distribuciones a *priori* para todos los parámetros para ejecutar el modelo en Stan y estimar la probabilidad a *posteriori*.

Para mejorar la velocidad de cómputo del modelo en Stan este submodelo longitudinal lo podemos escribir en formato vectorial:

$$\begin{cases} y_o \sim \text{Normal}\left(\mu_{p_o,i_o}, \sigma\right) \\ \boldsymbol{\mu} = (\boldsymbol{\chi} + \boldsymbol{\delta})\mathbf{J}_{\mathbf{1},\mathbf{N_i}} - (\boldsymbol{\delta}\mathbf{J}_{\mathbf{1},\mathbf{N_i}}) \circ e^{-\boldsymbol{\nu}\bar{\mathbf{t}}} + \sum_{f=1}^{N_f} \phi_f \{\mathbf{F}\}_f \\ \boldsymbol{\chi} = \alpha_\chi + \mathbf{C}_{\boldsymbol{\chi}}\boldsymbol{\beta}_{\boldsymbol{\chi}} + \sigma_\chi \widehat{\boldsymbol{\chi}} \\ \boldsymbol{\delta} = \alpha_\delta + \mathbf{C}_{\boldsymbol{\delta}}\boldsymbol{\beta}_{\boldsymbol{\delta}} + \sigma_\delta \widehat{\boldsymbol{\delta}} \\ \nu_p = \text{Normal}(\alpha_\nu + \mathbf{C}_{\boldsymbol{\nu} p,*}\boldsymbol{\beta}_{\boldsymbol{\nu}}, \sigma_\nu) \end{cases} \tag{5.6}$$

donde $\mathbf{J}_{\mathbf{1},\mathbf{N_i}}$ es la matriz de todo unos de una fila y N_i columnas —al multiplicar esta matriz por un vector se obtiene una matriz formada por ese vector repetido en cada columna— y el operador «$\circ$» denota la multiplicación elemento a elemento de dos matrices[16].

5.2 | Modelo de supervivencia

Para definir el modelo de supervivencia, como hemos visto en el apartado 2.4.3, tenemos que modelizar las funciones de riesgo $h(t)$ y de supervivencia $S(t)$.

[16]Este producto entre dos matrices de las mismas dimensiones $\mathbf{A} \in \mathbb{R}^{n,m}$ y $\mathbf{B} \in \mathbb{R}^{n,m}$ se denomina producto de Hadamard o producto de Schur y se define como

$$\mathbf{A} \circ \mathbf{B} = (A_{i,j}) \circ (B_{i,j}) = (A_{i,j}B_{i,j}) \in \mathbb{R}^{n \times m}.$$

El modelo que utilizaremos es un modelo de riesgos proporcionales con riesgo basal constante por intervalos, lo que nos permitirá calcular la integral de la función de supervivencia (ver ecuación (2.7)).

En este caso, en lugar de definir todas las distribuciones de probabilidad de las variables para determinar la función de verosimilitud (como en el modelo longitudinal) nos basaremos directamente en la función de verosimilitud del análisis de supervivencia (ecuación (2.8)) y, con ella, incrementaremosen Stan el logaritmo de la probabilidad total (ver apartado 2.5.3).

Suponiendo todos los pacientes independientes, la función de verosimilitud para el modelo de supervivencia es igual a la multiplicación de la función de distribución de densidad para cada observación, esto es

$$p(\mathbf{e}, \mathbf{t}|h, S) = \prod_{p=1}^{N_p} h(t_p)^{e_p} S(t_p),$$

por lo que, tomando logaritmos, llegamos a la ecuación

$$\log(p(\mathbf{e}, \mathbf{t}|h, S)) = \sum_{p=1}^{N_p} (e_p \log(h(t_p)) + \log(S(t_p))). \tag{5.7}$$

De esta manera, en las siguientes secciones modelizaremos los logaritmos de las funciones de riesgo y supervivencia, $\log(h(t))$ y $\log(S(t))$.

5.2.1. Función de riesgo h

Los tiempos que separan cada uno de los intervalos de nuestro modelo los denotaremos s_i. Así, suponiendo que tenemos N_i intervalos, estos vienen delimitados por los puntos $0 = s_1 < s_2 < \ldots < s_{N_i+1}$.

Siguiendo la notación del *joint model* (ver ecuación (2.10)), donde la matriz de covariables se denotará como $\mathbf{C_h}$ para diferenciarla del modelo longitudinal, con λ_i denotaremos la función de riesgo basal para cada intervalo i. El tiempo de supervivencia (o censura) t_p para cada paciente p se situará en un intervalo j concreto, esto es, $t_p \in [s_j, s_{j+1})$ para algún j, por lo que la función de riesgo se puede escribir como

$$h(t_p) = \lambda_j \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,j}).$$

Esta ecuación se puede generalizar a todos los intervalos utilizando una variable que nos indique qué intervalo es al que pertenece el tiempo t_p. Si llamamos $\mathbf{E} = (E_{p,i})$ a esta matriz (ver apartado 4.3.2) y la definimos como

$$E_{p,i} = \begin{cases} 1 & \text{si } t_p \in [s_i, s_{i+1}) \\ 0 & \text{en caso contrario} \end{cases}$$

entonces la ecuación para $h(t_p)$ se puede escribir como

$$h(t_p) = \prod_{i=1}^{N_i} (\lambda_i \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i}))^{E_{p,i}}.$$

Si tomamos logaritmos en la ecuación anterior y aplicamos sus propiedades esta ecuación se puede reescribir como

$$\begin{aligned}\log(h(t_p)) &= \sum_{i=1}^{N_i} E_{p,i} \log(\lambda_i \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})) = \\ &= \sum_{i=1}^{N_i} E_{p,i} \log(\lambda_i) + \boldsymbol{\gamma} C_{hp,*} \sum_{i=1}^{N_i} E_{p,i} + \gamma_\mu \sum_{i=1}^{N_i} E_{p,i}\mu_{p,i} = \\ &= \sum_{i=1}^{N_i} E_{p,i} \log(\lambda_i) + \boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \sum_{i=1}^{N_i} E_{p,i}\mu_{p,i}\end{aligned}$$

donde la última igualdad se obtiene por la condición $\sum_{i=1}^{N_i} E_{p,i} = 1$, ya que, para cada paciente el evento (o censura) solo se producirá en un intervalo.

Finalmente, esta ecuación se puede expresar en forma vectorial como

$$\log(\mathbf{h}) = \mathbf{E} \log(\boldsymbol{\lambda}) + \mathbf{C_h}\boldsymbol{\gamma} + \mathbf{E} \odot (\gamma_\mu \boldsymbol{\mu}) \tag{5.8}$$

donde el operador «$\odot$» denota el producto escalar entre las filas de dos matrices[17].

5.2.2. Función de supervivencia S

Análogamente a la función de riesgo del apartado anterior, también se puede obtener una ecuación cerrada para la función de supervivencia $S(t)$.

Teniendo en cuenta la ecuación (2.6) de la función de riesgo acumulada

$$-\log(S(t_p)) = \int_0^{t_p} h(u)\, du,$$

podemos separar la integral en cada uno de los intervalos determinados por los tiempos s_i mencionados en el apartado anterior. Así, simplificando los términos dentro de la exponencial

[17] Si se tienen dos matrices de las mismas dimensiones $\mathbf{A} \in \mathbb{R}^{n\times m}$ y $\mathbf{B} \in \mathbb{R}^{n\times m}$, este producto se define como

$$\mathbf{A} \odot \mathbf{B} = \begin{pmatrix} A_{1,*} \cdot B_{1,*} \\ A_{2,*} \cdot B_{2,*} \\ \vdots \\ A_{n,*} \cdot B_{n,*} \end{pmatrix} \in \mathbb{R}^n.$$

obtenemos

$$-\log(S(t_p)) = \int_{s_1=0}^{s_2} \lambda_1 \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})\, du + \cdots + \int_{s_j}^{t_p} \lambda_j \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})\, du,$$

y, utilizando que hemos planteado un modelo de riesgo constante por intervalos, cada integral se resuelve como la diferencia entre sus límites, por lo que

$$\begin{aligned} -\log(S(t_p)) =& \lambda_1 \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})(s_2 - s_1) + \cdots + \lambda_j \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})(t_p - s_j) = \\ =& \sum_{k=1}^{j-1} \lambda_k \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})(s_{k+1} - s_k) + \lambda_j \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})(t_p - s_j) \end{aligned}$$

En este punto, utilizando la matriz de eventos $\mathbf{E}$ (como hemos definido en el apartado anterior), el logaritmo de la supervivencia puede escribirse como

$$\begin{aligned} \log(S(t_p)) = -\sum_{i=1}^{N_i} E_{p,i} \Bigg[& \sum_{k=1}^{i-1} \lambda_k \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})(s_{k+1} - s_k) \\ & + \lambda_i \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i})(t_p - s_i) \Bigg]. \end{aligned}$$

Sin embargo, si definimos una matriz $\mathbf{T} = (T_{p,i})$ con los tiempos de supervivencia por intervalos de cada paciente (ver apartado 4.3.2), donde

$$T_{p,i} = \begin{cases} s_{i+1} - s_i & \text{si } t_p < s_i \\ t_p - s_i & \text{si } s_i \leq t_p < s_{i+1} \\ 0 & \text{si } t_p > s_{i+1} \end{cases}$$

entonces podemos reducir todos los sumandos dentro de una única suma

$$\log(S(t_p)) = -\sum_{i=1}^{N_i} \lambda_i \exp(\boldsymbol{\gamma} C_{hp,*} + \gamma_\mu \mu_{p,i}) T_{p,i}.$$

Finalmente, esta última ecuación se puede escribir en forma vectorial como

$$\log(\mathbf{S}) = \exp(\mathbf{C_h}\boldsymbol{\gamma}) \circ \left[(\mathbf{T} \circ (\boldsymbol{\gamma}_\mu \boldsymbol{\mu}))\boldsymbol{\lambda} \right] \tag{5.9}$$

donde, recordamos, el símbolo «∘» denota el producto elemento a elemento entre dos matrices.

5.2.3. Modelo de supervivencia final

De forma similar al modelo longitudinal, a partir de las ecuaciones (5.7) a (5.9) de riesgo y supervivencia, tenemos establecida la función de verosimilitud.

En este submodelo de supervivencia solo nos faltaría expresar en forma vectorial la ecuación (5.7) denotando como $\mathbf{e} = (e_p)$ al vector de eventos. Así, el modelo está determinado por

$$\begin{cases} \log(L) = \mathbf{e} \circ \log(\mathbf{h}) + \log(\mathbf{S}) \\ \log(\mathbf{S}) = \exp(\mathbf{C_h}\boldsymbol{\gamma}) \circ \left[(\mathbf{T} \circ (\boldsymbol{\gamma}_\mu \boldsymbol{\mu}))\boldsymbol{\lambda}\right] \\ \log(\mathbf{h}) = \mathbf{E}\log(\boldsymbol{\lambda}) + \mathbf{C_h}\boldsymbol{\gamma} + \mathbf{E} \odot (\gamma_\mu \boldsymbol{\mu}). \end{cases} \tag{5.10}$$

5.3 | Modelo completo final

5.3.1. Función de verosimilitud

A partir de la ecuación (5.6) para el modelo longitudinal y la ecuación (5.10) para el de supervivencia tenemos establecida la función de verosimilitud final de los datos,

$$p(\textit{datos}|\boldsymbol{\theta}) = p(\mathbf{y}, p_o, i_o, \mathbf{F}, \bar{\mathbf{t}}, \mathbf{C}_{\boldsymbol{\chi}}, \mathbf{C}_{\boldsymbol{\delta}}, \mathbf{C}_{\boldsymbol{\nu}}|\boldsymbol{\theta})p(\mathbf{e}, \mathbf{T}, \mathbf{E}, \mathbf{C_h}|\boldsymbol{\theta}). \tag{5.11}$$

5.3.2. Distribuciones a *priori*

Para las *priors* utilizamos distribuciones débilmente informativas (en inglés *weakly informative priors*) como se recomienda en artículos como Gabry, Simpson *et al.* (2019) o Schoot *et al.* (2021). Estas distribuciones permiten aportar cierta información de antemano al modelo pero sin ser muy restrictivo y dejando cierto grado de incertidumbre[18].

En la elección de las distribuciones a *priori* de los valores iniciales α_* de los parámetros χ, δ y ν utilizamos el conocimiento disponible sobre los valores de FEVI en pacientes con IC-FEr. Dejando suficiente variabilidad para explorar todo el espacio de parámetros establecemos distribuciones normales comenzando en unos valores de FEVI de 25 %, con un aumento de 25 % y una velocidad de 2 (que corresponde a una vida media de unos cuatro meses), lo que podemos escribir como

$$\begin{aligned} \alpha_\chi &\sim \text{Normal}(25, 20)\big|_{0 \le x \le 100} \\ \alpha_\delta &\sim \text{Normal}(30, 30) \\ \alpha_\nu &\sim \text{Normal}(2, 1)\big|_{x>0}, \end{aligned}$$

donde el truncamiento de α_χ refleja que los valores de FEVI deben estar entre 0 y 100 y el de α_ν que este debe ser positivo.

[18] Véase la página de recomendaciones de Stan `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`

Para los parámetros correspondientes a la parametrización no centrada, como hemos comentado en el apartado 5.1.3, establecemos normales tipificadas

$$\begin{aligned}\widehat{\chi}_p &\sim \text{Normal}(0,1), \qquad \text{para } p = 1,\dots,N_p\\ \widehat{\delta}_p &\sim \text{Normal}(0,1), \qquad \text{para } p = 1,\dots,N_p.\end{aligned}$$

Para establecer las *priors* de las varianzas seguimos la recomendación habitual de utilizar distribuciones de Cauchy positivas

$$\begin{aligned}\sigma &\sim \text{Cauchy}(0,4)\big|_{x>0}\\ \sigma_\chi &\sim \text{Cauchy}(0,4)\big|_{x>0}\\ \sigma_\delta &\sim \text{Cauchy}(0,4)\big|_{x>0}\\ \sigma_\nu &\sim \text{Cauchy}(0,4)\big|_{x>0}.\end{aligned}$$

Para las distribuciones a *priori* de los parámetros β_* y ϕ_*, definidos como el efecto de las covariables y de los fármacos sobre los parámetros principales del modelo, escogemos distribuciones normales con suficiente varianza

$$\begin{aligned}\beta_{\chi c} &\sim \text{Normal}(0,4), \qquad \text{para } c = 1,\dots,N_{c\chi}\\ \beta_{\delta c} &\sim \text{Normal}(0,4), \qquad \text{para } c = 1,\dots,N_{c\delta}\\ \beta_{\nu c} &\sim \text{Normal}(0,4), \qquad \text{para } c = 1,\dots,N_{c\nu}\\ \phi_f &\sim \text{Normal}(0,4), \qquad \text{para } f = 1,\dots,N_f.\end{aligned}$$

Análogamente, para los parámetros γ_* y γ_μ del modelo de supervivencia también establecemos distribuciones normales modificando las varianzas según los rangos de las variables. Por razones de computación, elegimos una distribución gamma para el riesgo basal en cada intervalo, ya que, es la que mejor funciona en nuestro modelo

$$\begin{aligned}\gamma_c &\sim \text{Normal}(0,2), \qquad \text{para } c = 1,\dots,N_{ch}\\ \gamma_\mu &\sim \text{Normal}(0,0.01),\\ \lambda_i &\sim \text{Gamma}(0.1,1.5), \quad \text{para } i = 1,\dots,N_i\end{aligned}$$

5.3.3. Distribución de probabilidad conjunta a *posteriori*

Establecida la función de verosimilitud final del modelo (ecuación (5.11)) y todas las distribuciones a *priori* de los parámetros (apartado 5.3.2), por el teorema de Bayes (ecuación (2.11)) tenemos especificada la distribución de probabilidad conjunta a *posteriori* de nuestro modelo,

$$p(\boldsymbol{\theta}|\mathit{datos}) \propto p(\mathit{datos}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{y}, p_o, i_o, \mathbf{F}, \bar{\mathbf{t}}, \mathbf{C}_{\boldsymbol{\chi}}, \mathbf{C}_{\boldsymbol{\delta}}, \mathbf{C}_{\boldsymbol{\nu}}|\boldsymbol{\theta})p(\mathbf{e}, \mathbf{T}, \mathbf{E}, \mathbf{C}_{\mathbf{h}}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Esta ecuación se puede desarrollar para cada variable y parámetro utilizando la regla de la cadena, estableciendo de esta forma las relaciones de dependencia entre todos los parámetros y observaciones, quedando como

$$\begin{aligned} p(\boldsymbol{\theta}|\textit{datos}) \propto\ & p(\mathbf{y}|p_o, i_o, \boldsymbol{\mu}, \sigma, \bar{\mathbf{t}}, \mathbf{F}, \boldsymbol{\phi}) p(\boldsymbol{\nu}|\alpha_\nu, \boldsymbol{\beta_\nu}, \mathbf{C_\nu}, \sigma_\nu) p(\mathbf{e}, \mathbf{T}, \mathbf{E}|\mathbf{h}, \mathbf{S}, \boldsymbol{\mu}, \mathbf{C_h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \gamma_\mu) \\ & \times p(\alpha_\chi) p(\boldsymbol{\beta_\chi}) p(\sigma_\chi) p(\widehat{\boldsymbol{\chi}}) p(\alpha_\delta) p(\boldsymbol{\beta_\delta}) p(\sigma_\delta) p(\widehat{\boldsymbol{\delta}}) p(\alpha_\nu) p(\boldsymbol{\beta_\nu}) p(\sigma_\nu) \\ & \times p(\sigma) p(\boldsymbol{\phi}) p(\boldsymbol{\lambda}) p(\boldsymbol{\gamma}) p(\gamma_\mu) \end{aligned}$$

donde los dos primeros factores se pueden desarrollar como

$$p(\mathbf{y}|p_o, i_o, \boldsymbol{\mu}, \sigma, \bar{\mathbf{t}}, \mathbf{F}, \boldsymbol{\phi}) = \prod_{o=1}^{N_o} p(y_o|p_o, i_o, \mu_{p_o,i_o}, \sigma, \bar{t}_{i_o}, \mathbf{F}, \boldsymbol{\phi}),$$

$$p(\boldsymbol{\nu}|\alpha_\nu, \boldsymbol{\beta_\nu}, \mathbf{C_\nu}, \sigma_\nu) = \prod_{p=1}^{N_p} p(\nu_p|\alpha_\nu, \boldsymbol{\beta_\nu}, \mathbf{C_\nu}, \sigma_\nu),$$

el tercer factor se corresponde con las ecuaciones (5.7) y (5.10), y las distribuciones correspondientes a las *priors* corresponden a

$$p(\boldsymbol{\beta_\chi}) = \prod_{c=1}^{N_{c\chi}} p(\beta_{\chi c}), \quad p(\widehat{\boldsymbol{\chi}}) = \prod_{p=1}^{N_p} p(\widehat{\chi}_p), \quad p(\boldsymbol{\beta_\delta}) = \prod_{c=1}^{N_{c\delta}} p(\beta_{\delta c}), \quad p(\widehat{\boldsymbol{\delta}}) = \prod_{p=1}^{N_p} p(\widehat{\delta}_p)$$

$$p(\boldsymbol{\beta_\nu}) = \prod_{c=1}^{N_{c\nu}} p(\beta_{\nu c}), \quad p(\boldsymbol{\phi}) = \prod_{f=1}^{N_f} p(\phi_f), \quad p(\boldsymbol{\lambda}) = \prod_{i=1}^{N_i} p(\lambda_i), \quad p(\boldsymbol{\gamma}) = \prod_{c=1}^{N_{ch}} p(\gamma_c).$$

Como la ecuación resultante es compleja y difícil de manejar, una forma de visualizar todas las dependencias establecidas en nuestro modelo es mediante un grafo acíclico dirigido o DAG (del inglés *directed acyclic graph*). Este gráfico representa cada variable y parámetro como un nodo, que se unen según su relación de dependencia (Höhna *et al.* 2014). El DAG de nuestro modelo se presenta en la figura 5.3.

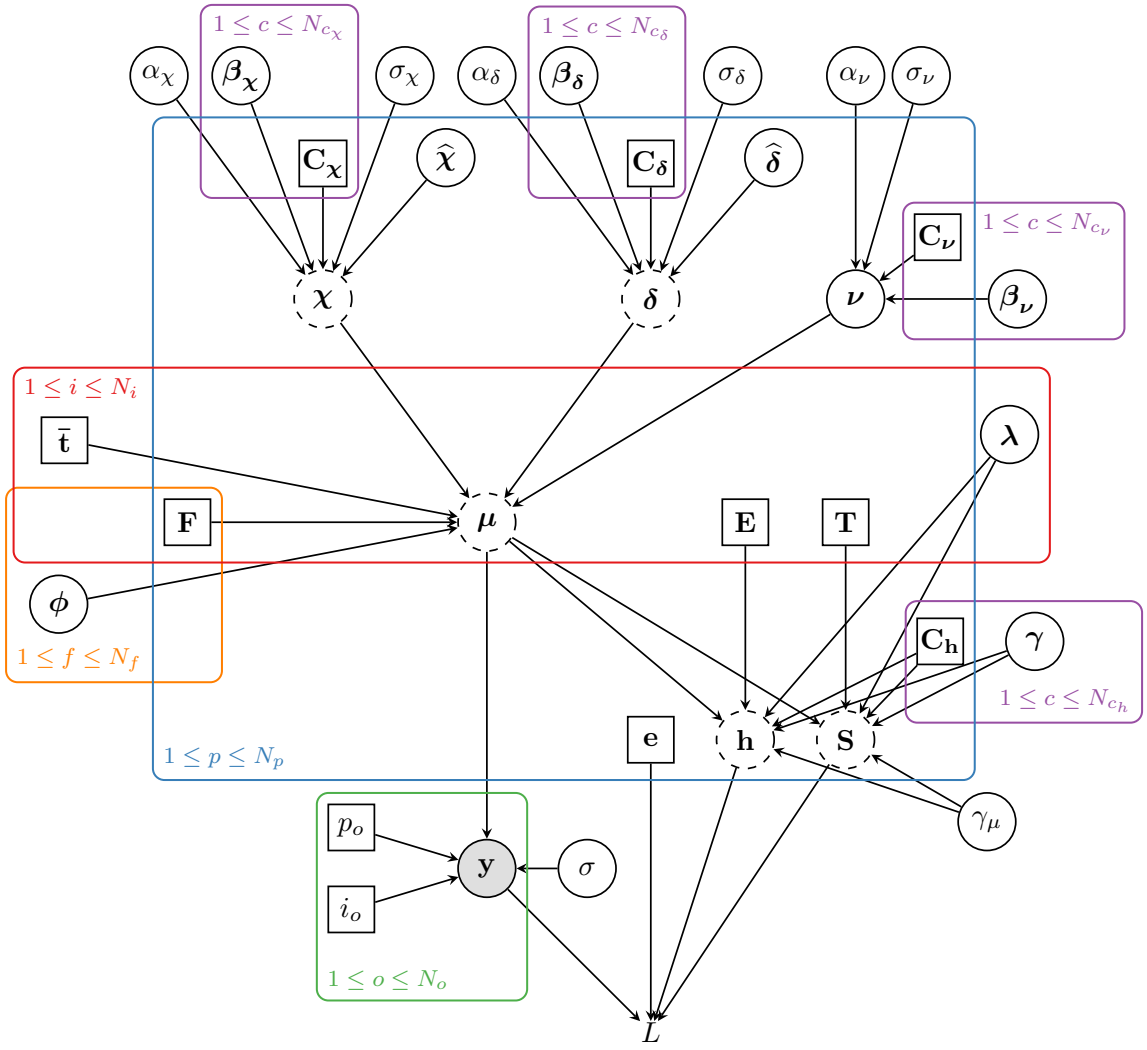

Figura 5.3. *DAG del* joint model *completo para FEVI y mortalidad. Siguiendo la nomenclatura utilizada en Höhna et al. (2014) las variables y parámetros se representan como nodos en el grafo y sus relaciones de dependencia con flechas. Para diferenciar los distintos tipos de variables y parámetros se utilizan cuatro estilos de nodo (exceptuando la verosimilitud L, para la que se utiliza un nodo vacío): (1) cuadrados para representar las constantes; (2) círculos completos para las variables aleatorias; (3) círculos discontinuos para representar las variables que vienen determinadas a partir de otras; y (4) círculos con fondo gris para las observaciones que tienen distribuciones de probabilidad. Para agrupar conjuntos de nodos que varían de la misma forma (por intervalos, por pacientes, etc.), estos se encierran en rectángulos coloreados que denominamos paneles. Los distintos colores que utilizamos para diferenciar los paneles son: rojo, para los intervalos i; naranja, para los fármaco f; violeta: para los paneles de covariables c de cada parámetro ($\boldsymbol{\chi}$, $\boldsymbol{\delta}$, $\boldsymbol{\nu}$ y $\mathbf{h}$); azul, para los pacientes p; y verde, para el panel de observaciones o.*

5.3.4. Código en Stan

Una vez establecido nuestro modelo completo, con verosimilitud y distribuciones a *priori*, estamos en condiciones de escribirlo en Stan para realizar el proceso de estimación. Para ello, transcribimos el DAG mostrado en la figura 5.3 a Stan como mostramos en el código 5.1, situando cada parámetro, variable o cálculo en los distintos bloques según corresponde y dejando líneas en blanco y comentarios para mayor legibilidad.

```
data {
  int<lower=0> N_obs;
  int<lower=0> N_pat;
  int<lower=0> N_int;
  int<lower=0> N_far;

  int<lower=0> N_cov_chi;
  int<lower=0> N_cov_delta;
  int<lower=0> N_cov_nu;
  int<lower=0> N_cov_h;

  vector[N_obs] y;
  row_vector[N_int] times_bar;
  matrix[N_pat, N_cov_chi] Cov_chi;
  matrix[N_pat, N_cov_delta] Cov_delta;
  matrix[N_pat, N_cov_nu] Cov_nu;
  matrix[N_pat, N_cov_h] Cov_h;

  array[N_far] matrix[N_pat, N_int] Far;

  vector[N_pat] event;
  matrix[N_pat, N_int] Events;
  matrix[N_pat, N_int] Times;

  array[N_obs] int ind_pat;
  array[N_obs] int ind_int;
}

transformed data {
}

parameters {
  // y
  real<lower=0> sigma;

  // chi
  real<lower=0, upper=100> alpha_chi;
  vector[N_cov_chi] beta_chi;
```

```
  vector[N_pat] chi_hat;
  real<lower=0> sigma_chi;

  // delta
  real alpha_delta;
  vector[N_cov_delta] beta_delta;
  vector[N_pat] delta_hat;
  real<lower=0> sigma_delta;

  // nu
  real<lower=0> alpha_nu;
  vector[N_cov_nu] beta_nu;
  real<lower=0> sigma_nu;
  vector<lower=0>[N_pat] nu;

  // F
  vector[N_far] phi;

  // survival
  vector<lower=0>[N_int] lambda;
  vector[N_cov_h] gamma;
  real gamma_mu;
}
transformed parameters {
  vector[N_pat] chi;
  vector[N_pat] delta;

  matrix[N_pat, N_int] mu;
  vector[N_pat] log_h;
  vector[N_pat] log_S;

  // parámetros evolución FEVI
  chi = alpha_chi + Cov_chi * beta_chi + sigma_chi * chi_hat;
  delta = alpha_delta + Cov_delta * beta_delta + sigma_delta * delta_hat;

  // evolución FEVI
  mu = rep_matrix(chi + delta, N_int) -
         rep_matrix(delta, N_int) .* exp(-nu * times_bar);
  for(f in 1:N_far){
    mu += phi[f] * Far[f];
  }

  // supervivencia
  log_h = Events * log(lambda) + Cov_h * gamma +
            rows_dot_product(Events, gamma_mu * mu);
  log_S = -exp(Cov_h * gamma) .* ((Times .* exp(gamma_mu*mu)) * lambda);
```

```
}

model {
  // priors
  alpha_chi ~ normal(25, 20) T[0, 100];
  alpha_delta ~ normal(30, 30);
  alpha_nu ~ normal(2, 1) T[0,];

  chi_hat ~ normal(0, 1);
  delta_hat ~ normal(0, 1);

  sigma ~ cauchy(0, 4);
  sigma_chi ~ cauchy(0, 4);
  sigma_delta ~ cauchy(0, 4);
  sigma_nu ~ cauchy(0, 4);

  beta_chi ~ normal(0, 4);
  beta_delta ~ normal(0, 4);
  beta_nu ~ normal(0, 4);
  phi ~ normal(0, 4);

  lambda ~ gamma(0.1, 1.5);
  gamma ~ normal(0, 2);
  gamma_mu ~ normal(0, 0.01);

  nu ~ normal(alpha_nu + Cov_nu * beta_nu, sigma_nu);

  // verosimilitud
  for(o in 1:N_obs){
    target += normal_lpdf(y[o] | mu[ind_pat[o], ind_int[o]], sigma);
  }
  target += sum(event .* log_h + log_S);
}

generated quantities {
}
```

Código 5.1. **Stan**. *Joint model* final para FEVI y mortalidad.

En las tablas 5.1 y 5.2 presentamos una breve descripción de las correspondencias entre los términos utilizados en las ecuaciones y en código Stan.

Tabla 5.1. *Correspondencia entre los términos utilizados para las observaciones y código Stan.*

Ecuación	**Stan**	**Descripción**
Tamaños de los datos		
$N_o \in \mathbb{N}$	`int<lower=0> N_obs;`	número de observaciones
$N_p \in \mathbb{N}$	`int<lower=0> N_pat;`	número de pacientes
$N_i \in \mathbb{N}$	`int<lower=0> N_int;`	número de intervalos
$N_{c\chi} \in \mathbb{N}$	`int<lower=0> N_cov_chi;`	número de covariables para $\boldsymbol{\chi}$
$N_{c\delta} \in \mathbb{N}$	`int<lower=0> N_cov_delta;`	número de covariables para $\boldsymbol{\delta}$
$N_{c\nu} \in \mathbb{N}$	`int<lower=0> N_cov_nu;`	número de covariables para $\boldsymbol{\nu}$
$N_{ch} \in \mathbb{N}$	`int<lower=0> N_cov_h;`	número de covariables para $\mathbf{h}$
$N_f \in \mathbb{N}$	`int<lower=0> N_far;`	número de fármacos
Índices		
$p_o \in \{\mathbb{N}\}^{N_o}$	`array[N_obs] int ind_pat;`	índices de pacientes
$i_o \in \{\mathbb{N}\}^{N_o}$	`array[N_obs] int ind_int;`	índices de intervalos
Covariables y fármacos		
$\mathbf{C}_{\boldsymbol{\chi}} \in \mathbb{R}^{N_p \times N_{c\chi}}$	`matrix[N_pat, N_cov_chi] Cov_chi;`	matriz de covariables para $\boldsymbol{\chi}$
$\mathbf{C}_{\boldsymbol{\delta}} \in \mathbb{R}^{N_p \times N_{c\delta}}$	`matrix[N_pat, N_cov_delta] Cov_delta;`	matriz de covariables para $\boldsymbol{\delta}$
$\mathbf{C}_{\boldsymbol{\nu}} \in \mathbb{R}^{N_p \times N_{c\nu}}$	`matrix[N_pat, N_cov_nu] Cov_nu;`	matriz de covariables para $\boldsymbol{\nu}$
$\mathbf{C}_{\mathbf{h}} \in \mathbb{R}^{N_p \times N_{ch}}$	`matrix[N_pat, N_cov_h] Cov_h;`	matriz de covariables para $\mathbf{h}$
$\mathbf{F} \in \{\mathbb{R}^{N_p \times N_i}\}^{N_f}$	`array[N_far] matrix[N_pat, N_int] Far;`	lista de fármacos por intervalos
Datos modelo longitudinal		
$\bar{\mathbf{t}} \in \mathbb{R}^{1 \times N_i}$	`row_vector[N_int] times_bar;`	tiempo medio de cada intervalo
$\mathbf{y} \in \mathbb{R}^{N_o}$	`vector[N_obs] y;`	valores de FEVI
Datos modelo de supervivencia		
$\mathbf{e} \in \mathbb{Z}_2^{N_p}$	`vector[N_pat] event;`	evento (muerte o TC)
$\mathbf{E} \in \mathbb{Z}_2^{N_p \times N_i}$	`matrix[N_pat, N_int] Events;`	matriz de eventos por intervalos
$\mathbf{T} \in \mathbb{R}^{N_p \times N_i}$	`matrix[N_pat, N_int] Times;`	matriz de supervivencia por intervalos

TABLA 5.2. *Correspondencia entre los términos utilizados para los parámetros y código Stan.*

Ecuación	**Stan**	**Descripción**
Parámetros modelo longitudinal		
$\boldsymbol{\mu} \in \mathbb{R}^{N_p \times N_i}$	`matrix[N_pat, N_int] mu;`	evolución media de la FEVI
$\sigma \in \mathbb{R}_{x>0}$	`real<lower=0> sigma;`	varianza de la FEVI
$\boldsymbol{\chi} \in \mathbb{R}^{N_p}$	`vector[N_pat] chi;`	valor inicial de la FEVI
$\boldsymbol{\delta} \in \mathbb{R}^{N_p}$	`vector[N_pat] delta;`	aumento de la FEVI
$\boldsymbol{\nu} \in \mathbb{R}^{N_p}_{x>0}$	`vector<lower=0>[N_pat] nu;`	velocidad de saturación de FEVI
$\boldsymbol{\phi} \in \mathbb{R}^{N_f}$	`vector[N_far] phi;`	efectos de los fármacos
Subparámetros de $\mathbf{x}$ *y* $\boldsymbol{\delta}$		
$\alpha_\chi \in \mathbb{R}_{0 \leq x \leq 100}$	`real<lower=0, upper=100> alpha_chi;`	valor inicial de $\boldsymbol{\chi}$
$\alpha_\delta \in \mathbb{R}$	`real alpha_delta;`	valor inicial de $\boldsymbol{\delta}$
$\alpha_\nu \in \mathbb{R}_{x>0}$	`real<lower=0> alpha_nu;`	valor inicial de $\boldsymbol{\nu}$
$\boldsymbol{\beta}_{\boldsymbol{\chi}} \in \mathbb{R}^{N_{c\chi}}$	`vector[N_cov_chi] beta_chi;`	efectos de las covariables sobre $\boldsymbol{\chi}$
$\boldsymbol{\beta}_{\boldsymbol{\delta}} \in \mathbb{R}^{N_{c\delta}}$	`vector[N_cov_delta] beta_delta;`	efectos de las covariables sobre $\boldsymbol{\delta}$
$\boldsymbol{\beta}_{\boldsymbol{\nu}} \in \mathbb{R}^{N_{c\nu}}$	`vector[N_cov_nu] beta_nu;`	efectos de las covariables sobre $\boldsymbol{\nu}$
$\sigma_\chi \in \mathbb{R}_{x>0}$	`real<lower=0> sigma_chi;`	varianza de $\boldsymbol{\chi}$
$\sigma_\delta \in \mathbb{R}_{x>0}$	`real<lower=0> sigma_delta;`	varianza de $\boldsymbol{\delta}$
$\sigma_\nu \in \mathbb{R}_{x>0}$	`real<lower=0> sigma_nu;`	varianza de $\boldsymbol{\nu}$
Parametrización no centrada		
$\widehat{\boldsymbol{\chi}} \in \mathbb{R}^{N_p}$	`vector[N_pat] chi_hat;`	parametrización no centrada de $\boldsymbol{\chi}$
$\widehat{\boldsymbol{\delta}} \in \mathbb{R}^{N_p}$	`vector[N_pat] delta_hat;`	parametrización no centrada de $\boldsymbol{\delta}$
Parámetros modelo de supervivencia		
$\log(\mathbf{h}) \in \mathbb{R}^{N_p}$	`vector[N_pat] log_h;`	logaritmo de la función de riesgo
$\log(\mathbf{S}) \in \mathbb{R}^{N_p}$	`vector[N_pat] log_S;`	logaritmo de la función de supervivencia
$\boldsymbol{\lambda} \in \mathbb{R}^{N_i}$	`vector<lower=0>[N_int] lambda;`	función de riesgo basal
$\boldsymbol{\gamma} \in \mathbb{R}^{N_{ch}}$	`vector[N_cov_h] gamma;`	efectos de las covariables sobre $\mathbf{h}$
$\gamma_\mu \in \mathbb{R}$	`real gamma_mu;`	efecto de la FEVI sobre la supervivencia

Capítulo 6

Metodología

6.1 | Variables del modelo

Para la estimación del modelo fijamos el número de intervalos en 20, de tal forma que haya un número similar de observaciones de la FEVI en todos.

Las covariables que utilizamos en el ajuste de los parámetros $\boldsymbol{\chi}$, $\boldsymbol{\delta}$ y $\mathbf{h}$ —y que conforman las matrices $\mathbf{C}_{\boldsymbol{\chi}}$, $\mathbf{C}_{\boldsymbol{\delta}}$ y $\mathbf{C}_{\mathbf{h}}$— son: sexo (mujer), edad inicial, años desde el diagnóstico de IC, VTD basal, etiología —cuyas categorías son: alcohólica, familiar, hipertensiva, isquémica, miocarditis, valvular y otras (resto menos frecuentes como congénita, periparto, neuromuscular o arrítmica); la etología idiopática se usa como la categoría de referencia por ser la más frecuente— IAM, HTA y DM. Por otro lado, para el parámetro $\boldsymbol{\nu}$ las únicas covariables que utilizamos —matriz $\mathbf{C}_{\boldsymbol{\nu}}$— son el sexo y la etiología isquémica (sí/no).

En la parte de tratamiento farmacológico nos centramos en las cuatro familias más importantes para el tratamiento de la IC: IECA, betabloqueantes, ARA-II y espironolactona. Consideramos que los tratamientos no tienen un efecto inmediato en los valores de FEVI, si no que su efecto se observa en los siguientes valores de FEVI observados en el seguimiento. Esto se incluye en el modelo considerando para cada intervalo el tratamiento farmacológico aplicado en el intervalo anterior.

6.2 | *Workflow* para estadística bayesiana

Como hemos comentado en el apartado 2.1 el análisis de datos se suele componer de varias partes: la depuración y procesamiento de los datos, la exploración y visualización de datos, la modelización e inferencia estadística, y la interpretación y comunicación de los resultados.

Cuando la estimación de un modelo se hace utilizando estadística bayesiana, al usar métodos computacionales para muestrear el espacio de probabilidad, se requiere de algunos pasos adicionales para evaluar que el modelo se ha ejecutado correctamente o detectar posibles problemas de computación o errores en el modelo. Este es un campo activo de investigación donde en los últimos años se han publicado algunas propuestas y recomendaciones en forma

de *workflows* (flujos de trabajo) con los pasos a seguir para el análisis, ver Gabry, Simpson *et al.* (2019), Betancourt (2020) y Gelman, Vehtari *et al.* (2020) y Schoot *et al.* (2021), entre otros.

En este trabajo seguiremos las recomendaciones principales de todos ellos: la elección y evaluación de la idoneidad de las distribuciones a *priori*, la especificación de diferentes opciones para la ejecución del modelo, la modificación de esas opciones para solucionar posibles problemas computacionales o de convergencia, o la evaluación del ajuste del modelo a partir de la distribución a *posteriori*.

6.3 | Distribuciones a *priori*

A la hora de establecer las *priors* de los parámetros se tiene cierta libertad de elección. Lo más recomendable, como hemos hecho en este trabajo, es establecer *priors* débilmente informativas (Schoot *et al.* 2021).

En el momento de escoger las distribuciones a *priori* es conveniente evaluar si las elecciones son muy restrictivas o demasiado vagas. Una forma de evaluarlo es mediante la simulación de datos sin tener en cuenta los datos reales observados. El hecho de que con un conjunto de *priors* podamos generar datos simulados similares a los reales es un buen indicador de que son apropiadas. Esto se conoce como *prior predictive checks* (Gabry, Simpson *et al.* 2019), y es lo que usaremos como indicador para evaluar nuestras distribuciones a *priori*.

6.4 | Ejecución del modelo

La estimación de nuestro modelo en Stan la realizaremos usando la interfaz CmdStanR para R, donde podemos especificar todas las opciones de ejecución del modelo: el número de iteraciones, el número de cadenas independientes, el número de cadenas a paralelizar (ejecución en paralelo para ahorrar tiempo de computación), los valores iniciales de los parámetros, etc.

El *script* para ejecutar el modelo desde R se muestra en el código 6.1. El fichero con el modelo final en Stan (código 5.1) lo hemos denominado `model_jmfevi.stan` y el archivo con los datos de entrada `pdata.json`. El modelo ajustado con todas las simulaciones que determinan las distribuciones a *posteriori* de los parámetros lo exportamos a un archivo que llamamos `cmdstanmcmc_jmfevi.RDS`.

```
library("cmdstanr")
mod <- cmdstan_model("model_jmfevi.stan")
fit <- mod$sample(data = "pdata.json",
                  chains = 4, parallel_chains = 4, seed = 78542, refresh = 100,
                  iter_warmup = 1500, iter_sampling = 3000,
                  adapt_delta = 0.95, init = 0.01,
                  output_dir = "samplingoutput")
```

```
fit$save_object(file = "cmdstanmcmc_jmfevi.RDS")
```

Código 6.1. **R**. *Script* de R para ejecutar un modelo de Stan con la interfaz CmdStanR.

La ejecución del modelo anterior se realizar directamente en R; sin embargo, como este proceso puede requerir varias horas de computación es recomendable ejecutar el proceso en segundo plano desde una línea de comandos como Bash. De esta manera, además, se generará un archivo `.Rout` en el que se guardarán todos los avisos y salidas que se imprimen por pantalla durante la ejecución del modelo.

La orden para ejecutar el código 6.1, cuyo archivo denominamos `cmdstanr_jmfevi.R`, se muestra en el código 6.2. La opción nohup previene que la ejecución se detenga cuando se cierra la sesión de usuario, siendo muy útil, por ejemplo, cuando se ejecuta el modelo en servidores compartidos.

```
nohup R CMD BATCH cmdstanr_jmfevi.R &
```

Código 6.2. **Bash**. Sentencia de Bash para la ejecución del *script* con el modelo.

6.5 | Diagnóstico del modelo

Uno de los métodos más comunes para evaluar la convergencia de nuestro modelo es ejecutar varias cadenas de Markov independientes y comprobar si en todas ellas se obtienen distribuciones similares. Esta comprobación se puede visualizar representando gráficamente las trayectorias de las diferentes cadenas, conocidas como *trace plots*. Junto a estos gráficos es habitual calcular algunos estadísticos como $\hat{R}$ (o *R-hat*) o el tamaño muestral efectivo, N_{eff}.

El estadístico $\hat{R}$ se calcula como un ratio entre varianzas de las cadenas. Si las cadenas convergen a una distribución de equilibrio común las varianzas serán similares para todas las cadenas y este estadístico tendrá un valor cercano a 1 (Vehtari *et al.* 2021). Por otro lado, el tamaño muestral efectivo representa el número de muestras independientes obtenidas de la distribución a *posteriori* para cada parámetro de interés; en este caso, se considera mejor cuanto mayor sea el ratio N_{eff}/N (Gabry y Mahr 2022).

Sobre otros lenguajes probabilísticos para la inferencia bayesiana, Stan tiene la ventaja de ser muy conservador, avisando ante cualquier sospecha de problema durante la ejecución del modelo[19]. Además, por defecto, se calculan algunos estadísticos como el E-BFMI (estimación bayesiana de la fracción de información faltante, del inglés *estimated bayesian fraction missing information*) que nos indica problemas con el algoritmo NUTS utilizado por Stan (más detalles en Betancourt 2016, 2017).

En este trabajo seguiremos las guías que se pueden encontrar en los casos de estudio del paquete bayesplot[20] (Gabry y Mahr 2022) y en las utilidades para el diagnóstico de CmdStan[21]

[19] En la web `https://mc-stan.org/misc/warnings.html` se puede ver la lista de avisos que puede lanzar Stan, qué significan y cómo solucionarlos.

[20] Disponibles en la web `https://mc-stan.org/bayesplot/articles/`.

[21] Más información en la guía de usuario disponible en `https://mc-stan.org/docs/cmdstan-guide/diagnose.html`.

(del que depende CmdStanR), que se basan en las recomendaciones de Gelman, J. B. Carlin *et al.* (2013), Betancourt (2017) y Carpenter *et al.* (2017) y Vehtari *et al.* (2021). Aunque los puntos de corte para algunos estadísticos pueden variar entre las distintas referencias, entre estas recomendaciones están: utilizar al menos cuatro cadenas independientes, desechar las iteraciones iniciales para la inferencia (conocidas como *warmup*, donde la distribución todavía se está explorando), considerar como aceptables los valores de E-BFMI mayores a 0.3, los de $\hat{R} < 1.1$, los valores N_{eff} mayores a 400 o $N_{\text{eff}}/N \geq 0.001$.

6.6 | *Software* y *hardware* utilizado

En todo este documento hemos utilizado principalmente los lenguajes de programación R v.4.2.1 (R Core Team 2022) y Stan v.2.29 (Stan Development Team 2021). Puntualmente también hemos hecho uso del intérprete de línea de comandos Bash (Free Software Foundation 2019).

La lectura de los datos, la fusión de todas las tablas recopiladas, su depuración y procesamiento lo hemos realizado con R utilizando los paquetes del tidyverse (Wickham, Averick *et al.* 2019): readxl (Wickham y Bryan 2022), dplyr (Wickham, François *et al.* 2022), tidyr (Wickham y Girlich 2022) y lubridate (Grolemund y Wickham 2011); además de otros como survival (Therneau 2022). En los pasos iniciales, para convertir las bases de datos mdb (formato del programa de bases de datos Access) a formato csv hemos hecho uso, además, de la librería para la línea de comandos MDBtools.

Los análisis descriptivos y exploratorios los hemos realizado con los paquetes tableone (Yoshida y Bartel 2022) y ggplot2 (Wickham 2016), mientras que las simulaciones de datos las hemos realizado con las funciones básicas de probabilidad de R y con los paquetes LaplaceDemon (Statisticat y LLC. 2021) y simsurv (Brilleman, Wolfe *et al.* 2021).

Finalmente, para la estimación e inferencia del modelo final hemos utilizado la interfaz CmdStanR (Gabry y Češnovar 2022) para interactuar con él desde el propio entorno de R. Para el manejo de las distribuciones a *posteriori* obtenidas del modelo usamos los paquetes posterior (Bürkner *et al.* 2022) y bayesplot (Gabry y Mahr 2022).

La ejecución y estimación del modelo completo la hemos realizado en un ordenador con sistema operativo GNU/Linux Kubuntu 20.04.4 LTS de 64 bits, núcleo Linux 5.4.0, 32GiB de memoria RAM y CPU 6×Intel®Core™i5-9600K a 3.70GHz.

Parte IV

Resultados

Capítulo 7

Análisis descriptivos y exploratorios

Partiendo de la base de datos completa de insuficiencia cardíaca de uso clínico en el HCUVA —que comprende unos 713 pacientes con ecocardiografías desde el año 1995 hasta el 2017— para este estudio se han seleccionado los 251 pacientes que inicialmente parten con una FEVI reducida (IC-FEr, FEVI $\leq 40\,\%$), que tienen al menos dos ecocardiografías disponibles y que no tienen valores perdidos en las variables de nuestro modelo.

Para estudiar la evolución temporal de la FEVI, se han analizado 1504 ecocardiografías, con una media de 5.99 ± 3.67 por paciente (mediana=5.00, RIQ=[3.00, 8.00]) y una media de 6.08 ± 4.53 (mediana=5.12, RIQ=[2.30, 8.89]) años de seguimiento para la FEVI.

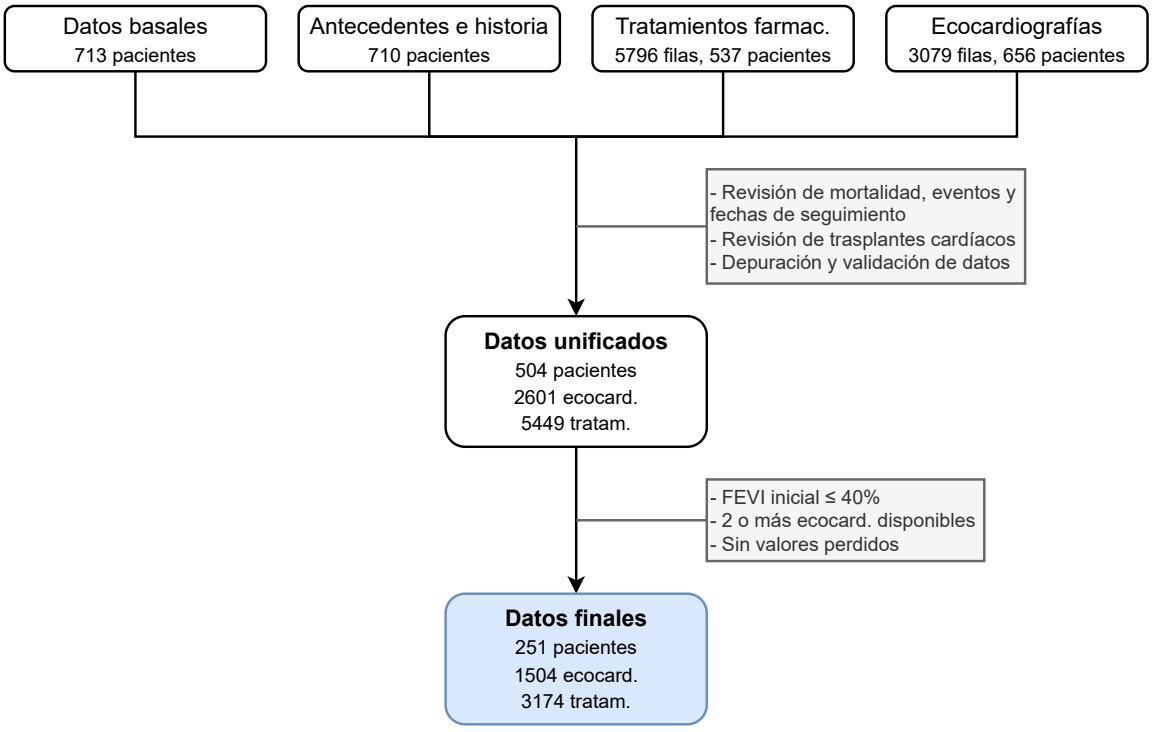

Figura 7.1. *Diagrama de flujo CONSORT de los datos utilizados donde se detalla el número de pacientes, de ecocardiografías y de registros de los tratamientos farmacológicos disponibles en cada paso antes de la selección de datos finales. *farmac.: farmacológicos; ecocard.: ecocardiografías; tratam.: tratamientos farmacológicos.*

7.1 | Descriptivos

Esta base datos de pacientes, consta mayoritariamente de hombres —197 (78.5 %) frente a 54 (21.5 %) de mujeres— con una edad en torno a los 52 años y una FEVI inicial de 27 % y, como se observa en el resumen de estadísticos descriptivos de la tabla 7.1, la etiología más común es la idiopática (espontánea o desconocida) con casi un 47 % de todos los casos.

Los fármacos iniciales más comunes son los IECA (con un 67 %) y los betabloqueantes (con un 64 %).

De la muestra total de 251 pacientes, casi la mitad muere o es trasplantada con un tiempo de seguimiento para muerte de casi 9 años.

Tabla 7.1. *Tabla de descriptivos basal. Se representa media±SD o n(%), según el caso.*

Variable	**Total (n=251)**
Edad inicial, años	52.08±14.10
Mujer	54 (21.5 %)
FEVI inicial, %	27.02±7.74
VTD inicial, ml	194.54±64.92
Etiología	
Idiopática	117 (46.6 %)
Isquémica	61 (24.3 %)
Hipertensiva	17 (6.8 %)
Familiar	16 (6.4 %)
Valvular	14 (5.6 %)
Alcohólica	13 (5.2 %)
Miocarditis	7 (2.8 %)
Otras	6 (2.4 %)
Historia	
Clase NYHA, III-IV	121 (48.2 %)
Infarto agudo de miocardio (IAM)	48 (19.1 %)
Insuficiencia renal (IR)	38 (15.1 %)
Hipertensión arterial (HTA)	115 (45.8 %)
Diabetes *mellitus* (DM)	63 (25.1 %)
Hipercolesterolemia	74 (29.5 %)
Fumador	106 (42.2 %)
Bebedor	242 (96.4 %)
Tratamiento farmacológico inicial	
IECA	168 (66.9 %)
ARA-II	50 (19.9 %)
Betabloqueantes	160 (63.7 %)
Espironolactona	121 (48.2 %)
Mortalidad	
Estado final	
Fallecido	91 (36.3 %)
Trasplantado (TC)	23 (9.2 %)
Vivo	137 (54.6 %)
Tiempo de seguimiento, años	8.94±4.66

7.2 | Exploratorios

Si observamos la evolución de los valores de FEVI para cada paciente, la figura 7.2 destaca las evoluciones de nueve de ellos, por lo general, estas no se aproximan del todo a una recta —en cuyo caso podría plantearse un modelo lineal—. Más bien se observa un aumento de la FEVI en los primeros años, manteniéndose estable en algunos casos y decayendo en otros, lo que sugiere utilizar un modelo no lineal para aproximar mejor el comportamiento estocástico de los datos, tal y como hemos propuesto en el apartado 5.1.

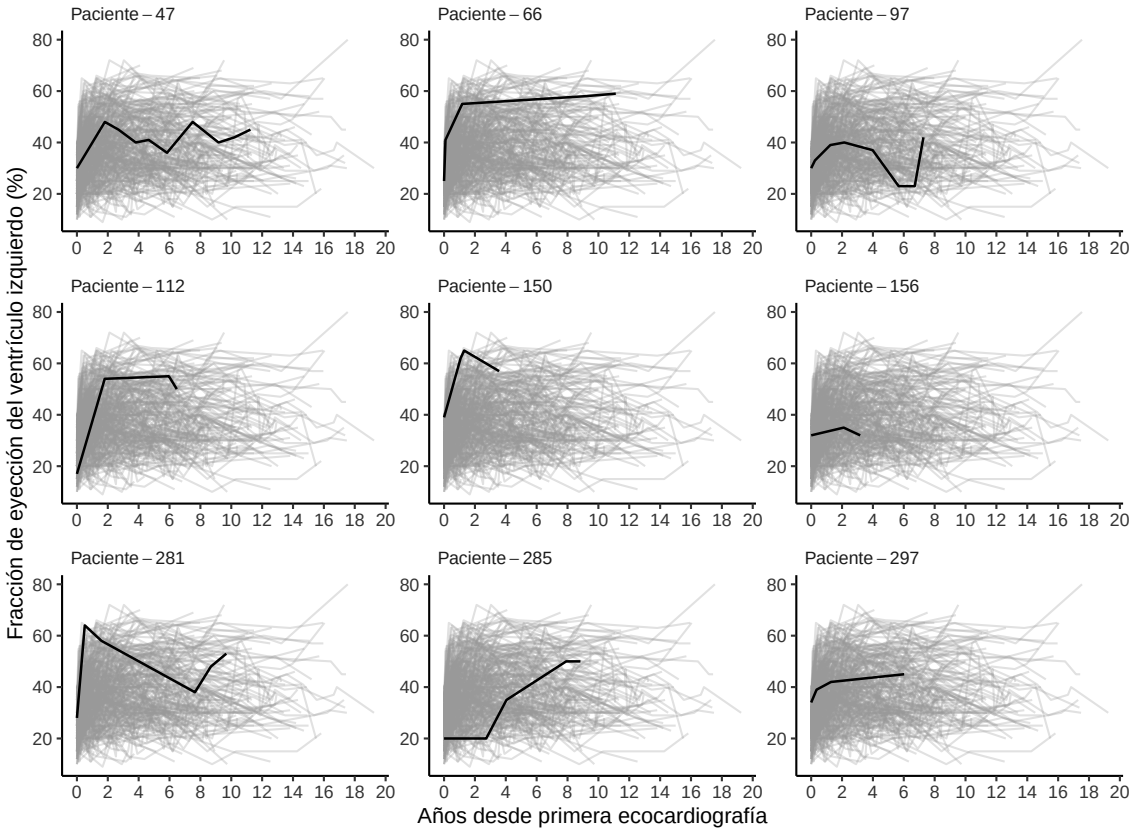

Figura 7.2. *Gráfico exploratorio de la evolución de la FEVI en nueve pacientes seleccionados aleatoriamente. En el eje OX se representa el tiempo (en años) desde la primera ecocardiografía; en el eje OY, la fracción de eyección del ventrículo izquierdo (FEVI, %). Las líneas en negro representan la evolución del paciente destacado sobre el resto de pacientes (líneas en gris).*

7.3 | Datos simulados

Una vez elegidas las probabilidades a *priori* de los parámetros (ver apartado 5.3.2), mediante un proceso de simulación, se pueden generar datos sintéticos en el rango de los datos observados (ver figura 7.3). Repitiendo la simulación (réplicas simuladas) se puede analizar diversas situaciones que podrían darse en la realidad y comparar gráficamente los datos simulados con los observados (figura 7.4).

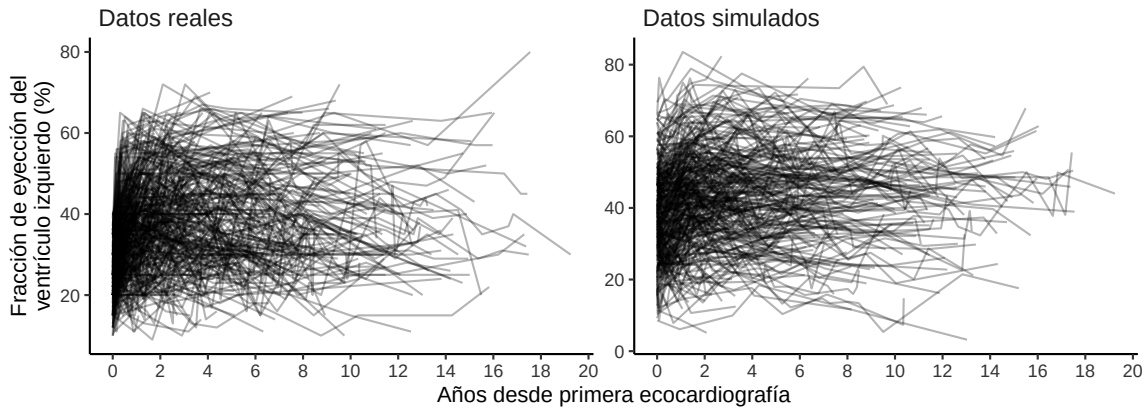

Figura 7.3. *Gráfico exploratorio de la evolución de la FEVI con los datos reales (izquierda) y los datos simulados (derecha) utilizando distribuciones a* priori *débilmente informativas. En el eje OX se representa el tiempo (en años) desde la primera ecocardiografía; en el eje OY, la fracción de eyección del ventrículo izquierdo (FEVI, %). Las líneas representan la evolución de cada paciente.*

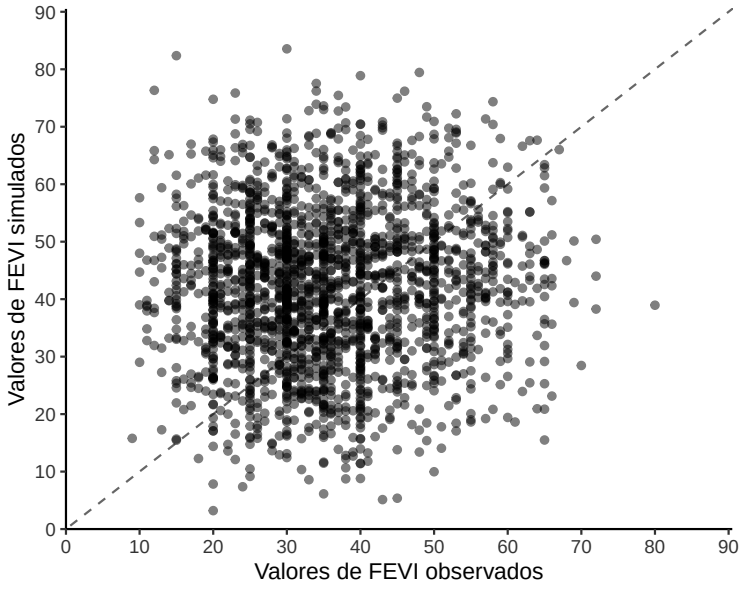

Figura 7.4. *Gráfico exploratorio de los valores de la FEVI reales (eje OX) frente a simulados (eje OY) utilizando distribuciones a* priori *débilmente informativas. La recta gris discontinua representa la función identidad (mismo valor en x que en y).*

Capítulo 8

Resultados del modelo

8.1 | Ajuste y diagnóstico del modelo

La computación para la estimación del modelo es inferior a una hora (ver tabla 8.1) ejecutándolo con cuatro cadenas de Markov en paralelo, 2000 iteraciones de calentamiento (o *warmup*), 6000 de muestreo (o *sampling*) y utilizando los datos reales.

TABLA 8.1. *Tiempos de ejecución del modelo (en minutos) por cadena e iteración (warmup o sampling).*

cadena	*warmup*	*sampling*	total
1	14.7	35.2	50.0
2	13.7	36.3	50.0
3	15.0	38.2	53.2
4	14.4	35.3	49.7

Durante la ejecución no se producen grandes problemas de computación. No aparecen problemas de divergencias en ninguna de las cuatro cadenas, los valores de *E-BFMI* se mantienen por encima de 0.3 (0.33, 0.33, 0.31 y 0.39, para cada cadena, respectivamente), los valores de N_{eff}/N son mayores de 0.001 y los de $\hat{R}$ menores de 1.05. Los únicos parámetros para los que se obtienen unos valores de N_{eff}/N algo más bajos que en el resto (menores a 0.1) son los relativos al parámetro ν, aunque conviene recordar que estos puntos de corte son orientativos (ver apartado 6.5).

En los *trace plots* representados en la figura 8.1 no se observan comportamientos extraños ni puntos en los que se concentren los parámetros de nuestro modelo.

8.2 | Evolución de la FEVI y factores predictores

Tal y como hemos representado la evolución temporal de la FEVI en nuestro modelo (ver apartado 5.1 y figura 5.1) esta viene determinada por tres parámetros que representan el punto inicial (χ), la velocidad de cambio (ν) y el aumento que se produce (δ).

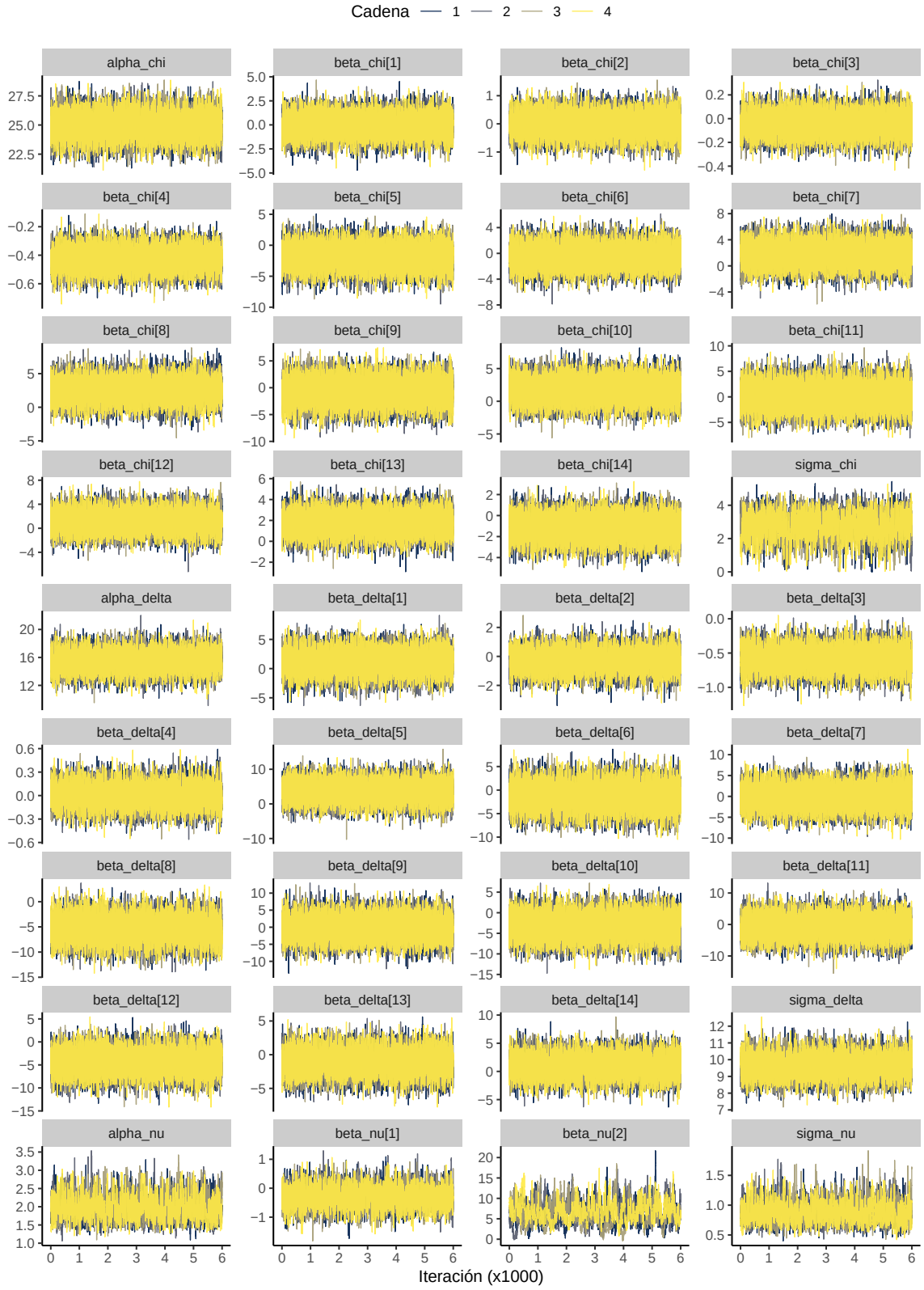

Figura 8.1. *Gráficos* traceplots *de los parámetros α, β y σ de χ, δ y ν. Se representan los valores (eje OY) de cada parámetro (paneles) en cada iteración (eje OX, en miles) para cada una de las cuatro cadenas independientes (colores).*

χ Valor inicial de la FEVI
δ Aumento de la FEVI hasta el punto de saturación
ν Velocidad de cambio

Cada uno de estos parámetros ha sido ajustado con un modelo multinivel con un valor α_* (denominado intercepto) y las distintas covariables lo aumentan o disminuyen un valor β_*. Recordemos que, al haber centrado las covariables, los interceptos α_* representan el punto inicial para un paciente hombre, con ninguna historia asociada (sin IAM, HTA, etc.) con etiología idiopática y con las variables numéricas en sus valores medios.

8.2.1. Valor inicial, χ

En la figura 8.2 se ha presentado el gráfico de coeficientes con las estimaciones de los parámetros relativos al valor inicial, χ, con: las medias, los intervalos de credibilidad al 95 % (denotados como IC95) y la probabilidad de cada parámetro de ser positivo $P(\cdot > 0)$ (denotado como PP).

Los pacientes comienzan con un valor medio de FEVI de 25.0 % (IC95 de [22.6, 26.8]; ver parámetro α_χ en la figura 8.2).

Este valor inicial de FEVI parece que no se ve afectado por el sexo, la edad o el tiempo desde el diagnóstico. Sí se estima un efecto negativo en el caso del VTD y uno positivo en la etiología isquémica. Para el VTD, por cada aumento de 10 ml el valor inicial de la FEVI se estima que es 0.43 % menor (IC95 de [-0.57, -0.28], PP<0.01; ver β_{χ_4} en la figura 8.2). Por otro lado, para la etiología isquémica se estima que el valor inicial de la FEVI es 2.55 % mayor (IC95 de [-0.64, 5.71], PP=0.94; ver β_{χ_8} en la figura 8.2).

De entre las variables de la historia clínica de los pacientes solo se observa un efecto destacable en la HTA, para el que se estima un aumento de 1.75 % (con un IC95 de [-0.27, 3.76], PP=0.95; ver $\beta_{\chi_{13}}$ en la figura 8.2). Sin embargo, el IAM y la DM no parecen afectar al valor inicial de la FEVI de los pacientes.

8.2.2. Aumento, δ

En la figura 8.3 se han presentado el gráfico de coeficientes de los parámetros relativos al aumento producido en los valores de la FEVI, δ.

El aumento de FEVI medio hasta el nivel de saturación se estima en 15.5 % (IC95 de [12.7, 18.3], PP>0.99; ver α_δ en la figura 8.3).

Igual que para el valor inicial χ, ni la edad ni el sexo parecen tener un efecto sobre el aumento δ. Sin embargo, sí pasa a observarse un descenso de este valor con el tiempo desde el diagnóstico. Por cada 10 años se estima un efecto de -0.59 (IC95 de [-0.91, -0.26], PP>0.99; ver $\beta_{\delta 3}$ en la figura 8.3). El VTD, por el contrario, pasa a no afectar al aumento de la FEVI.

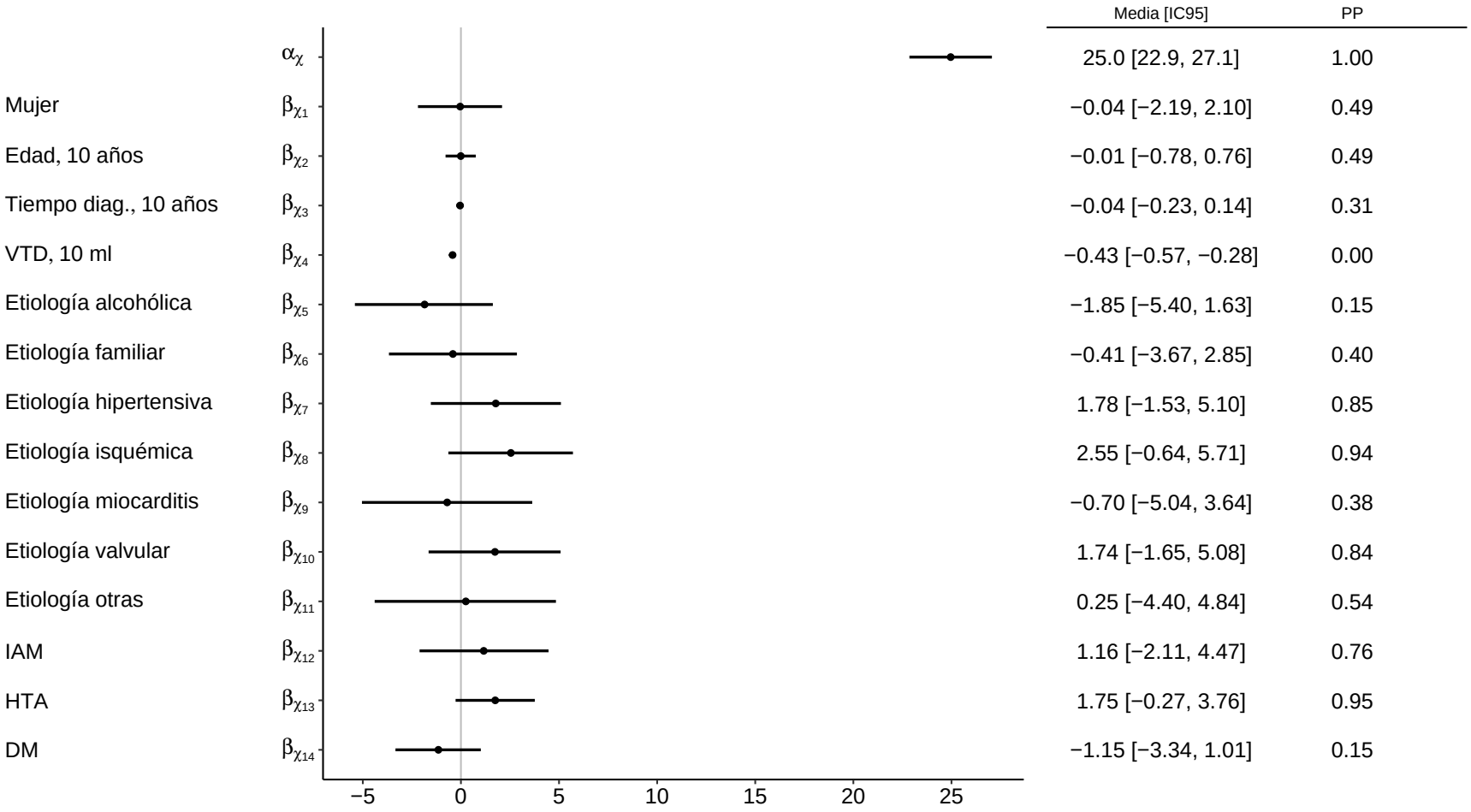

Figura 8.2. *Estimaciones de los parámetros que definen el valor inicial de la FEVI, χ. La correspondencia entre los parámetros y las covariables que representan se sitúa a la izquierda. En la parte central se representa el gráfico de coeficientes con su IC95. A la derecha se incluyen los resultados de las estimaciones y la PP de cada parámetro.*

De todas las etiologías, la que más destaca es la etiología isquémica, para la que se estima que el aumento de FEVI es 5.32 % menor respecto a la idiopática (IC95 de [-9.95, -0.75], PP=0.01; ver $\beta_{\delta 8}$ en la figura 8.3).

Para las variables recuperadas de la historia clínica del paciente solo destaca el efecto negativo del IAM sobre el aumento de la FEVI. Para el IAM se estima un efecto de -4.38 (IC95 de [-9.18, 0.39], PP=0.04; ver $\beta_{\delta 12}$ en la figura 8.3). Sin embargo, la HTA y la DM no parecen tener un efecto claro sobre el aumento.

8.2.3. Velocidad de saturación, ν

En la figura 8.4.A se resumen los resultados correspondientes a la velocidad de saturación, ν, que representa lo rápido que mejoran los pacientes sus valores de FEVI. El valor inicial para ν se estima en 1.98 (IC95 de [1.46, 2.65]; ver α_ν en la figura 8.4.A). Mientras que el sexo parece no afectar a la velocidad de mejora, sí se observa un efecto claro en la etiología isquémica respecto al resto, siendo la estimación de 6.38 (IC95 de [1.49, 12.1], PP>0.99; ver $\beta_{\nu 2}$ de la figura 8.4.A).

Una de las ventajas de la estadística bayesiana es que, debido a que disponemos de la distribución de a *posteriori* de los parámetros a través de simulaciones, podemos operar con ellos y obtener las distribuciones a *posteriori* para nuevas cantidades. De esta manera, a partir de los parámetros que determinan ν podemos calcular la vida media ($\log(2)/\nu$) para cada grupo de sexo y etiología isquémica.

Observar que, cuando los pacientes no tienen etiología isquémica, la vida media se sitúa para ambos sexos sobre los 0.4 años (en torno a 5 meses). Los tiempos se estiman en 0.36 años para los hombres (IC95 de [0.26, 0.47]; ver figura 8.4.B) y de 0.40 años para las mujeres (IC95 de [0.28, 0.64]; ver figura 8.4.B).

En presencia de etiología isquémica, en cambio, podemos ver que los tiempos son mucho más cortos, estando sobre los 0.1 años (algo más de un mes). Las estimaciones en este caso son de 0.09 años para los hombres (IC95 de [0.05, 0.20]; ver figura 8.4.B) y de 0.10 años para las mujeres (IC95 de [0.05, 0.22]; ver figura 8.4.B).

8.3 | Efecto de los fármacos sobre la FEVI, ϕ

Como habíamos mencionado en el apartado 6.1, para el estudio de los fármacos nos hemos centrado en el efecto que producen sobre los niveles de la FEVI. Concretamente, el efecto se modeló por intervalos, considerando que el tratamiento con un fármaco en un intervalo de tiempo afecta a los niveles de FEVI en el intervalo posterior.

La figura figura 8.4.C muestra los resultados obtenidos para las cuatro familias de fármacos que hemos estudiado con nuestro modelo. El único que parece destacar positivamente sobre el resto son los betabloqueantes. Por otra parte, destaca negativamente la espironolactona.

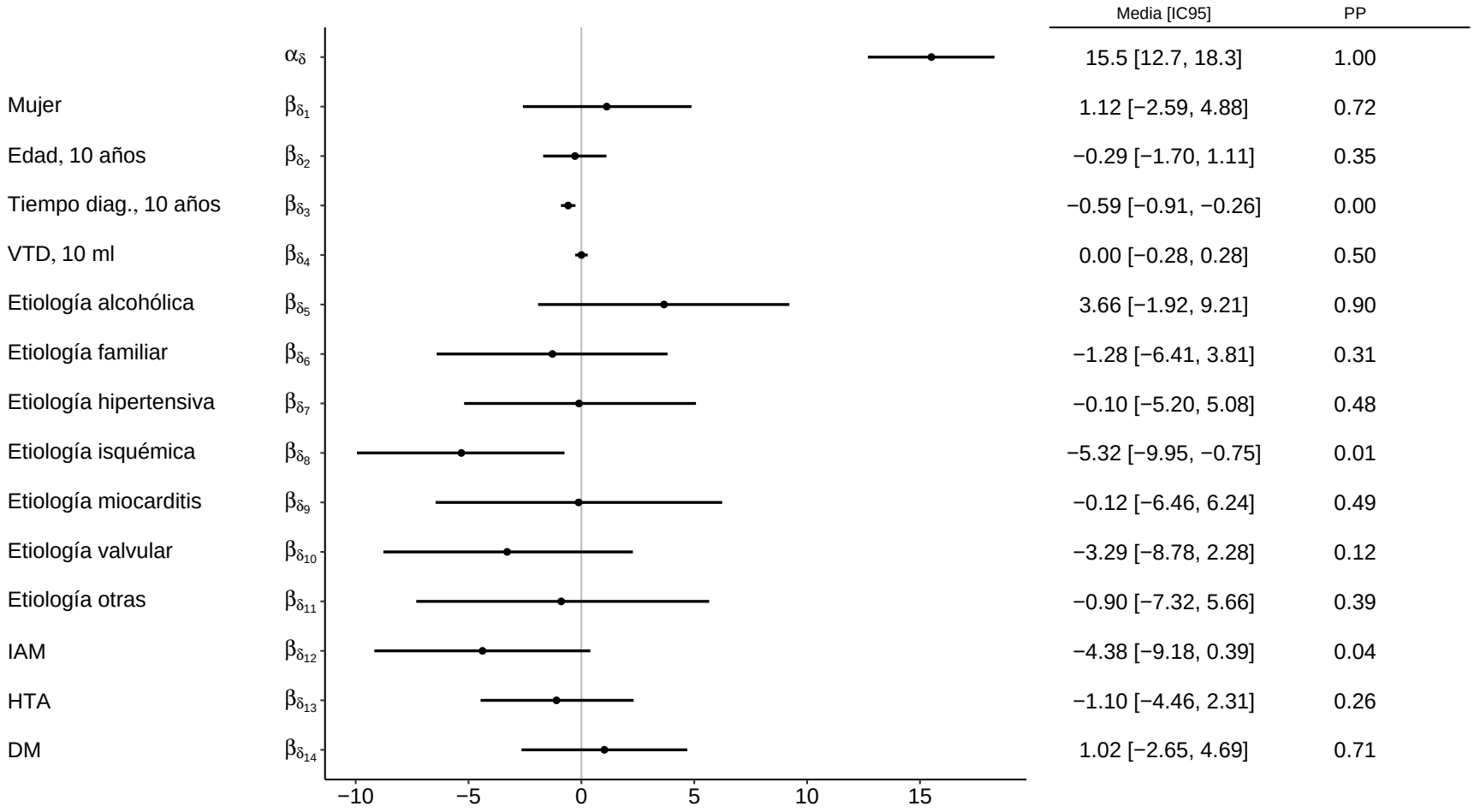

Figura 8.3. *Estimaciones de los parámetros que definen el aumento de la FEVI, δ. La correspondencia entre los parámetros y las covariables que representan se sitúa a la izquierda. En la parte central se representa el gráfico de coeficientes con su IC95. A la derecha se incluyen los resultados de las estimaciones y la PP de cada parámetro.*

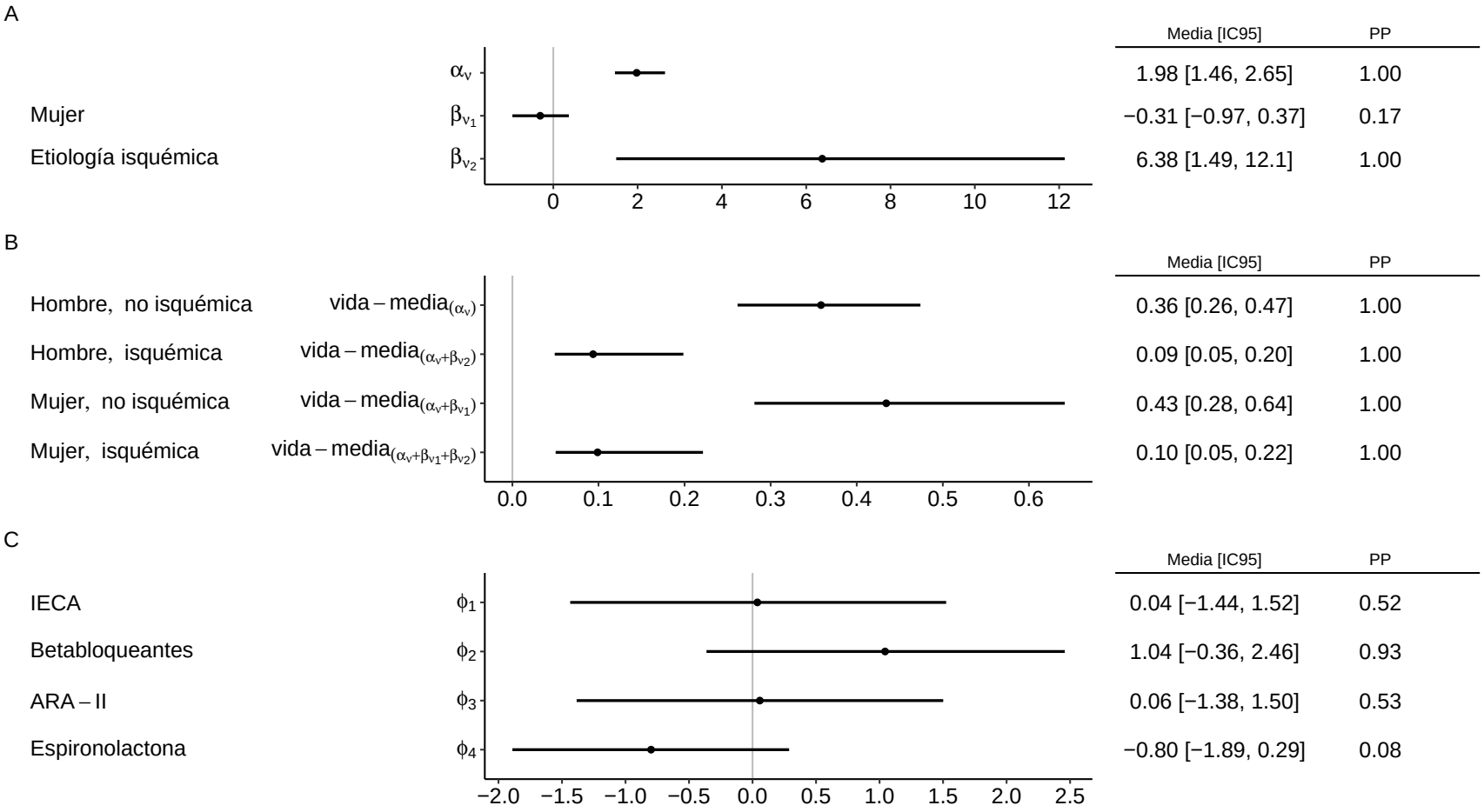

Figura 8.4. *Estimaciones de los parámetros que determinan la velocidad de saturación ν (A), las vidas medias para cada grupo de pacientes (B) y los efectos de los tratamientos ϕ (C). La correspondencia entre los parámetros y las covariables que representan se sitúa a la izquierda. En la parte central se representa el gráfico de coeficientes con su IC95. A la derecha se incluyen los resultados de las estimaciones y la PP de cada parámetro.*

Los IECA y los ARA-II resultan en efectos similares en los valores posteriores de la FEVI, en ambos casos se estima un aumento de alrededor de un 0.5 % (0.04 y 0.06, respectivamente; ver ϕ_1 y ϕ_2 en la figura 8.4.C).

La administración de betabloqueantes en un determinado intervalo parece que se asocia con un aumento de 1.04 % en los valores de la FEVI en el intervalo posterior (IC95 de [-0.36, 2.46], PP=0.93; ver ϕ_3 en la figura 8.4.C). Por el contrario, el efecto estimado para la espironolactona es una reducción de un 0.8 % (IC al 95 % de [-1.89, 0.29], PP=0.03; ver ϕ_4 en la figura 8.4.C).

8.4 | Efectos sobre la mortalidad

Los resultados correspondientes al análisis de supervivencia se han mostrado en el gráfico de coeficientes de la figura 8.5, donde se representan los HR, los IC95 y las PP para cada parámetro.

8.4.1. Efecto de las covariables, γ

Como es de esperar, el hecho de ser mujer se relaciona con una menor mortalidad y la edad inicial con una mayor mortalidad (ver figura 8.5.A). En concreto, para las mujeres se estima una reducción del 38 % ($1 - 0.62$) en el riesgo de muerte (HR=0.62, con IC95 de [0.35, 1.00], PP=0.03; ver el parámetro γ_1 en la figura 8.5.A), mientras que por cada 10 años adicionales en la edad inicial se estima un aumento del 48 % (HR=1.48, con IC95 de [1.21, 1.79], PP>0.99; ver γ_2 en la figura 8.5.A).

Aunque es un efecto pequeño, también se observa un aumento del riesgo para el tiempo desde el diagnóstico, con un aumento del riesgo estimado en un 4 % por cada 10 años (HR=1.04, con IC95 de [1.00, 1.07], PP=0.98; ver γ_3 en la figura 8.5.A).

Para el VTD no se observa una asociación clara con la mortalidad. Tampoco para las etiologías más frecuentes; solo se obtienen asociaciones con una mayor mortalidad en la valvular (HR=2.33, con IC95 de [0.99, 4.39], PP=0.97; ver γ_{10} en la figura 8.5.A), en la miocarditis (HR=3.89, con IC95 de [0.74, 10.8], PP=0.94; ver γ_9 en la figura 8.5.A) y en el resto menos frecuentes, que hemos catalogado como «otras» (HR=6.86, con IC95 de [1.74, 16.6], PP>0.99; ver γ_{11} en la figura 8.5.A).

Respecto a los antecedentes, mientras que para el IAM no se obtiene un efecto destacable sí se observa una asociación de la HTA (HR=1.69, IC95 de [1.08, 2.55], PP=0.99) y la DM (HR=1.49, IC95 de [0.92, 2.25], PP=0.95) (ver los parámetros γ_{13} y γ_{14}, respectivamente, de la figura 8.5.A).

8.4.2. Efecto de la FEVI, γ_μ

Los valores longitudinales de la FEVI se asocian claramente con una reducción del riesgo de muerte. La estimación para esta asociación —que hemos establecido mediante el parámetro γ_μ en la parte de supervivencia del modelo— es que por cada aumento de 10 % en la FEVI se produce un reducción del riesgo de muerte de un 18 % (HR=0.82, con IC95 de [0.71, 0.94], PP<0.01; ver γ_μ en la figura 8.5.B).

8.5 | Remodelado cardíaco

Los resultados obtenidos sobre el remodelado cardíaco, mediante la medida del VTD, han reportado que se asocia con unos valores iniciales menores de FEVI (ver figura 8.2). No se observa, sin embargo, un efecto sobre el aumento posterior de la FEVI (ver figura 8.3).

Respecto a la mortalidad, aunque nuestros resultados parecen indicar una ligera asociación, la incertidumbre no es lo suficientemente baja para poder afirmarlo (ver figura 8.5).

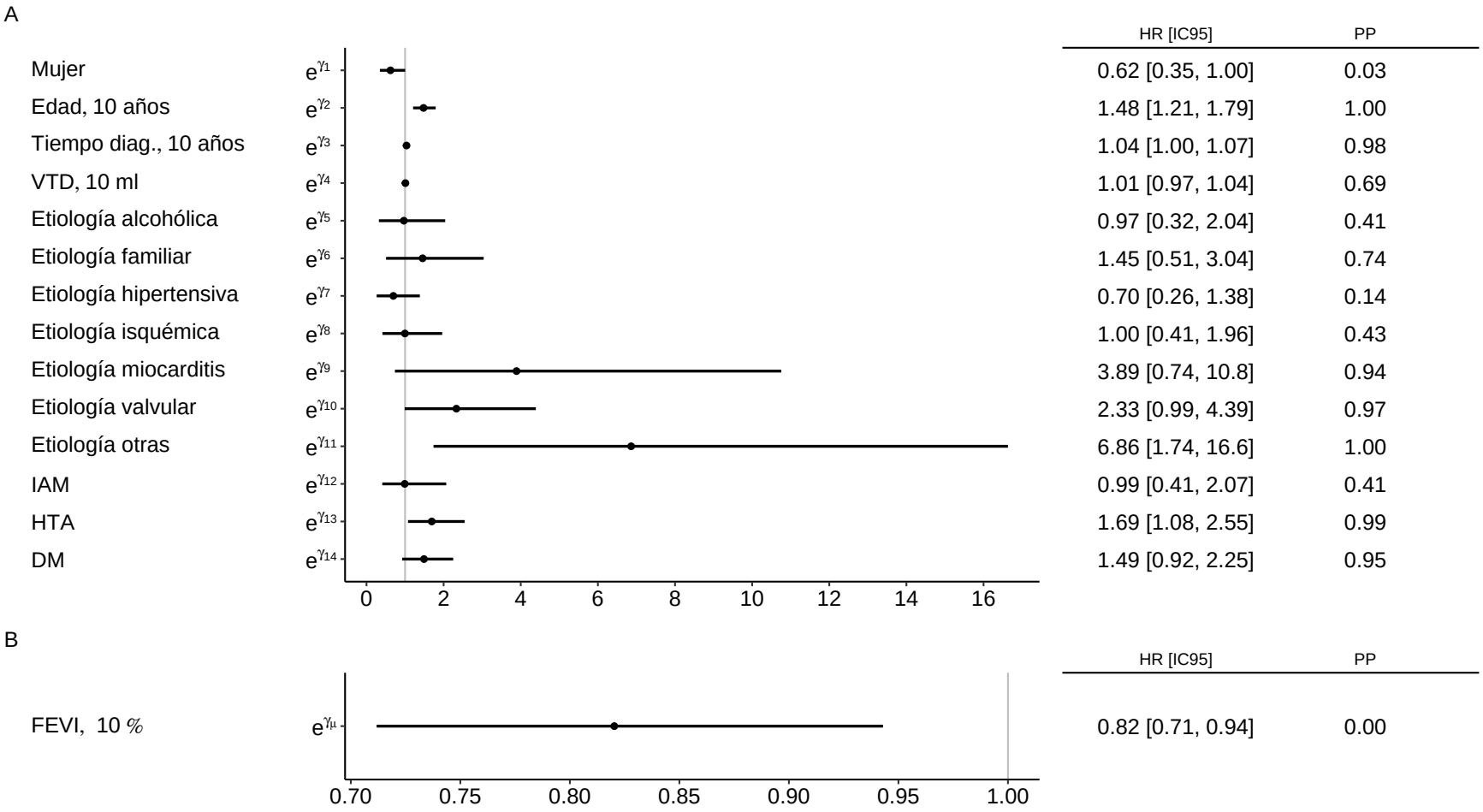

Figura 8.5. *Estimaciones de los parámetros que determinan el efecto de las variables γ (A) y el efecto de los valores de la FEVI γ_μ (B) sobre la mortalidad. La correspondencia entre los parámetros y las variables que representan se sitúa a la izquierda. En la parte central se representa el gráfico con los HR y su IC95. A la derecha se puede ver la tabla con los detalles de las estimaciones de los HR y la PP de cada parámetro.*

Parte V

Discusión y conclusiones

Capítulo 9

Discusión

9.1 | Aportación del modelo a la estadística médica

En enfermedades y síndromes complejos, como la IC, son múltiples las variables clínicas y los factores que afectan tanto a su pronóstico como a su desarrollo y evolución (McDonagh *et al.* 2022). En estos casos, para tener una visión global y más completa de la patología se requieren técnicas y modelos más avanzados que permiten describir las posibles relaciones que puedan existir en la realidad (Claggett *et al.* 2013).

La inferencia bayesiana tiene una dificultad inicial mayor que la frecuentista, mucho más conocida y utilizada, y requiere un cambio de mentalidad y proceder. No se hace necesario dedicar tiempo a la búsqueda de un modelo que se ajuste bien a los datos sino que hay que dedicarlo a la exploración de los datos y a estudiar cómo se han podido generar (McElreath 2020). Así, en la práctica, además de tener algunas ventajas como una mejor estimación de los parámetros en modelos multinivel, el enfoque bayesiano ofrece esta flexibilidad en la modelización estadística.

En este trabajo, haciendo uso de la estadística bayesiana, hemos desarrollado e implementado en el lenguaje probabilístico Stan un modelo estadístico, que nos ha permitido, de forma conjunta:

1. describir mejor la evolución longitudinal de la FEVI en pacientes con IC-FEr,
2. estimar el efecto que tienen características clínicas como la cardiopatía isquémica, el remodelado cardíaco o el tratamiento farmacológico en la mejora de la FEVI, y
3. estudiar la asociación entre la mejora en los valores de la FEVI y el pronóstico en términos de mortalidad.

El primero de estos puntos lo hemos conseguido estableciendo una función para la evolución que tiene en cuenta: el valor inicial de la FEVI, el aumento total hasta un punto de saturación (al que se aproximan los valores con el tiempo) y la velocidad a la que se produce este aumento. La elección de esta función se basa tanto en la exploración de nuestros propios datos como en la evolución que se puede observar en otros trabajos como Lupón *et al.* (2018).

Los puntos segundo y tercero los hemos conseguido mediante la implementación de un *joint model* para datos longitudinales y de supervivencia constante por intervalos.

Aunque de forma separada se podrían haber planteado otras formas de análisis, e. g. utilizando paquetes y librerías disponibles para estimar *joint models* (Rizopoulos 2010; Goodrich *et al.* 2022), en nuestro modelo lo hemos planteado de forma conjunta y utilizamos una función que, en nuestra opinión, describe mejor la evolución que se observa de la FEVI.

Para estudiar trayectorias no lineales de evolución de un marcador, un método habitual es la regresión local (también conocida como LOESS, del inglés *locally estimated scatterplot smoothing*). Ejemplos en los que se utiliza esta técnica son Lupón *et al.* (2018), Julián *et al.* (2020) o K. Wang *et al.* (2021), donde, al igual que nosotros, el marcador estudiado es la FEVI. Este tipo de regresión ofrece una gran flexibilidad para ajustar curvas no lineales a los datos pero, dado que es una técnica no paramétrica, no es fácil de representar matemáticamente ni de interpretar más allá de la curva obtenida. Por ello, se suele combinar con otros métodos de regresión como la regresión multinivel, como se hace en los tres artículos recién citados.

Con nuestro modelo se ajusta una regresión no lineal paramétrica con la forma específica de evolución de la FEVI, lo que permite considerar el efecto de las covariables, los tratamientos y la supervivencia a la vez.

Adicionalmente, este trabajo también refleja la utilidad de la inferencia bayesiana en el análisis de datos clínicos. Este documento y nuestro modelo, que está disponible en el repositorio `https://codeberg.org/alvarohv/jmfevi-stan`, puede servir de base que modificar y ajustar la evolución de otro tipo de variables o adaptarse para estimar otros modelos.

9.2 | Aportación del modelo a nivel clínico

Los resultados que obtenemos del ajuste de nuestro modelo a los datos disponibles proporcionan unos valores iniciales para la FEVI de un 25 % [22.9, 27.1] y un aumento medio de un 15.5 % [12.7, 18.3]. Estos valores de inicio concuerdan con los valores de la FEVI que se observan en estudios como el CHAMP-HF (DeVore *et al.* 2020), donde se estiman en un 29 ± 7.9 %. El aumento, al estar definido como la mejora de la FEVI hasta el punto de saturación, es más difícil de comparar con otros estudios. En el mencionado CHAMP-HF se estima un aumento al seguimiento de 6 ± 12.3 %. Este seguimiento lo establecen entre las 9 semanas y los 2 años. Así, una posible comparación que podríamos hacer sería respecto a la mitad de nuestro aumento, que se alcanza en la vida media y se estima en torno a 1 y 5 meses. En este caso, nuestra estimación para el aumento sería de $15.5/2 = 7.75$ %, algo superior al del estudio CHAMP-HF, pero en un rango esperable.

Respecto a los antecedentes y la historia clínica, nuestros resultados muestran que el IAM se asocia con un aumento menor de la FEVI (-4.38 % [-9.18, 0.39]) y la HTA con unos valores iniciales mayores (+1.75 % [-0.27, 3.76]), al igual que el tiempo desde el diagnóstico (-0.59 % [-0.91, -0.26]).

La etiología isquémica merece una mención aparte, ya que, se podría asociar con unos valores iniciales mayores (2.55 % [-0.64, 5.71]) y, al mismo tiempo, con un aumento menor de la FEVI (-5.32 % [-9.95, -0.75]) y con una mayor velocidad (alrededor de un mes en alcanzar la mitad del aumento). En conjunto, estos resultados sugieren que la evolución en los pacientes con esta etiología es más suave o constante que el resto.

En relación al remodelado cardíaco, evaluado a través del VTD, los resultados nos muestran que puede estar relacionado con unos valores iniciales menores de la FEVI (-0.43 % [-0.57, -0.28] por cada 10 ml), no asociándose con el aumento de los valores posteriores.

Respecto a los tratamientos farmacológicos, nuestros resultados no muestran un efecto destacado de ninguno de ellos sobre el resto. Solo los betabloqueantes parecen tener un efecto algo mayor que los demás y la espironolactona un efecto menor. Esto, conviene aclarar, no significa que los fármacos no tengan un efecto sobre los valores de FEVI. Los efectos que se pueden calcular estadísticamente siempre se establecen en comparación a una referencia. En nuestra población, todos los pacientes reciben algún tipo de tratamiento farmacológico durante su evolución, lo cual dificulta poder diferenciar el efecto concreto de alguno de ellos. Cabe señalar, además, que nuestro tamaño muestral relativamente pequeño también puede ser un condicionante. En nuestro modelo tampoco hemos establecido relaciones entre estos fármacos con otras variables clínicas o diferenciado la posibilidad de que el efecto a corto plazo no sea el mismo que a largo plazo.

Aunque nuestros resultados para los tratamientos no sean concluyentes, los efectos que obtenemos sobre la FEVI son consistentes con los publicados en otros trabajos y ensayos clínicos como el el SOLVD (Greenberg *et al.* 1995), el CARMEN (Remme *et al.* 2004) o el REVERT (Colucci *et al.* 2007). La principal conclusión que se obtiene tanto en el estudio CARMEN como en el REVERT es un beneficio de los betabloqueantes en el remodelado cardíaco inverso, donde se observa un aumento claro de la FEVI a partir de los 6 meses (comparado con el grupo placebo). En el CARMEN, además, los IECA no mostraron beneficio, coincidiendo con el SOLVD, donde no se observan cambios en la FEVI a los 12 meses. Respecto a los antialdosterónicos como la espironolactona, la evidencia clínica es limitada y, en algunos casos, inconsistente (Udelson *et al.* 2010; Vizzardi *et al.* 2010).

Los resultados que obtenemos de mortalidad, en muchos casos, son esperables, como el menor riesgo de muerte estimado para las mujeres o el mayor riesgo según la edad inicial de los pacientes. También se asocian con un mayor riesgo la HTA, la DM o el tiempo desde el diagnóstico. Dentro de las distintas etiologías, destaca un mayor riesgo de muerte de la valvular junto con la miocarditis y el resto (otras). Estas últimas etiologías coincide que son las menos frecuentes, por lo que puede ser indicativo de que no están tan estudiadas como el resto o una limitación de nuestro trabajo para estimar sus efectos con más precisión (hecho que coincide también con la mayor incertidumbre que obtenemos en estos casos) y que se requiere de estudios posteriores para mejorar su estimación.

Por último, se ha obtenido una clara asociación entre la mejora de los valores de la FEVI y

un menor riesgo de mortalidad. Por cada mejora de un 10 % en la FEVI el riesgo de muerte se reduce un 18 %. Esta reducción de riesgo es importante, sobre todo si la comparamos con la reducción esperable en las mujeres, un 38 %, o el aumento debido a la edad, un 48 %.

9.3 | Limitaciones y líneas futuras de trabajo

Una de las principales limitaciones de nuestro estudio es el bajo número de pacientes final que analizamos. Esto hace que no podamos obtener mejores estimaciones para algunos parámetros y que no podamos establecer más interacciones o un mayor número de efectos en nuestro modelo.

Para estudiar el efecto de los distintos tratamientos farmacológicos, como hemos comentado en el apartado anterior, el tamaño muestral es un factor importante. Además de no poder estimar con más seguridad los valores de los efectos en nuestro modelo, tampoco nos ha permitido modelizar posibles interacciones entre los fármacos ni efectos a corto y largo plazo.

Otra de las limitaciones de nuestro modelo ha sido en la parte de supervivencia, dado que para evaluar el pronóstico de los pacientes, nos hemos centrado solamente en mortalidad total. La información relativa a causas de muerte no estaban disponibles, tampoco hemos utilizado la información de reingresos de los pacientes en el seguimiento. Esto se podría plantear como nuevas líneas de trabajo. Revisando y completando las causas de muerte podríamos evaluar en nuestro modelo el evento «muerte por IC» o, de forma similar, el evento «reingreso». Para ello, un punto importante a tener en cuenta es que estos eventos pueden no ocurrir por producirse otro evento que lo impida (e. g. un paciente puede no tener un reingreso por otra causa de muerte). Este fenómeno se denomina «eventos competitivos» y para modelizarlos habría que plantear un modelo que los tenga en cuenta como la regresión de Fine-Gray (Pintilie 2011).

Otra línea importante de trabajo para el futuro es investigar más profundamente el papel del remodelado cardíaco en la mejora de los pacientes y su pronóstico. De la misma manera que la FEVI juega un papel central en nuestro modelo, ya que es la variable central para la que estimamos la evolución, lo mismo podríamos hacer para el VTD. Así, en lugar de estimar solo el efecto de los valores iniciales del VTD sobre los valores de la FEVI, se podría modelizar la evolución conjunta de la FEVI y del VTD. Esto conllevaría pasar de un modelo univariante a un modelo multivariante, con dos variables longitudinales dependientes del resto y en el que tendríamos que tener en cuenta la posible correlación existente entre ambos marcadores a lo largo del tiempo. Esta modelización multivariante para estimar la evolución conjunta de la FEVI y del VTD se podría hacer generalizando la función utilizada en el apartado 5.1 a dos dimensiones como se plantea en trabajos como Vatiwutipong y Phewchean (2019) y Oravecz *et al.* (2016).

Capítulo 10

Conclusiones

1. El modelo propuesto en este trabajo permite describir la evolución a largo plazo de un parámetro relevante de función cardíaca, la FEVI, en pacientes que presentan valores reducidos, tomando en consideración la mortalidad y su interacción con los tratamientos farmacológicos recibidos y variables clínicas del paciente y su enfermedad.
2. La estimación bayesiana del modelo propuesto permite determinar los valores iniciales de la FEVI, su aumento hasta el nivel de saturación y la velocidad de dicho aumento.
3. Los valores iniciales de la FEVI, estimados en un 25 %, presentan una mejoría temprana que se satura antes de los 6 meses y que alcanza un incremento medio de un 15.5 %.
4. La evolución de la FEVI en el tiempo se relaciona directamente con la mortalidad de los pacientes.
5. Las variables que principalmente afectan a la evolución de la FEVI son el antecedente de infarto de miocardio, la etiología isquémica y el tiempo desde el diagnóstico inicial.
6. El remodelado cardíaco, evaluado como grado de dilatación ventricular al final de la diástole, afecta a los valores iniciales de la FEVI pero no a su evolución posterior, y lo hacen en sentido negativo.
7. De entre todos los fármacos estudiados, los betabloqueantes muestran un efecto positivo superior al resto, aunque no se encontraron diferencias concluyentes entre cada uno de ellos.
8. Este trabajo de tesis aporta un modelo de estudio, aplicable a variables biológicas o médicas con trascendencia clínica, cuyo estudio a lo largo del tiempo es relevante y que están influidas por diferentes características de los pacientes o intervenciones terapéuticas a lo largo del tiempo.

Parte VI

Publicaciones

Capítulo 11

Publicaciones

El trabajado desarrollado en esta tesis por el autor, Álvaro Hernández Vicente, ha dado lugar a la presentación de un póster en el congreso internacional de insuficiencia cardíaca organizado por la Sociedad Europea de Cardiología y a una comunicación oral en las jornadas científicas del IMIB-Arrixaca, estando pendiente la publicación de un artículo científico con los resultados finales.

- **A. Hernandez-Vicente**, C. Marti-Gomez, D. Saura-Espin, M. J. Oliva-Sandoval, F. Pastor-Perez, A. Riquelme-Perez, C. Sanchez-Perez, M. C. Asensio-Lopez, M. D. Espinosa, M. Gomez-Molina, C. Caro, J. M. Vivo-Molina, M. Franco-Nicolas, F. Sanchez-Cabo y D. A. Pascual-Figal (2022). «Longitudinal analysis of ejection fraction in heart failure with reduced ejection fraction patients: pharmacological treatment effects». En: *Heart Failure 2022 and the World Congress on Acute Heart Failure*. Vol. 24. S2. Póster. Heart Failure Association of the ESC. Madrid, Spain: Wiley, pág. 154. DOI: 10.1002/ejhf.2569
- **Á. Hernández Vicente**, C. Martí Gómez-Aldaravi, F. Sánchez-Cabo, G. De La Morena Valenzuela, F. Pastor Perez, J. M. Vivo Molina, M. Franco Nicolás y D. A. Pascual Figal (2020). «Análisis longitudinal de la fracción de eyección del ventrículo izquierdo enpacientes con insuficiencia cardíaca y fracción reducida: evolución y efectode los tratamientos». En: *V Jornadas Científicas del IMIB-Arrixaca*. Comunicación oral. Instituto Murciano de Investigación Biosanitaria Virgen de la Arrixaca. Murcia, Spain

Además, la formación y experiencia adquiridas por el autor durante la realización de esta tesis han contribuido a numerosas comunicaciones a congresos y a la publicación de los siguientes artículos en revistas científicas:

- L. Mohamed-Salem, J. J. Santos-Mateo, J. Sanchez-Serna, **Á. Hernández-Vicente**, R. Reyes-Marle, M. I. Castellón Sánchez, M. A. Claver-Valderas, E. Gonzalez-Vioque, F. J. Haro-del Moral, P. García-Pavía y D. A. Pascual-Figal (2018). «Prevalence of wild type ATTR assessed as myocardial uptake in bone scan in the elderly population». En: *International Journal of Cardiology* 270, págs. 192-196. DOI: 10.1016/j.ijcard.2018.06.006
- M. T. Pérez-Martínez, J. Lacunza-Ruiz, J. García de Lara, J. A. Noguera-Velasco, A. Lax, **Á. Hernández-Vicente**, M. C. Asensio-López, J. L. Januzzi, B. Ibañez y D. A. Pascual-Figal (2018). «Noncardiac production of soluble ST2 in ST-segment elevation myocardial

infarction». En: *Journal of the American College of Cardiology* 72.12, págs. 1429-1430. DOI: 10.1016/j.jacc.2018.06.062

- L. Fernández-Gassó, L. Hernando-Arizaleta, J. A. Palomar-Rodríguez, M. V. Abellán-Pérez, **Á. Hernández-Vicente** y D. A. Pascual-Figal (2019b). «Population-based study of first hospitalizations for heart failure and the interaction between readmissions and survival». En: *Revista Española de Cardiología (English Edition)* 72.9, págs. 740-748. DOI: 10.1016/j.rec.2018.08.014
- D. A. Pascual-Figal, A. Bayes-Genis, M. C. Asensio-Lopez, **A. Hernández-Vicente**, I. Garrido-Bravo, F. Pastor-Perez, J. Díez, B. Ibáñez y A. Lax (2019). «The Interleukin-1 Axis and Risk of Death in Patients With Acutely Decompensated Heart Failure». En: *Journal of the American College of Cardiology* 73.9, págs. 1016-1025. DOI: 10.1016/j.jacc.2018.11.054
- J. Sanchez-Serna, M. Martinez-Villanueva, **Á. Hernández-Vicente**, M. C. Asensio-Lopez, J. A. Noguera y D. A. Pascual-Figal (2019). «Galectina-3 as a biomarker of acute kidney injury risk in patients with decompensated heart failure». En: *Revista Clínica Española (English Edition)* 219.6, págs. 315-319. DOI: 10.1016/j.rceng.2019.02.005
- M. d. C. Asensio Lopez, A. Lax, **A. Hernandez Vicente**, E. Saura Guillen, A. Hernandez-Martinez, M. J. Fernandez del Palacio, A. Bayes-Genis y D. A. Pascual Figal (2020). «Empagliflozin improves post-infarction cardiac remodeling through GTP enzyme cyclohydrolase 1 and irrespective of diabetes status». En: *Scientific Reports* 10.1. DOI: 10.1038/s41598-020-70454-8
- J. Sanchez-Serna, **A. Hernandez-Vicente**, I. P. Garrido-Bravo, F. Pastor-Perez, J. A. Noguera-Velasco, T. Casas-Pina, A. I. Rodriguez-Serrano, J. Núñez y D. Pascual-Figal (2020). «Impact of pre-hospital renal function on the detection of acute kidney injury in acute decompensated heart failure». En: *European Journal of Internal Medicine* 77, págs. 66-72. DOI: 10.1016/j.ejim.2020.02.028
- A. Aimo, D. A. Pascual-Figal, A. Barison, G. Cediel, **Á. Hernández Vicente**, L. F. Saccaro, M. Emdin y A. Bayes-Genis (2021). «Colchicine for the treatment of coronary artery disease». En: *Trends in Cardiovascular Medicine* 31.8, págs. 497-504. DOI: 10.1016/j.tcm.2020.10.007
- D. A. Pascual-Figal, A. Bayes-Genis, M. Díez-Díez, **Á. Hernández-Vicente**, D. Vázquez-Andrés, J. de la Barrera, E. Vazquez, A. Quintas, M. A. Zuriaga, M. C. Asensio-López, A. Dopazo, F. Sánchez-Cabo y J. J. Fuster (2021). «Clonal Hematopoiesis and Risk of Progression of Heart Failure With Reduced Left Ventricular Ejection Fraction». En: *Journal of the American College of Cardiology* 77.14, págs. 1747-1759. DOI: 10.1016/j.jacc.2021.02.028
- D. A. Pascual-Figal, A. E. Roura-Piloto, E. Moral-Escudero, E. Bernal, H. Albendin-Iglesias, M. T. Pérez-Martínez, J. A. Noguera-Velasco, I. Cebreiros-López, **Á. Hernández-Vicente**,

D. Vázquez-Andrés, C. Sánchez-Pérez, A. Khan, F. Sánchez-Cabo y E. García-Vázquez (2021). «Colchicine in Recently Hospitalized Patients with COVID-19: A Randomized Controlled Trial (COL-COVID)». En: *International Journal of General Medicine* Volume 14, págs. 5517-5526. DOI: 10.2147/ijgm.s329810

- J. A. Ros-Lucas, D. A. Pascual-Figal, J. A. Noguera-Velasco, **Á. Hernández-Vicente**, I. Cebreiros-López, M. Arnaldos-Carrillo, I. M. Martínez-Ardil, E. García-Vázquez, M. Aparicio-Vicente, E. Solana-Martínez, S. Y. Ruiz-Martínez, L. Fernández-Mula, R. Andujar-Espinosa, B. Fernández-Suarez, M. D. Sánchez-Caro, C. Peñalver-Mellado y F. J. Ruiz-López (2022). «CA 15-3 prognostic biomarker in SARS-CoV-2 pneumonia». En: *Scientific Reports* 12.1. DOI: 10.1038/s41598-022-10726-7

- S. Sánchez-Cámara, M. C. Asensio-López, M. Royo-Villanova, F. Soler, R. Jara-Rubio, J. F. Garrido-Peñalver, E. Pinar, **Á. Hernández-Vicente**, J. A. Hurtado, A. Lax y D. A. Pascual-Figal (2022). «Critical warm ischemia time point for cardiac donation after circulatory death». En: *American Journal of Transplantation* 22.5, págs. 1321-1328. DOI: 10.1111/ajt.16987

Apéndices

Apéndice A

Notación

Tipografía y código

En este documento usaremos la *letra cursiva* para denotar palabras en un idioma distinto del castellano (además de para los títulos de citas bibliográficas) y la `letra mecanográfica` para el código, funciones, nombres de variables, etc. Las partes de código amplias que necesiten más de una línea se representarán en cajas sombreadas. Al pie se especificará el lenguaje en el que están escritas —Stan (Stan Development Team 2021) o R (R Core Team 2022), principalmente— y la descripción. Por ejemplo:

```
data {
  int<lower=0> N;
  vector[N] y;
}
parameters {
  real theta;
  real<lower=0> sigma;
}
model {
  y ~ normal(theta, sigma);
}
```

Código A.1. **Stan**. Ejemplo de caja de código en Stan.

```
y <- rnorm(n = 100, mean = 1, sigma = 0)
```

Código A.2. **R**. Ejemplo de caja de código en R.

Para las variables y líneas del código se ha utilizado el inglés, por lo que se han usado, en la medida de lo posible, palabras con el mayor parecido con términos en español (como *observations*, *patients* o *events*).

Notación matemática

Respecto a la notación matemática de los modelos, como norma general, se han utilizado letras griegas para los parámetros (estimaciones) y letras latinas para las observaciones (datos).

Los vectores y matrices, para diferenciarlos de los escalares (números), se representan en **letra negrita**. Las matrices, además, irán en mayúscula.

En el margen derecho de las fórmulas matemáticas se pondrán las numeraciones o las anotaciones, con color gris y delimitadas por paréntesis.

$$y = 1 + 2x \qquad \text{(esto es una anotación)}$$

Conjuntos de números

Tal y como recomienda la Real Academia Española en su *Ortografía de la lengua española* (2010, pág. 666) utilizaremos el punto (en lugar de la coma) como separador decimal.

Para indicar que un número es un entero positivo (esto es, sin decimales) diremos que «pertenece al conjunto de los números naturales», que escribiremos como $a \in \mathbb{N}$. Todos los conjuntos de números utilizados a lo largo de este documento se especifican en la tabla A.1.

Tabla A.1. *Notación y descripción de los conjuntos de números utilizados.*

$\mathbb{N}$	$\{0, 1, 2, 3, \ldots\}$	Números naturales (incluyendo el cero)
$\mathbb{Z}$	$\{\ldots, -2, -1, 0, 1, 2 \ldots\}$	Números enteros
$\mathbb{Z}_2$	$\{0, 1\}$	Números 0 y 1 (utilizados en matrices lógicas)
$\mathbb{R}$	e. g. $2.03, -1.99, 34.5, \ldots$	Números reales (los utilizados habitualmente, positivos o negativos, con o sin decimales)

Cuando necesitemos especificar el rango de un número utilizaremos subíndices. Por ejemplo, con $\mathbb{R}_{x>0}$ denotaremos al conjunto de los números reales positivos y con $\mathbb{R}_{a<x<b}$ representaremos los números reales comprendidos entre a y b.

Vectores y matrices

Utilizaremos superíndices sobre el conjunto de números (ver tabla A.1) para especificar las dimensiones de vectores y matrices. Así, un vector $\mathbf{b}$ de n números reales se representará como $\mathbf{b} \in \mathbb{R}^n$ y una matriz $\mathbf{B}$ de n filas y m columnas ($n \times m$) como $\mathbf{B} \in \mathbb{R}^{n \times m}$.

Para especificar elementos concretos tanto de vectores como de matrices se usarán subíndices. Así, los elementos de la matriz $\mathbf{B} \in \mathbb{R}^{n \times m}$ se escribirán como $B_{i,j} \in \mathbb{R}$ con $i = 1, \ldots, n$ y $j = 1, \ldots, m$, esto es, el elemento de la matriz situado en la fila 3 y la columna 5 se representará como $B_{3,5}$. En general, escribiremos

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = (b_i) \in \mathbb{R}^n, \quad \mathbf{B} = \begin{pmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,m} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n,1} & B_{n,2} & \cdots & B_{n,m} \end{pmatrix} = (B_{i,j}) \in \mathbb{R}^{n \times m}.$$

Cuando queramos mencionar todos los elementos de una fila o columna utilizaremos el asterisco. Así, con $B_{p,*}$ denotaremos todos los elementos de la fila p de la matriz $\mathbf{B}$.

La matriz traspuesta de $\mathbf{B}$ —esto es, la reordenación de los elementos de filas y columnas de una matriz a columnas y filas— se representará como $\mathbf{B}^\top = (B_{j,i})$.

Una matriz especial que utilizaremos en este documento es aquella formada completamente por unos (conocida como matriz de unos o matriz de todo unos). La denotaremos como $\mathbf{J_{n,m}}$ donde n y m será el número de filas y columnas que la conforman.

$$\mathbf{J_{n,m}} = \begin{pmatrix} J_{1,1} & J_{1,2} & \cdots & J_{1,m} \\ J_{2,1} & J_{2,2} & \cdots & J_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ J_{n,1} & J_{n,2} & \cdots & J_{n,m} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

Como caso especial, dado que pertenece más a la informática que a las matemáticas, se representarán las listas (o *arrays*) de elementos entre llaves. Así, por ejemplo, representaremos los elementos de una lista formada por p matrices $\mathbf{B}_k \in \mathbb{R}^{n\times m}$ como $\{B_{i,j}\}_k$, para cada $k = 1, \ldots, p$.

Álgebra lineal

El producto escalar para multiplicar los elementos de dos vectores de las mismas dimensiones $\mathbf{a} \in \mathbb{R}^n$ y $\mathbf{b} \in \mathbb{R}^n$ se denotará con el operador punto «·»,

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

La multiplicación estándar de dos matrices $\mathbf{A} \in \mathbb{R}^{n\times m}$ y $\mathbf{B} \in \mathbb{R}^{m\times p}$ —donde, recordemos, cada elemento viene determinado por una fila de $\mathbf{A}$ y una columna de $\mathbf{B}$— se denotará con como $\mathbf{AB} \in \mathbb{R}^{n\times p}$, manteniendo la misma notación cuando se trate del producto de un escalar por un vector, un escalar por una matriz, un vector por una matriz, etc.

Integrales y derivadas

La notación para las integrales y derivadas será también la estándar, usando la comilla simple (o apóstrofo) para denotar a esta última,

$$F(x) = \int_a^x f(u)\, du, \qquad F'(x) = f(x).$$

Operaciones y funciones generales

Para representar sumas (o multiplicaciones) de los elementos a_i de un vector $\mathbf{a} \in \mathbb{R}^n$ cualquiera utilizaremos la notación estándar con sumatorios $\sum$ (o productorios $\prod$),

$$\begin{aligned}\sum_{i=1}^{n} a_i &= a_1 + a_2 + \cdots + a_n,\\ \prod_{i=1}^{n} a_i &= a_1 \cdot a_2 \cdot \ldots \cdot a_n.\end{aligned}$$

Para representar el límite de una función $f(x)$ cuando una variable ε se aproxima (o «tiende») a un valor c utilizaremos también la notación habitual,

$$\lim_{\varepsilon \to c} f(x).$$

La función exponencial, para facilitar la lectura (con la notación de superíndices el tamaño se reduce demasiado), se representará como $\exp$, esto es, $\exp(x) = e^x$.

Cuando una función f se aplique a una matriz $\mathbf{B}$ denotará que se aplica a cada elemento de la matriz. Por ejemplo,

$$f(\mathbf{B}) = \begin{pmatrix} f(B_{1,1}) & f(B_{1,2}) & \cdots & f(B_{1,m}) \\ f(B_{2,1}) & f(B_{2,2}) & \cdots & f(B_{2,m}) \\ \vdots & \vdots & \ddots & \vdots \\ f(B_{n,1}) & f(B_{n,2}) & \cdots & f(B_{n,m}) \end{pmatrix}.$$

Estadística y probabilidad

Para la notación estadística y de probabilidad se usará como referencia el libro de Gelman, J. B. Carlin *et al.* 2013, que es la usada también en las guías de usuario y manuales de referencia del lenguaje Stan que se utilizará en este documento (Stan Development Team 2021).

Así, se utilizará la notación $p(\cdot)$ para denotar la función de distribución de probabilidad —que usaremos como sinónimo de función de densidad de probabilidad (o función de masa de probabilidad, para el caso discreto)— de una variable aleatoria. Para evitar posibles confusiones, para denotar la probabilidad de un evento concreto se utilizará $P(\cdot)$. Por ejemplo, $P(y > 0)$ denotará la probabilidad de que la variable y sea positiva.

La probabilidad conjunta de dos variables x e y se escribirá como $p(x, y)$ mientras que para la probabilidad condicional (probabilidad de una variable conocida la otra) se usará la forma habitual $p(y|x)$.

Para indicar que una variable sigue una cierta distribución de probabilidad se usará el operador «$\sim$». Por ejemplo, para decir que una variable y sigue una distribución normal con

parámetros μ y σ se escribirá

$$y \sim \text{Normal}(\mu, \sigma),$$

lo que sería equivalente a escribir

$$p(y) = p(y|\mu, \sigma) = \text{Normal}(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Para indicar que una distribución está truncada a un subconjunto de valores A se utilizará la notación matemática habitual para la restricción. Así, una distribución normal truncada a valores positivos se escribirá como

$$\text{Normal}(\mu, \sigma)\big|_{\mathbb{R}_{x>0}}.$$

Con el símbolo «$\propto$» se denotará «proporcional a». Así, si se tiene una distribución $p(y)$ que es proporcional a otra $p(x)$ —esto es, que $p(y) = k\,p(x)$ (siendo k una constante cualquiera)— entonces esta relación se podrá escribir como $p(y) \propto p(x)$.

En algunos casos muy puntuales, utilizaremos la función de distribución acumulada de una variable aleatoria Y, que representa la probabilidad que esta variable Y sea menor o igual a un valor c, que denotaremos como $F_Y(c)$,

$$F_Y(c) = P(Y \leq c).$$

Apéndice B

Código y *software*

El código principal donde se define el modelo que desarrollamos en este trabajo utilizando el lenguaje probabilístico Stan, así como el código con el que lo ejecutamos y estimamos con R y CmdStanR, se presenta en este documento (ver capítulos 5 y 6).

Todo el código que hemos utilizado para la depuración y el procesamiento de los datos, el desarrollo, la estimación, el diagnóstico y la evaluación del modelo, la simulación de datos y la generación de gráficos y tablas se encuentra disponible en el repositorio público y abierto de Codeberg: `https://codeberg.org/alvarohv/jmfevi-stan`.

El presente documento ha sido escrito y elaborado con el sistema para composición de textos LaTeX (Lamport 1994).

Bibliografía y referencias

Allignol, A. y A. Latouche (2022). *CRAN task view: survival analysis*. Version 2022-03-07. URL: https://CRAN.R-project.org/view=Survival.

Amrhein, V., S. Greenland y B. McShane (2019). «Scientists rise up against statistical significance». En: *Nature* 567.7748, págs. 305-307. DOI: 10.1038/d41586-019-00857-9.

Baker, M. (2016). «1,500 scientists lift the lid on reproducibility». En: *Nature* 533, págs. 452-454. DOI: 10.1038/533452a.

Bates, D., M. Mächler, B. Bolker y S. Walker (2015). «Fitting linear mixed-effects models using lme4». En: *Journal of Statistical Software* 67.1. DOI: 10.18637/jss.v067.i01.

Benjamin, E. J. *et al.* (2018). «Heart disease and stroke statistics—2018 update: a report from the American Heart Association». En: *Circulation* 137.12. DOI: 10.1161/cir.0000000000000558.

Betancourt, M. (2016). *Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo*. DOI: 10.48550/ARXIV.1604.00695.

— (2017). *A conceptual introduction to Hamiltonian Monte Carlo*. DOI: 10.48550/ARXIV.1701.02434.

— (2020). *Towards a principled bayesian workflow (RStan)*. URL: https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html.

Blitzstein, J. K. y J. Hwang (2019). *Introduction to probability*. 2.ª ed. Boca Raton: Chapman and Hall/CRC. ISBN: 9781138369917. DOI: 10.1201/9780429428357.

Bozkurt, B. *et al.* (2021). «Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure». En: *European Journal of Heart Failure* 23.3, págs. 352-380. DOI: 10.1002/ejhf.2115.

Bradburn, M. J., T. G. Clark, S. B. Love y D. G. Altman (2003). «Survival analysis part II: multivariate data analysis – an introduction to concepts and methods». En: *British Journal of Cancer* 89.3, págs. 431-436. DOI: 10.1038/sj.bjc.6601119.

Brilleman, S. L., M. J. Crowther *et al.* (2018). «Joint longitudinal and time-to-event models for multilevel hierarchical data». En: *Statistical Methods in Medical Research* 28.12, págs. 3502-3515. DOI: 10.1177/0962280218808821.

Brilleman, S. L., R. Wolfe, M. Moreno-Betancur y M. J. Crowther (2021). «Simulating survival data using the simsurv R package». En: *Journal of Statistical Software* 97.3. DOI: 10.18637/jss.v097.i03.

Bürkner, P.-C. (2017). «brms: an R package for bayesian multilevel models using Stan». En: *Journal of Statistical Software* 80.1, págs. 1-28. DOI: 10.18637/jss.v080.i01.

Bürkner, P.-C., J. Gabry, M. Kay y A. Vehtari (2022). *posterior: tools for working with posterior distributions*. R package version 1.2.2. URL: https://mc-stan.org/posterior/.

Camm, A. J., T. F. Lüscher, G. Maurer y P. W. Serruys, eds. (2018). *ESC CardioMed*. 3.ª ed. Oxford, New York: Oxford University Press. ISBN: 9780198784906. DOI: 10.1093/med/9780198784906.001.0001.

Carpenter, B. *et al.* (2017). «Stan: a probabilistic programming language». En: *Journal of Statistical Software* 76.1. DOI: 10.18637/jss.v076.i01.

Chang, W. (2022). *R6: encapsulated classes with reference semantics*. R package version 2.5.1. URL: https://r6.r-lib.org.

Chaudhry, S.-P. y G. C. Stewart (2016). «Advanced heart failure». En: *Heart Failure Clinics* 12.3, págs. 323-333. DOI: 10.1016/j.hfc.2016.03.001.

Chioncel, O. *et al.* (2017). «Epidemiology and one-year outcomes in patients with chronic heart failure and preserved, mid-range and reduced ejection fraction: an analysis of the ESC Heart Failure Long-Term Registry». En: *European Journal of Heart Failure* 19.12, págs. 1574-1585. DOI: 10.1002/ejhf.813.

Claggett, B., L.-J. Wei y M. A. Pfeffer (2013). «Moving beyond our comfort zone». En: *European Heart Journal* 34.12, págs. 869-871. DOI: 10.1093/eurheartj/ehs485.

Clark, T. G., M. J. Bradburn, S. B. Love y D. G. Altman (2003). «Survival analysis part I: basic concepts and first analyses». En: *British Journal of Cancer* 89.2, págs. 232-238. DOI: 10.1038/sj.bjc.6601118.

Cohen, J. (1994). «The earth is round (p<.05)». En: *American Psychologist* 49.12, págs. 997-1003. DOI: 10.1037/0003-066X.49.12.997.

Cohn, J. N., R. Ferrari y N. Sharpe (2000). «Cardiac remodeling—concepts and clinical implications: a consensus paper from an international forum on cardiac remodeling». En: *Journal of the American College of Cardiology* 35.3, págs. 569-582. DOI: 10.1016/s0735-1097(99)00630-0.

Collett, D. (2015). *Modelling survival data in medical research*. 3.ª ed. Boca Raton: Chapman and Hall/CRC, pág. 548. ISBN: 9780429196294. DOI: 10.1201/b18041.

Colucci, W. S. *et al.* (2007). «Metoprolol reverses left ventricular remodeling in patients with asymptomatic systolic dysfunction». En: *Circulation* 116.1, págs. 49-56. DOI: 10.1161/circulationaha.106.666016.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, New York: Cambridge University Press, pág. 236. ISBN: 9780511813559. DOI: 10.1017/cbo9780511813559.

Delgado, J. F. *et al.* (2014). «Costes sanitarios y no sanitarios de personas que padecen insuficiencia cardiaca crónica sintomática en España». En: *Revista Española de Cardiología* 67.8, págs. 643-650. DOI: 10.1016/j.recesp.2013.12.016.

DeVore, A. D. *et al.* (2020). «Improvement in left ventricular ejection fraction in outpatients with heart failure with reduced ejection fraction». En: *Circulation: Heart Failure* 13.7. DOI: 10.1161/circheartfailure.119.006833.

Donoho, D. (2017). «50 years of data science». En: *Journal of Computational and Graphical Statistics* 26.4, págs. 745-766. DOI: 10.1080/10618600.2017.1384734.

Fanaroff, A. C., R. M. Califf, S. Windecker, S. C. Smith y R. D. Lopes (2019). «Levels of evidence supporting American College of Cardiology/American Heart Association and European Society of Cardiology guidelines, 20082018». En: *JAMA* 321.11, pág. 1069. DOI: 10.1001/jama.2019.1122.

Feldman, D. S. y P. Mohacsi, eds. (2019). *Heart failure*. Cham, Switzerland: Springer International Publishing. DOI: 10.1007/978-3-319-98184-0.

Fernández-Gassó, L., L. Hernando-Arizaleta, J. A. Palomar-Rodríguez, M. V. Abellán-Pérez, Á. Hernández-Vicente y D. A. Pascual-Figal (2019a). «Estudio poblacional de la primera hospitalización por insuficiencia cardiaca y la interacción entre los reingresos y la supervivencia». En: *Revista Española de Cardiología* 72.9, págs. 740-748. DOI: 10.1016/j.recesp.2018.05.037.

Fonarow, G. C., C. W. Yancy, A. F. Hernandez, E. D. Peterson, J. A. Spertus y P. A. Heidenreich (2011). «Potential impact of optimal implementation of evidence-based heart failure therapies on mortality». En: *American Heart Journal* 161.6, págs. 1024-1030. DOI: 10.1016/j.ahj.2011.01.027.

Free Software Foundation (2019). *GNU bash*. Unix shell version 5.0.17. URL: http://www.gnu.org/software/bash.

Frieden, T. R. (2017). «Evidence for health decision making—beyond randomized, controlled trials». En: *New England Journal of Medicine* 377.5. Ed. por J. M. Drazen, D. P. Harrington, J. J. V. McMurray, J. H. Ware y J. Woodcock, págs. 465-475. DOI: 10.1056/nejmra1614394.

Gabry, J. y R. Češnovar (2022). *CmdStanR*. R package version 0.5.2. URL: https://mc-stan.org/cmdstanr/.

Gabry, J. y T. Mahr (2022). *bayesplot: plotting for bayesian models*. R package version 1.9.0. URL: https://mc-stan.org/bayesplot/.

Gabry, J., D. Simpson, A. Vehtari, M. Betancourt y A. Gelman (2019). «Visualization in bayesian workflow». En: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182.2, págs. 389-402. DOI: 10.1111/rssa.12378.

Gallin, J. I., F. P. Ognibene y L. L. Johnson, eds. (2017). *Principles and practice of clinical research*. 4.ª ed. London: Elsevier, pág. 864. ISBN: 978-0-12-849905-4. DOI: 10.1016/c2014-0-03999-8.

Gelman, A. y J. Carlin (2017). «Some natural solutions to the p-value communication problem—and why they won't work». En: *Journal of the American Statistical Association* 112.519, págs. 899-901. DOI: 10.1080/01621459.2017.1311263. eprint: https://doi.org/10.1080/01621459.2017.1311263. URL: https://doi.org/10.1080/01621459.2017.1311263.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari y D. B. Rubin (2013). *Bayesian data analysis*. 3.ª ed. New York: Chapman and Hall/CRC. ISBN: 9781439840955. DOI: 10.1201/b16018.

Gelman, A. y J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, New York: Cambridge University Press, pág. 648. ISBN: 9780521867061. DOI: 10.1017/cbo9780511790942.

Gelman, A., A. Vehtari *et al.* (2020). «Bayesian workflow». En: DOI: https://doi.org/10.48550/arXiv.2011.01808.

Goldstein, H. (2010). *Multilevel statistical models*. 4.ª ed. Chichester, West Sussex: John Wiley & Sons, Ltd. ISBN: 9780470748657. DOI: 10.1002/9780470973394.

Goodrich, B., J. Gabry, I. Ali y S. Brilleman (2022). *rstanarm: bayesian applied regression modeling via Stan.* R package version 2.21.3. URL: https://mc-stan.org/rstanarm/.

Greenberg, B., M. A. Quinones, C. Koilpillai, M. Limacher, D. Shindler, C. Benedict y B. Shelton (1995). «Effects of long-term enalapril therapy on cardiac structure and function in patients with left ventricular dysfunction». En: *Circulation* 91.10, págs. 2573-2581. DOI: 10.1161/01.cir.91.10.2573.

Grimes, D. A. y K. F. Schulz (2002). «An overview of clinical research: the lay of the land». En: *The Lancet* 359.9300, págs. 57-61. DOI: 10.1016/s0140-6736(02)07283-5.

Grolemund, G. y H. Wickham (2011). «Dates and times made easy with lubridate». En: *Journal of Statistical Software* 40.3, págs. 1-25. URL: https://www.jstatsoft.org/v40/i03/.

Hand, D. J. (2014). «Wonderful examples, but let's not close our eyes». En: *Statistical Science* 29.1. DOI: 10.1214/13-sts446.

Hauer, E. (2004). «The harm done by tests of significance». En: *Accident Analysis & Prevention* 36.3, págs. 495-500. DOI: https://doi.org/10.1016/S0001-4575(03)00036-8. URL: https://www.sciencedirect.com/science/article/pii/S0001457503000368.

Held, L. y D. Sabanés Bové (2014). *Applied statistical inference. likelihood and bayes*. Berlin: Springer Berlin Heidelberg. ISBN: 978-3-642-37886-7. DOI: 10.1007/978-3-642-37887-4.

Hickey, G. L., P. Philipson, A. Jorgensen y R. Kolamunnage-Dona (2018). «Joint models of longitudinal and time-to-event data with more than one event time outcome: a review». En: *The International Journal of Biostatistics* 14.1. DOI: 10.1515/ijb-2017-0047.

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist y J. P. Huelsenbeck (2014). «Probabilistic graphical model representation in phylogenetics». En: *Systematic Biology* 63.5, págs. 753-771. DOI: 10.1093/sysbio/syu039.

Hoijtink, H., I. Klugkist y P. A. Boelen, eds. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer New York. ISBN: 978-0-387-09612-4. DOI: 10.1007/978-0-387-09612-4.

Ibrahim, J. G., M.-H. Chen y D. Sinha (2001). *Bayesian survival analysis*. New York: Springer, pág. 481. ISBN: 9780387952772. DOI: 10.1007/978-1-4757-3447-8.

Ibrahim, J. G., H. Chu y L. M. Chen (2010). «Basic concepts and methods for joint models of longitudinal and survival data». En: *Journal of Clinical Oncology* 28.16, págs. 2796-2801. DOI: 10.1200/jco.2009.25.0654.

Instituto Nacional de Estadística (2022). *Defunciones según la mausa de muerte. Defunciones por causas (lista reducida) por sexo y grupos de edad*. URL: https://www.ine.es/jaxiT3/Tabla.htm?t=7947 (visitado 13-07-2022).

Ioannidis, J. P. A. (2005). «Why most published research findings are false». En: *PLOS Medicine* 2.8. DOI: 10.1371/journal.pmed.0020124.

Julián, M. T. *et al.* (2020). «Long-term LVEF trajectories in patients with type 2 diabetes and heart failure: diabetic cardiomyopathy may underlie functional decline». En: *Cardiovascular Diabetology* 19.1. DOI: 10.1186/s12933-020-01011-w.

Jurgens, C. Y. *et al.* (2022). «State of the science: the relevance of symptoms in cardiovascular disease and research: a scientific statement from the American Heart Association». En: *Circulation* 146.0, págs. 00-00. DOI: 10.1161/cir.0000000000001089.

Kleinbaum, D. G. (2012). *Survival analysis. a self-learning text*. New York: Springer, pág. 699. ISBN: 9781441966452. DOI: 10.1007/978-1-4419-6646-9.

Konstam, M. A., D. G. Kramer, A. R. Patel, M. S. Maron y J. E. Udelson (2011). «Left ventricular remodeling in heart failure». En: *JACC: Cardiovascular Imaging* 4.1, págs. 98-108. DOI: 10.1016/j.jcmg.2010.10.008.

Lambert, B. (2018). *Student's guide to bayesian statistics*. Los Ángeles: SAGE Publications Ltd, pág. 520. ISBN: 9781473916364. URL: https://uk.sagepub.com/en-gb/eur/book/student%E2%80%99s-guide-bayesian-statistics.

Lamport, L. (1994). *LaTeX. a document preparation system*. 2.ª ed. Addison-Wesley Series on Tools and Techniques for Computer Typesetting. Boston: Addison-Wesley Professional, pág. 288. ISBN: 9780201529838.

Law, C. W., K. Zeglinski, X. Dong, M. Alhamdoosh, G. K. Smyth y M. E. Ritchie (2020). «A guide to creating design matrices for gene expression experiments». En: *F1000Research* 9, pág. 1444. DOI: 10.12688/f1000research.27893.1.

Leek, J., B. B. McShane, A. Gelman, D. Colquhoun, M. B. Nuijten y S. N. Goodman (2017). «Five ways to fix statistics». En: *Nature* 551.7682, págs. 557-559. DOI: 10.1038/d41586-017-07522-z.

Lesyuk, W., C. Kriza y P. Kolominsky-Rabas (2018). «Cost-of-illness studies in heart failure: a systematic review 2004–2016». En: *BMC Cardiovascular Disorders* 18.1. DOI: 10.1186/s12872-018-0815-3.

Leung, K.-M., R. M. Elashoff y A. A. Afifi (1997). «Censoring issues in survival analysis». En: *Annual Review of Public Health* 18.1, págs. 83-104. DOI: 10.1146/annurev.publhealth.18.1.83.

Lindgren, F. y H. Rue (2015). «Bayesian spatial modelling with R-INLA». En: *Journal of Statistical Software* 63.19. DOI: 10.18637/jss.v063.i19.

Lu, T. (2017). «Mixed-effects location and scale Tobit joint models for heterogeneous longitudinal data with skewness, detection limits, and measurement errors». En: *Statistical Methods in Medical Research* 27.12, págs. 3525-3543. DOI: 10.1177/0962280217704225.

Lunn, D., C. Jackson, N. Best, A. Thomas y D. Spiegelhalter (2012). *The BUGS book. a practical introduction to bayesian analysis*. New York: Chapman and Hall/CRC, pág. 399. ISBN: 9781466586666. DOI: 10.1201/b13613.

Lupón, J. *et al.* (2018). «Dynamic trajectories of left ventricular ejection fraction in heart failure». En: *Journal of the American College of Cardiology* 72.6, págs. 591-601. DOI: 10.1016/j.jacc.2018.05.042.

Maller, R. A., G. Müller y A. Szimayer (2009). «Ornstein–Uhlenbeck processes and extensions». En: *Handbook of Financial Time Series*. Ed. por T. Mikosch, J. P. Kreiß, R. A. Davis y T. G. Andersen. Springer Berlin Heidelberg, págs. 421-437. DOI: 10.1007/978-3-540-71297-8_18.

Matthews, R. (2021). «The p-value statement, five years on». En: *Significance* 18.2, págs. 16-19. DOI: 10.1111/1740-9713.01505.

McDonagh, T. A. (2011). *Oxford textbook of heart failure*. Oxford, New York: Oxford University Press, pág. 641. ISBN: 9780199577729.

McDonagh, T. A. *et al.* (2022). «Guía ESC 2021 sobre el diagnóstico y tratamiento de la insuficiencia cardiaca aguda y crónica». En: *Revista Española de Cardiología*. DOI: 10.1016/j.recesp.2021.11.027.

McElreath, R. (2020). *Statistical rethinking. a bayesian course with examples in R and Stan*. 2.ª ed. Boca Raton: Chapman and Hall/CRC, pág. 598. ISBN: 9780429029608. DOI: 10.1201/9780429029608.

McMurray, J. J. V., M. Packer *et al.* (2014). «Angiotensin–neprilysin inhibition versus enalapril in heart failure». En: *New England Journal of Medicine* 371.11, págs. 993-1004. DOI: 10.1056/nejmoa1409077.

McMurray, J. J. V., S. D. Solomon *et al.* (2019). «Dapagliflozin in patients with heart failure and reduced ejection fraction». En: *New England Journal of Medicine* 381.21, págs. 1995-2008. DOI: 10.1056/nejmoa1911303.

Molenberghs, G. y G. Verbeke (2000). *Linear mixed models for longitudinal data*. New York: Springer New York, pág. 608. ISBN: 9780387950273. DOI: 10.1007/978-1-4419-0300-6.

Murphy, K. P. (2012). *Machine learning. a probabilistic perspective*. Adaptive Computation and Machine Learning series. Cambridge, Massachussets: MIT Press. ISBN: 9780262018029. URL: https://mitpress.mit.edu/books/machine-learning-1.

Noorden, R. V., B. Maher y R. Nuzzo (2014). «The top 100 papers». En: *Nature* 514.7524, págs. 550-553. DOI: 10.1038/514550a.

Nuzzo, R. (2014). «Scientific method: statistical errors». En: *Nature* 506.7487, págs. 150-152. DOI: 10.1038/506150a.

O'Connor, C. M. (2019). «A call for change: level of statistical significance». En: *JACC: Heart Failure* 7.11, págs. 993-994. DOI: 10.1016/j.jchf.2019.09.005.

Oliva, J., N. Jorgensen y J. M. Rodríguez (2010). «Carga socioeconómica de la insuficiencia cardíaca: revisión de los estudios de coste de la enfermedad». En: *PharmacoEconomics Spanish Research Articles* 7.2, págs. 68-79. DOI: 10.1007/bf03321475.

Oravecz, Z., F. Tuerlinckx y J. Vandekerckhove (2016). «Bayesian data analysis with the bivariate hierarchical Ornstein-Uhlenbeck process model». En: *Multivariate Behavioral Research* 51.1, págs. 106-119. DOI: 10.1080/00273171.2015.1110512.

Packer, M. *et al.* (2020). «Cardiovascular and renal outcomes with Empagliflozin in heart failure». En: *New England Journal of Medicine* 383.15, págs. 1413-1424. DOI: 10.1056/nejmoa2022190.

Papageorgiou, G., K. Mauff, A. Tomer y D. Rizopoulos (2019). «An overview of joint modeling of time-to-event and longitudinal outcomes». En: *Annual Review of Statistics and Its Application* 6.1, págs. 223-240. DOI: 10.1146/annurev-statistics-030718-105048.

Peng, R. y E. Matsui (2016). *The art of data science. a guide for anyone who works with data*. Victoria, British Colombia: Lulu PR. 170 págs. ISBN: 1365061469. URL: https://leanpub.com/artofdatascience.

Philipson, P., I. Sousa, P. J. Diggle, P. Williamson, R. Kolamunnage-Dona, R. Henderson y G. L. Hickey (2018). *joineR: joint modelling of repeated measurements and time-to-event data*. R package version 1.2.6. URL: https://github.com/graemeleehickey/joineR/.

Pinheiro, J. C. y D. M. Bates (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer. DOI: 10.1007/b98882.

Pintilie, M. (2011). «Análisis de riesgos competitivos». En: *Revista Española de Cardiología* 64.7, págs. 599-605. DOI: 10.1016/j.recesp.2011.03.017.

Plummer, M. (2003). «JAGS: a program for analysis of bayesian graphical models using Gibbs sampling». En: *Proceedings of the 3rd international workshop on distributed statistical computing*. Vol. 124. 125.10. Vienna, Austria, págs. 1-10.

— (2022). *rjags: bayesian graphical models using MCMC*. R package version 4-13. URL: https://CRAN.R-project.org/package=rjags.

R Core Team (2022). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org.

Real Academia Española (2010). *Ortografía de la lengua española*. Madrid: Espasa Calpe. ISBN: 9788467034264. URL: https://aplica.rae.es/orweb/cgi-bin/v.cgi?i=UMthSrkshUdyWiO,.

Remme, W. J. *et al.* (2004). «The benefits of early combination treatment of carvedilol and an ACE-inhibitor in mild heart failure and left ventricular systolic dysfunction. The carvedilol and ACE-inhibitor remodelling mild heart failure evaluation trial (CARMEN)». En: *Cardiovascular Drugs and Therapy* 18.1, págs. 57-66. DOI: 10.1023/b:card.0000025756.32499.6f.

Riet, E. E. S. van, A. W. Hoes, A. Limburg, M. A. J. Landman, H. van der Hoeven y F. H. Rutten (2014). «Prevalence of unrecognized heart failure in older persons with shortness of breath on exertion». En: *European Journal of Heart Failure* 16.7, págs. 772-777. DOI: 10.1002/ejhf.110.

Rizopoulos, D. (2010). «JM: an R package for the joint modelling of longitudinal and time-to-event data». En: *Journal of Statistical Software* 35.9. DOI: 10.18637/jss.v035.i09.

— (2012). *Joint models for longitudinal and time-to-event data. with applications in R*. Chapman and Hall / CRC biostatistics series. Boca Raton: Chapman and Hall/CRC. ISBN: 9781439872864. DOI: 10.1201/b12208.

— (2016). «The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC». En: *Journal of Statistical Software* 72.7. DOI: 10.18637/jss.v072.i07.

Sayago-Silva, I., F. García-López y J. Segovia-Cubero (2013). «Epidemiología de la insuficiencia cardiaca en España en los últimos 20 años». En: *Revista Española de Cardiología* 66.8, págs. 649-656. DOI: 10.1016/j.recesp.2013.03.014.

Schoot, R. van de *et al.* (2021). «Bayesian statistics and modelling». En: *Nature Reviews Methods Primers* 1.1, págs. 1-26. DOI: 10.1038/s43586-020-00001-2.

Schutt, R. y C. O'Neil (2013). *Doing data science*. Sebastopol: O'Reilly. ISBN: 9781449358655. URL: http://www.oreilly.com/library/view/doing-data-science/9781449363871.

Sicras-Mainar, A., A. Sicras-Navarro, B. Palacios, L. Varela y J. F. Delgado (2022). «Epidemiología y tratamiento de la insuficiencia cardiaca en españa: estudio PATHWAYS-HF». En: *Revista Española de Cardiología* 75.1, págs. 31-38. DOI: 10.1016/j.recesp.2020.09.014.

Spiegelhalter, D., A. Thomas, N. Best y D. Lunn (2014). *OpenBUGS user manual*. Version 3.2.3. URL: https://www.mrc-bsu.cam.ac.uk/software/bugs/openbugs/.

Stan Development Team (2021). *Stan modeling language users guide and reference manual, v2.29*. URL: https://mc-stan.org.

— (2022). *RStan: the R interface to Stan*. R package version 2.21.5. URL: https://mc-stan.org/rstan/.

Statisticat y LLC. (2021). *LaplacesDemon: complete environment for bayesian inference*. R package version 16.1.6. URL: https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software.

Suits, D. B. (1957). «Use of dummy variables in regression equations». En: *Journal of the American Statistical Association* 52.280, págs. 548-551. DOI: 10.1080/01621459.1957.10501412.

Therneau, T. M. (2022). *A package for survival analysis in R*. R package version 3.3-1. URL: https://CRAN.R-project.org/package=survival.

Therneau, T. M. y P. M. Grambsch (2000). *Modeling survival data: extending the Cox model*. New York: Springer New York. ISBN: 0-387-98784-3. DOI: 10.1007/978-1-4757-3294-8.

Torp-Pedersen, C. *et al.* (2019). «'Real-world' observational studies in arrhythmia research: data sources, methodology, and interpretation. A position document from European Heart Rhythm Association (EHRA), endorsed by Heart Rhythm Society (HRS), Asia-Pacific HRS (APHRS), and Latin America HRS (LAHRS)». En: *EP Europace* 22.5, págs. 831-832. DOI: 10.1093/europace/euz210.

Udelson, J. E., A. M. Feldman, B. Greenberg, B. Pitt, R. Mukherjee, H. A. Solomon y M. A. Konstam (2010). «Randomized, double-blind, multicenter, placebo-controlled study evaluating the effect of aldosterone antagonism with eplerenone on ventricular remodeling in patients with mild-to-moderate heart failure and left ventricular systolic dysfunction». En: *Circulation: Heart Failure* 3.3, págs. 347-353. DOI: 10.1161/circheartfailure.109.906909.

Uhlenbeck, G. E. y L. S. Ornstein (1930). «On the theory of the brownian motion». En: *Physical Review* 36.5, págs. 823-841. DOI: 10.1103/physrev.36.823.

Vallverdú, J. (2016). *Bayesians versus frequentists. a philosophical debate on statistical reasoning*. SpringerBriefs in Statistics. Heidelberg: Springer Berlin, Heidelberg. ISBN: 9783662486368. DOI: 10.1007/978-3-662-48638-2.

Vatiwutipong, P. y N. Phewchean (2019). «Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process». En: *Advances in Difference Equations* 2019.1. DOI: 10.1186/s13662-019-2214-1.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter y P.-C. Bürkner (2021). «Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC (with discussion)». En: *Bayesian Analysis* 16.2. DOI: 10.1214/20-ba1221.

Velazquez, E. J. *et al.* (2019). «Angiotensin–neprilysin inhibition in acute decompensated heart failure». En: *New England Journal of Medicine* 380.6, págs. 539-548. DOI: 10.1056/nejmoa1812851.

Vizzardi, E. *et al.* (2010). «Effect of spironolactone on left ventricular ejection fraction and volumes in patients with class I or II heart failure». En: *The American Journal of Cardiology* 106.9, págs. 1292-1296. DOI: 10.1016/j.amjcard.2010.06.052.

Wagenmakers, E. J., M. Lee, T. Lodewyckx y G. J. Iverson (2008). «Bayesian versus frequentist inference». En: *Bayesian Evaluation of Informative Hypotheses*. Ed. por H. Hoijtink, I. Klugkist y P. A. Boelen. Springer New York, págs. 181-207. DOI: 10.1007/978-0-387-09612-4_9.

Wandt, B., L. Bojo, K. Tolagen y B. Wranne (1999). «Echocardiographic assessment of ejection fraction in left ventricular hypertrophy». En: *Heart* 82.2, págs. 192-198. DOI: 10.1136/hrt.82.2.192.

Wang, K., E. Youngson, J. A. Bakal, J. Thomas, F. A. McAlister y G. Y. Oudit (2021). «Cardiac reverse remodelling and health status in patients with chronic heart failure». En: *ESC Heart Failure* 8.4, págs. 3106-3118. DOI: 10.1002/ehf2.13417.

Wang, Y. y J. M. G. Taylor (2001). «Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome». En: *Journal of the American Statistical Association* 96.455, págs. 895-905. DOI: 10.1198/016214501753208591.

Wasserstein, R. L., A. L. Schirm y N. A. Lazar (2019). «Moving to a world beyond "p<0.05"». En: *The American Statistician* 73.sup1, págs. 1-19. DOI: 10.1080/00031305.2019.1583913.

Wickham, H. (2014). «Tidy data». En: *Journal of Statistical Software* 59.10. DOI: 10.18637/jss.v059.i10.

— (2016). *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, H., M. Averick *et al.* (2019). «Welcome to the tidyverse». En: *Journal of Open Source Software* 4.43, pág. 1686. DOI: 10.21105/joss.01686.

Wickham, H. y J. Bryan (2022). *readxl: read excel files*. R package version 1.4.0. URL: https://CRAN.R-project.org/package=readxl.

Wickham, H., R. François, L. Henry y K. Müller (2022). *dplyr: a grammar of data manipulation*. R package version 1.0.9. URL: https://CRAN.R-project.org/package=dplyr.

Wickham, H. y M. Girlich (2022). *tidyr: tidy messy data*. R package version 1.2.0. URL: https://CRAN.R-project.org/package=tidyr.

Williamson, P., R. Kolamunnage-Dona, P. Philipson y A. G. Marson (2008). «Joint modelling of longitudinal and competing risks data». En: *Statistics in Medicine* 27.30, págs. 6426-6438. DOI: 10.1002/sim.3451.

Yoshida, K. y A. Bartel (2022). *tableone: create 'table 1' to describe baseline characteristics with or without propensity score weights*. R package version 0.13.2. URL: https://CRAN.R-project.org/package=tableone.

Zyl, C. J. J. van (2018). «Frequentist and bayesian inference: a conceptual primer». En: *New Ideas in Psychology* 51, págs. 44-49. DOI: 10.1016/j.newideapsych.2018.06.004.

Índice de figuras

Índice de tablas

Índice de códigos