Modelos multiestado bayesianos en análisis de supervivencia con efectos espaciales y curación

Fran Llopis Cardona

Directores:
Carmen Armero Cervera
Gabriel Sanfélix Gimeno

Programa de Doctorado en Estadística y Optimización
Universitat de València
Octubre 2024

Esta tesis doctoral ha sido financiada por el Instituto de Salud Carlos III y cofinanciada por el Fondo Social Europeo, FI19/00190.

Agradecimientos

Esta tesis no hubiese sido posible sin el apoyo de todas las personas con las que he tenido la inmensa suerte de poder contar a lo largo de estos años.

En primer lugar, como no podría ser de otra forma me gustaría dar las gracias a mis padres. Por vuestro amor incondicional y apoyo incansable, por enseñarme a ser feliz y estar siempre dispuestos a decirme lo orgullosos que estáis de mí. Sois desde luego mis mayores referentes de lo que quiero ser como persona. Ni en mil textos como este podría explicar lo importantes que sois para mí ni agradecer todo lo que hacéis por mí.

Gracias Paola, mi persona especial, la compañera con la que he compartido tantos momentos buenos, con la que he crecido y a la que he visto crecer durante estos 6 años, juntos, como espero que podamos seguir, construyendo nuestro futuro, juntos. Gracias por apoyarme y estar siempre ahí en los momentos más delicados, con tu luz y alegría, haciendo tan fácil nuestra vida.

También quiero dar las gracias a todos aquellos amigos que de una forma u otra habéis estado conmigo y me habéis ayudado a lo largo de este proceso. Poder contar con vosotros siempre que me ha hecho falta, compartir mis mejores y mis peores sentimientos, así como muchos momentos divertidos, os lo agradezco de corazón.

Finalmente, no puedo acabar este apartado de agradecimientos sin mencionar a mis dos directores, Carmen y Gabriel, dos personas esenciales, sin las que este trabajo no hubiese sido posible. Me habéis enseñado, acompañado y ayudado muchísmo durante estos años, no solo desde una perspectiva profesional sino también personal, humana. Gracias Carmen por haber sido mi tutora durante todo este tiempo, haciéndome disfrutar verdaderamente de la estadística y transmitiéndome unos valores que me acompañarán siempre, como estadístico y como persona. Gracias Gabriel por motivarme a descubrir una mejor versión de mí mismo, proponiendo retos y cuestiones interesantes, que inevitablemente despiertan mi curiosidad y me hacen querer aprender.

A todos vosotros, gracias.

Resumen

Esta tesis doctoral ha tenido como objetivo principal desarrollar metodologías estadísticas dentro del marco de los modelos multiestado que permitan profundizar en el análisis de datos de supervivencia multivariantes desde una perspectiva propia de la inferencia bayesiana. Los modelos multiestado son modelos pertenecientes al ámbito de los procesos estocásticos que se centran en analizar las trayectorias definidas por la ocurrencia de múltiples eventos de interés a lo largo del tiempo. Estos modelos se construyen en base a una serie de estados entre los que los individuos se desplazan, a través de unas determinadas transiciones. Así pues, el objeto de estudio de estos modelos es el tiempo que tardan los individuos en avanzar entre los distintos estados, definiéndose un tiempo para cada transición posible. Los modelos multiestado suponen una generalización de un gran número de escenarios relativos al análisis de supervivencia, desde el caso univariante con un único tiempo de supervivencia, hasta modelos con múltiples tiempos y eventos como los modelos de riesgos competitivos.

El análisis de supervivencia resulta especialmente útil en epidemiología, ya que ofrece métodos y herramientas para el estudio de tiempos de supervivencia habitualmente relacionados con la muerte, el desarrollo y progresión de una enfermedad, o la curación de la misma, entre otros. Esta investigación se ha desarrollado precisamente motivada por un estudio con datos del mundo real sobre fractura de cadera recurrente, que a lo largo de esta memoria, ilustrará y acompañará las distintas propuestas metodológicas que se presentan. El Capítulo 2 está especialmente dedicado a introducir un tipo de modelo multiestado, el modelo de enfermedad-muerte, así como a presentar los datos del mundo real que serán objeto de análisis. En este estudio se dispone de la cohorte PREV2FO, una cohorte poblacional formada por pacientes mayores de 65 años que fueron dados de alta tras ser hospitalizados a causa de una fractura de cadera, entre el 1 de enero de 2008 y el 31 de diciembre de 2015, en la Comunitat Valenciana. Tras esta fractura inicial se sigue a los pacientes en el tiempo hasta la muerte o fin de estudio (31 de diciembre de 2016), registrando las fracturas de cadera recurrentes que tienen lugar durante el seguimiento. Así pues, los modelos multiestado se presentan como una opción natural para nuestro estudio, en el que se considerarán tres estados, fractura inicial, refractura y muerte, y tres transiciones: de fractura a refractura, de fractura a muerte y de refractura a muerte. Estos modelos se abordan en todo momento desde una perspectiva bayesiana, presentando en términos probabilísticos resultados relativos a la supervivencia y la progresión de los pacientes de la cohorte, como lo son los *hazard ratios*, las funciones de incidencia

acumulada y las probabilidades de transición.

Además del interés que pueda tener la aplicación de los modelos mencionados al estudio real, tanto desde un punto de vista estadístico como epidemiológico, este plantea cuestiones que han motivado desarrollos y avances metodológicos. En el Capítulo 3 se propone un modelo de enfermedad-muerte bayesiano que combina esta modelización multiestado con información de datos espaciales, permitiendo analizar diferencias geográficas en las distintas transiciones que componen el modelo multiestado. Para ello se define un vector de efectos aleatorios que sigue una versión multivariante del modelo de Leroux, el cual considera, por un lado, una posible correlación entre transiciones, y por otro, la correlación espacial entre regiones. La aplicación de este modelo se ilustra mediante los datos de la cohorte PREV2FO, siendo posible identificar diferencias entre regiones por lo que respecta al riesgo de refractura, muerte sin refractura y muerte tras refractura, así como diferencias en las incidencias acumuladas y probabilidades de transición. La inferencia bayesiana se realiza mediante la aproximación anidada integrada de Laplace (INLA).

Otras cuestiones relativas al análisis de supervivencia, como son la curación y la cero-inflación, también son susceptibles de generalización mediante modelos multiestado. Así pues, en el Capítulo 4 se propone un marco metodológico conjunto en el que se presenta un modelo de enfermedad-muerte general que incluye cero-inflación y curación. Un caso particular de cero-inflación, el asociado a la muerte al inicio del seguimiento, se presenta con mayor detalle y se ilustra su aplicación al estudio de la muerte intrahospitalaria tras fractura de cadera. Por otro lado, la curación, entendida como la imposibilidad de experimentar la enfermedad de interés, independientemente del tiempo de seguimiento, es tratada como una variable latente, dada su naturaleza semiobservable. La presencia de valores faltantes en esta variable es abordada de forma natural por la inferencia bayesiana, sin embargo, INLA no permite su tratamiento directamente. Es por esta razón que se propone un algoritmo basado en muestreo de Gibbs y las estimaciones obtenidas con INLA, resolviendo esta dificultad. También se analizan la precisión de las estimaciones y la convergencia del método.

Finalmente, cabe mencionar que las metodologías planteadas en esta memoria ofrecen un marco estadístico con el que dar respuesta a preguntas científicas concretas del mundo real, no viéndose limitadas al estudio de fractura de cadera con el que se han ilustrado. Además, se muestra cómo los modelos multiestado, desde una perspectiva bayesiana, pueden adaptarse para incluir nuevos elementos y ampliar el conocimiento que se tiene de cuestiones prácticas, al mismo tiempo que se plantean desafíos metodológicos interesantes desde una perspectiva estadística.

Índice

Índice de figuras

Índice de Tablas

Capítulo 1

Conceptos básicos: inferencia bayesiana, supervivencia y estadística espacial

1.1. Introducción

En este primer capítulo se introducirán algunos de los conceptos fundamentales que aparecerán a lo largo de toda esta memoria y que definen el marco sobre el que se ha construido toda la tesis. Empezaremos introduciendo los elementos básicos que definen la inferencia bayesiana, perspectiva desde la cual se ha realizado todo el trabajo de esta memoria. Posteriormente, se presentarán los métodos y técnicas propias del análisis de supervivencia en general, para después proceder al estudio, específicamente, de los modelos multiestado. Finalmente, se introducirán la curación y los modelos espaciales, elementos adicionales que enriquecerán los modelos multiestado y cuya inclusión en estos supone la principal aportación metodológica de este trabajo.

1.2. Inferencia bayesiana

La inferencia bayesiana es una metodología estadística basada en una concepción de la probabilidad que permite expresar en términos probabilísticos la incertidumbre asociada a cualquier elemento de un modelo estadístico, pudiendo considerarse una definición de la probabilidad más amplia en comparación con la perspectiva frecuentista. En el paradigma bayesiano no solo se dotaría de probabilidad a las variables aleatorias sino que cualquier elemento tradicionalmente considerado invariable, como los parámetros desconocidos de un modelo, sería susceptible de presentar variabilidad expresada en términos de una distribución de probabilidad.

Como su nombre indica, la estadística bayesiana tiene como eje principal el teorema de Bayes. En su forma más simple este teorema define una relación entre las probabilidades condicionadas de dos sucesos. En particular, dados dos sucesos A y B dentro de

un espacio de probabilidad Ω, se cumple que

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{1.1}$$

Esta relación relativamente sencilla es sobre la que se ha construido todo el paradigma y marco metodológico bayesiano.

Traslademos este teorema a un proceso inferencial. Empecemos suponiendo que se dispone de una variable de interés Y continua, con una función de densidad (de probabilidad en el caso discreto) que depende de un conjunto de parámetros desconocidos $\boldsymbol{\theta}$, de forma que $Y \mid \boldsymbol{\theta} \sim f(y \mid \boldsymbol{\theta})$. Cabe destacar que en caso de conocerse estos parámetros se tendría perfectamente cuantificada la variabilidad asociada a la variable de interés. Sin embargo, y dado que no es el caso, estos parámetros desconocidos son precisamente el objeto de la inferencia, es decir, los valores que se querrían estimar o sobre los que se querría obtener información. Para ello, se obtiene una muestra de n individuos en los que se observa o mide la variable Y, obteniendo así observaciones y_i para cada individuo i, recogidas en conjunto en el vector $\boldsymbol{y} = (y_1, \cdots, y_n)$. La conexión entre el modelo definido por la función de densidad y los datos recogidos en la muestra viene dada por la función de verosimilitud, $L(\boldsymbol{\theta})$, que no es más que la función de densidad conjunta asociada a las observaciones, es decir,

$$L(\boldsymbol{\theta}) = f(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}), \tag{1.2}$$

verificándose la última igualdad siempre y cuando se asuma independecia condicional entre las distintas observaciones.

Esta verosimilitud es la que contiene la información de la muestra, de los datos observados, que aportarán la información necesaria para conocer Y, a través del conocimiento que se obtenga sobre $\boldsymbol{\theta}$. Esta información obtenida acerca de los parámetros se puede expresar en términos de una distribución de probabilidad, dado que así lo contempla la perspectiva bayesiana. De esta forma, y considerando que es a partir de los datos de donde se obtiene nueva información de $\boldsymbol{\theta}$, buscaremos obtener una distribución de probabilidad para los parámetros que dependerá de los datos disponibles, es decir, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$. El uso de π en lugar de f es únicamente para denotar que hablamos de una distribución de probabilidad de unos parámetros. Como se puede observar, esta expresión es muy similar a la función de verosimilitud, siendo el resultado de intercambiar $\boldsymbol{y}$ y $\boldsymbol{\theta}$, lo que nos lleva al teorema de Bayes una vez más:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})} \propto f(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.3}$$

Esta formulación del teorema de Bayes es la propia de cualquier proceso inferencial bayesiano e incluye todos los elementos que participan en el mismo.

- $\pi(\boldsymbol{\theta})$ es la denominada distribución *a priori* de los parámetros, ya que no depende

de los datos recogidos en la muestra. Recoge el conocimiento previo que se tiene acerca de $\boldsymbol{\theta}$.

- $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ es la función de verosimilitud, que combina la información distribucional de Y con la muestra $\boldsymbol{y}$.
- $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ es la distribución *a posteriori* de los parámetros, ya que se obtiene a partir de la previa y tras añadir la información incluída en la verosimilitud. Su obtención es el objetivo principal de la inferencia bayesiana.
- $f(\boldsymbol{y})$ es la distribución predictiva previa, también denominada verosimilitud marginal, obtenida tras integrar (en el marco continuo) la función de verosimilitud con respecto a los parámetros. Precisamente esta integral no es, en general, fácil de obtener, lo que en muchos casos conlleva la necesidad de emplear métodos numéricos.

Dada la dificultad de obtener $f(\boldsymbol{y})$ resulta habitual formular el teorema de Bayes eliminando este término, indicando pues la proporcionalidad entre el resto de elementos.

Así pues, la estadística bayesiana plantea un proceso de obtención de conocimiento definido de manera secuencial: se parte de una distribución *a priori*de los parámetros que recoge la información previa de la que se dispone, a partir de una muestra se obtiene la función de verosimilitud, y mediante el teorema de Bayes se combinan ambos elementos para actualizar la información, obteniendo la distribución *a posteriori*.

El primer paso, como se ha indicado anteriormente, es la especificación de una distribución *a priori* de los parámetros, $\pi(\boldsymbol{\theta})$, en la que se recoge el conocimiento previo del que se dispone acerca de los mismos. Esta distribución puede ser más o menos informativa. Resulta habitual emplear distribuciones poco informativas tanto si no se dispone de información previa como si se busca que el peso de la inferencia recaiga sobre los datos. Además, en muchos casos, se busca que la elección de la distribución previa no altere en gran medida el resultado, lo que supondría una pérdida de credibilidad de la inferencia realizada. La especificación de las distribuciones *a priori* es un tema que ha sido ampliamente discutido y cuyo estudio genera interés por si mismo [1, 2, 3].

En segundo lugar, se dispone de la información contenida en la muestra, $\boldsymbol{y}$, que como se ha mencionado anteriormente, es incluída en el proceso inferencial a través de la función de verosimilitud. Esta información combinada con la distribución previa de los parámetros lleva tras la aplicación del teorema de Bayes a la obtención de la distribución *a posteriori*, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$. Esta distribución describe el comportamiento y variabilidad que se espera en los parámetros, $\boldsymbol{\theta}$, contando con la información que proveen los datos, y por tanto aportando información sobre la propia variable de interés Y, la cual depende intrínsecamente de estos parámetros.

Cabe destacar, que si bien se ha dicho que el objetivo del proceso inferencial será principalmente obtener esta distribución, esto no significa que sea siempre $\pi(\boldsymbol{\theta})$ la medida última obtenida. De hecho, como se verá en secciones posteriores ya centradas en supervivencia, la distribución *a posteriori* de los parámetros se empleará para obtener las distribuciones de otras medidas de supervivencia relevantes.

1.2.1. Modelización en el marco bayesiano

Una vez descrito el proceso de inferencia desde la perspectiva bayesiana, resulta necesario profundizar en la modelización estadística dentro del marco bayesiano. Si bien para formular el teorema de Bayes como en la Equación 1.3 se ha considerado una variable respuesta Y, que depende de unos parámetros $\boldsymbol{\theta}$, este escenario no es más que un caso particular, relativamente simple. Así pues, existen formulaciones más generales para un modelo estadístico bayesiano, dentro de las que se encontraría el caso mencionado anteriormente.

Supongamos que se dispone de múltiples variables de interés, en lugar de tener una única como en el caso anterior, denotadas como $\boldsymbol{Y} = (Y_1, \cdots, Y_m)$. Será importante también considerar covariables para asegurar un buen ajuste del modelo, por lo que se considerará el vector de covariables $\boldsymbol{X} = (X_1, \cdots, X_l)$. Estas covariables no se dotarán de distribución de probabilidad, ya que se considerarán valores fijos observados. Por otro lado, vamos a incluir en la formulación de este modelo más general los dos elementos diferenciadores que son los que en capítulos posteriores permitirán profundizar en la modelización de la supervivencia que da lugar a esta tesis: una variable latente, Z, que será una variable no observada o parcialmente observada, y un vector de efectos aleatorios $\boldsymbol{\psi}$, que recogerá las diferencias entre individuos o grupos de ellos. Finalmente, como antes, $\boldsymbol{\theta}$ representará el vector de parámetros e hiperparámetros del modelo, de forma que todos estos elementos definen un modelo general dado por la expresión:

$$f(\boldsymbol{y}, z, \boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{y} \mid \boldsymbol{x}, z, \boldsymbol{\psi}, \boldsymbol{\theta}) f(z \mid \boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{\psi} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}). \tag{1.4}$$

Como se observa en 1.4, el modelo general se divide en distintos submodelos separados, debido a la asunción de independencia condicional entre los distintos elementos que componen el modelo. Así pues, por un lado se tendrá un modelo para las variables respuesta $f(\boldsymbol{y} \mid \boldsymbol{x}, z, \boldsymbol{\psi}, \boldsymbol{\theta})$ que se verá influido por todos los elementos mencionados, un modelo para la variable latente, $f(Z \mid \boldsymbol{x}, \boldsymbol{\theta})$, un modelo para los efectos aleatorios, $f(\boldsymbol{\psi} \mid \boldsymbol{\theta})$, y por último, una distribución *a priori* para los parámetros e hiperparámetros, $\pi(\boldsymbol{\theta})$.

Si bien este modelo general contempla la inclusión de todos estos elementos, se podrán realizar análisis en los que no participe alguno de ellos. De hecho, a lo largo de esta tesis se plantearán metodologías separadas para incluir una variable latente o efectos aleatorios, aunque serán fácilmente combinables dada la separabilidad, especificada anteriormente, entre los elementos del modelo.

En particular, ya desde la sección siguiente 1.3, introductoria al análisis de supervivencia, las variables $\boldsymbol{Y}$ serán tiempos de supervivencia, pudiendo incluirse un único tiempo (en un modelo univariante) o varios tiempos (riesgos competitivos y modelos multiestado). En el capítulo 2 se emplearán modelos de supervivencia multivariantes, en concreto modelos multiestado, en los que únicamente se tomarán en consideración covariables y parámetros, es decir, un modelo multiestado de la forma $f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$. En el capítulo 3 se añadirá un vector de efectos aleatorios, los cuales tendrán una estructura de correlación espacial. El modelo se dividirá en dos partes, uno para los tiempos

donde se incluyan estos efectos aleatorios, $f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\psi}, \boldsymbol{\theta})$, y otro para los propios efectos $f(\boldsymbol{\psi} \mid \boldsymbol{\theta})$, este último siguiendo una distribución normal multivariante, en cuya matriz de correlación se incluirán relaciones entre los tiempos y entre unas determinadas unidades espaciales. En el capítulo 4 se dejará de lado los efectos aleatorios y en su lugar se definirá una metodología para trabajar con la denominada curación como variable latente. El modelo se dividirá pues en $f(\boldsymbol{y} \mid \boldsymbol{x}, z, \boldsymbol{\theta})$ y $f(z \mid \boldsymbol{x}, \boldsymbol{\theta})$, de manera que para la variable de curación, Z, se definirá un modelo de regresión logística.

1.3. Análisis de supervivencia

Si en la sección anterior se trataba una parte esencial de esta tesis, que es el enfoque bayesiano como marco metodológico, en esta sección se introducirá lo que podríamos considerar como el verdadero eje vertebrador de este trabajo: el análisis de supervivencia. Cuando nos referimos a supervivencia lo hacemos pensando precisamente en un tiempo, el tiempo de supervivencia, y en cuántos individuos han sobrevivido tras este tiempo. Así pues, las variables principales y objeto de estudio en esta sección introductoria, así como a lo largo de todo el documento, serán tiempos, tiempos de supervivencia o, alternativamente, tiempos hasta la observación de determinados eventos de interés. La inferencia respecto a estos tiempos se realizará empleando técnicas propias del análisis de supervivencia, así como otras propias de los procesos estocásticos, siempre desde una perspectiva bayesiana. La introducción de los modelos será gradual: en primer lugar, el caso univariante, donde se modelizará un solo tiempo, a la vez que se introducen conceptos básicos y transversales del análisis de supervivencia; a continuación, se presentarán los modelos de riesgos competitivos como propuesta para el abordaje de situaciones con múltiples eventos; y se finalizará con la definición del marco de los modelos multiestado, el cual supone la generalización de los casos anteriores.

1.3.1. Conceptos básicos

El objeto de estudio en el análisis de supervivencia es lo que se denomina tiempo de supervivencia, definido como el tiempo que un individuo pasa sin manifestar un determinado evento de interés. Equivalentemente se puede entender como el tiempo hasta el evento. Este evento puede representar un amplio abanico de situaciones, condiciones médicas, enfermedades, etc.

En términos de notación estadística, supongamos que tenemos una variable aleatoria positiva T que es el tiempo de supervivencia. Este tiempo, que consideraremos continuo durante todo el documento pero que podría ser analizado como una variable discreta, tendrá asociada una función de densidad, $f(t)$, para $t \geq 0$. Tendrá también una función de distribución acumulada, $F(t) = P(T \leq t)$, que por definición será la probabilidad de que el tiempo de supervivencia sea menor que un tiempo t, y por tanto, la probabilidad de mostrar el evento antes de ese instante de tiempo t. Con esto, se puede definir una función alternativa, que es la función de supervivencia, $S(t)$, la cual toma la expresión

$$S(t) = P(T > t) = 1 - F(t) = 1 - \int_0^t f(u)\mathrm{d}u. \tag{1.5}$$

La función de supervivencia, como su nombre indica, puede interpretarse en una población como la proporción de individuos que sobreviven, sin evento, hasta el instante t. En cuanto a su valor, si la función $F(t)$, al ser una función de distribución acumulada, es una función creciente que varía entre 0 y 1, la función de supervivencia mostrará justo el comportamiento contrario. La función $S(t)$ es continua, con un valor máximo de 1 en $t = 0$, y decrece con el tiempo hasta alcanzar 0, cuando todos los individuos han experimentado el evento, $\lim_{t\to\infty} S(t) = 0$.

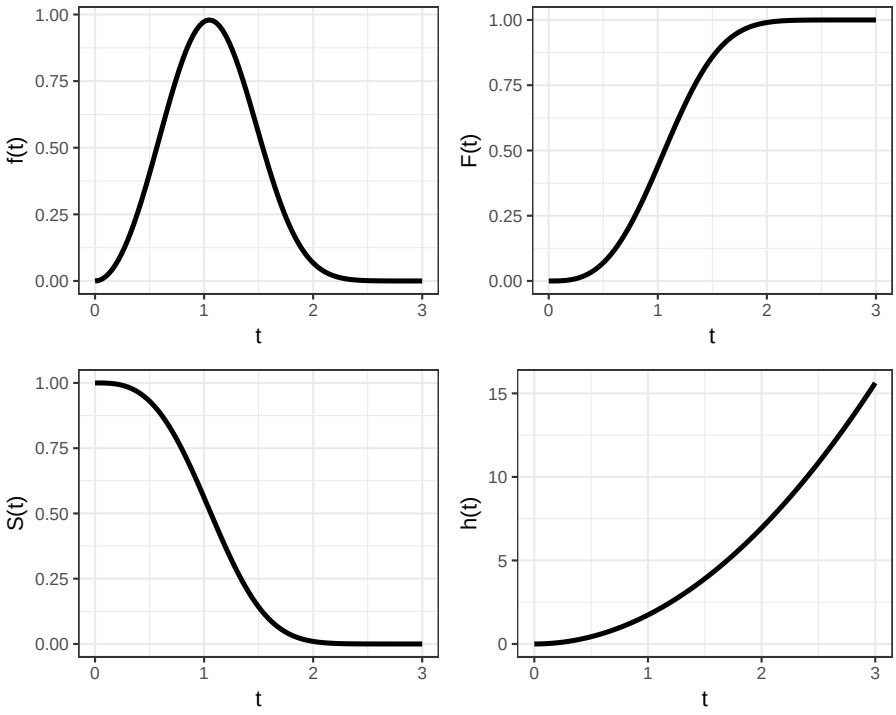

Figura 1.1: Ejemplos de función de densidad, $f(t)$, función de distribución acumulada, $F(t)$, función de supervivencia, $S(t)$, y función de riesgo, $h(t)$.

Todas las funciones mencionadas con anterioridad, $f(t)$, $F(t)$ y $S(t)$, son funciones que definen el compartamiento aleatorio del tiempo de supervivencia de forma única, de manera que con conocer una de ellas ya es suficiente como para caracterizar el tiempo de supervivencia y su distribución de probabilidad. Sin embargo, en supervivencia se suele emplear una cuarta función que también caracteriza el tiempo: la función de riesgo, $h(t)$, cuya definición formal vendría dada por la expresión

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} log(S(t)). \tag{1.6}$$

Esta función de riesgo, puede interpretarse, a grandes rasgos, como la probabilidad instantánea de observarse el evento en t, dado que no se ha observado hasta ese instante, si bien no es una función de probabilidad como tal. Guarda relación directa con el resto de funciones a través de las igualdades presentadas y, en particular, de la última formulación como la derivada del logaritmo de la función de supervivencia, cambiada de signo, se deduce que la función de supervivencia puede obtenerse partiendo de la función de riesgo mediante la expresión

$$S(t) = \exp\{-\int_0^t h(u)\mathrm{d}u\}. \tag{1.7}$$

Así pues, en análisis de supervivencia lo más habitual es abordar la modelización en términos de la función de riesgo, $h(t)$, en lugar de trabajar con la función de densidad como tal. En esta función de riesgo se basa, precisamente, el modelo más empleado en este campo, que es el modelo de riesgos proporcionales de Cox[4], el cual permite incluir covariables para estudiar su efecto en el riesgo de evento. El modelo se construye a partir de dos elementos: una función de riesgo basal, $h_0(t)$, que indicaría el riesgo base que tendría un individuo de experimentar el evento, y un término de regresión, donde se incluyen las covariables, $\boldsymbol{x}$, así como sus posibles efectos, $\boldsymbol{\beta}$, incrementando o disminuyendo este riesgo base.

$$h(t) = h_0(t)\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}. \tag{1.8}$$

Como se aprecia en la equación 1.8, la función de riesgo no es más que el producto de la función de riesgo basal y la exponencial del término de regresión, el cual es constante respecto del tiempo; de ahí su nombre de riesgos proporcionales. Este modelo se define como un modelo semiparamétrico, de forma que en el paradigma frecuentista no requiere asumir una distribución de probabilidad para el tiempo T, ya que la función de riesgo basal puede estimarse mediante métodos no paramétricos, siendo esta una de sus mayores ventajas. Desde la perspectiva bayesiana, esta función de riesgo basal requiere ser modelizada, para lo cual existen múltiples aproximaciones. Desde los modelos paramétricos, como por ejemplo la distribución Weibull, sin duda el modelo tradicional más usado en aplicaciones relacionadas con la biometría, hasta modelizaciones dotadas de una mayor flexibilidad, como las obtenidas a partir de funciones constantes por partes o B-splines [5].

Por otro lado, en estos modelos el efecto de las covariables se mide en términos relativos mediante los denominados *hazard ratios* (HR). Dados dos individuos con valores de las covariables x_1 y x_2, definimos el HR como el cociente entre los riesgos de estos individuos, $HR_{x_2\,vs\,x_1}(t) = h(t \mid x_2)/h(t \mid x_1)$. Mediante esta medida se puede cuantificar el riesgo adicional resultante de tener como covariables x_2 frente a x_1. Habitualmente, se suelen fijar los valores de todas las covariables menos una, obteniendo así una medida del efecto de esta última sobre el riesgo. En particular, en los modelos de riesgos proporcionales los *hazard ratios* resultantes son constantes respecto al tiempo. Sin pérdida de generalidad, al incluir una única covariable en el modelo de riesgos proporcionales, se

obtiene que

$$HR_{x_2\, vs\, x_1} = \frac{h(t \mid x_2)}{h(t \mid x_1)} = \frac{h_0(t)\exp\{\beta x_2\}}{h_0(t)\exp\{\beta x_1\}} = \exp\beta(x_2 - x_1). \tag{1.9}$$

Censura y verosimilitud

La particularidad de los datos de supervivencia es que para disponer del tiempo de supervivencia de todos los individuos se requiere que se haya observado el evento en todos ellos. Sin embargo, en general, no siempre se van a poder observar estos tiempos debido a que el tiempo de seguimiento no será suficiente como para que todos los individuos experimenten el evento de interés. Es lo que en análisis de supervivencia se conoce como censura por la derecha. En nuestro caso, la censura por la derecha vendrá dada por la existencia de un límite superior del intervalo de observación, fijo y establecido previamente, como el marcado por la disponibilidad de seguimiento hasta un año determinado. Cabe mencionar que existen otros tipos de censura, como la censura por la izquierda cuando el estudio empieza después de que el individuo muestre el evento, o la censura por intervalos en el caso de que no se conozca exactamente el instante en que tiene lugar el evento, sino un intervalo. Todos los modelos presentados en este trabajo contemplan la censura por la derecha, aunque podrían generalizarse a otros tipos de censura.

Esta censura tendrá un efecto en la verosimilitud, dado que para cada individuo la información que se tendrá será diferente en función de si su tiempo se ha observado o no. Para aquellos individuos que presentaron el evento en cuestión la verosimilitd será el valor de la función de densidad en el instante de tiempo observado, $f(t)$, como es habitual. Por otro lado, de los individuos con tiempos censurados se sabe que han sobrevivido al menos hasta un determinado tiempo t, incluyéndose como verosimilitud la función de supervivencia, $S(t)$. Así pues, sea C la variable indicadora de censura, con valor 1 en caso de observarse el evento y 0 en caso de ser un tiempo censurado, puede definirse la función de verosimilitud, $L(\boldsymbol{\theta})$, para un modelo de supervivencia con censura como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i)^{c_i} S(t_i)^{1-c_i} = \prod_{i=1}^{n} h(t_i)^{c_i} S(t_i), \tag{1.10}$$

donde t_i es el tiempo de supervivencia observado para el individuo i, coincidiendo con el tiempo hasta el evento o siendo un tiempo censurado, y c_i es el valor de la variable de censura para ese mismo individuo. Notar que resultará habitual emplear la segunda igualdad en 1.10, ya que será $h(t)$ la función a modelizar.

1.3.2. Modelos de riesgos competitivos

Los modelos de riesgos competitivos son modelos estadísticos que se emplean en análisis de supervivencia cuando existen múltiples eventos posibles. Estos modelos pueden ser aplicados en aquellos estudios en los que los individuos no tienen un único evento de

interés, sino que este compite con otros eventos, o aquellos en los que el evento a estudiar podría no observarse debido a otros eventos, como es el caso del riesgo competitivo de muerte.

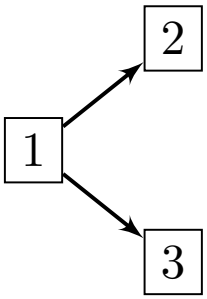

Figura 1.2: Diagrama de un modelo de riesgos competitivos con dos eventos.

Es en la investigación clínica y epidemiológica donde los modelos de riesgos competitivos son especialmente útiles y donde se han utilizado ampliamente, con el fin de obtener y evaluar tasas y riesgos asociados a distintas enfermedades (Lau et al., 2009; Andersen et al., 2012; Wei et al., 2018; Schuster et al., 2020)[6, 7, 8, 9]. Así pues, existen diferentes métodos y perspectivas desde las que estudiar este tipo de modelos. Las dos metodologías más populares se basan en realidad en dos planteamientos distintos, en función de cuál es la medida sobre la que recae la modelización. Por un lado, el modelo de Fine y Gray [10] busca modelizar la función de incidencia acumulada (también denominada subdistribución) planteando un modelo de regresión que incluye el efecto de unas determinadas covariables sobre la propia incidencia. Alternativamente, en los modelos de riesgos específicos se estudia el efecto de las covariables sobre cada una de las funciones de riesgo asociadas a cada uno de los eventos. A lo largo de este trabajo nos centraremos en los modelos de riesgos específicos ya que no tienen los problemas de identificabilidad de los primeros y se conectan de manera natural e intuitiva con el marco de los modelos multiestado.

Los individuos partirán desde un punto inicial común y se les seguirá en el tiempo hasta la ocurrencia del primer evento. Se asume que este evento no permitirá observar el resto de eventos, censurando su observación. Sin pérdida de generalidad, se define un modelo de riesgos competitivos con un estado inicial, 1, y dos posibles eventos de interés, los cuales se denotan como los estados 2 y 3 (Figura 1.2). En términos estadísticos, se denotará por T_{12} y T_{13} al tiempo hasta los eventos 2 y 3, respectivamente, así como $T = \text{mín}\{T_{12}, T_{13}\}$, el tiempo hasta el primer evento. Cabe destacar que la notación empleada es la propia de los modelos multiestado, si bien tiene sentido en este caso dado que los modelos de riesgos competitivos encajan perfectamente dentro del marco multiestado.

Los dos conceptos clave dentro de este ámbito son las funciones de riesgo específicas y las funciones de incidencia acumulada, ambas totalmente ligadas entre sí.

La función de riesgo específico de experimentar el evento j, $j = 2, 3$, en presencia del

otro evento competitivo se define como

$$h_{1j}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta = j \mid T \geq t)}{\Delta t}, \quad t > 0, \tag{1.11}$$

indicando $\delta = j$ la ocurrencia de ese evento j en particular. Las funciones de riesgo se definen en términos de probabilidades condicionales, sin embargo no son probabilidades como tal ya que pueden tomar valores mayores a 1. De forma intuitiva y aproximada, estos riesgos se pueden interpretar como la probabilidad instantánea de que un individuo que no ha experimentado el evento de interés antes de t, lo experimente en una pequeña ventana temporal.

Por otro lado, las funciones de incidencia acumulada son medidas que aportan mucha más información y de una forma más interpretable, ya que indican la probabilidad de haber sufrido un determinado evento en un instante concreto. En particular, para un evento j, se define como la probabilidad de fallo debido a la ocurrencia del evento j antes de un tiempo t, $F_{1j}(t) = P(T \leq t, \delta = j)$, cuya expresión se deriva a partir de las funciones de riesgo como

$$F_{1j}(t) = P(T \leq t, \delta = j) = \int_0^t h_{1j}(u) S(u) \, \mathrm{d}u \,, \tag{1.12}$$

donde $S(t)$ es la función de supervivencia que indica la probabilidad de no sufrir ningún evento antes de t.

$$S(t) = \exp\Big\{ - \int_0^t \Big(h_{12}(u) + h_{13}(u) \Big) \mathrm{d}u \Big\} \tag{1.13}$$

Por definición, las funciones de incidencia acumulada indican la proporción de individuos que han sufrido un evento en un momento dado, por lo que junto con la probabilidad de supervivencia se obtiene que $F_{12}(t) + F_{13}(t) = 1 - S(t)$. En concreto, también se asume que todos los individuos llegarían a experimentar uno u otro evento si se les diese el tiempo necesario, es decir, $\lim_{t \to 0} S(t) = 1$, como verifica cualquier función denominada función de supervivencia. De esta forma, nos encontramos también con que $\lim_{t \to 0} F_{1j}(t) = P(\delta = j)$, es decir, que cuando el tiempo t es suficientemente grande la función de incidencia acumulada para el evento j coincidirá precisamente con la probabilidad que tendría esa población de sufrir el evento en cuestión.

1.3.3. Modelos multiestado

Los modelos multiestado son modelos estocásticos complejos centrados en analizar las trayectorias definidas por la ocurrencia de múltiples eventos de interés a lo largo del tiempo. Estos modelos generalizan un gran número de escenarios propios del análisis de supervivencia, desde modelos unidimensionales con un único tiempo hasta un determinado evento, hasta modelos con múltiples eventos como los modelos de riesgos competitivos o de eventos repetidos. Dentro del marco de los modelos multiestado los eventos son tratados como estados de un proceso estocástico, mientras que su ocurrencia se define como transiciones entre un estado de partida y uno de llegada. La incertidumbre

asociada a cada una de las transiciones se modeliza mediante las llamadas probabilidades de transición, o equivalentemente, empleando las intensidades de transición. Estas últimas son análogas a las funciones de riesgo procedentes del análisis de supervivencia. Por su construcción, los modelos multiestado resultan especialmente útiles en investigación clínica y epidemiología ya que proporcionan naturalmente un marco con el que analizar el historial y evolución de enfermedades con patrones complejos.

Desde una perspectiva probabilística, un modelo multiestado se define como un proceso estocástico $\{Z(t), t > 0\}$ función de un tiempo continuo t, que toma como posibles valores los diferentes estados en los que pueden encontrarse los individuos. En particular, tenemos un espacio de estados $S = \{1, 2, 3, \cdots\}$ y un conjunto Ω que estaría formado por todas las transiciones que serían posibles en este proceso, $\Omega = \{1 \to 2,\ 1 \to 3, \cdots\}$.

El marco metodológico de los modelos multiestado es el propio de los procesos estocásticos en tiempo continuo y espacio de estados finito, este último, como ya se ha mencionado con anterioridad, definido por los distintos eventos de interés objeto de estudio. Así pues, y debido a su tratamiento estocástico, los conceptos y procedimientos que se emplearán serán los propios de este campo. En este sentido, resulta habitual expresar el comportamiento probabilístico del modelo multiestado en términos de las llamadas probabilidades de transición. De esta manera es posible representar la evolución del proceso tomando como referencia dos instantes de tiempo diferentes mediante estas probabilidades condicionadas que toman la forma

$$p_{ij}(s,t) = P(Z(t) = j \mid Z(s) = i), \tag{1.14}$$

es decir, la probabilidad de estar en el estado j en tiempo t ($Z(t) = j$) dado que el mismo individuo se encontraba en el estado i en tiempo s ($Z(s) = i$), para $s < t$.

Las probabilidades de transición nos ofrecen una información intuitiva y fácilmente interpretable sobre el problema de interés, aunque su modelización no es en general sencilla. Por esta razón, la modelización estadística normalmente se centra en las intensidades de transición, que a pesar de ser menos intuitivas resultan más fáciles de modelizar. Tanto las probabilidades como las intensidades de transición son conceptos relacionados y las dos caras de una misma realidad, de forma que unas se derivan de las otras. Así pues, las intensidades pueden ser interpretadas como el riesgo instantáneo de avanzar a un estado j desde un estado i en el que se está actualmente, definidas en términos generales a partir de las probabilidades de transición como:

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t}. \tag{1.15}$$

Es importante destacar que existe una equivalencia natural entre algunos conceptos del mundo de los procesos estocásticos y del análisis de supervivencia. Al fin y al cabo, un tiempo de transición desde un estado i a uno j, T_{ij} puede verse como el tiempo de supervivencia entre un evento inicial i y la entrada en el estado j, la cual sería el propio evento de interés. Profundizando en esta equivalencia, podemos decir que las intensidades de transición del marco estocástico (2.2) serían análogas a las funciones de riesgo del mundo de la supervivencia, es decir,

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T_{ij} < t + \Delta t \mid T_{ij} \geq t)}{\Delta t}. \tag{1.16}$$

Así pues, al contrario de las probabilidades de transición, la modelización de las intensidades puede abordarse de manera natural empleando, por ejemplo, modelos de riesgos proporcionales de Cox, en los que incluir términos de regresión para estudiar el efecto de covariables, efectos aleatorios u otros elementos susceptibles de aportar información acerca del tiempo considerado como variable de interés.

1.3.4. Modelos de curación

Los modelos de curación son un tipo especial de modelos de supervivencia en los que se asume la existencia de una fracción de individuos que nunca experimentará el evento de interés [11]. Estos modelos son especialmente útiles en epidemiología para el estudio de enfermedades, siempre y cuando sea razonable plantear la curación de las mismas. En particular, nos centraremos en los modelos de curación de tipo mixtura, un subgrupo dentro de los modelos de curación en los que la población de estudio se define como combinación de dos grupos de individuos claramente diferenciados, curados y no curados.

En términos estadísticos, podemos decir que los individuos curados presentan tiempos de supervivencia "infinitos", $T \to \infty$, de forma que por mucho que se siguiese a estos individuos a lo largo del tiempo no se esperaría observar el evento. Si bien este fenómeno recuerda a la censura por la derecha, se puede afirmar que es sustancialmente distinto. Cuando hablamos de censura realmente pensamos que sí que se observaría el evento en caso de tener un tiempo de seguimiento suficiente, es decir, que a largo plazo todos los individuos experimentarían dicho evento, o equivalentemente, ningún individuo podría sobrevivir, $\lim_{t\to\infty} S(t) = 0$. No obstante, en los modelos de curación de tipo mixtura se asume que habrá un grupo de pacientes curados que sí que sobreviviría, es decir, que $\lim_{t\to\infty} S(t) = \pi$, donde π es la proporción de individuos curados, $\pi \in [0, 1]$. Esta función de supervivencia (Figura 1.3) no será propia, pero se construirá a partir de una que sí que lo es, pudiendo tomar la expresión [12]:

$$S(t) = \pi + (1 - \pi)G(t) \tag{1.17}$$

donde $G(t)$ es la función de supervivencia propia, también llamada distribución de latencia, que describe el comportamiento de los individuos no curados. Dado que para este grupo de individuos se espera que todos muestren el evento, aunque la censura impida su observación, se cumple que $\lim_{t\to\infty} G(t) = 0$. Esta distribución de latencia podría modelizarse tanto de forma paramétrica como con modelos más flexibles, de la misma forma que se modeliza la supervivencia habitualmente.

El problema relevante a nivel estadístico que plantean este tipo de modelos es precisamente la identificación de los pacientes curados y no curados. Sin definir bien los grupos no resulta posible asociarles uno u otro comportamiento, una verosimilitud que

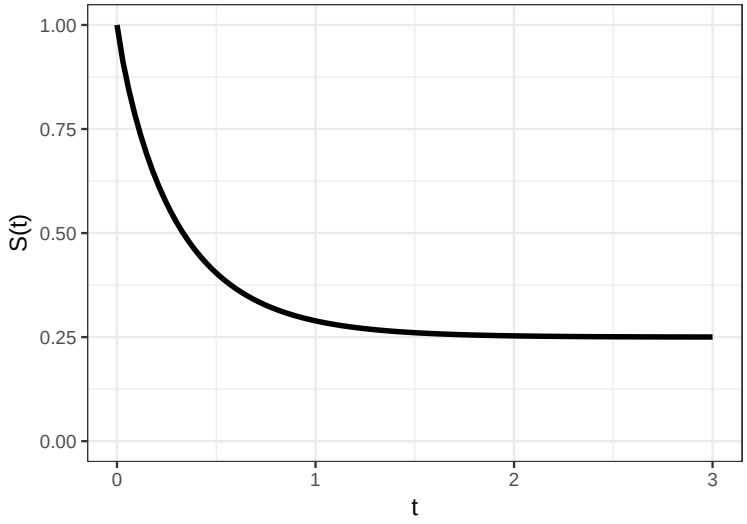

Figura 1.3: Función de supervivencia en un modelo de curación de tipo mixtura.

se corresponda con la del grupo al que pertencen. Respecto a los pacientes no curados tendremos algo de información, sabiendo con certeza que aquellos individuos en los que se observó el evento no estaban curados. Sin embargo, los individuos con tiempos censurados podrían ser tanto de un grupo como de otro, siendo necesario emplear otro tipo de técnicas, como podría ser su tratamiento como variables latentes.

1.4. Estadística espacial

La estadística espacial puede definirse como el conjunto de procedimientos estadísticos con los que se estudia la distribución de eventos o procesos que tienen lugar sobre un determinado espacio geográfico y de los cuales se conoce su ubicación. La forma que toma esta información espacial es la que define la naturaleza de los datos espaciales, llevando al desarrollo y utilización de distintas metodologías en función del tipo de datos. En términos generales, hay tres tipos bien establecidos, cada uno de los cuales define un campo en sí mismo dentro de la estadística espacial: datos geoestadísticos, datos de patrones puntuales y datos espaciales agregados. Brevemente, los datos geoestadísticos se obtienen como resultado de la observación de una cantidad de interés continua sobre una región, registrando sus valores en algunos puntos. Así pues, un objetivo frecuente sería obtener estimaciones y predicciones de las cantidades de interés para toda una región del espacio, extendiendo la información registrada puntualmente. En segundo lugar, los patrones puntuales no son más que una disposición de puntos a lo largo del espacio que indican la presencia de individuos. En lugar de focalizarse en una medida de interés, el objetivo principal suele ser estudiar el patrón espacial que genera esos datos. Finalmente, por datos agregados o de red (*lattice data* en inglés) nos referimos a datos que son resultado de conteos o de cantidades de interés agregadas, las cuales se estudian sobre

un número finito de regiones que dividen el espacio (ejemplo en Figura 1.4). Nos centraremos en este último tipo de datos, presentando algunos de los modelos empleados para su análisis y sirviendo de base para la construcción del modelo multiestado con efectos espaciales que se presentará en el Capítulo 3.

Como se ha mencionado, la información que contienen los datos agregados, por lo general, se suele expresar como datos de conteos, que indican cuántos individuos cumplen una cierta condición o muestran una determinada característica en cada una de las regiones que componen el espacio de estudio. En epidemiología, estos conteos se refieren a menudo a la presencia de una determinada enfermedad en la región, siendo el objetivo principal estudiar el riesgo de padecerla, así como las diferencias que existen entre las distintas áreas. Dada su naturaleza como datos de conteos el modelo de Poisson es con diferencia el más utilizado para su modelización, de forma que si denotamos por Y_i al número observado de individuos que muestran la enfermedad de interés en la región i, $i = 1, \cdots, n$, podemos definir un modelo de Poisson de la forma:

$$\begin{aligned} Y_i &\sim Po(E_i\theta_i) \\ log(\theta_i) &= \boldsymbol{x}_i^T\boldsymbol{\beta} + b_i \end{aligned} \tag{1.18}$$

donde E_i es el número esperado de individuos enfermos en la región i, $\boldsymbol{x}_i$ es un vector de covariables que incluye información agregada de las características de esta región i, y b_i es el efecto aleatorio de la propia región i sobre el riesgo relativo de enfermedad, θ_i. Este efecto aleatorio b_i será sobre el que se definirán las distribuciones de probabilidad y las distintas estructuras de correlación espacial entre regiones.

Figura 1.4: Mapa de la Comunitat Valenciana, España, dividido en Áreas de Salud.

Modelo normal

Una primera opción para modelizar estos efectos aleatorios es asumir que cada uno de ellos sigue una distribución normal iid., es decir, $b_i \sim N(0, \sigma^2)$. En este modelo considera la existencia de una heterogeneidad de las regiones con una distribución común e independiente, con varianza σ^2. Así pues, no incluye información espacial de ningún tipo.

Modelos CAR

Los modelos autoregresivos condicionales, en inglés *Conditional autoregressive* (CAR) models [13], pueden utilizarse de forma natural para modelizar efectos aleatorios que consideren asociación espacial entre las observaciones de distintas regiones. En particular, estos modelos asumen para los efectos aleatorios una distribución de probabilidad normal pero condicionada a los efectos de sus vecinos. A su vez, la relación de vecindad, por lo general, se entiende como relación de adyacencia, aunque otras modelizaciones podrían definir relaciones alternativas.

La primera posibilidad dentro de estos modelos es la distribución CAR intrínseca, la cual para un efecto aleatorio b_i se define como

$$b_i \mid b_{\sim i} \sim N\left(\frac{\sum_{j=1}^{n} w_{ij} b_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\sigma^2}{\sum_{j=1}^{n} w_{ij}}\right) \tag{1.19}$$

donde $b_{\sim i}$ representa los valores de los efectos aleatorios de los vecinos de i, excluyendo al propio i, w_{ij} es el elemento de la fila i y columna j de la matriz de adyacencia W, que toma como valores $w_{ij} = 1$ si i y j son vecinos y $w_{ij} = 0$ en caso contrario, y σ^2 es un parámetro de varianza desconocido. Alternativamente, si denotamos por $\boldsymbol{b} = \{b_i\}_{i=1}^{n}$ al vector de efectos aleatorios, podemos obtener la distribución conjunta de este vector, la cual se puede demostrar mediante la aplicación del Lema de Brook [14] que es

$$\boldsymbol{b} \sim N(0, \Sigma), \qquad \Sigma = \sigma^2 (D - W)^{-1} \tag{1.20}$$

donde $D = \text{diag}\{d_1, \cdots, d_n\}$ es la matriz diagonal cuyos elementos no cero son el número de vecinos de cada una de las regiones, d_i, para cada región i y siguiendo una distribución normal multivariante de dimensión n.

No obstante, esta distribución CAR intrínseca es una distribución impropia, lo que hace que no sea la mejor de las opciones. Convertirla en una distribución propia resulta sencillo, únicamente siendo necesario añadir un parámetro adicional, α, obteniendo un modelo CAR propio (también conocido como modelo de Cressie [15]), en el que el vector de efectos aleatorios seguiría una distribución

$$\boldsymbol{b} \sim N(0, \Sigma), \qquad \Sigma = \sigma^2 (D - \alpha W)^{-1} \tag{1.21}$$

Para mantener esta distribución propia, los valores del parámetro α deberían restringirse en función de los valores propios de la matriz $D^{-1}W$, aunque también suelen

restringirse en el intervalo $(0, 1)$. De esta forma, $\alpha = 1$ puede interpretarse como una correlación espacial total entre regiones, volviendo al modelo CAR intrínseco que era impropio. Por otro lado, $\alpha = 0$ implicaría la no existencia de correlación espacial, es decir, un escenario totalmente independiente por lo que respecta a los efectos aleatorios.

Modelo de Besag, York y Mollié

El siguiente modelo que presentaremos es el modelo de Besag, York y Mollié (BYM)[16], el cual puede considerarse el modelo más utilizado dentro del mapeo de enfermedades. Bajo esta modelización los efectos aleatorios se definen como la suma de dos efectos diferenciados, un efecto normal iid. que representa la heterogeneidad entre las regiones, v_i, y un efecto espacialmente estructurado, u_i, para cada región i:

$$\begin{aligned} \boldsymbol{b} &= \boldsymbol{u} + \boldsymbol{v} \\ \boldsymbol{u} &\sim N(0, \sigma_u^2 (D - W)^{-1}) \\ \boldsymbol{v} &\sim N(0, \sigma_v^2 I) \end{aligned} \tag{1.22}$$

Cabe destacar que, a pesar de su popularidad, el modelo presenta ciertos problemas de identificabilidad, ya que a partir de los datos solo es posible obtener información de la suma de los dos efectos y no de cada uno de ellos por separado. Esto ha llevado a algunos autores a estudiar y abordar esta cuestión, buscando versiones alternativas de este modelo [17].

Modelo de Leroux

Por último, se presentará el modelo de Leroux [18], el cual se define como una mixtura de un modelo con efectos aleatorios independientes y otro modelo de Besag impropio. En su formulación se pueden diferenciar también dos fuentes de variabilidad, al igual que en el modelo de BYM, dadas por los parámetros σ^2 para la dispersión de los datos y γ para cuantificar la fuerza de la correlación espacial. No obstante, el modelo de Leroux no presentaría los problemas de identificabilidad que sí presentaba el modelo de BYM. Bajo este modelo los efectos aleatorios siguen una distribución

$$\boldsymbol{b} \sim N(0, \Sigma), \qquad \Sigma = \sigma^2 [(1 - \gamma) I + \gamma (D - W)]^{-1}. \tag{1.23}$$

La interpretación del parámetro γ sería muy similar a la del parámetro α del modelo CAR propio, de forma que un valor de $\gamma = 0$ definiría un modelo independiente, sin correlación espacial entre regiones, mientras que $\gamma = 1$ implicaría un escenario con una correlación espacial total, equivalente a un modelo CAR intrínseco (que sería impropio). La ventaja principal del modelo de Leroux frente al CAR propio reside en que, si bien con $\alpha = 0$ se obtiene un escenario sin correlación espacial, los efectos aleatorios bajo el modelo CAR propio siguen dependiendo del número de vecinos de cada región, mientras que en el modelo de Leroux, si $\gamma = 0$, los efectos dependen únicamente de σ^2.

1.5. Estructura de la tesis

En este primer capítulo se han introducido muchos de los conceptos y elementos básicos que han sido esenciales para el desarrollo de todo el trabajo realizado en el marco de esta tesis. A continuación, se presentará a grandes rasgos el contenido y estructura de este trabajo. En primer lugar, en el Capítulo 2 se aplicarán modelos multiestado bayesianos, en particular, modelos de enfermedad-muerte, a un estudio de fractura recurrente de cadera y muerte. En este estudio se analizarán los datos de la cohorte PREV2FO, formada por pacientes mayores de 65 años con una fractura de cadera previa, los cuales fueron ingresados en los hospitales públicos de la Comunitat Valenciana. Se estimarán funciones de riesgo, incidencias acumuladas y probabilidades de transición relacionadas con refractura y muerte, con el objetivo de ilustrar la aplicación de estos modelos en un estudio epidemiológico con datos de la vida real. El Capítulo 3 se centrará en definir una metodología, desde una perspectiva bayesiana, para estudiar las diferencias entre regiones dentro del marco de los modelos multiestado. Para ello se propondrá un modelo de enfermedad-muerte con efectos aleatorios espaciales, cuya distribución se definirá a partir de una versión multivariante del modelo de Leroux. La metodología propuesta se ilustrará con una aplicación a los mismos datos de fractura de cadera, la cohorte PREV2FO, incluyendo información de carácter espacial relativa a las Áreas de Salud que componen la Comunitat Valenciana. Esta metodología utilizará la aproximación anidada integrada de Laplace (INLA) para realizar la inferencia. En el Capítulo 4 se explorará la extensión del modelo de enfermedad-muerte para incluir curación y cero-inflación, integrando estos elementos como estados dentro del modelo. Para el tratamiento de la curación como variable latente se definirá un algoritmo que combinará INLA y muestreo de Gibbs. La aplicación de este algoritmo se ilustrará mediante datos simulados, buscando valorar su efectividad a la hora de identificar subpoblaciones de pacientes curados y estimar su perfil de supervivencia. Por otro lado, se empleará un modelo de enfermedad-muerte cero-inflado con datos de la cohorte PREV2FO que incluya, además de las distintas transiciones a refractura y muerte, la muerte intrahospitalaria. En el Capítulo 5 se discutirán algunos de los resultados y conclusiones que se derivan de nuestro trabajo, así como las posibles líneas de investigación futuras. Finalmente, cabe mencionar que la memoria incluye un apéndice con el código de `R` que se ha empleado, con el fin de facilitar la reproducibilidad de los resultados presentados.

Capítulo 2

Modelos multiestado bayesianos en epidemiología

2.1. Introducción

La epidemiología es la disciplina científica dedicada al estudio de la incidencia y evolución de las enfermedades o condiciones de salud en poblaciones humanas. Aunque se pueden realizar estudios epidemiológicos con múltiples objetivos y desde perspectivas muy diferentes, el tiempo es, en muchos casos, un elemento importante, e incluso esencial. El análisis de supervivencia, como se ha comentado en el capítulo anterior, aporta metodologías y herramientas para el estudio de estos tiempos, y por tanto, también para el tratamiento de datos epidemiológicos basados en información temporal. En particular, en supervivencia nos referimos a tiempo transcurrido hasta la ocurrencia de uno o varios eventos de interés que, en la investigación epidemiológica, suelen estar relacionados con la muerte, la curación de una enfermedad, la progresión positiva o negativa de una enfermedad, o la aparición de efectos adversos, entre otros.

Los modelos multiestado definen un marco metodológico muy útil en epidemiología, incorporando de forma natural todos estos eventos que, en conjunto, permiten definir una historia clínica completa de los individuos. Los eventos son incluidos como estados dentro de un proceso estocástico. Estos estados generalmente representan diferentes condiciones de enfermedad y/o salud, y los individuos pueden moverse a través de ellos a lo largo del tiempo. Así, los tiempos de supervivencia relevantes serán tiempos entre los estados que, desde una perspectiva metodológica, pueden analizarse mediante procesos estocásticos y procedimientos de supervivencia (Ver Andersen y Keiding[19]). Las probabilidades de transición serán el resultado principal de estos modelos. Se pueden utilizar para evaluar la progresión de un individuo entre estados en función de la historia clínica, lo cual es especialmente útil de cara a la atención médica individualizada. A pesar de su enorme utilidad, los modelos multiestado aún no son muy populares en el mundo de la epidemiología [20, 21, 22], donde predominan los modelos de riesgos competitivos.

El contenido de este capítulo se basa en dos de nuestros artículos, el primero (Llopis-

Cardona et al. 2020)[23] publicado en *Journal of Bone and Mineral Research* y el segundo (Llopis-Cardona et al. 2023)[24] publicado en *BMC Medical Research Methodology*. En ambos artículos se propone un marco metodológico bayesiano para aplicar los modelos bayesianos multiestado en el ámbito de la investigación biomédica, en concreto, en el estudio de la fractura de cadera osteoporótica. En el primer artículo, el foco se pone en la estimación de cantidades de interés desde el punto de vista clínico, como lo son las incidencias acumuladas de refractura y muerte, o la probabilidad de transición de muerte tras refractura. En el segundo, se enfatiza la utilidad y potencial de estos modelos para estudiar procesos epidemiológicos en los que se consideren diferentes condiciones de salud, cuya ocurrencia puede ser secuencial en el tiempo.

Este capítulo estará estructurado de la siguiente manera. Primero, se presentará el modelo de enfermedad-muerte, un modelo multiestado con tres estados, dos de ellos transitorios y uno terminal. Además, describiremos cómo se calculan las probabilidades de transición en términos de las funciones de riesgo del marco de supervivencia. A continuación, presentaremos un estudio basado en datos de la cohorte PREV2FO, que comprende pacientes de 65 años o más dados de alta tras una hospitalización por fractura de cadera osteoporótica entre 2008 y 2015. Utilizaremos un modelo de enfermedad-muerte para analizar la evolución de estos pacientes a lo largo del tiempo, con especial interés en la aparición de una posterior fractura de cadera (refractura) y la muerte. Se realizará con un enfoque bayesiano, obteniendo distribuciones *a posteriori* de las incidencias acumuladas de refractura y muerte, así como de múltiples probabilidades de transición. El capítulo finalizará con algunas observaciones finales.

2.2. Modelo de enfermedad-muerte

Los modelos multiestado pueden dar lugar a múltiples escenarios y estructuras dependiendo del número de estados a considerar y cómo estos se relacionan mutuamente. En esta sección nos centraremos en un tipo particular de modelo multiestado conocido como *illness-death model* (modelo de enfermedad-muerte) o *disability model*. Este modelo consta de tres estados, cada uno representando diferentes condiciones, generalmente relativas a la salud aunque extrapolable a otros campos de estudio. Los individuos se encontrarían inicialmente en un estado inicial 1, pudiendo por ejemplo indicar estar sano o cumplir un criterio de inclusión en el estudio; el estado 2 sería un estado intermedio, asociado con la enfermedad o recaída; y el estado 3 sería un estado final, como la muerte del paciente. Así pues, tras su entrada en el estudio, estando en el estado 1, los individuos pueden directamente progresar hasta enfermedad o muerte. Sin embargo, también es posible que un paciente fallezca tras padecer la enfermedad, pasando del estado 2 al 3. Así pues nos enfrentamos a una estructura en la que existen dos eventos competitivos, enfermedad y muerte, pero uno de ellos, el estado de enfermedad, dotado de una naturaleza transitoria, la cual permite acceder al evento terminal que sería la muerte (Figura 2.1).

Como caso particular de modelo multiestado, un modelo de enfermedad-muerte no es sino un proceso estocástico $\{Z(t), t > 0\}$, dotado de un espacio de estados $S = \{1, 2, 3\}$

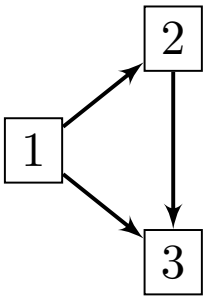

Figura 2.1: Diagrama de un modelo de enfermedad-muerte.

y con tres posibles transiciones, $\Omega = \{1 \to 2,\ 1 \to 3,\ 2 \to 3\}$. Así pues, existen diferentes posibles formulaciones por lo que respecta a la estructura y propiedades del proceso estocástico. Dos características primordiales a considerar serán: la homogeneidad del proceso y la propiedad markoviana. Un proceso se considera homogéneo si la probabilidad de transición depende únicamente de la diferencia entre el instante inicial s y el t, es decir, que mientras haya transcurrido el mismo tiempo, la probabilidad de transición será la misma, independientemente del instante inicial, $p_{ij}(s,t) = p_{ij}(s+h, t+h)$, $s+h \geq 0$. Por otro lado, la propiedad markoviana se refiere a la dependencia de las probabilidades de transición respecto al estado actual, olvidando aquello ocurrido antes de llegar al estado actual. En nuestro caso, consideraremos que el proceso será homogéneo, pero asumiremos una estructura semimarkoviana [25]. Así, si bien la evolución desde el estado inicial a los estados de enfermedad o muerte sí que dependerán únicamente del estado actual (el inicial), la transición de enfermedad a muerte dependerá además del tiempo que se ha estado en el estado inicial antes de pasar al estado de enfermedad.

Por lo que respecta al comportamiento aleatorio del modelo de enfermedad-muerte, este vendrá determinado por la distribución inicial del proceso, $\mathrm{P}(Z(0) = i)\ \forall i \in S$. En nuestro modelo todos los individuos estarán en el estado inicial 1 al principio del estudio, $t = 0$, es decir, $\mathrm{P}(Z(0) = 1) = 1$, y las probabilidades de transición entre estados se definirán como:

$$\begin{aligned} &p_{1j}(s,t) = \mathrm{P}(Z(t) = j \mid Z(s) = 1), \quad s \leq t,\ j = 2,3, \\ &p_{23}(s, t \mid t_{12}) = \mathrm{P}(Z(t) = 3 \mid Z(s) = 2, T_{12} = t_{12}), \quad t_{12} \leq s \leq t, \end{aligned} \tag{2.1}$$

donde T_{12} es el tiempo que transcurre en el estado 1 hasta que se alcanza el estado 2.

Por otro lado, estarán las intensidades de transición o equivalentemente funciones de riesgo, que tomarán la forma

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t}. \tag{2.2}$$

En particular, debido a la asunción de una condición semimarkoviana, a la intensidad de la transition $2 \to 3$ se le tendrá que añadir la condición $T_{12} = t_{12}$.

En la práctica, y dado que el peso de la modelización recaerá sobre las intensidades de transición, se utilizarán otras relaciones que enlacen ambas medidas[35], derivándose así las probabilidades de transición a partir de las funciones de riesgo como:

$$\begin{aligned}
p_{11}(s,t) &= \exp\Big\{-\int_s^t \Big(h_{12}(u)+h_{13}(u)\Big)\mathrm{d}u\Big\}\\
p_{22}(s,t \mid t_{12}) &= \exp\Big\{-\int_s^t h_{23}(u-t_{12}|t_{12})\mathrm{d}u\Big\}\\
p_{12}(s,t) &= \int_s^t p_{11}(s,u)h_{12}(u)p_{22}(u,t|u)\mathrm{d}u\\
p_{13}(s,t) &= 1-p_{11}(s,t)-p_{12}(s,t)\\
p_{23}(s,t \mid t_{12}) &= 1-p_{22}(s,t)\\
p_{33}(s,t) &= 1.
\end{aligned} \tag{2.3}$$

Algunas de las probabilidades más relevantes que se pueden obtener mediante estas expresiones son: $p_{11}(0,t)$, la probabilidad de permanencia en el estado inicial en un instante t y por tanto la no ocurrencia de cualquier evento en ese periodo; $p_{12}(0,t)$ y $p_{13}(0,t)$, que proporcionan información acerca de la probabilidad de estar en los estados 2 y 3 en tiempo t, respectivamente; y $p_{22}(s,t \mid t_{12})$, la cual quantifica la permanencia en el estado de enfermedad en t, sabiendo que en s aún se estaba en ese estado y que se enfermó en t_{12}.

Por otro lado, dado que un modelo de riesgos competitivos con dos eventos es un caso particular de modelo de enfermedad-muerte, en el cual no se considera la transición de enfermedad a muerte, las funciones de incidencia acumulada, presentadas en secciones anteriores, también se pueden emplear para obtener información de la proporción de individuos que ha accedido al estado de enfermedad o al estado de muerte sin enfermedad.

Finalmente, cabe mencionar que todas las probabilidades y funciones mostradas se definen en una escala de tiempo desde el instante incial, de manera que t_{12}, s, y t son tiempos desde el momento en el que se entra en el estudio, en el estado 1. Sin embargo, la probabilidad de transición $p_{23}(s,t \mid t_{12})$ puede reescalarse fijando $s = t_{12}$, es decir, empezando a contar desde el mismo instante en que se alcanza el estado de enfermedad. En consecuencia, esta probabilidad dependerá únicamente de $t-t_{12}$, que es precisamente el tiempo de transición desde 2 a 3. Alternativamente, se podría fijar otros puntos iniciales como $s = t_{12}+1$ o $s = t_{12}+2$, obteniendo de esta forma la probabilidad de muerte para aquellos que sobrevivieron durante 1 y 2 años tras entrar en el estado de enfermedad, respectivamente.

2.3. Caso práctico: progresión tras fractura de cadera.

2.3.1. Cohorte PREV2FO

Se dispone de una cohorte poblacional formada por 34491 pacientes mayores de 65 años que fueron dados de alta tras ser hospitalizados a causa de una fractura de

cadera. Estas fracturas ocurrieron específicamente durante el periodo de estudio entre el 1 de enero de 2008 y el 31 de diciembre de 2015. No obstante, con el fin de asegurar un mínimo seguimiento de 1 año para todos los individuos, se dispone de información relativa a nuevos eventos, ya sea fracturas o fallecimiento, hasta el 31 de diciembre de 2016, considerado como fecha de fin de estudio.

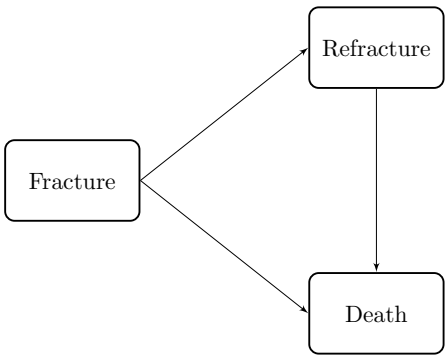

Figura 2.2: Diagrama del modelo de enfermedad-muerte empleado para el estudio de la fractura de cadera recurrente.

El estudio se llevó a cabo en la Comunitat Valenciana, una de las Comunidades Autónomas de España con alrededor de 5 millones de habitantes, suponiendo cerca del 10 % de la población de este país. El Sistema Valenciano de Salud (SVS) proporciona un servicio de asistencia sanitaria universal (exceptuando el copago farmacéutico) al 97 % de la poblacion valenciana, y está formado por una amplia red de hospitales públicos, centros de atención primaria y otros recursos públicos, bajo la gestión autónoma del Gobierno de la Comunidad.

Como se ha mencionado con anterioridad, se incluyeron todos los pacientes dados de alta con vida de los hospitales del SVS tras sufrir una fractura de cadera (códigos de la Clasificación Internacional de Enfermedades, novena revisión, modificación clínica [CIE-9 MC]: 820.xx y 733.14), con 65 o más años en el momento del alta, entre el 1 de enero de 2008 y el 31 de diciembre de 2015, excluyendo a aquellos individuos con diagnósticos de fractura múltiple, accidente de tráfico o cáncer de huesos. Se consideraron otros criterios de exclusión tales como no ser residente en la Comunidad o no disponer de cobertura farmacéutica, dado que estas situaciones podrían suponer problemas en el seguimiento de estos pacientes.

Los datos se obtuvieron de la Valencia Health System Integrated Database (VID). Esta base de datos constituye un extenso conjunto de información, procedente de múltiples fuentes, relativa a los habitantes de la Comunitat Valenciana, incluyendo desde datos sociodemográficos y administrativos (sexo, edad, nacionalidad, etc.), hasta diagnósticos, procedimientos, prescripción y dispensación farmacéutica, utilización de servicios hospitalarios, ambulatorios u otros servicios públicos de salud. Si bien la recogida de información no se inició simultáneamente en lo que se refiere a cada una de estas bases de datos, VID integra toda esta información desde 2008. Enlazar la información de las

distintas fuentes ha sido posible mediante un número de identificación personal único (número de tarjeta sanitaria).

Por lo que respecta a los eventos analizados tras la fractura, se consideraron hospitalizaciones por fractura de cadera recurrente (siguiendo los mismos criterios empleados para la definición de la fractura índice) y muerte por cualquier causa tras la fecha índice. Se siguió a los pacientes en el tiempo tras la fractura índice censurando únicamente debido a fallecimiento o fin de estudio (31 de diciembre de 2016). Así pues, las variables resultantes de interés serían el tiempo hasta una fractura recurrente, el tiempo hasta la muerte, y el tiempo desde la refractura hasta la muerte.

Este estudio se puede definir en términos de un modelo multiestado, en concreto un modelo de enfermedad-muerte, que estaría formado por tres estados: fractura inicial, por la que los pacientes entran en la cohorte, refractura y muerte. Se definirían también tres transiciones, dos desde la fractura hasta la refractura o la muerte, con carácter competitivo, y una última transición para la muerte tras la refractura 2.2. De esta manera, la fractura de cadera recurrente se considera tanto un estado de transición entre la fractura índice y la muerte, como un punto final para el cálculo de la incidencia acumulada de refractura. En este modelo, los pacientes fallecidos antes de padecer una refractura se consideran observaciones censuradas de cara a la estimación de los riesgos e incidencias de fractura de cadera recurrente.

2.3.2. Consideraciones éticas

El Estudio PREV2FO fue aprobado por el Comité Ético de la Dirección General de Salud Pública y Centro Superior de Investigación en Salud Pública (reuniones del 31 de mayo de 2013 y del 29 de enero de 2016). El Estudio PREV2FO2, que amplía la cohorte poblacional, fue aprobado por el Comité Ético de Investigación con Medicamentos del Hospital Clínico Universitario de Valencia (reunión del 29 de noviembre de 2018).

2.3.3. Modelización e inferencia bayesiana

En este estudio se han modelizado las posibles trayectorias descritas por los pacientes del estudio PREV2FO mediante un modelo de enfermedad-muerte. De acuerdo con lo mencionado con anterioridad, el estado inicial de todos los pacientes del estudio es una fractura de cadera de la que han sido dados de alta del hospital, que marcaría la entrada en el estudio. Los dos eventos restantes de interés serían una nueva fractura de cadera, definida como el estado de refractura, y la muerte, con su respectivo estado. En realidad, para la estimación de la incidencia de fracturas recurrentes de cadera y muerte tras una fractura de cadera osteoporótica, uno de los principales objetivos del estudio, podrían emplearse modelos de riesgos competitivos. No obstante, con una modelización de riesgos específicos, esto supondría considerar la condición de refractura como un evento terminal que compite con el riesgo de muerte. Este no sería el caso en nuestro estudio dado que la condición de refractura se plantea como un estado transitorio hacia la muerte (Ver Figura 2.2).

Los tiempos de supervivencia relevantes y por tanto a definir en el marco de esta modelización son: T_{FD}, el tiempo desde el alta tras la fractura índice (F) hasta la muerte (D), T_{FR}, el tiempo desde el alta tras la fractura índice hasta la refractura (R), y T_{RD}, el tiempo desde la refractura hasta la muerte.

Se han utilizado modelos de riesgos proporcionales de Cox [4] para modelizar el comportamiento aleatorio de estos tiempos T_{FR}, T_{FD}, y T_{RD}. En particular, empleando funciones de riesgo basales Weibull e incluyendo el sexo y la edad al alta como covariables en el término de regresión. Cabe destacar que la transición de refractura a muerte se ha modelizado teniendo en cuenta el momento en que los pacientes se refracturaron, definiendo así un modelo de semi-Markov para esta transición.

Dentro del modelo, cada transición consta de cuatro parámetros desconocidos. Dos parámetros son la forma y la escala de las distribuciones Weibull basales, $\alpha^{(ij)}$ y $\lambda^{(ij)}$, respectivamente, para la transición del estado i al j. Los otros dos parámetros serían los coeficientes de regresión asociados a las covariables sexo y edad: $\beta^{(ij)}_{\text{Wo}}$, definido como el coeficiente asociado a ser mujer, ya que los hombres serían el grupo de referencia incluido en el riesgo basal, y $\beta^{(ij)}_{\text{Age}}$, respectivamente. Finalmente, considerando estos parámetros, el modelo de enfermedad-muerte se puede definir en términos de funciones de riesgo como:

$$
\begin{aligned}
h^{FR}(t) &= h_0^{FR}(t)\exp\{\beta^{FR}_{\text{Wo}}\text{I}_{\text{Wo}} + \beta^{FR}_{\text{Age}}\text{Age}\}\\
h^{FD}(t) &= h_0^{FD}(t)\exp\{\beta^{FD}_{\text{Wo}}\text{I}_{\text{Wo}} + \beta^{FD}_{\text{Age}}\text{Age}\}\\
h^{RD}(t-t_{FR} \mid T_{FR}=t_{FR}) &= h_0^{RD}(t-t_{FR} \mid T_{FR}=t_{FR})\exp\{\beta^{RD}_{\text{Wo}}\text{I}_{\text{Wo}} + \beta^{RD}_{\text{Age}}\text{Age}\},
\end{aligned}
\tag{2.4}
$$

donde I_{Wo} es la variable indicadora que toma el valor 1 para las mujeres y 0 para los hombres. La edad se incluyó como covariable en nuestro modelo centrada en la media, para mejorar la convergencia de los métodos computacionales MCMC posteriores. Por último, h_0 son las funciones de riesgo basal que, como se ha especificado previamente, serían las propias de una distribución Weibull y tomarían la forma

$$
\begin{aligned}
h_0^{FR}(t) &= \alpha^{(FR)}\lambda^{(FR)}t^{\alpha^{(FR)}-1},\\
h_0^{FD}(t) &= \alpha^{(FD)}\lambda^{(FD)}t^{\alpha^{(FD)}-1},\\
h_0^{RD}(t-t_{FR} \mid T_{FR}=t_{FR}) &= \alpha^{(RD)}\lambda^{(RD)}(t-t_{FR})^{\alpha^{(RD)}-1}.
\end{aligned}
\tag{2.5}
$$

En nuestro análisis, utilizamos un enfoque bayesiano para estimar los parámetros del modelo de enfermedad-muerte, siguiendo los pasos habituales del proceso inferencial bayesiano: especificación de una distribución previa para los parámetros del modelo, construcción de la función de verosimilitud y cálculo de la distribución *a posteriori* mediante el teorema de Bayes.

En primer lugar, para las distribuciones *a priori* se ha asumido un escenario previo de independencia y no informativo. Para ello se seleccionaron distribuciones previas normales amplias para los coeficientes de regresión, $\boldsymbol{\beta}$, y distribuciones Gamma para

los parámetros de forma, α, y escala, λ, de la función de riesgo basal Weibull [26]. De esta forma, si representamos mediante $\boldsymbol{\theta}$ al vector que incluye todos estos parámetros e hiperparámetros, lo que hemos hecho es definir una distribución *a priori* $\pi(\boldsymbol{\theta})$.

Las distribuciones posteriores se estimaron empleando métodos MCMC, de forma que mediante la muestra simulada con estos métodos se obtiene una aproximación de esta distribución *a posteriori* de los parámetros, $\pi(\boldsymbol{\theta} \mid \mathcal{D})$. Los cálculos se realizaron utilizando `JAGS` [27] y `R` [28]. Finalmente, a partir de esta información posterior podemos aproximar también la distribución de cualquier resultado de interés que sea función de los parámetros del modelo. Por ejemplo, en el modelo de enfermedad-muerte, hablamos de la distribución *a posteriori* de las incidencias acumuladas, $\pi(F_{1j}(t) \mid \mathcal{D})$, o de las probabilidades de transición, $\pi(p_{ij}(s,t) \mid \mathcal{D})$. Dado que esas distribuciones contienen información completa, en términos probabilísticos, sobre esas medidas de interés, se pueden calcular sus medias, medianas, cuantiles e intervalos de credibilidad *a posteriori*.

2.3.4. Estimación de riesgos e incidencias

La distribución *a posteriori* aproximada de los parámetros del modelo enfermedad-muerte se ha resumido en la Tabla 2.1 a través de su media *a posteriori* y su desviación típica. También se han analizado los datos mediante un modelo de riesgos competitivos, sin observarse diferencias entre los resultados de ambos modelos para las dos transiciones comunes, de fractura a refractura y de fractura a muerte. Esto es razonable debido a la separabilidad de las transiciones en la verosimilitud bajo las hipótesis de independencia condicional, como se indicaba en la sección 1.3.2. Sin embargo, como es lógico, el modelo enfermedad-muerte proporciona información adicional ya que incluye la transición de la refractura a la muerte.

Tabla 2.1: Resumen de las distribuciones *a posteriori* aproximadas de los parámetros del modelo de enfermedad-muerte.

Transición	Parámetro	Media	SD
De F a R	$\alpha^{(FR)}$	0.9198	0.0157
	$\lambda^{(FR)}$	0.0279	0.0013
	$\beta_{Wo}^{(FR)}$	0.0262	0.0486
	$\beta_{Age}^{(FR)}$	0.0244	0.0030
De F a D	$\alpha^{(FD)}$	0.7759	0.0051
	$\lambda^{(FD)}$	0.3311	0.0050
	$\beta_{Wo}^{(FD)}$	-0.5092	0.0166
	$\beta_{Age}^{(FD)}$	0.0705	0.0012
De R a D	$\alpha^{(RD)}$	0.6234	0.0154
	$\lambda^{(RD)}$	0.5769	0.0329
	$\beta_{Wo}^{(RD)}$	-0.6127	0.0655
	$\beta_{Age}^{(RD)}$	0.0498	0.0046

Tabla 2.2: Hazard ratios (HR) *a posteriori* de las covariables sexo y edad para cada transición, estimados con el modelo de enfermedad-muerte. Medias e intervalos de credibilidad 0.95 *a posteriori*.

Transición	Covariable	Media	IC 0.95
De F a R	Sexo	1.03	(0.94, 1.12)
	Edad	1.02	(1.02, 1.03)
De F a D	Sexo	0.60	(0.58, 0.62)
	Edad	1.07	(1.07, 1.08)
De R a D	Sexo	0.54	(0.48, 0.61)
	Edad	1.05	(1.04, 1.06)

Se observa que la media *a posteriori* de los coeficientes asociados a la edad es positiva para las tres transiciones, lo que indica un incremento en el riesgo correspondiente a cada evento con la edad, siendo mayor ese efecto para el tiempo desde la fractura hasta la muerte. Así pues, se obtienen hazard ratios que no solo son mayores que 1 de media, indicando el mencionado incremento, sino que se puede afirmar con probabilidad 0.95 que ese será el sentido del efecto 2.2. Por otro lado, no existen diferencias entre mujeres y hombres por lo que respecta al riesgo de refractura, lo que se deduce del intervalo de credibilidad del HR de (0.94, 1.12). Sin embargo, ser mujer se asocia a un menor riesgo de muerte, tanto tras la fractura inicial como tras sufrir una refractura.

La función de incidencia acumulada de refractura (muerte) evalúa la probabilidad de que una persona se refracture (muera) antes de un instante de tiempo determinado. Ambas probabilidades dependen de los parámetros del modelo y, en consecuencia, su distribución *a posteriori* se deriva de la distribución *a posteriori* calculada previamente para los parámetros.

En general, la probabilidad de refractura aumenta con la edad (Tabla 2.3), siendo las mujeres las que tienen mayor incidencia acumulada (incidencias anuales del 1.96 % a los 70 años y del 2.80 % a los 90 años para las mujeres, mientras que para los hombres son del 1.86 % y 2.45 %, respectivamente). En cuanto a la muerte sin refractura, la incidencia aumenta con la edad y se ha estimado mayor para los hombres (incidencias al año del 7.36 % a los 70 años y del 26.73 % a los 90 años entre las mujeres, frente a un 11.94 % y un 40.34 % para los hombres de 70 y 90 años, respectivamente). Algo similar ocurre con la muerte tras la refractura, aunque con mayores incidencias que en los individuos únicamente con la fractura inicial. Cabe destacar que si bien se aprecian diferencias entre hombres y mujeres en la incidencia acumulada de refractura, según los resultados de nuestro modelo, estas diferencias se justificarían debido a una mayor mortalidad de los hombres, y no a una diferencia en el riesgo de refractura, como se ha mencionado anteriormente al observar los HR.

Tabla 2.3: Incidencias acumuladas al año de refractura, muerte sin refractura y muerte tras refractura.

Transición	Sexo	Edad	Media	IC 0.95
De F a R	Mujeres	70	1.96	(1.76, 2.16)
		80	2.39	(2.24, 2.55)
		90	2.80	(2.60, 3.00)
	Hombres	70	1.86	(1.65, 2.09)
		80	2.21	(2.01, 2.42)
		90	2.45	(2.21, 2.72)
De F a D	Mujeres	70	7.36	(7.05, 7.67)
		80	14.30	(13.95, 14.63)
		90	26.72	(26.17, 27.26)
	Hombres	70	11.95	(11.43, 12.46)
		80	22.63	(22.02, 23.29)
		90	40.35	(39.42, 41.36)
De R a D	Mujeres	70	14.81	(12.81, 16.95)
		80	23.15	(21.51, 24.83)
		90	35.16	(32.79, 37.54)
	Hombres	70	25.50	(22.02, 29.43)
		80	38.49	(35.02, 42.07)
		90	55.04	(50.55, 59.51)

*La tabla contiene media e intervalo de credibilidad 0.95 *a posteriori* en porcentaje.
*La incidencia acumulada de muerte tras refractura equivale a la probabilidad de transición de refractura a muerte desde $s = 0$.

2.3.5. Probabilidades de transición

La primera de las probabilidades de transición estimadas y mostradas en la Figura 2.3 es la probabilidad de seguir libre de eventos, o la probabilidad de permanecer en el estado inicial (fractura) sin progresión de ningún tipo, p_{FF}, ya sea refractura o muerte. Las mujeres muestran menos eventos que los hombres; se estimaron probabilidades medias de permanecer libre de eventos después de 5 años del 51.69 % y 36.12 %, respectivamente, para pacientes de 80 años. Además, cuanto más mayores son los pacientes, más probabilidades hay de que progresen, ya que tanto el riesgo de muerte como el de refractura aumentan con la edad, como se deduce de la Tabla 2.2.

En segundo lugar, se estimó la probabilidad de transición de fractura a refractura (p_{FR} de la Figura 2.3), la cual debe ser interpretada como la probabilidad de encontrarse en el estado de refractura en un momento dado. Así pues, los pacientes incluidos en esta probabilidad no solo deben haber progresado al estado de refractura sino que deben haber sobrevivido hasta ese momento (en caso contrario se encontrarían en el estado de

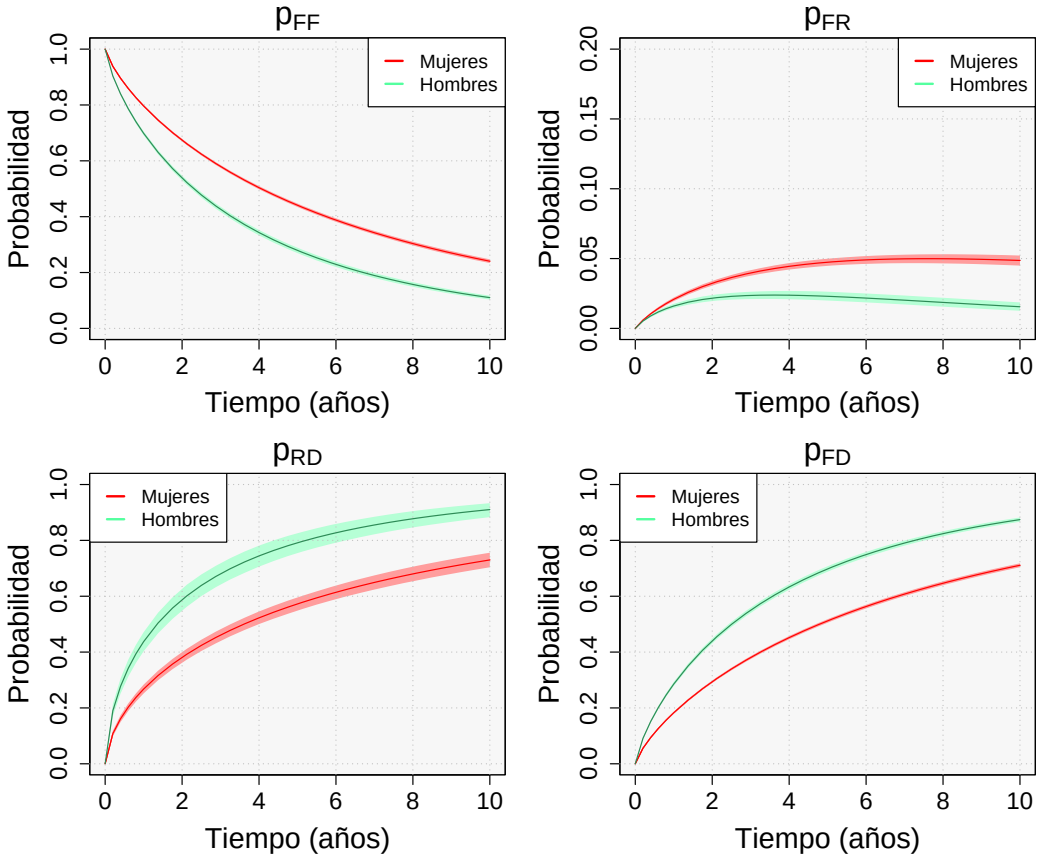

Figura 2.3: Media *a posteriori* e intervalos de credibilidad 0.95 de algunas probabilidades de transición relevantes: probabilidad de seguir libre de eventos (p_{FF}) desde el alta tras la primera fractura; probabilidad de transición del estado de fractura inicial al estado de refractura (p_{FR}) contando desde la primera; probabilidad de muerte tras refractura (p_{RD}) desde el alta tras la refractura; y probabilidad total de muerte (p_{FD}) contando desde la primera fractura. Seguimiento de 10 años, según sexo y para pacientes de 80 años.

muerte y no en el de refractura).

Las mujeres mostraron probabilidades de transición a refractura medias más altas que los hombres. Para los pacientes con 80 años en el momento del alta, se esperaría que, un año después de la primera fractura, el 2.02 % de las mujeres hubiesen sufrido una segunda fractura y se mantuviesen vivas, frente al 1.62 % de los hombres (5.16 % y 2.84 % después de 5 años, respectivamente). Esto se deriva de las tasas de mortalidad más altas que habría entre los hombres: un mayor riesgo competitivo de muerte resultaría en menos refracturas, mientras que mayores probabilidades de muerte después de la refractura disminuirían la permanencia de los hombres en el estado de refractura. Por la misma razón, a medida que aumenta la edad, la probabilidad de transición disminuye

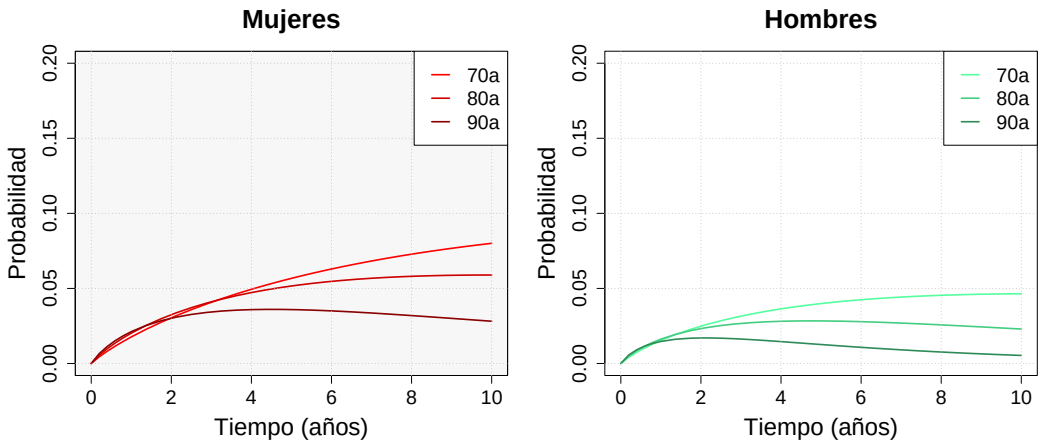

Figura 2.4: Media *a posteriori* de la probabilidad de transición del estado de fractura inicial al estado de refractura (p_{FR}) para un seguimiento de hasta 10 años tras el alta, según sexo y edad.

(Figura 2.4). Las tendencias temporales esperadas también difieren entre los grupos de edad. Como la mortalidad entre los pacientes de 70 años es bastante baja en comparación con las poblaciones de mayor edad, esperamos que la proporción de pacientes en el estado de refractura aumente con el tiempo. Por el contrario, la proporción de pacientes de 90 años con refractura y aún vivos disminuye después del segundo año para los hombres y después del quinto para las mujeres.

Finalmente, se estimaron la probabilidad de muerte tras la refractura, p_{RD}, y la probabilidad total de muerte, p_{FD}, que incluye muertes por ambas transiciones, sin refractura y después de la misma. En términos generales, la probabilidad de transición de muerte tras refractura muestra un patrón similar al de la muerte sin refractura, es decir, aumenta con la edad y es mayor para los hombres: probabilidades a un año del 14.81 % teniendo una edad de 70 años y del 35.16 % con 90 años, para las mujeres, mientras que para los hombres de 70 y 90 años se situarían en el 25.50 % y el 55.04 %, respectivamente. En cuanto a la probabilidad total de muerte, se estimó una mayor mortalidad para los hombres que para las mujeres, aumentando esta con la edad. Un año después de la fractura inicial, se estimaron probabilidades de muerte del 7.55 % y del 27.39 % para las mujeres de 70 y 90 años, respectivamente, mientras que para los hombres sería de un 12.27 % y un 41.34 %. Cabe destacar que esas probabilidades totales de muerte son estimables después de considerar la transición de refractura a muerte, como hace el modelo de enfermedad-muerte.

Probabilidades condicionadas

Es importante tener en cuenta que cada probabilidad de transición hasta ahora se ha calculado a partir de la ocurrencia del evento ($s = 0$). En otras palabras, estimamos la

probabilidad de muerte un año después de la refractura y la probabilidad total de muerte un año después de la fractura inicial. Sin embargo, el potencial de las probabilidades de transición es mucho mayor ya que su tiempo de inicio se puede definir en cualquier instante durante el periodo de estudio y, por lo tanto, no solo a partir del momento en que ocurre el evento. En particular, podemos estimar esas probabilidades de transición después de un cierto periodo de tiempo sin eventos. Por ejemplo, podemos considerar una situación libre de eventos un año después de ($s = 1$) del inicio del proceso. Por un lado, se podría estimar la probabilidad de muerte después de la refractura en t, considerando que no se observó ningún evento, es decir, que el individuo no murió, durante ese año siguiente a la refractura. Por otro lado estaría la probabilidad total de muerte en t, dado que no se observó ningún evento durante el año posterior a la fractura inicial. Se puede observar que sobrevivir un año tras el alta (tanto para fractura como para refractura) disminuye la probabilidad de una posterior muerte.

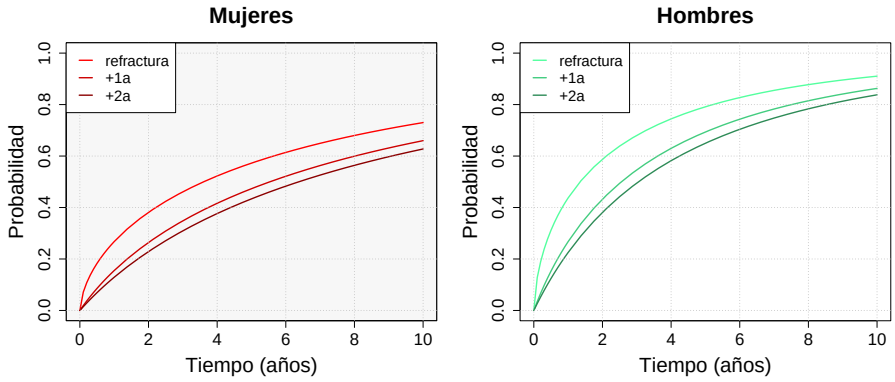

Figura 2.5: Media *a posteriori* de la probabilidad de muerte tras refractura (p_{RD}) con seguimiento de 10 años, contando desde diferentes instantes iniciales: alta tras refractura ($s = 1$), asumiendo que los pacientes sobreviven al menos 1 año tras la refractura ($s = 2$), y asumiendo una supervivencia mínima de 2 años ($s = 2$). Según sexo, para pacientes con una edad de 80 años.

Las probabilidades de muerte tras refractura a un año para aquellos pacientes que no tuvieron ningún evento durante el año posterior a la refractura (es decir, después de un total de 2 años desde el alta tras la fractura recurrente) se estimaron en el 8.30 % a los 70 y el 20.88 % a los 90 años para las mujeres, en comparación al 14.78 % a los 70 y el 35.10 % a los 90 años para los hombres.

Respecto a la mortalidad total, estimamos probabilidades totales de muerte a un año para aquellos pacientes libres de eventos durante el año posterior a la primera fractura (es decir, después de un total de 2 años tras la primera fractura) del 5.47 % a los 70 y del 20.50 % a los 90 años para las mujeres, siendo del 8.95 % a los 70 y del 31.77 % a los 90 años para los hombres.

Con fines ilustrativos, también representamos la variación en la probabilidad de muerte después de la refractura para aquellos pacientes que sobreviven 1 año y 2 años desde

que fueron dados de alta contando desde la refractura (Figura 2.5). En esta figura se muestra que, cuanto más tiempo sobreviven los pacientes después de la refractura, menos probabilidades tienen de fallecer. Además, las diferencias entre esas probabilidades de muerte disminuyen según aumenta el tiempo conocido de supervivencia de los pacientes.

Riesgos competitivos vs enfermedad-muerte

Nuestro modelo de enfermedad-muerte proporciona información relevante sobre la transición de refractura a muerte, la cual el modelo de riesgos competitivos no puede abordar. La Figura 2.6 muestra la media *a posteriori* de la incidencia acumulada de refractura y la media *a posteriori* de la probabilidad de transición del estado inicial de fractura al estado de refractura, $p_{FR}(0,t)$. Ambos resultados se definen en relación a un tiempo específico t. La curva superior representa la acumulación total de refracturas (incidencia acumulada) ocurridas antes del tiempo t, mientras que la curva inferior se refiere a la probabilidad de que un paciente se encuentre en el estado de refractura en el momento t y, por tanto, refracturado y vivo. Así pues, debe entenderse que en el momento t el área sombreada oscura muestra los pacientes que se han refracturado y siguen vivos, al contrario del área sombreada clara que muestra los pacientes refracturados que han muerto.

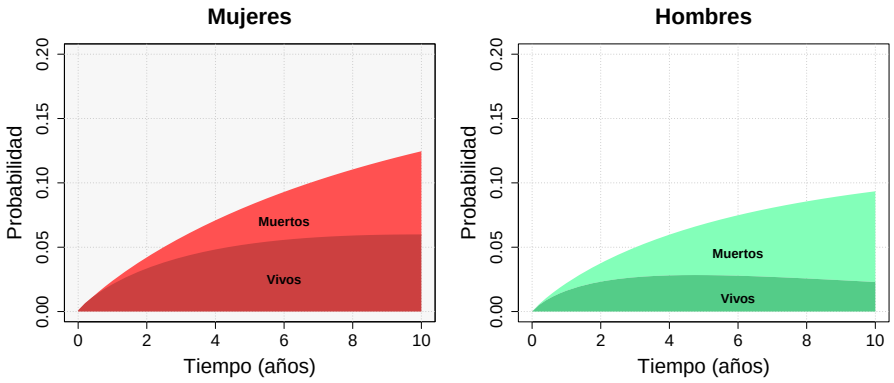

Figura 2.6: Media *a posteriori* de la incidencia acumulada de refractura (curva superior) y de la probabilidad de transición de fractura a refractura (curva inferior), según sexo y para pacientes de 80 años.

2.4. Discusión

En este capítulo se ha mostrado cómo los modelos multiestado, y en particular los modelos de enfermedad-muerte, pueden ser especialmente útiles cuando se trata de analizar escenarios de supervivencia en los que, además de la epidemiología relativa a la enfermedad, interviene la muerte como un evento que compite con la misma. Esto se ha ilustrado con datos de un estudio de una cohorte real de pacientes con fractura de

cadera, para el cual el modelo de enfermedad-muerte ha proporcionado no sólo la misma información que aportaría un modelo de riesgos competitivos, sino también información sobre la transición de la refractura, enfermedad en general, a la muerte.

Las probabilidades de transición que se obtienen a partir de los modelos multiestado se plantean como resultados naturales, informativos y dinámicos. Así pues, proporcionan evidencia clara sobre la progresión y las trayectorias que siguen los pacientes desde el mismo momento en que ingresan al estudio. Estas probabilidades permiten incrementar el conocimiento que se tiene sobre una determinada enfermedad y la mortalidad de los individuos que la padecen, por lo que podrían emplearse para encontrar áreas de mejora en el manejo de los pacientes y mejorar así su tratamiento.

Por otro lado, se debe hacer una consideración importante acerca de la estimación de la incidencia de muerte en la población. Aunque la incidencia acumulada de refractura puede estimarse igualmente a partir de modelos de enfermedad-muerte y de riesgos competitivos, ese no es el caso con la incidencia de muerte. El modelo de riesgos competitivos, definido como en el apartado 1.3.2, solo tiene en cuenta la muerte en pacientes sin refractura, ya que en este modelo los eventos de refractura y muerte se censuran mutuamente. Sin embargo, al incluir la muerte tras la refractura como se ha hecho mediante el modelo enfermedad-muerte, la probabilidad total de muerte se convierte en un resultado natural definido como una probabilidad de transición.

En este estudio hemos proporcionado estimaciones de algunas probabilidades de transición que se han seleccionado según su importancia clínica pero también con el fin de ilustrar su potencial. Sin embargo, se han excluido muchas otras posibilidades, ya que en realidad podríamos estimar infinidad de probabilidades de transición solo con modificar los intervalos de tiempo o condicionando a diferentes tiempos sin evento. Y es que estos modelos permiten formular y responder una amplia gama de preguntas científicas, proporcionando una mayor comprensión de problemas del mundo real, de una manera muy flexible. Por ejemplo, con relativa facilidad se podría incluir otro estado para una segunda refractura, o transiciones reversibles, en caso de que tuvieran sentido desde el punto de vista epidemiológico.

Finalmente, es importante mencionar que el enfoque bayesiano facilita todo el proceso inferencial, permitiendo estimar distribuciones *a posteriori* de cada medida que sea función de los parámetros del modelo, como *hazard ratios*, incidencias acumuladas y probabilidades de transición. Con este enfoque fuertemente probabilístico se consigue una inferencia estadística más natural e intuitiva.

Todas estas ventajas, siempre bajo los supuestos y consideraciones anteriormente mencionados, hacen del modelo de enfermedad-muerte, y en particular desde una perspectiva bayesiana, una herramienta/método relevante en aquellos escenarios con más de un evento, o en los que la muerte debe incluirse necesariamente (por ejemplo, en poblaciones de edad avanzada).

Capítulo 3

Modelos multiestado con efectos espaciales

3.1. Introducción

El contenido de este capítulo se basa en nuestro artículo (Llopis-Cardona et al. 2023), publicado en *Statistical Methods in Medical Research* [29], en el cual se presenta un marco conjunto multiestado y espacial. Esta extensión espacial de los modelos de supervivencia mostrados en el capítulo anterior es posible ya que, al ser modelos de regresión, pueden incluir no sólo covariables sino también efectos latentes que expliquen aquella heterogeneidad no explicada por las características propias de la población objetivo. Concretamente, la existencia de diferencias espaciales entre regiones es especialmente común en los estudios epidemiológicos. En muchos casos, serían factores de riesgo no controlados los que podrían llevar a que los individuos de algunas regiones presentasen mayores o menores riesgos de enfermedad, creando esta heterogeneidad. Cabe mencionar que en nuestro trabajo trataremos con un número finito de subregiones que forman parte de una region mayor que constituye el espacio de estudio. El parecido entre regiones cercanas, vecinas, será lo que permita o no hablar de correlación espacial.

En este sentido, como ya se ha comentado en el Capítulo 1, en la literatura estadística existe una gran variedad de modelos para evaluar la correlación espacial. Los modelos autorregresivos condicionales (CAR) [13] y sus variantes, todas ellas basadas en una definición de correlación construida a partir de una relación de vecindad, se han utilizado ampliamente en el mapeo de enfermedades. En particular, el modelo propuesto por Besag, York y Mollié (1991) [16] se ha postulado como la principal opción durante las últimas décadas, principalmente para abordar datos de conteos asumiendo un proceso de Poisson. Leroux et al. (1999) [18] propusieron una especificación alternativa para la matriz de precisión de los efectos aleatorios que distingue mejor entre dependencia espacial y dispersión. Según este modelo, los efectos aleatorios se definen como una mixtura de un escenario independiente y otro espacialmente correlacionado. Por otro lado, algunos autores analizaron el comportamiento de los modelos espaciales dentro

del marco de la supervivencia como Banerjee, Wall y Carlin (2003) [40], comparando diferentes modelos: sin efectos aleatorios (normalmente denominados fragilidades en el contexto de supervivencia), modelos con fragilidad no espacial y modelos con fragilidad CAR.

En cuanto a los modelos de enfermedad-muerte, no sólo se puede modelar la correlación espacial, sino también la correlación entre los efectos sobre las tres transiciones, dando como resultado un modelo multivariante para efectos aleatorios. En este sentido, Carlin y Banerjee (2003) [41] propusieron un modelo CAR multivariante para tiempos de supervivencia espacialmente correlacionados. Sin embargo, y a pesar de su interés, existen pocos estudios que consideren componentes espaciales en el marco de los modelos de enfermedad-muerte. El trabajo de investigación más notable en esta dirección es Nathoo y Dean (2007) [42] en el que se proponen varias estructuras para los efectos aleatorios de las regiones, con especial atención a la comparación de diferentes funciones de riesgo basal, como distribuciones de Weibull, exponenciales por partes y B-splines cúbicos.

En nuestro artículo, así como en este capítulo, proponemos, en primer lugar, un marco metodológico bayesiano para abordar los efectos aleatorios correlacionados espacialmente dentro de un modelo de enfermedad-muerte. En particular, se utilizará una versión multivariante del modelo de Leroux para modelizar conjuntamente correlación espacial y correlación entre los tiempos de supervivencia de cada transición. Para ello se propone la utilización de la aproximación anidada integrada de Laplace (INLA) como método empleado para aproximar las distribuciones *a posteriori* fruto del proceso bayesiano de inferencia. A continuación, aplicaremos este modelo al mismo estudio del mundo real que se ha analizado en el capítulo anterior, ampliando los resultados obtenidos. En este nuevo análisis de la cohorte PREV2FO, además de las características individuales, se incluyen las Áreas de Salud a las que pertenecen los pacientes, precisamente para evaluar las diferencias geográficas relativas a los riesgos y probabilidades de transición. Por último, se analizará la sensibilidad de los resultados por lo que respecta al método con INLA y a las condiciones establecidas en las distribucions *a priori*.

3.2. Modelo enfermedad-muerte espacial bayesiano

En esta sección se propone un modelo espacial bayesiano de enfermedad-muerte para datos de áreas, constituidos por un conjunto finito de regiones cuya relación se define por la vecindad geográfica entre ellas. Este modelo de enfermedad-muerte incluye la modelización conjunta de los tres tiempos de supervivencia que intervienen en dicho modelo, así como una estructura espacial basada en el modelo de Leroux [18], conectando los procesos de supervivencia en cada una de las áreas que dividen el dominio espacial. Dada su formulación bayesiana, la notación que se empleará de aquí en adelante considera que todas las probabilidades y conceptos derivados son condicionales, ya que todos los parámetros e hiperparámetros de los que dependen tienen distribución de probabilidad.

3.2.1. Modelización

Sea $h_{ij}^{(k)}(t \mid \boldsymbol{\theta})$ la función de riesgo para el tiempo de supervivencia T_{ij}, asociado a la transición $i \to j \in C$, en el momento t, para un individuo de la región k, $k = 1, \ldots, K$ que expresamos mediante el modelo de Cox [4]

$$\begin{aligned} h_{ij}^{(k)}(t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= h_{ij,0}(t \mid \boldsymbol{\theta}) \exp\{\eta_{ij}^{(k)}\}, \\ \eta_{ij}^{(k)} &= \boldsymbol{x}'\boldsymbol{\beta}_{ij} + b_{ij}^{(k)}, \end{aligned} \tag{3.1}$$

donde $\boldsymbol{\theta}$ es el vector de parámetros e hiperparámetros del modelo, $\boldsymbol{\psi}$ el vector que incluye todos los efectos aleatorios, $h_{ij,0}(t \mid \boldsymbol{\theta})$ la función de riesgo basal, y η_{ij} un término de regresión definido a partir de ciertas covariables $\boldsymbol{x} = (x_1, \ldots, x_L)'$, un vector de coeficientes de regresión $\boldsymbol{\beta}_{ij} = (\beta_{ij,1}, \ldots, \beta_{ij,L})'$ y un efecto aleatorio, $b_{ij}^{(k)}$, asociado a la transición $i \to j$ en la región k. Para las funciones de riesgo basal se propone un enfoque completamente paramétrico a través de funciones de riesgo basal Weibull, $h_{ij,0}(t \mid \boldsymbol{\theta}) = \alpha_{ij}\lambda_{ij}t^{\alpha_{ij}-1}$.

Los efectos aleatorios para una región k genérica que aparecen en (3.1), $b_{ij}^{(k)}$, indican una dependencia entre las diferentes transiciones del modelo así como entre las distintas áreas vecinas a la región k. Sea $\boldsymbol{B}$ la matriz en la cual se incluyen todos estos efectos aleatorios:

$$\mathbf{B} = \begin{pmatrix} b_{12}^{(1)} & b_{13}^{(1)} & b_{23}^{(1)} \\ \vdots & \vdots & \vdots \\ b_{12}^{(k)} & b_{13}^{(k)} & b_{23}^{(k)} \\ \vdots & \vdots & \vdots \\ b_{12}^{(K)} & b_{13}^{(K)} & b_{23}^{(K)} \end{pmatrix}, \tag{3.2}$$

de manera que $\mathbf{B}(, i)$ es la iésima columna de la matriz $\mathbf{B}$, y $vec(\mathbf{B}) = [\mathbf{B}(, 1)', \mathbf{B}(, 2)', \mathbf{B}(, 3)']'$ es un vector columna de dimensión $(3K \times 1)$ el cual incluye cada una de las columnas de esta matriz $\mathbf{B}$. Para este vector de efectos aleatorios, $vec(\mathbf{B})$, asumimos un *Gaussian Markov random field* (GMRF) multivariante y condicional, con un vector de medias cuyos elementos son todos 0 y con una matriz de varianzas-covarianzas Σ

$$(vec(\mathbf{B}) \mid \boldsymbol{\Sigma}) \sim N_{3K}(0, \boldsymbol{\Sigma}). \tag{3.3}$$

Cabe mencionar que $vec(\mathbf{B})$ es precisamente el conjunto de efectos aleatorios que anteriormente y de forma genérica se ha representado como $\boldsymbol{\psi}$.

La estructura de $\boldsymbol{\Sigma}$ incluye tanto dependencia multivariante entre las transiciones del modelo de enfermedad-muerte para una misma área geográfica (entre columnas de $\boldsymbol{B}$, *between*) como dependencia espacial entre áreas dentro de cada transición (dentro de las columnas de $\boldsymbol{B}$, *within*) tomando la expresión

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{between} \otimes \boldsymbol{\Sigma}_{within}, \tag{3.4}$$

donde $\otimes$ se refiere al producto de Kronecker.

La matriz que contiene la variabilidad y dependencia espacial, $\mathbf{\Sigma}_{within}$, tiene una gama más amplia de opciones en cuanto a modelos: desde el escenario independiente más simple hasta los modelos autorregresivos condicionales (CAR) [13] y sus variantes (CAR intrínseco, CAR propio, Besag York & Mollié, modelo de Leroux). Durante las últimas décadas, la propuesta de Besag, York & Mollié (BYM) [16] ha sido ampliamente utilizada en la literatura epidemiológica para el mapeo de enfermedades. Parte de su popularidad reside en su interpretabilidad, ya que consiste en un efecto aleatorio que considera la correlación espacial entre regiones según una estructura de vecindad, y un efecto aleatorio no estructurado que toma en cuenta la heterogeneidad entre regiones. Sin embargo, con el modelo BYM sólo resulta identificable la suma de ambos tipos de efectos aleatorios, no siendo posible identificar ambos efectos aleatorios por separado [43]. El modelo de Leroux [18] evita este problema al considerar un único efecto aleatorio que se define en términos de una mixtura de elementos que indicarían o independencia o dependecia espacial, siendo posible evaluar la relevancia o intensidad de cada uno de ellos a partir de un parámetro γ de la siguiente manera

$$\mathbf{\Sigma}_{within} = \Big(\tau\left[(1-\gamma)\mathbf{I} + \gamma(\mathbf{D}-\mathbf{W})\right]\Big)^{-1}, \tag{3.5}$$

donde τ es un hiperparámetro de dispersión, $\mathbf{I}$ la matriz identidad, $\mathbf{D}$ una matriz diagonal cuyos elementos distintos de cero en la diagonal son el número de vecinos del área correspondiente, $\mathbf{W}$ una matriz de adyacencia, cuyos valores son $\mathbf{W}_{kl} = 1$ si las áreas k y l son vecinas, $k \neq l$, y 0 en caso contrario; y $\gamma \in [0, 1)$ es un hiperparámetro que puede entenderse como un parámetro de correlación espacial, el cual determina cómo se combinan las matrices $\mathbf{I}$ y $\mathbf{D} - \mathbf{W}$. Así pues, un valor de $\gamma = 0$ se simplifica a un modelo de efectos aleatorios independiente sin relación espacial entre áreas, mientras que $\gamma = 1$ se corresponde con un modelo CAR intrínseco.

Por otro lado, la matriz $\mathbf{\Sigma}_{between}$ se define como una matriz de varianzas-covarianzas que indica la correlación entre los tiempos de las tres transiciones

$$\mathbf{\Sigma}_{between} = \begin{pmatrix} \dfrac{1}{\tau_{12}} & \dfrac{\rho_{(12)(13)}}{\sqrt{\tau_{12}\tau_{13}}} & \dfrac{\rho_{(12)(23)}}{\sqrt{\tau_{12}\tau_{23}}} \\ \dfrac{\rho_{(12)(13)}}{\sqrt{\tau_{12}\tau_{13}}} & \dfrac{1}{\tau_{13}} & \dfrac{\rho_{(13)(23)}}{\sqrt{\tau_{13}\tau_{23}}} \\ \dfrac{\rho_{(12)(23)}}{\sqrt{\tau_{12}\tau_{23}}} & \dfrac{\rho_{(13)(23)}}{\sqrt{\tau_{13}\tau_{23}}} & \dfrac{1}{\tau_{23}} \end{pmatrix}. \tag{3.6}$$

En esta matriz se incluyen dos tipos de hiperparámetros: τ_{ij}, que es la precision marginal de los efectos aleatorios asociados a la transición $i \to j$, y $\rho_{(ij)(i'j')}$, que se define como la correlación entre los efectos aleatorios de dos transiciones $i \to j$ e $i' \to j'$.

Resulta importante notar que por razones de identificabilidad se fija un valor de $\tau = 1$ para el hiperparámetro de precisión del modelo de Leroux en 3.5, de forma que la matriz de covarianza $\mathbf{\Sigma}_{between}$ será la que capture la dispersión.

3.2.2. Inferencia bayesiana e información previa

Se ha considerado un enfoque bayesiano basado en la *integrated nested Laplace approximation* (INLA) [44] para estimar la distribución *a posteriori* de todas las medidas con incertidumbre del modelo. Como se ha mencionado en las secciones introductorias, la inferencia bayesiana combina el conocimiento previo de todos los parámetros e hiperparámetros del modelo desconocidos, $\boldsymbol{\theta}$, expresado en términos probabilísticos a través de la distribución previa, con la función de verosimilitud obtenida de los datos, $\mathcal{D}$, mediante el teorema de Bayes. Como resultado se deriva la distribución *a posteriori* conjunta $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ de los parámetros e hiperparámetros. A medida que los modelos se vuelven más complejos, resulta más difícil encontrar una expresión analítica para esas distribuciones posteriores, por lo que se requieren métodos computacionales para abordarlas. Los procedimientos más populares son los métodos de cadena de Markov Monte Carlo (MCMC) [45], que en la mayoría de casos implican grandes cantidades de tiempo de cálculo para asegurar la convergencia de las estimaciones. Alternativamente, INLA se postula como una opción rápida y precisa, la cual utiliza aproximaciones de Laplace para obtener la distribución marginal posterior aproximada de los parámetros, hiperparámetros y términos latentes del modelo. Los modelos de supervivencia, incluidos los modelos de riesgos proporcionales de Cox, pueden adaptarse e implementarse en INLA, ya que pueden expresarse en términos de modelos del tipo GMRF [46]. En particular, los modelos de riesgos competitivos [47], y los modelos de enfermedad-muerte como una extensión de ellos, pueden abordarse utilizando INLA. Además, permite la inclusión de efectos aleatorios gaussianos en el término de regresión del modelo de riesgos proporcionales de Cox y, como consecuencia, el modelo espacial de enfermedad-muerte propuesto resulta naturalmente accesible mediante INLA.

Como hemos mencionado, necesitamos completar el modelo bayesiano con una distribución previa para los parámetros e hiperparámetros del modelo, $\boldsymbol{\theta}$. Para ello se ha considerado un marco de independencia previa entre los diferentes elementos de $\boldsymbol{\theta}$. Respecto a los parámetros de forma α_{ij} de las funciones de riesgo basal en 3.1, se ha asumido que seguían una distribución *a priori* PC (*penalized complexity prior*) como se describe en la documentación de INLA (consulte `inla.doc("pc.alphaw")` para obtener una definición detallada). Estas distribuciones previas PC consideran el exponencial como modelo base, o lo que es lo mismo, un modelo de Weibull con $\alpha_{ij} = 1$, penalizando pues la desviación respecto a este modelo exponencial. Así pues, se preferiría el modelo de Weibull más general solo en caso de que exista suficiente evidencia que lo respalde [48]. Por otro lado, el parámetro de escala para las intensidades de transición basales, λ_{ij}, no tiene una distribución previa por sí mismo sino que se considera el intercepto $\beta_{ij,0} = log(\lambda_{ij})$, el cual sigue, de la misma forma que los coeficientes de regresión $\beta_{ij,l}$, una distribución gaussiana con media 0 y precisión 0.001.

Respecto a la matriz de covarianzas que incluye la correlación entre transiciones, $\boldsymbol{\Sigma}_{between}$, se supondrá que sigue una distribución Wishart inversa, o equivalentemente, que la matriz de precisión, $\boldsymbol{\Sigma}^{-1}_{between}$, sigue una distribución Wishart. A grandes rasgos, se puede considerar que la distribución Wishart es una generalización multivariada de

la distribución gamma [49]. Esta distribución resulta especialmente relevante cuando se modelizan efectos aleatorios normales correlacionados entre ellos, ya que es una distribución previa conjugada para la matriz de precisión cuando se trata con distribuciones normales multivariantes, siendo pues la opción más común al hacer inferencia sobre matrices de covarianza. En nuestro caso, los valores previos dados para los parámetros de la Wishart serán los proporcionados por defecto en la especificación de INLA para el modelo de efectos aleatorios correlacionados, es decir, $\mathbf{\Sigma}_{between}^{-1} \sim \text{Wishart}_3(\nu = 7, \mathbf{R} = \mathbf{I})$, donde $\nu = 7$ son los grados de libertad. El uso de la matriz identidad como matriz de escala de la distribución Wishart resulta habitual cuando se busca especificar una distribución *a priori* poco informativa. Esta consideración, sin embargo, se debe abordar con cierta relatividad. De hecho, dado que la elección de previas no informativas siempre es un tema relevante en inferencia bayesiana, es natural que varios autores hayan discutido su idoneidad y se hayan propuesto algunas alternativas. Por ejemplo, esta especificación previa podría no ser apropiada en presencia de parámetros con varianzas pequeñas, lo que resultaría en una distribución previa fuertemente informativa [50]. Alternativamente, existe la posibilidad de descomponer, mediante estrategias de separación, la matriz de covarianza en componentes relativos a varianza y correlación, siendo posible especificar distribuciones previas para cada una de las componentes de manera separada. Así, por ejemplo, se puede suponer que la matriz de correlación obtenida tras esta separación sigue una distribución Wishart inversa [51], o una Lewandowski-Kurowicka-Joe [52], entre otras. Respecto a las varianzas, existen muchas posibles previas positivas, tales como distribuciones normales truncadas, seminormales, semiCauchy o uniformes [49].

Por último, se asume un escenario previo no informativo uniforme, $\text{U}(0, 1)$, para el parámetro de mixtura γ del modelo de Leroux. Notar que para definir el modelo multivariante de Leroux para efectos aleatorios se ha utilizado el efecto latente `rgeneric`. Mediante este mecanismo, se pueden implementar efectos latentes en INLA vía `R` [53]. Cabe mencionar que, a pesar de que no se aplicasen específicamente a modelos de supervivencia, otros autores ya habían definido, empleando este método, efectos aleatorios con correlación espacial en su versión multivariante, como los efectos latentes CAR intrínsecos multivariantes, recopilados en el paquete `INLAMSM` para `R` [54].

3.2.3. Medidas *a posteriori*

La distribución *a posteriori*, $\pi(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$, contiene toda la información actualizada sobre el comportamiento aleatorio de la población definido por el modelo de enfermedad-muerte. Sin embargo, por sí misma, proporciona evidencia práctica poco clara sobre el estado clínico o el pronóstico de un paciente en el tiempo. Existen, sin embargo, medidas compuestas, como las distribuciones de los tiempos de estancia, las probabilidades de transición y ocupación, y las funciones de incidencia acumulada [55, 56], que permiten combinar la información relativa a la evolución temporal, contenida en los parámetros del modelo de enfermedad-muerte, junto con la variación del riesgo entre regiones. Estas medidas son especialmente relevantes para proveer conocimiento útil desde una perspectiva clínica o epidemiológica, mientras que desde un punto de vista estadístico bayesiano,

la inferencia *a posteriori* sobre esas medidas es muy sencilla, ya que estas se definen como funciones de los parámetros y efectos aleatorios $g(\boldsymbol{\theta}, \boldsymbol{\psi})$. El procedimiento para estimarlas es el siguiente: primero, tomamos muestras de la distribución *a posteriori* conjunta, obteniendo muestras de parámetros, hiperparámetros y efectos aleatorios; en segundo lugar, para cada uno de los valores de la muestra, $(\boldsymbol{\theta}, \boldsymbol{\psi})_i$, calculamos el valor de la medida de interés a partir de su expresión como función de estos elementos, $g(\boldsymbol{\theta}, \boldsymbol{\psi})_i$, de manera que se obtiene una muestra de la medida en cuestión y, en consecuencia, una aproximación de su distribución *a posteriori*; finalmente, podemos resumir esta distribución mediante la media muestral, la mediana o intervalos de credibilidad, siendo estos aproximaciones de la media *a posteriori*, la mediana *a posteriori* y los intervalos de credibilidad *a posteriori* de la medida en cuestión.

Cabe mencionar que, por defecto, INLA proporciona muestras obtenidas a partir de aproximaciones de las distribuciones *a posteriori* marginales y no de las conjuntas. Esto no resulta problemático en caso de querer analizar cada uno de los elementos (parámetros, efectos, etc.) por separado, pero no es apropiado para el cálculo de medidas que combinen varios de estos elementos. Esto es así ya que, al inferir sobre las medidas de interés empleando muestras de las distribuciones marginales, no se tienen en cuenta las posibles correlaciones entre los componentes del campo latente. Afortunadamente, INLA también permite obtener muestras de la distribución *a posteriori* conjunta [53, 57], siendo el método empleado en este trabajo. A continuación, presentaremos algunas de las medidas o resultados mencionados anteriormente y discutiremos su estimación *a posteriori*.

3.2.3.1. Tiempos de ocupación

El tiempo de ocupación del estado i para un individuo se refiere al tiempo que este individuo permanece ese mismo estado sin abandonarlo. Es posible que el tiempo de ocupación más importante dentro de los modelos de enfermedad-muerte sea el tiempo en el estado inicial. Este puede definirse en términos de una función de supervivencia condicional, para cada área k, como

$$\begin{aligned} S_1^{(k)}(t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= P(T^{(k)} > t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \\ &= \exp\Big(- \int_0^t (h_{12}^{(k)}(u \mid \boldsymbol{\theta}, \boldsymbol{\psi}) + h_{13}^{(k)}(u \mid \boldsymbol{\theta}, \boldsymbol{\psi}))\, \mathrm{d}u \Big) \end{aligned} \tag{3.7}$$

donde $T^{(k)} = \min\{T_{12}^{(k)}, T_{13}^{(k)}\}$. Dado que el tiempo de ocupación del estado 1 es una medida que depende de $(\boldsymbol{\theta}, \boldsymbol{\psi})$, ya que se expresa como función de estos, la distribución *a posteriori* consiguiente, $\pi(S_1^{(k)}(t \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \mid \mathcal{D})$, $\forall t$, podrá aproximarse fácilmente a partir de una muestra simulada de la distribución *a posteriori* $\pi(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$.

3.2.3.2. Probabilidades de transición y ocupación

Las probabilidades de transición son funciones que dependen de los parámetros y efectos aleatorios a través de las funciones de riesgo correspondientes. Así pues, su distribución *a posteriori*, $\pi(p_{ij}^{(k)}(s,t \mid \boldsymbol{\theta},\boldsymbol{\psi} \mid \mathcal{D})$, para individuos de la región k, también podrá obtenerse a partir de una muestra simulada de $\pi(\boldsymbol{\theta},\boldsymbol{\psi} \mid \mathcal{D})$. Por lo que respecta a las probabilidades de ocupación, estas se refieren a la probabilidad asociada a la permanencia en cada uno de los estados en un tiempo dado t. Lejos de ser un concepto distinto, pueden expresarse como probabilidades transición donde el estado inicial y el final son el mismo, $\pi(p_{ii}^{(k)}(0,t \mid \boldsymbol{\theta},\boldsymbol{\psi} \mid \mathcal{D})$, por lo que su distribución posterior puede obtenerse de manera análoga al resto de probabilidades de transición.

3.2.3.3. Funciones de incidencia acumulada

Si bien las funciones de incidencia acumulada son medidas que se usan con mayor frecuencia en estudios basados en riesgos competitivos, ámbito del que provienen, pueden ser muy útiles en los modelos de enfermedad-muerte. Estas funciones expresan cómo los individuos se acumulan en un determinado estado, a diferencia de las probabilidades de transición que són medidas más dinámicas. Es por eso que las incidencias acumuladas pueden ser especialmente relevantes cuando la enfermedad tiene interés por sí misma y no solo como un estado intermedio entre el inicio y la muerte. Se definen de manera equivalente al marco competitivo, a partir de los tiempos de supervivencia $T_{12}^{(k)}$ y $T_{13}^{(k)}$ como

$$F_{12}^{(k)}(t \mid \boldsymbol{\theta},\boldsymbol{\psi}) = P(T^{(k)} \leq t, \delta^{(k)} = 1 \mid \boldsymbol{\theta},\boldsymbol{\psi}), \tag{3.8}$$

$$F_{13}^{(k)}(t \mid \boldsymbol{\theta},\boldsymbol{\psi}) = P(T^{(k)} \leq t, \delta^{(k)} = 0 \mid \boldsymbol{\theta},\boldsymbol{\psi}), \tag{3.9}$$

donde $\delta^{(k)}$ es una variable indicadora con valor 1 si $T_{12}^{(k)} < T_{13}^{(k)}$ y 0 en caso contrario. Pueden interpretarse como la probabilidad en un instante t de haber pasado directamente del estado 1 a un estado j, $j = 2,3$, conservando un sentido de acumulación como su nombre sugiere. La incidencia acumulada asociada al estado de enfermedad, 2, resulta especialmente informativa ya que, en esencia, indica qué proporción de individuos se espera que desarrollen la enfermedad. Al mismo tiempo, puede compararse directamente con la probabilidad de transición de 1 a 2, la cual indica la tasa esperada de pacientes que sufren la enfermedad y se mantienen con vida. Las funciones de incidencia acumulada no dejan de ser funciones de $(\boldsymbol{\theta},\boldsymbol{\psi})$ siguiendo la expresión

$$F_{1j}^{(k)}(t \mid \boldsymbol{\theta},\boldsymbol{\psi}) = \int_0^t h_{1j}^{(k)}(s) \exp\Big\{ -\int_0^s (h_{12}^{(k)}(u) + h_{13}^{(k)}(u)\mathrm{d}u\Big\}\,\mathrm{d}s,\ j = 1,2. \tag{3.10}$$

Como consecuencia, su distribución *a posteriori*, $\pi(F_{1j}^{(k)}(t \mid \boldsymbol{\theta},\boldsymbol{\psi}) \mid \mathcal{D})$, también podrá aproximarse con una muestra simulada de la distribución *a posteriori* $\pi(\boldsymbol{\theta},\boldsymbol{\psi} \mid \mathcal{D})$.

3.3. Aplicación al estudio de fractura de cadera

Los modelos multiestado, el modelo de enfermedad-muerte entre ellos, tienen su principal campo de aplicación en estudios clínicos en los que se consideran enfermedades no terminales y su progresión, eventos repetidos o poblaciones en las que existe un riesgo competitivo de muerte considerable. En esta sección ilustraremos la aplicación del modelo de enfermedad-muerte espacial desde una perspectiva bayesiana a un estudio de fractura de cadera recurrente, basado en la cohorte PREV2FO. Este análisis supone una ampliación tanto técnica como en términos de resultados respecto a los análisis efectuados en el capítulo anterior.

3.3.1. Cohorte PREV2FO

Al igual que en el capítulo anterior, analizamos la cohorte PREV2FO, esta vez con información espacial. La Comunitat Valenciana está dividida en 24 Áreas de Salud, cada una de las cuales se corresponde con el área de influencia administrativa de un hospital público del SVS. La consideración en el análisis de estas Áreas de Salud será la mayor diferencia respecto a los análisis previos, para cuyo tratamiento se requerirá emplear técnicas propias del análisis espacial como las incluidas en el modelo de enfermedad-muerte espacial propuesto.

Los pacientes de la cohorte fueron seguidos después de la fractura índice hasta la muerte o fin de estudio (31 de diciembre de 2016), teniendo en cuenta las fracturas de cadera recurrentes durante el período de seguimiento. La Figura 2.1 del capítulo anterior resulta igualmente válida en esta sección, mostrando el proceso de supervivencia de este estudio como un modelo de enfermedad-muerte, el cual cuenta con un estado inicial de alta tras una primera fractura de cadera (F), un estado intermedio que representa el alta tras una refractura (R), y el estado de muerte (D). Los tiempos de transición entre estados fueron censurados por la derecha debido a fin de estudio o muerte.

Cabe mencionar que desde el punto de vista clínico existe la posibilidad de sufrir más de una refractura. Sin embargo, se ha descartado esta posibilidad en nuestro modelo porque solo le ocurría a un número reducido de pacientes de nuestro estudio. En concreto, de los 2532 pacientes con refractura, solo 26 de ellos presentaban una segunda refractura y solo uno con una tercera. Debido a la complejidad del modelo y la información limitada acerca de las refracturas posteriores, se decidió no incluir esta información en el modelo, por lo que la definición exacta de nuestro estado de refractura sería “tener al menos una refractura”, entrando en el estado con la primera de ellas. No obstante, el modelo planteado es fácilmente generalizable y, en caso de que se contase con un mayor número de pacientes con una segunda o tercera refractura, podríamos agregar los estados correspondientes y así representar esas transiciones adicionales.

Para definir un perfil básico del paciente se han considerado como covariables el sexo, la edad al alta y el Área de Salud en la que tuvo lugar la hospitalización. En el estudio participaron 34491 pacientes dados de alta vivos después de una fractura de cadera, de los cuales 25807 (74.8 %) eran mujeres y 8684 (25.2 %) hombres. La edad se incluyó

como predictor continuo y centrado en la media. La edad media en el momento de la primera fractura fue de 83.4 años (RIC: 79.0-88.3). Se siguió a los pacientes un tiempo en mediana de 5.0 años (RIC: 3.0-7.0 años). Por grupo de edad, el 12.4 % de los pacientes eran menores de 75 años, el 43.6 % entre 75 y 85 años, el 40.6 % entre 85 y 94 años de edad, y el 3.4 % tenía más de 95 años.

Mediante el modelo propuesto en la sección anterior, se modelizará cada uno de los tiempos de supervivencia, es decir, desde F a R, desde F a D, y desde R a D, para cada una de las Áreas de Salud k, $k = 1, \ldots, 24$, $T_{FR}^{(k)}$, $T_{FD}^{(k)}$ y $T_{RD}^{(k)}$.

3.3.2. Distribución *a posteriori*

A continuación presentaremos la distribución *a posteriori* aproximada de forma secuencial: primero los parámetros, luego los hiperparámetros y, finalmente, cada uno de los efectos aleatorios. En la Tabla 3.1 se resume la información que se obtiene al aproximar la distribución marginal *a posteriori* de cada uno de los parámetros del modelo. Las estimaciones de los parámetros de forma de las funciones de riesgo basal, α_{FD}, α_{FR} y α_{RD}, indican riesgos decrecientes con el tiempo, especialmente por lo que respecta a los riesgos de muerte sin y tras refractura. α_{FR} es el más cercano a 1, que es el valor umbral que define el comportamiento de las funciones de riesgo de Weibull, presentando un riesgo casi constante en el tiempo. No se observaron diferencias relevantes entre mujeres y hombres del riesgo de fractura recurrente de cadera ($\mathrm{E}(\beta_{FR,Mujer} \mid \mathcal{D}) = 0.021$), mientras que las mujeres sí mostraron menores riesgos de muerte en comparación con los hombres ($\mathrm{E}(\beta_{FD,Mujer} \mid \mathcal{D}) < 0$, $\mathrm{E}(\beta_{RD,Mujer} \mid \mathcal{D}) < 0$). La edad se identificó como un factor que incrementa el riesgo en todas las transiciones.

Tabla 3.1: Resumen de la distribución *a posteriori* aproximada de los parámetros del modelo de enfermedad-muerte con efectos aleatorios Leroux multivariante. Parámetros relativos a las transiciones: parámetros de forma y escala de las distribuciones Weibull y coeficientes de regresión.

Transición	Parámetro	Media	Mediana	SD	2.5 %	97.5 %
De F a R	α_{FR}	0.921	0.921	0.016	0.891	0.953
	λ_{FR}	0.028	0.027	0.005	0.018	0.040
	$\beta_{FR,\,Mujer}$	0.021	0.021	0.050	-0.076	0.119
	$\beta_{FR,\,Edad}$	0.024	0.024	0.003	0.018	0.030
De F a D	α_{FD}	0.776	0.776	0.005	0.766	0.786
	λ_{FD}	0.335	0.331	0.054	0.238	0.460
	$\beta_{FD,\,Mujer}$	-0.510	-0.510	0.017	-0.543	-0.477
	$\beta_{FD,\,Edad}$	0.070	0.070	0.001	0.068	0.073
De R a D	α_{RD}	0.628	0.628	0.016	0.597	0.659
	λ_{RD}	0.593	0.579	0.131	0.374	0.897
	$\beta_{RD,\,Mujer}$	-0.634	-0.634	0.065	-0.761	-0.505
	$\beta_{RD,\,Edad}$	0.049	0.049	0.005	0.040	0.059

La Figura 3.1 muestra la media *a posteriori* de la función de riesgo basal asociada a cada uno de los tres tiempos de transición. Estas funciones de riesgo basal son, de hecho, las funciones de riesgo calculadas para los valores de referencia de los predictores: hombres, de edad promedio, de un Área de Salud con efecto aleatorio igual a 0. En este grupo en particular, se puede observar cómo el riesgo estimado de muerte después de una fractura recurrente de cadera es mayor que el de muerte sin esta refractura. Por otro lado, la intensidad de la transición de fractura a refractura es notablemente menor que en las transiciones al estado de muerte. Estas comparaciones entre transiciones no presentarán grandes variaciones al considerar las covariables, dadas las estimaciones de los efectos de las mismas, por lo que se mantendrán en general las conclusiones en mujeres y para otras edades. Por último, se observa también que las funciones basales sugieren mayores riesgos durante el primer año, incluido el riesgo de refractura, aunque este último no se puede apreciar gráficamente. En cualquier caso, se podía esperar algo parecido debido a la naturaleza decreciente que se deducía de las estimaciones. Esto da como resultado un aumento inicial más pronunciado en la incidencia acumulada de esos eventos, así como mayores aumentos o disminuciones en las respectivas probabilidades de transición durante los primeros meses de seguimiento.

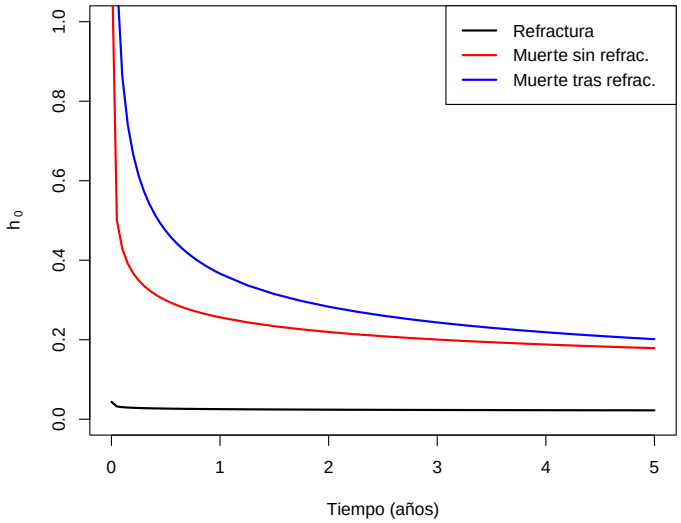

Figura 3.1: Media *a posteriori* de las funciones de riesgo basal para cada transición: de F a R, de F a D y de R a D. El eje horizontal indica el tiempo en años desde la fractura inicial para las transiciones $F \to R$ y $F \to D$, y el tiempo desde la refractura para la transición $R \to D$.

La Tabla 3.2 presenta un resumen de las distribuciones marginales *a posteriori* aproximadas de los hiperparámetros del modelo de enfermedad-muerte espacial, todos ellos relacionados con la variabilidad bien entre los tiempos de transición o bien dentro de las diferentes Áreas de Salud de la Comunitat Valenciana. La estimación del parámetro γ del modelo de Leroux de 0.841 indica que la mixtura entre un escenario independiente

Tabla 3.2: Resumen de la distribución *a posteriori* aproximada de los hiperparámetros del modelo de enfermedad-muerte con efectos aleatorios Leroux multivariante. Parámetro γ del modelo de Leroux, precisiones de los efectos aleatorios y correlaciones entre tiempos de transición.

Parámetro	Media	Mediana	SD	2.5 %	97.5 %
γ	0.841	0.862	0.101	0.591	0.973
τ_{FR}	14.257	13.581	4.620	7.197	25.185
τ_{FD}	19.896	19.228	5.595	10.915	32.737
τ_{RD}	11.743	11.137	4.181	5.386	21.625
$\rho_{(FR)(FD)}$	-0.044	-0.047	0.181	-0.388	0.315
$\rho_{(FR)(RD)}$	-0.076	-0.078	0.178	-0.415	0.275
$\rho_{(FD)(RD)}$	0.109	0.111	0.164	-0.217	0.423

y un modelo CAR intrínseco se inclina hacia el segundo (Figura 3.2). Este hecho se ve confirmado con la estimación de un intervalo de credibilidad del 95 % que excluye valores más bajos de este hiperparámetro y, por tanto, sugiere la existencia de una correlación espacial entre áreas relevante. Los parámetros de correlación entre transiciones muestran distribuciones *a posteriori* que no solo contienen al 0 sino que estan centradas en el mismo 0, indicando que los parámetros de correlación son poco relevantes; lo que se traduce en un escenario con transiciones no correlacionadas. Aún en este escenario no correlacionado, el valor más alto se estimó para la correlación entre muerte sin refractura y muerte después de refractura, $\rho_{(FD)(RD)}$, mostrando una ligera correlación entre ambos tipos de mortalidad. La incertidumbre acerca de los efectos aleatorios viene dada por los parámetros de precisión τ. Cabe destacar que mayores estimaciones de las precisiones indican una menor variabilidad entre los efectos aleatorios. Así, aunque la magnitud de las tres es muy similar, el orden de los efectos aleatorios de menor a mayor incertidumbre sería: muerte sin refractura, refractura y muerte tras refractura.

La Figura 3.3 muestra la media *a posteriori* de los efectos aleatorios asociados a cada tiempo de transición y Área de Salud de la Comunitat Valenciana. Las Áreas de Salud coloreadas en marrón indican un mayor riesgo de experimentar el evento de interés en comparación con la media global dada por los efectos fijos. Las áreas sombreadas en verde indican lo contrario. Los efectos aleatorios asociados a los tres tiempos de supervivencia de una misma Área de Salud no siempre se comportan de la misma forma. Podemos observar algunas áreas con efectos aleatorios positivos en los tres tiempos de supervivencia considerados, pero también algunos casos donde los efectos muestran comportamientos que apuntan a una relación inversa entre ellos. Concretamente, destacan algunas áreas con un patrón espacial particular. Es el caso de Requena-Utiel (el Área de Salud más occidental) y Dénia (situada en el cabo al este de la Comunitat Valenciana). La primera muestra un menor riesgo de fractura recurrente de cadera y mayores riesgos de muerte sin refractura y tras ella. La segunda, sin embargo, muestra el escenario opuesto, con un mayor riesgo de refractura y menor mortalidad. Ambos casos ilustran una asociación negativa entre el riesgo de refractura y mortalidad, mientras que se de-

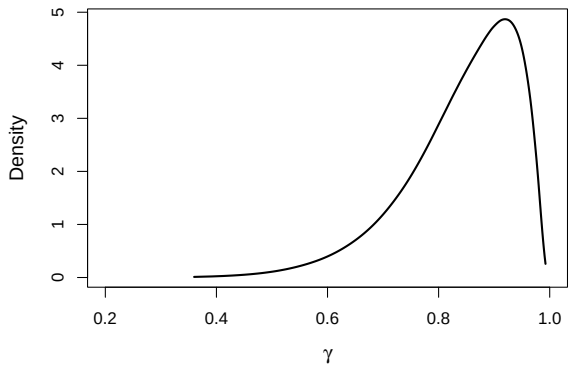

Figura 3.2: Distribución *a posteriori* aproximada del parámetro γ del modelo de enfermedad-muerte con efectos aleatorios Leroux multivariante.

duce una asociación positiva entre ambos riesgos de muerte. No obstante, cabe recordar que el modelo no aporta evidencia sobre la existencia de ningún tipo de correlación que permita generalizar lo observado para estas áreas.

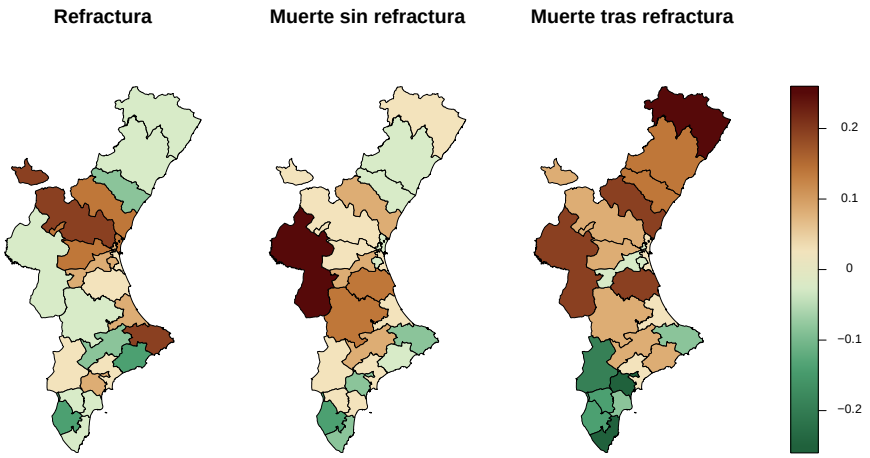

Figura 3.3: Media *a posteriori* de los efectos aleatorios específicos de cada Área de Salud de la Comunitat Valenciana del modelo de enfermedad-muerte con efectos aleatorios Leroux multivariante.

3.3.3. Medidas del proceso de fractura de cadera

El análisis de las estimaciones proporcionadas por la distribución *a posteriori* $\pi(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$ contiene la información completa por lo que respecta a las diferencias en el riesgo de cada evento de interés. Sin embargo, estas estimaciones aportan por sí mismas una evidencia poco clara sobre cuál será el pronóstico de un paciente que tuvo una fractura de cadera en una Área de Salud en particular o cuál sería la evolución esperada de los distintos tiempos de transición en la población objetivo. Para poder obtener este tipo de información se requiere combinar parámetros y efectos aleatorios en funciones más complejas, las cuales aporten tanto esta noción de evolución temporal, propia del marco de enfermedad-muerte, como la variación de riesgo entre áreas. Con este fin, se han obtenido las distribuciones *a posteriori* de funciones de incidencia acumulada y probabilidades de transición.

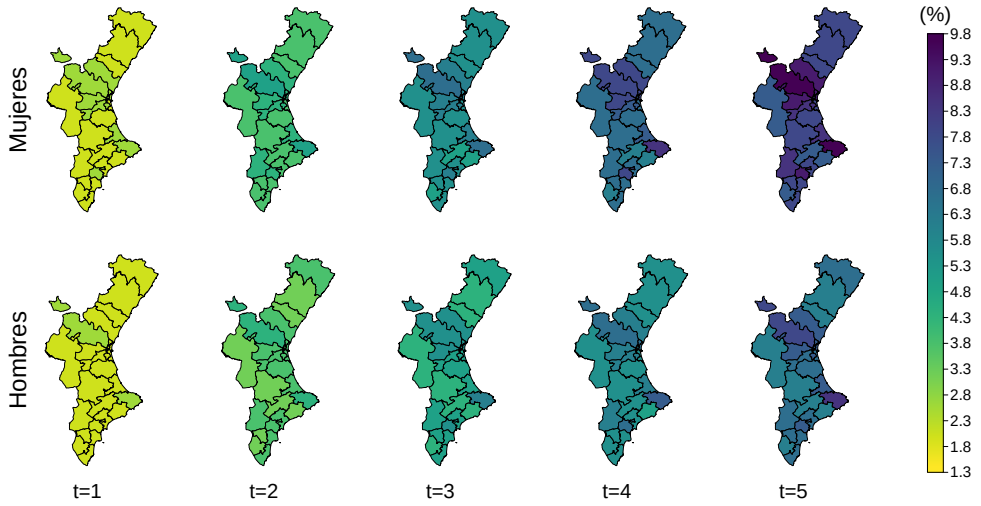

Figura 3.4: Media *a posteriori* de la incidencia acumulada de refractura en mujeres y hombres de 80 años, $t = 1, 2, \ldots, 5$ años después de la fractura índice, por Área de Salud.

La incidencia acumulada de refractura de cadera en el instante t puede interpretarse como la probabilidad de haber tenido una fractura recurrente de cadera antes de ese instante t. Esto solo sería posible para aquellos pacientes que hubiesen sobrevivido el tiempo suficiente, ya que en caso de fallecer se censuraría su observación. De hecho, por cómo se define esta medida, un mayor riesgo de muerte conduce a la observación de menos refracturas. Es por esta razón que dos Áreas con el mismo riesgo de refractura podrían presentar diferentes incidencias de refractura en función del riesgo de muerte. En la Figura 3.4 se muestra la media *a posteriori* de la incidencia acumulada de refractura para mujeres y hombres de 80 años, en las diferentes Áreas de Salud de la Comunitat Valenciana, $t = 1, 2, \ldots, 5$ años después de una primera fractura de cadera. En términos generales, se estima una mayor incidencia de fractura de cadera recurrente

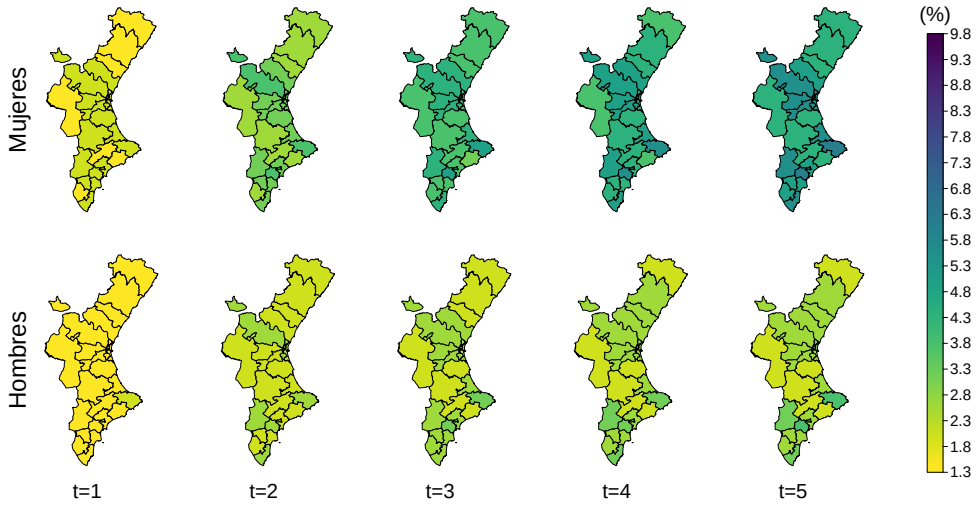

Figura 3.5: Media *a posteriori* de la probabilidad de transición de fractura a refractura (p_{FR}) en mujeres y hombres de 80 años, $t = 1, 2, \ldots, 5$ años después de la fractura índice, por Área de Salud.

en aquellas regiones con mayor riesgo de refractura, como se podía esperar. Esas diferencias se vuelven más visibles después de algunos años desde la fractura inicial. El Área de Salud de Requena-Utiel (la región más occidental) muestra una incidencia especialmente baja a pesar de su riesgo no tan bajo, lo que puede relacionarse con ser la región con mayor riesgo de muerte sin refractura. Los hombres muestran una menor incidencia de refractura debido a su mayor riesgo de muerte, ya que no encontramos diferencias en el riesgo de refractura en comparación con las mujeres. En general, los hombres alcanzan los mismos valores de incidencia que las mujeres con un retraso de entre 1 o 2 años, aproximadamente.

Respecto a las probabilidades de transición de fractura a refractura (Figura 3.5), también son mayores para las mujeres, ya que tienen más probabilidades de sufrir una refractura que los hombres, o equivalentemente, mayores incidencias acumuladas de refractura. Muestra una tendencia creciente durante los 2 y los 4 años posteriores a la fractura inicial en hombres y mujeres, respectivamente. Pasado este tiempo, esta probabilidad de refracturarse y seguir vivo se mantiene estable en el tiempo, ya que el número de pacientes con riesgo de refractura disminuye y la mortalidad tras la refractura compensa el total de nuevas refracturas.

La mortalidad es mayor en los hombres, tanto para la probabilidad total de muerte, como para únicamente la muerte tras la refractura. Las mujeres alcanzan a los 4 años, aproximadamente, la misma mortalidad que muestran los hombres a los 2 años (Figura 3.6). Esta diferencia es aún mayor en el caso de la muerte tras refractura, alcanzando las mujeres, a los 5 años desde la refractura, las mismas tasas de mortalidad que los

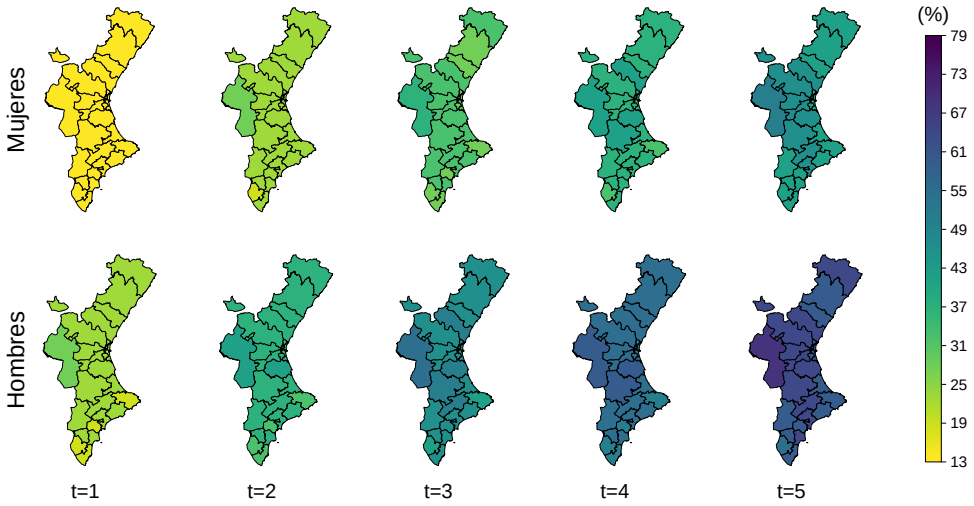

Figura 3.6: Media *a posteriori* de la probabilidad total de muerte (p_{FD}) en mujeres y hombres de 80 años, $t = 1, 2, \ldots, 5$ años después de la fractura índice, por Área de Salud.

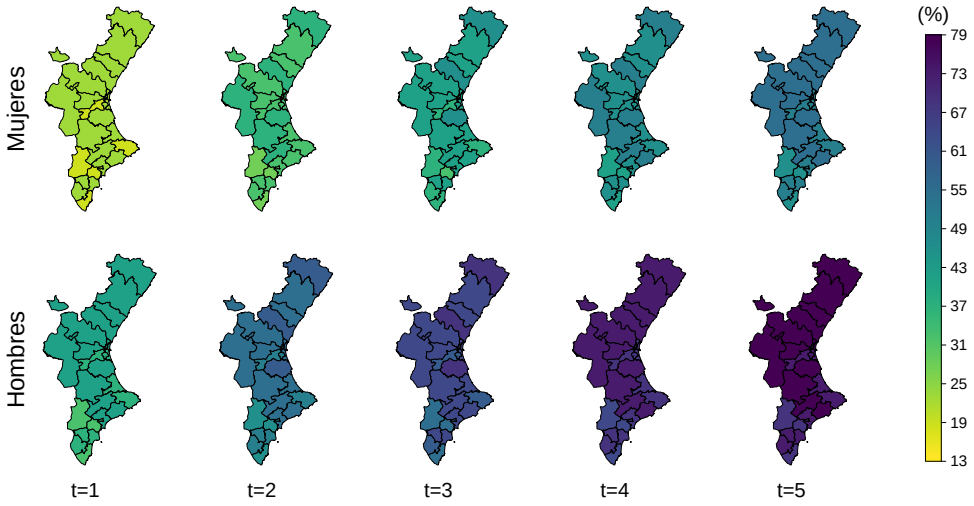

Figura 3.7: Media *a posteriori* de la probabilidad de muerte tras refractura (p_{RD}) en mujeres y hombres de 80 años, $t = 1, 2, \ldots, 5$ años después de la refractura, por Área de Salud.

hombres presentan solo 2 años después de la misma (Figura 3.7).

El número de pacientes que mueren después de una refractura representa una fracción pequeña de la mortalidad total. En particular, la incidencia acumulada de refractura

indica que menos del 10 % de las mujeres experimentan una refractura 5 años después de la fractura inicial (incluso menor en los hombres). Como consecuencia, el patrón espacial de la probabilidad total de muerte (Figura 3.6) es muy similar al mostrado por los efectos aleatorios en el riesgo de muerte sin refractura (Figura 3.3).

Finalmente, la probabilidad de muerte es mayor para aquellos pacientes con refractura. La mortalidad 1 año después de la refractura es similar a la esperada 2 años después de la fractura índice. Su patrón espacial también es diferente con respecto al mostrado por la mortalidad total, y es idéntico al mostrado por los respectivos efectos aleatorios sobre la transición de la refractura a la muerte en la Figura 3.3. Esto se debe a que la probabilidad de muerte después de la refractura es la única que depende exclusivamente de una intensidad de transición, en particular, la intensidad de la transición de la refractura a la muerte.

3.4. Estabilidad de la inferencia

La estabilidad de la inferencia *a posteriori* realizada, por lo que respecta a posibles cambios en la especificación de los diferentes elementos del modelo bayesiano, es una cuestión muy importante en un análisis bayesiano [58]. En esta sección se pretende analizar, de forma breve y no exhaustiva, la sensibilidad de los resultados obtenidos mediante el modelo de enfermedad-muerte espacial propuesto. Para ello se abordará el tema desde dos perspectivas. En primer lugar, se efectuará una comparación entre los métodos INLA y MCMC. En segundo lugar, se observará cómo afecta a los resultados la variación de algunos valores cuya especificación tuvo lugar previamente a la aplicación del modelo.

3.4.1. INLA vs JAGS

Para la comparación entre los métodos MCMC, JAGS específicamente, e INLA se ha empleado un modelo bayesiano de enfermedad-muerte dotado de una menor complejidad, así como un espacio geográfico que se ha reducido a un grupo de cinco Áreas de Salud vecinas entre ellas. El modelo, al igual que el propuesto en secciones anteriores, incluye efectos aleatorios gaussianos con correlación entre transiciones, aunque no considera la existencia de correlación espacial entre las áreas. Como tal, la parte del modelo de enfermedad-muerte permanece sin cambios, con intensidades de transición Weibull. Son los efectos aleatorios los que se verán modificados para reflejar este escenario de independencia espacial, siguiendo una distribución:

$$(b_k^{FR}, b_k^{FD}, b_k^{RD})^T \sim N(0, \mathbf{\Sigma}_{between}), \qquad k = 1, \cdots, 5 \tag{3.11}$$

donde b_k^{ij} es el efecto aleatorio correspondiente al Área de Salud k sobre la transición $i \to j$, y $\mathbf{\Sigma}_{between}$ es la matriz de correlación entre transiciones, la cual se define como:

$$\boldsymbol{\Sigma}_{between} = \begin{pmatrix} \frac{1}{\tau_{FR}} & \frac{\rho_{(FR)(FD)}}{\sqrt{\tau_{FR}\tau_{FD}}} & \frac{\rho_{(FR)(RD)}}{\sqrt{\tau_{FR}\tau_{RD}}} \\ \frac{\rho_{(FR)(FD)}}{\sqrt{\tau_{FR}\tau_{FD}}} & \frac{1}{\tau_{FD}} & \frac{\rho_{(FD)(RD)}}{\sqrt{\tau_{FD}\tau_{RD}}} \\ \frac{\rho_{(FR)(RD)}}{\sqrt{\tau_{FR}\tau_{RD}}} & \frac{\rho_{(FD)(RD)}}{\sqrt{\tau_{FD}\tau_{RD}}} & \frac{1}{\tau_{RD}} \end{pmatrix}. \tag{3.12}$$

Se ha empleado distribuciones *a priori*no informativas. En concreto, previas Gamma(0.01,0.01) para los parámetros de forma α y distribuciones normales, N(0,0.001) (escrita en términos de precisión), para tanto interceptos como efectos fijos de las covariables. Para la matriz de precision, $\boldsymbol{\Sigma}^{-1}_{between}$, se asumió una distribución *a priori*Wishart con $\nu = 7$ grados de libertad y con la matriz identidad como matriz de escala.

Además, se ha aplicado un post-barrido a los efectos aleatorios obtenidos con JAGS, lo que permite obtener interceptos y efectos aleatorios identificables [59, 60].

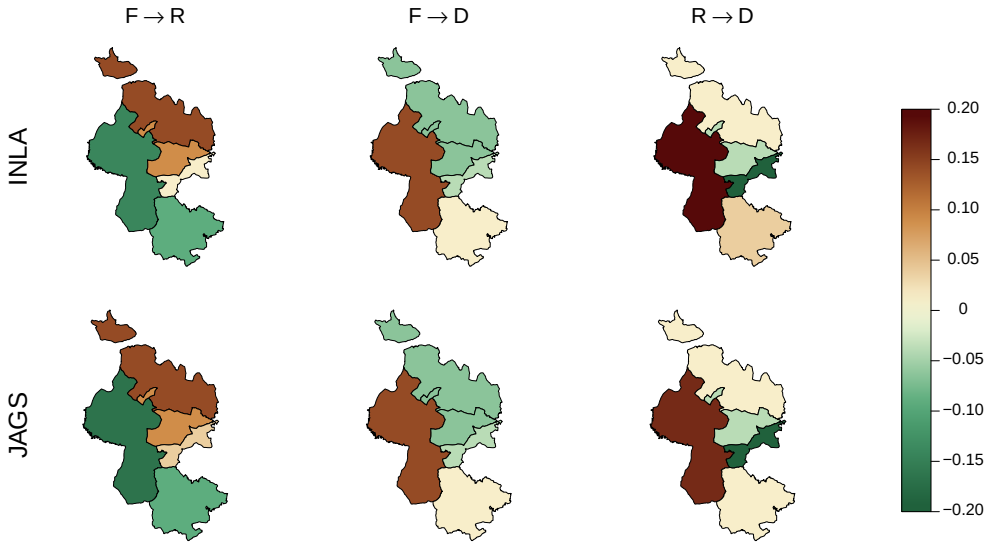

Figura 3.8: Media *a posteriori* de los efectos aleatorios de 5 Áreas de Salud de la Comunitat Valenciana, empleando un modelo de enfermedad-muerte con efectos aleatorios gaussianos que considera correlación entre transiciones, utilizando INLA y JAGS (MCMC).

Las estimaciones *a posteriori* de los parámetros y los efectos aleatorios obtenidas mediante INLA y JAGS fueron similares (Figuras 3.8 y 3.9) y, dado que en ambos procedimientos se toman muestras de la distribución *a posteriori* conjunta, pueden esperarse también resultados similares entre ambos métodos para las incidencias acumuladas, probabilidades de transición o en general, medidas de interés *a posteriori*. En resumen,

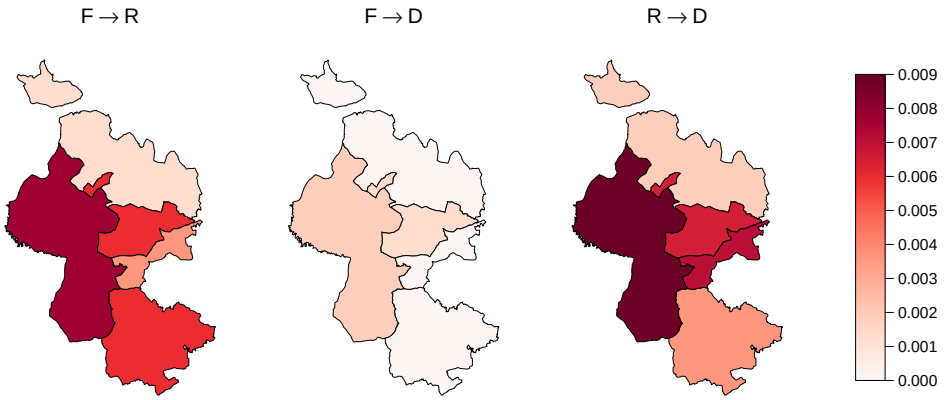

Figura 3.9: Diferencias absolutas en la media *a posteriori* estimada de los efectos aleatorios, utilizando INLA y JAGS (MCMC).

encontramos que INLA es un método razonable para la inferencia bayesiana con este tipo de modelos, lo que permite, entre otras cosas, beneficiarse de la mayor velocidad de estimación de INLA.

JAGS: diagnóstico

Para asegurar una buena mixtura y convergencia de las cadenas de Markov, buscando una buena aproximación de la distribución *a posteriori* objetivo mediante las muestras simuladas, se ha empleado tres cadenas de 5000 iteraciones (cada una con un periodo de adaptación de 1000 y un *burn-in* de 500 iteraciones). En concreto, la convergencia de este método se evaluó gráficamente y mediante el estimador $\hat{R}$ de Gelman y Rubin. Para ello se han obtenido estimaciones puntuales y los límites superiores del intervalo de confianza de dicho estimador, para cada parámetro y efecto latente del modelo, resumidos en la Tabla 3.3. Se puede observar que incluso en el peor de los casos se esperan valores del estimador menores a 1.1, lo que indica una convergencia suficientemente buena.

	Mín.	Q_1	Mediana	Media	Q_3	Máx.
Est. puntual	1.000	1.001	1.003	1.005	1.009	1.023
L. Superior IC	1.000	1.002	1.008	1.017	1.026	1.067

Tabla 3.3: Resumen de las estimaciones del factor de reducción potencial de escala de Gelman y Rubin para las muestras obtenidas con JAGS. Estimaciones puntuales y límites superiores de su intervalo de confianza.

MCSE de las medidas *a posteriori*

También se han estimado los errores estándar de Monte Carlo (MCSE) [61] para las muestras de las medidas de interés *a posteriori*. Como se ha mencionado con anterioridad, primero simulamos valores de los parámetros de la distribución *a posteriori* conjunta obtenida con INLA, y a partir de ellas obtenemos muestras de otras medidas, como incidencias acumuladas o probabilidades de transición. Es por este procedimiento que son susceptibles a la incertidumbre de Monte Carlo y se debe, por tanto, proporcionar el MCSE de esas muestras, de manera análoga a lo que se haría cuando se simula a partir de MCMC u otros métodos que funcionan con muestras aleatorias. Se ha calculado estos errores para cada Área de Salud, $k = 1, \cdots, 24$, para cada $t = 1, \cdots, 5$, para cada cantidad de interés, y tanto para mujeres como para hombres. La Tabla 3.4 resume los resultados. Puede observarse que los errores estándar de Monte Carlo son relativamente bajos comparados con las desviaciones estándar *a posteriori*, lo que se traduce en buenas estimaciones de la media *a posteriori* para estas medidas de interés (Figuras 3.4-3.7).

Sexo	Medida	Mín.	Q_1	Mediana	Media	Q_3	Máx.
Mujeres	F_{12}	1.82	2.06	2.22	2.22	2.36	2.92
	p_{FR}	1.86	2.11	2.26	2.26	2.36	2.99
	p_{FD}	1.78	2.13	2.34	2.40	2.61	3.61
	p_{RD}	1.91	2.10	2.27	2.27	2.46	2.68
Hombres	F_{12}	1.71	2.11	2.26	2.26	2.40	3.03
	p_{FR}	1.46	2.17	2.31	2.30	2.45	3.16
	p_{FD}	1.77	2.06	2.25	2.27	2.43	3.28
	p_{RD}	1.81	2.18	2.32	2.31	2.47	2.78

Tabla 3.4: Errores estándar de Monte Carlo como porcentaje de la desviación estándar *a posteriori*.

3.4.2. Análisis de sensibilidad

En este apartado nos centraremos en la inferencia *a posteriori* con INLA del modelo completo de enfermedad-muerte, propuesto en este capítulo, pero poniendo el foco en la sensibilidad respecto a variaciones de las condiciones previas. Para ello se realizarán separadamente modificaciones en los valores de algunos hiperparámetros que definen las distribuciones *a priori*: se fijará el parámetro de correlación espacial, γ, tomando distintos valores para representar escenarios más o menos correlados, y se variarán los grados de libertad, ν, de la distribución Wishart asociada a la matriz de precisión.

Fijando la correlación espacial, γ

En primer lugar, consideramos el parámetro γ, que es una medida relativa de dependencia espacial. En nuestro modelo, γ es un parámetro desconocido, objeto de inferencia y, por tanto, estimado mediante el propio modelo y a partir de los datos. Recordemos

que se estimó una media *a posteriori* de 0.841, lo que indicaría una gran correlación espacial entre los efectos aleatorios de las distintas regiones. Como valores fijos de este hiperparámetro se han considerado tres escenarios razonablemente diferentes, con valores para γ de 0, 0.5 y 0.99, no pudiendo alcanzarse el límite superior de 1 debido al rango de $\gamma \in [0, 1)$. Las estimaciones de los efectos aleatorios bajo estas condiciones difieren ligeramente de los resultados obtenidos al asumir un γ desconocido (Figura 3.10), mientras que las estimaciones de los parámetros se mantienen también estables (Tabla 3.5). Las mayores diferencias se observaron en el escenario sin correlación espacial, $\gamma = 0$, aunque siguen siendo relativamente pequeñas. Por otro lado, se observan mayores diferencias en las transiciones de fractura a refractura, $F \rightarrow R$, y de refractura a muerte, $R \rightarrow D$, pudiendo justificarse, respectivamente, por un menor número de eventos (más observaciones censuradas) y un menor tamaño muestral (menos pacientes refracturados).

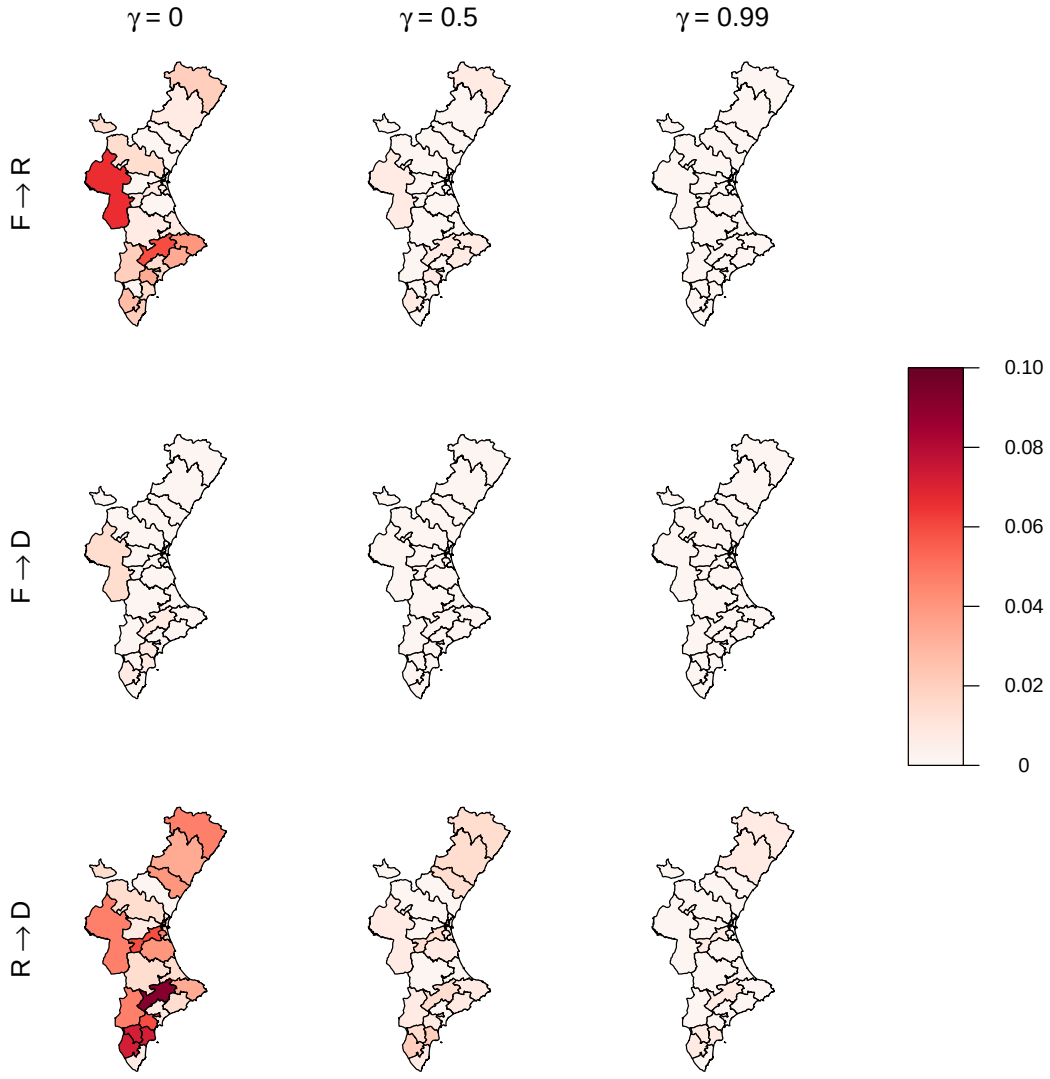

Figura 3.10: Diferencias absolutas en la media *a posteriori* estimada de los efectos aleatorios para los valores fijos de $\gamma = 0, 0.5, 0.99$ en comparación con un γ desconocido (media *a posteriori* estimada de 0.841 de acuerdo con el análisis principal), aproximadas mediante INLA.

	$\gamma = 0$		$\gamma = 0.5$		$\gamma = 0.99$		γ desconocido	
Parámetro	Media(IC 0.95)	Mediana	Media(IC 0.95)	Mediana	Media(CI 0.95)	Mediana	Media(IC 0.95)	Mediana
β^{FR}	-3.603 (-3.739, -3.468)	-3.603	-3.598 (-3.773, -3.424)	-3.598	-3.596 (-4.764, -2.430)	-3.596	-3.598 (-3.993, -3.204)	-3.597
β^{FD}	-1.104 (-1.196, -1.013)	-1.104	-1.105 (-1.234, -0.976)	-1.105	-1.105 (-2.094, -0.115)	-1.105	-1.106 (-1.436, -0.775)	-1.106
β^{RD}	-0.537 (-0.694, -0.381)	-0.537	-0.543 (-0.744, -0.343)	-0.543	-0.545 (-1.808, 0.719)	-0.546	-0.545 (-0.984, -0.107)	-0.545
$\beta_{FR,\,Mujer}$	0.021 (-0.076, 0.119)	0.021	0.021 (-0.076, 0.119)	0.021	0.021 (-0.076, 0.119)	0.020	0.021 (-0.076, 0.119)	0.021
$\beta_{FD,\,Mujer}$	-0.510 (-0.542, -0.477)	-0.510	-0.510 (-0.542, -0.477)	-0.510	-0.510 (-0.542, -0.477)	-0.510	-0.510 (-0.543, -0.477)	-0.510
$\beta_{RD,\,Mujer}$	-0.633 (-0.760, -0.504)	-0.634	-0.633 (-0.760, -0.505)	-0.634	-0.634 (-0.760, -0.505)	-0.634	-0.634 (-0.761, -0.505)	-0.634
$\beta_{FR,\,Edad}$	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024
$\beta_{FD,\,Edad}$	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070
$\beta_{RD,\,Edad}$	0.050 (0.040, 0.059)	0.050	0.050 (0.040, 0.059)	0.049	0.049 (0.040, 0.058)	0.049	0.049 (0.040, 0.059)	0.049
α^{FR}	0.922 (0.892, 0.953)	0.922	0.922 (0.892, 0.953)	0.921	0.922 (0.890, 0.955)	0.922	0.921 (0.891, 0.953)	0.921
α^{FD}	0.776 (0.766, 0.786)	0.776	0.776 (0.766, 0.787)	0.776	0.776 (0.766, 0.786)	0.776	0.776 (0.766, 0.786)	0.776
α^{RD}	0.629 (0.597, 0.662)	0.628	0.629 (0.598, 0.659)	0.629	0.628 (0.598, 0.659)	0.628	0.628 (0.597, 0.659)	0.628
γ	-	-	-	-	-	-	0.841 (0.591, 0.973)	0.862
τ^{FR}	19.756 (10.710, 32.414)	19.139	15.807 (8.211, 26.395)	15.302	14.499 (7.291, 25.747)	13.792	14.257 (7.197, 25.185)	13.581
τ^{FD}	24.442 (13.471, 39.911)	23.666	21.411 (11.768, 34.839)	20.764	19.362 (10.260, 33.176)	18.535	19.896 (10.915, 32.737)	19.228
τ^{RD}	16.412 (9.136, 27.109)	15.810	12.505 (6.220, 21.904)	11.969	12.041 (5.594, 22.992)	11.248	11.743 (5.386, 21.625)	11.137
$\rho^{(FR)(FD)}$	-0.013 (-0.343, 0.297)	-0.008	-0.058 (-0.384, 0.268)	-0.058	-0.036 (-0.377, 0.326)	-0.041	-0.044 (-0.388, 0.315)	-0.047
$\rho^{(FR)(RD)}$	-0.013 (-0.352, 0.333)	-0.015	-0.110 (-0.461, 0.246)	-0.110	-0.070 (-0.417, 0.291)	-0.073	-0.076 (-0.415, 0.275)	-0.078
$\rho^{(FD)(RD)}$	0.059 (-0.273, 0.359)	0.066	0.113 (-0.220, 0.434)	0.115	0.116 (-0.247, 0.439)	0.123	0.109 (-0.217, 0.423)	0.111

Tabla 3.5: Estimaciones *a posteriori* de los parámetros del modelo de enfermedad-muerte espacial para valores fijos $\gamma = 0, 0.5, 0.99$, y con γ desconocido y estimado por el modelo, mediante INLA.

Modificando los grados de libertad, ν

En este caso, fijaremos en 6, 8, 9 y 10 los grados de libertad de la distribución Wishart que se emplea como distribución *a priori*de la matriz de precisión, $\Sigma^{-1}_{between}$, que incluye la correlación entre transiciones. La comparación se realizará con los resultados obtenidos para $\nu = 7$, el valor de referencia que se ha utilizado en el análisis principal. Nuevamente, las estimaciones *a posteriori* mostraron pocas diferencias, tanto efectos aleatorios (Figura 3.11) como parámetros (Tabla 3.6). Además, la transición de refractura a muerte vuelve a mostrar una mayor variabilidad, a causa del menor tamáño muestral.

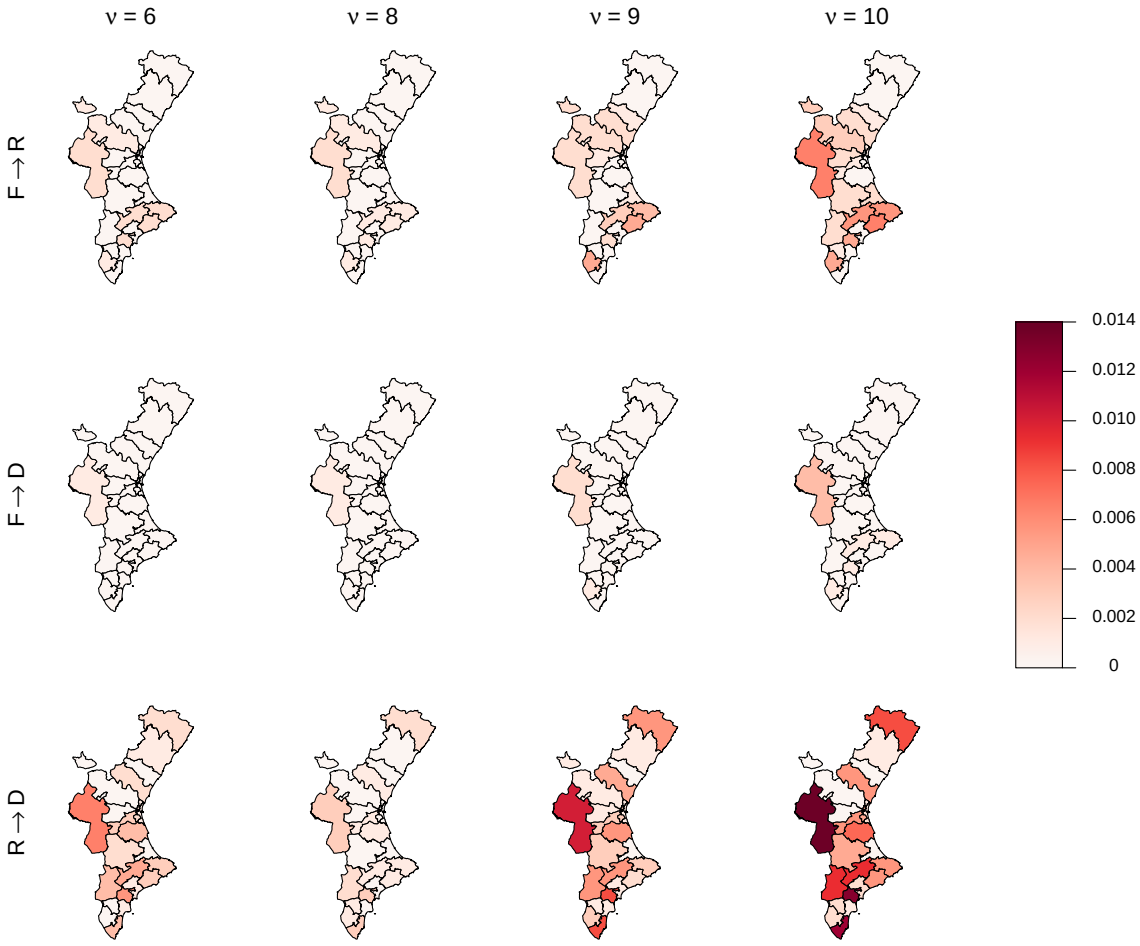

Figura 3.11: Diferencias absolutas en media *a posteriori* estimada de los efectos aleatorios para $\nu = 6, 8, 9, 10$ en comparación con $\nu = 7$ (el valor de referencia del análisis principal), aproximadas mediante INLA.

	$\nu = 6$		$\nu = 7$		$\nu = 8$		$\nu = 9$		$\nu = 10$	
Parámetro	Media(IC 0.95)	Mediana	Media(IC 0.95)	Mediana	Media(IC 0.95)	Mediana	Media(IC 0.95)	Mediana	Media(IC 0.95)	Mediana
β^{FR}	-3.598 (-3.954, -3.243)	-3.598	-3.598 (-3.993, -3.204)	-3.597	-3.598 (-3.913, -3.284)	-3.598	-3.597 (-3.945, -3.251)	-3.597	-3.597 (-3.878, -3.319)	-3.597
β^{FD}	-1.106 (-1.402, -0.809)	-1.106	-1.106 (-1.436, -0.775)	-1.106	-1.106 (-1.358, -0.854)	-1.105	-1.105 (-1.394, -0.818)	-1.105	-1.105 (-1.332, -0.879)	-1.105
β^{RD}	-0.544 (-0.946, -0.142)	-0.544	-0.545 (-0.984, -0.107)	-0.545	-0.545 (-0.891, -0.199)	-0.545	-0.545 (-0.924, -0.166)	-0.545	-0.546 (-0.852, -0.237)	-0.546
$\beta_{FR,\,Mujer}$	0.021 (-0.076, 0.119)	0.021	0.021 (-0.076, 0.119)	0.021	0.021 (-0.076, 0.119)	0.020	0.021 (-0.076, 0.119)	0.020	0.021 (-0.076, 0.119)	0.020
$\beta_{FD,\,Mujer}$	-0.510 (-0.543, -0.477)	-0.510	-0.510 (-0.543, -0.477)	-0.510	-0.510 (-0.543, -0.477)	-0.510	-0.510 (-0.543, -0.477)	-0.510	-0.510 (-0.542, -0.477)	-0.510
$\beta_{RD,\,Mujer}$	-0.634 (-0.761, -0.505)	-0.634	-0.634 (-0.761, -0.505)	-0.634	-0.633 (-0.760, -0.504)	-0.634	-0.633 (-0.760, -0.504)	-0.633	-0.633 (-0.759, -0.504)	-0.633
$\beta_{FR,\,Edad}$	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024	0.024 (0.018, 0.030)	0.024
$\beta_{FD,\,Edad}$	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070	0.070 (0.068, 0.073)	0.070
$\beta_{RD,\,Edad}$	0.049 (0.040, 0.059)	0.049	0.049 (0.040, 0.059)	0.049	0.049 (0.040, 0.059)	0.049	0.049 (0.040, 0.059)	0.049	0.049 (0.040, 0.059)	0.049
α^{FR}	0.922 (0.891, 0.953)	0.922	0.921 (0.891, 0.953)	0.921	0.922 (0.891, 0.953)	0.923	0.922 (0.891, 0.953)	0.922	0.923 (0.891, 0.956)	0.923
α^{FD}	0.776 (0.766, 0.786)	0.776	0.776 (0.766, 0.786)	0.776	0.776 (0.766, 0.786)	0.776	0.776 (0.766, 0.786)	0.776	0.776 (0.766, 0.786)	0.776
α^{RD}	0.629 (0.598, 0.661)	0.629	0.628 (0.597, 0.659)	0.628	0.628 (0.599, 0.659)	0.628	0.628 (0.598, 0.660)	0.628	0.629 (0.602, 0.658)	0.628
γ	0.814 (0.612, 0.942)	0.826	0.841 (0.591, 0.973)	0.862	0.810 (0.654, 0.916)	0.819	0.810 (0.556, 0.959)	0.828	0.800 (0.689, 0.879)	0.805
τ^{FR}	14.008 (6.921, 24.437)	13.436	14.257 (7.197, 25.185)	13.581	15.376 (8.155, 26.349)	14.716	16.220 (8.394, 27.121)	15.701	17.198 (9.565, 27.858)	16.677
τ^{FD}	19.401 (10.731, 31.805)	18.752	19.896 (10.915, 32.737)	19.228	21.030 (12.024, 33.543)	20.424	21.073 (11.809, 34.583)	20.327	22.757 (13.116, 36.496)	22.038
τ^{RD}	11.402 (5.580, 20.102)	10.908	11.743 (5.386, 21.625)	11.137	12.818 (6.600, 22.117)	12.283	12.998 (6.188, 23.091)	12.439	13.872 (8.029, 21.567)	13.560
$\rho^{(FR)(FD)}$	-0.048 (-0.367, 0.293)	-0.053	-0.044 (-0.388, 0.315)	-0.047	-0.058 (-0.383, 0.273)	-0.059	-0.061 (-0.399, 0.286)	-0.063	-0.044 (-0.307, 0.223)	-0.046
$\rho^{(FR)(RD)}$	-0.104 (-0.474, 0.281)	-0.107	-0.076 (-0.415, 0.275)	-0.078	-0.072 (-0.415, 0.289)	-0.075	-0.119 (-0.453, 0.226)	-0.121	-0.089 (-0.371, 0.185)	-0.087
$\rho^{(FD)(RD)}$	0.114 (-0.254, 0.451)	0.120	0.109 (-0.217, 0.423)	0.111	0.115 (-0.238, 0.450)	0.118	0.147 (-0.188, 0.476)	0.146	0.149 (-0.144, 0.458)	0.145

Tabla 3.6: Estimaciones *a posteriori* de los parámetros del modelo de enfermedad-muerte espacial para $\nu = 6, 7, 8, 9, 10$, mediante INLA. El valor de referencia $\nu = 7$ es el empleado en el análisis principal.

3.5. Discusión

Durante este capítulo se ha querido mostrar como el potencial y la utilidad de los modelos de enfermedad-muerte aumentan considerablemente al combinarlos con información espacial, en particular, modelizada mediante unos efectos aleatorios multivariantes asociados al conjunto de unidades espaciales. En este marco conjunto es posible estudiar tanto las diferencias entre regiones como la progresión de los individuos a lo largo del tiempo. A pesar de que la correlación entre transiciones no se estimó relevante en nuestro estudio de fractura de cadera, el hecho de modelarla conjuntamente con la correlación espacial abre muchas posibilidades. En nuestro caso, empleamos un modelo basado en una estructura de vecindad, como es el modelo de Leroux, el cual en sí mismo ya define un extenso número de escenarios en función de la realidad clínica y epidemiológica. Además, la validez de nuestro análisis no se limitaría a los modelos de enefrmedad-muerte, siendo posible trabajar con modelos multiestado más complejos, con estados y transiciones adicionales. El procedimiento sería análogo ya que incluir los efectos aleatorios en las intensidades de transición mediante un modelo de riesgos proporcionales de Cox es algo transversal a la mayoría de modelos empleados en supervivencia.

Algunas extensiones de nuestro trabajo con respecto a la censura y el truncamiento [62] podrían implementarse también sin muchas dificultades. Los pacientes con fractura de cadera requieren hospitalización y es por eso que sabemos exactamente cuándo ocurren los acontecimientos. Como resultado, se dispone únicamente de censura administrativa (por fin de estudio) o muerte. Sin embargo, otros tipos de censura podrían surgir en otras aplicaciones de salud que involucrasen supervivencia. Es el caso de la censura por intervalos la cual es frecuente en estudios en los que las visitas médicas son la única forma de conocer el estado de los pacientes [71]. Así un paciente podría alcanzar un estado de enfermedad (estado 2) entre dos visitas consecutivas, de forma que sólo tendríamos un intervalo de tiempo en el que ocurrió el evento de interés. Por otro lado, en nuestro análisis hemos considerado como escala temporal el tiempo transcurrido desde la fractura. Los pacientes ingresan al proceso en ese momento y se les realiza un seguimiento posterior. Se podría haber planteado, por ejemplo, un estudio alternativo donde la escala de tiempo estuviese marcada por la edad del paciente [64], con el tiempo cero determinado por la edad de 65 años. Esta aproximación generaría datos truncados a la izquierda para todos aquellos pacientes cuya entrada en el estudio y, por ende, cuya fractura de cadera, tuvo lugar con una edad superior a los 65 años. Ambos escenarios planteados serían posible en INLA, dado que sus funciones para los modelos Weibull de supervivencia permiten, además de manejar datos censurados por la derecha como es nuestro caso, trabajar con censura por intervalos y truncamiento por la izquierda.

En nuestro estudio, la asunción de funciones de riesgo de Weibull parece ser bastante consistente. Como se ha comentado a lo largo de este trabajo, los riesgos Weibull sólo pueden aumentar o disminuir, por definición. Por lo que respecta a las fracturas recurrentes de cadera y la muerte, la disminución de los riesgos que se ha estimado parece razonable, ya que se sabe que la incidencia de esos eventos es particularmente alta durante el primer año después de una fractura. Sin embargo, después de algún tiempo, se

podría llegar a esperar que el riesgo de muerte aumentase, mostrando tal vez una curva suave en forma de U, lo cual no encajaría con las distribuciones de Weibull. Así pues, especificaciones más flexibles, como aquellas basadas en funciones por partes o splines cúbicos [38], podrían ser una buena alternativa para explorar en trabajos futuros.

El uso de la aproximación anidada integrada de Laplace (INLA) puede considerarse otro punto fuerte de este trabajo. El tiempo de cálculo se reduce enormemente en comparación con los métodos MCMC, y la introducción de efectos aleatorios gaussianos en los términos de regresión resulta completamente natural en INLA. En este capítulo, hemos podido comparar, aunque con algunas simplificaciones, los resultados obtenidos mediante INLA y JAGS, pudiendo ver ligeras diferencias entre ambos métodos. Esto nos permite decantarnos por un análisis con INLA que, además de ser rápido, presenta resultados razonables y creíbles. Cabe mencionar que, a pesar de sus vantajas, todavía no es una opción popular a la hora de evaluar modelos multiestado, aunque esto podría cambiar en un futuro próximo. La modelización de procesos multiestado mediante INLA ofrece infinidad de posibilidades aún inexploradas que seguro que serán de gran interés.

Finalmente, nos gustaría remarcar que la posibilidad de evaluar las probabilidades de transición y las incidencias acumuladas, en lugar de analizar únicamente las estimaciones de los efectos aleatorios, proporciona una comprensión más profunda del problema clínico o epidemiológico. De hecho, la evaluación de riesgos por sí sola rara vez proporciona información predictiva sobre el pronóstico de los pacientes, sino que se centra más en valorar cómo las diferentes características de los mismos tienen efecto o no sobre estos riesgos. Por el contrario, expresar las trayectorias temporales en términos de probabilidad permite obtener resultados dinámicos e interpretables cuya información derivada puede ser determinante para los médicos y los responsables de las políticas sanitarias.

Capítulo 4

Modelos multiestado con curación

4.1. Introducción

Como se ha presentado en capítulos anteriores de esta tesis, el estudio del tiempo de supervivencia o tiempo hasta el evento es esencial en epidemiología ya que permite comprender la dinámica subyacente sobre cómo se desarrollan las enfermedades o condiciones de salud en una población a lo largo del tiempo. El análisis de supervivencia tiene como objetivo proporcionar conocimientos relevantes sobre lo que es más probable que suceda, expresado en términos probabilísticos, así como identificar características a nivel individual que sean factores de riesgo, es decir, que hagan que los individuos sean más o menos propensos a experimentar un evento [65]. En muchos escenarios del mundo real no hay un único evento de interés sino varios, los cuales pueden ocurrir secuencialmente o competir entre ellos (riesgos competitivos), por ejemplo. Los modelos multiestado definen un marco general que comprende esos escenarios multievento [66]. Son una clase de modelos estocásticos en los que los individuos pueden moverse entre ciertos estados. Estos modelos multiestado, y en particular los modelos de enfermedad-muerte, pueden entenderse como modelos de supervivencia multivariantes, lo que implica que muchas de las extensiones del análisis de supervivencia unidimensional pueden incluirse en el marco multiestado para reflejar escenarios más complejos. Por ejemplo, en el capítulo anterior ya hemos visto cómo era posible incluir efectos aleatorios que diesen información de carácter espacial. Por otro lado, los modelos de mixturas, en los que el tiempo de supervivencia se describe empleando varias distribuciones de probabilidad, resultan especialmente interesantes y, de añadirse al marco de los modelos multiestado, pueden abrir el campo a nuevas preguntas científicas relevantes. En este capítulo nos centraremos en dos tipos particulares de modelos de mixtura: los modelos de regresión cero-inflados y los modelos de curación de tipo mixtura.

Los modelos cero-inflados consideran que el proceso a partir del cual se generan los ceros es diferente al que genera el resto de valores [67]. Se utilizan principalmente para analizar datos de conteos mediante modelos de Poisson, concretamente cuando existe un exceso de ceros que el proceso de Poisson no puede explicar adecuadamente. La cero-inflación, sin embargo, también encaja a la perfección en estudios de supervivencia en

los que los ceros surgen debido a la existencia de una subpoblación en la que no se podrá observar el evento de interés [68]. Por ejemplo, si los pacientes mueren antes del seguimiento tendrán un tiempo observado con valor cero, no siendo posible la observación de ningún evento posterior. Una posibilidad sería excluir a aquellos individuos que no pueden ser seguidos en el tiempo, aunque en ocasiones puede ser relevante conservar esta probabilidad asociada a los ceros o profundizar en los factores que modifican esa probabilidad.

Los modelos de curación de tipo mixtura, por otro lado, se centran en problemas de supervivencia en los que se espera que una fracción de individuos no experimente el evento de interés, independientemente de cuál sea el tiempo de seguimiento [69, 70, 71]. En la práctica, esto resulta en tiempos censurados por la derecha, pero siendo esta censura esencialmente diferente de la que resulta de un seguimiento limitado, pérdida de seguimiento o eventos competitivos. También es diferente de la cero-inflación, ya que un paciente con un tiempo cero tiene un evento en el tiempo cero, mientras que un paciente curado nunca experimentará el evento principal, incluso después de un período de tiempo "infinito". Así pues, por definición, sólo será posible identificar quiénes son los pacientes no curados (con evento), mientras que aquellos con tiempos censurados podrían ser tanto curados como no curados, lo que resulta en una variable de curación parcialmente observada. Los modelos de curación se han empleado en diferentes campos, como la economía y las finanzas [72, 73], pero se aplican principalmente en entornos epidemiológicos. En particular, resultan de especial relevancia en estudios centrados en la progresión tras un cáncer, en los que una fracción de pacientes curados no mostrará una recaída [74, 75, 76]. Por otro lado, la muerte es un resultado probable que podría impedir la observación de recurrencias. Como consecuencia, un enfoque que combine modelos de curación y modelos de enfermedad-muerte será particularmente útil para estos estudios sobre cáncer, al considerar precisamente el riesgo competitivo de muerte.

En cuanto a la inferencia, los enfoques bayesianos resultan especialmente atractivos debido a la naturaleza faltante de la variable indicadora de la curación. De hecho, desde una perspectiva de inferencia bayesiana, las variables, los parámetros y los valores faltantes se tratan como variables aleatorias, lo que hace que la inferencia sea natural y coherente. Los métodos de Markov Chain Monte Carlo (MCMC) [77] son uno de los algoritmos más populares para realizar análisis bayesianos. A pesar de su precisión, pueden ser computacionalmente intensivos a medida que los modelos se vuelven más complejos o con bases de datos más grandes. Como resultado, resulta interesante explorar métodos más rápidos para los que evaluar su precisión y rendimiento. La aproximación anidada integrada de Laplace (INLA) [78] es una buena opción por lo que respecta al tiempo de cálculo y precisión, si bien no puede ajustar naturalmente los modelos de curación de tipo mixtura. En Lázaro et al. (2020) [79] se propone un algoritmo que combina INLA y muestreo de Gibbs para poder ajustar este tipo de modelos, a la vez que se reduce notablemente la cantidad de tiempo dedicado en comparación con MCMC.

Algunos autores han propuesto métodos que extienden los modelos de curación a escenarios de supervivencia más complejos. En Yu et al. (2007) se evaluó la curación utilizando modelos de mixtura en entornos de supervivencia multivariantes, extendiendo

esos modelos a datos de supervivencia bivariados [80]. En Basu et al. (2010) se propuso un enfoque bayesiano para evaluar la curación en modelos de riesgos competitivos, con aplicación a la supervivencia tras sufrir un cáncer de mama [81]. Además, en Conlon et al. (2013) [82] y Beesley et al. (2019) [83] se usaron modelos multiestado con una proporción de pacientes curados, el primero proponiendo un análisis bayesiano basado en métodos MCMC y el segundo utilizando un algoritmo EM para ajustar este tipo de modelos.

Con todo esto, el objetivo de este trabajo es proponer una metodología para ajustar modelos de enfermedad-muerte con cero-inflación y curación, basada en un algoritmo que combina INLA y muestreo de Gibbs, generalizando y extendiendo la clase de modelos con los que se puede utilizar INLA. En este capítulo, se definirá primero la estructura del modelo de enfermedad-muerte completo, con-cero inflación y curación, para pasar posteriormente a su contextualización dentro del marco de la inferencia bayesiana y al algoritmo central de esta metodología. Finalmente, aplicaremos una particularización de nuestro modelo, considerando solo cero-inflación, a un estudio del mundo real que involucra progresión tras fractura de cadera, mientras que la metodología completa será analizada, junto con la precisión de sus estimaciones, mediante conjuntos de datos simulados que representan escenarios reales.

4.2. Modelo multiestado bayesiano con curación y cero-inflado

4.2.1. Modelo de enfermedad-muerte

La metodología propuesta se basa en los modelos de enfermedad-muerte. No profundizaremos mucho en su definición, ya que han sido el objeto de estudio a lo largo de todo este trabajo, pero a grandes rasgos podemos decir que son modelos multiestado que constan de tres estados: un estado inicial, un estado de enfermedad y un estado de muerte (Figura 4.1). En este modelo, los individuos entran en el estado inicial, a partir del cual pueden avanzar a los estados siguientes de enfermedad o muerte, en caso de que presenten alguno de estos eventos, o quedarse en el mismo estado inicial, en caso contrario. La progresión del estado de enfermedad al de muerte también queda reflejada en este modelo, una cuestión esencial que distingue esta clase de modelos de los modelos de riesgos competitivos.

Este modelo representa la realidad y el diseño de una gran cantidad de estudios epidemiológicos, en los que junto a la enfermedad se presenta un riesgo de muerte importante que debe ser tenido en consideración. Sin embargo, no deja de ser un modelo relativamente sencillo que deja fuera algunas cuestiones que pueden ser relevantes desde el punto de vista clínico, como es el caso de la cero-inflación y la curación. Dos elementos relevantes que caracterizan a los individuos de la población de estudio y que limitan de una u otra manera su desplazamiento a través de los distintos estados. El modelo que incluye todos estos elementos se muestra en la Figura 4.2.

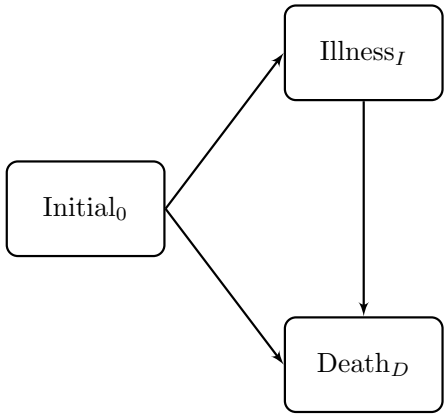

Figura 4.1: Modelo de enfermedad-muerte con tres estados: inicial, enfermedad y muerte.

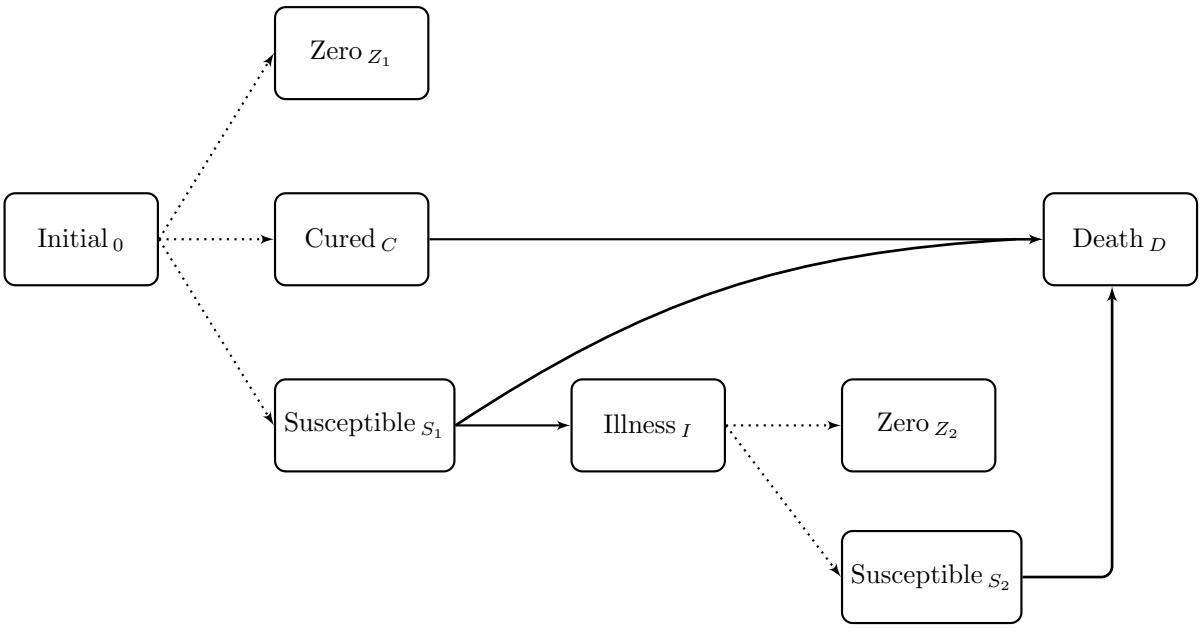

Figura 4.2: Modelo de enfermedad-muerte con cero-inflación y curación.

En primer lugar, trataremos la cero-inflación, es decir, la presencia de individuos cuyo tiempo de transición observado es precisamente 0. Estos pacientes entran en el estudio en el instante inicial $t = 0$ pero no se dispone de ninguna información temporal posterior, ya que de hecho su seguimiento resulta imposible. De forma general, se puede entender estos ceros como la existencia de un evento o una condición anterior al seguimiento, que impide la observación de la enfermedad o la muerte, aunque también puede representar un escenario de muerte inicial, como podría ser la muerte intrahospitalaria. Este último caso resulta de especial interés, teniendo un mismo evento, la muerte, que se puede dar

tanto al inicio, con tiempos 0, como a lo largo del seguimiento. En términos estadísticos, sea T el tiempo de muerte, lo que tenemos es que

$$P(T \leq t) = \begin{cases} \rho_0, & t = 0 \\ \rho_0 + (1 - \rho_0)P(T' \leq t), & t > 0 \end{cases} \tag{4.1}$$

donde ρ_0 es la probabilidad de muerte en $t = 0$ y T' denota el tiempo hasta la muerte en los individuos con $t > 0$, el cual seguirá una distribución de probabilidad propia del análisis de supervivencia. El caso cero-inflado dentro de los modelos de enfermedad-muerte se ilustrará mediante un caso práctico en secciones posteriores.

Por otro lado, en cuanto a curación se refiere, en determinadas enfermedades puede darse el caso de que un grupo de pacientes no presenten nunca más la enfermedad, pudiendo considerarse curados de la misma. Tiene particular interés en aquellos escenarios en los que el estado inicial es el que en sí mismo representa la recuperación tras la enfermedad, mientras que el estado de enfermedad se entiende como una recurrencia o recaída. Así pues, un individuo que efectivamente se cura no presentará ninguna recurrencia, independientemente del tiempo de seguimiento.

En general, los modelos de curación pueden definirse como una mixtura cuya expresión es similar a la cero-inflación pero con implicaciones e interpretaciones totalmente distintas. Sea T el tiempo hasta la enfermedad, su distribución de probabilidad se definirá en términos de la función de supervivencia como

$$S(t) = P(T > t) = \rho_C + (1 - \rho_C)P(T' > t) \tag{4.2}$$

donde ρ_C es la proporción de individuos curados y T' es el tiempo de supervivencia para el grupo de individuos no curados, al cual se dotará de una distribución de probabilidad temporal. Bajo esta modelización se considera que un grupo de individuos nunca experimentará la enfermedad, de forma que $\lim_{t\to\infty} S(t) = \rho_C$.

Trasladando estos elementos al modelo de enfermedad-muerte, lo que tendremos son subpoblaciones con sus particularidades en lo que a transiciones respecta: los individuos cero (Z_1) no avanzarán a ningún estado, los individuos curados (C) no podrán experimentar la enfermedad, y un último grupo de individuos, ni ceros ni curados, al que se denominará grupo de susceptibles (S_1), podrá presentar cualquier transición desde el estado inicial tanto a enfermedad como a muerte. Esto se traduce en la partición del estado inicial en tres estados previos al seguimiento, definidos en $t = 0$, a partir de los cuales se construirá el modelo. Esta asignación a uno u otro grupo se realizará asumiendo un modelo multinomial.

Cabe mencionar que, dado que el estado de enfermedad puede verse como un nuevo punto de partida, es posible discutir y definir los mismos tres estados (cero, curado y susceptible) después de la enfermedad, repitiendo la misma estructura de forma recursiva. Para simplificar, tras la enfermedad solo se hará distinción entre individuos con tiempos cero (Z_2) y susceptibles (S_2), lo que permitirá comparar la presencia de ceros en el estado inicial y después de la enfermedad.

En total son cuatro las posibles transiciones, una que partirá del estado de susceptible (S_1) y llegará al estado de enfermedad (I), y tres que llegarán al estado de muerte (D)

desde los estados de curado (C), susceptible (S_1) y susceptible después de la enfermedad (S_2), denotadas por: $C \to D$, $S_1 \to I$, $S_1 \to D$ y $S_2 \to D$. Como en los modelos multiestado presentados en capítulos anteriores, las transiciones se definen mediante intensidades de transición, $h_{ij}(t)$, que son equivalentes a las funciones de riesgo propias del análisis de supervivencia.

4.2.2. Tiempos de transición

4.2.2.1. Tiempo hasta muerte en curados

Sea T_{CD} el tiempo desde el estado inicial hasta la muerte para un paciente curado. Definimos la función de riesgo para la transición de la curación a la muerte en el momento t, $h_{CD}(t)$, mediante la expresión:

$$h_{CD}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_{CD} < t + \Delta t \mid T_{CD} \geq t)}{\Delta t}, \quad t > 0. \tag{4.3}$$

A partir de esa función de riesgo podemos definir la función de supervivencia respectiva, que es de hecho la probabilidad que tendría un individuo curado de estar vivo en un momento específico. Como los individuos curados sólo pueden pasar al estado de muerte, la función de supervivencia se deriva únicamente de esa función de riesgo:

$$S_{CD}(t) = P(T_{CD} > t) = \exp\Big\{ - \int_0^t h_{CD}(u)du \Big\}. \tag{4.4}$$

Alternativamente, también es posible calcular la función de distribución, $F_{CD}(t) = P(T_{CD} \leq t)$, que toma como valor el complementario de la función de supervivencia, es decir, $F_{CD}(t) = 1 - S_{CD}(t)$, y se interpreta como la probabilidad de muerte de los individuos curados antes o en el instante t. Esta función es equivalente a las funciones de incidencia acumulada de los modelos de riesgos competitivos, siendo este un escenario particular con un solo evento. Es por esto que a lo largo del capítulo nos referiremos indistintamente a ella como incidencia acumulada de muerte para los curados, probabilidad de transición de curado a muerte o, simplemente, probabilidad de muerte entre los individuos curados.

4.2.2.2. Tiempo hasta enfermedad o muerte en susceptibles

De manera análoga, denotamos T_{S_1I} y T_{S_1D}, respectivamente, como el tiempo desde el estado inicial hasta la enfermedad y hasta la muerte sin enfermedad, para un individuo susceptible. Ambos tiempos de transición tienen sus funciones de riesgo o intensidad: $h_{S_1I}(t)$, la función de riesgo para la transición de susceptible a enfermedad en el momento t, y $h_{S_1D}(t)$, la función de riesgo para la transición de susceptible a muerte en el momento t.

$$h_{S_1I}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_{S_1I} < t + \Delta t \mid T_{S_1I} \geq t)}{\Delta t}, \quad t > 0 \tag{4.5}$$

$$h_{S_1D}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_{S_1D} < t + \Delta t \mid T_{S_1D} \geq t)}{\Delta t}, \quad t > 0 \tag{4.6}$$

Aunque cada función de riesgo tiene una función de supervivencia asociada, estas no son especialmente útiles por sí mismas ya que no tienen en cuenta la censura que resulta de considerar un escenario de riesgos competitivos. Sí que podremos, en cambio, definir una probabilidad libre de eventos, es decir, la probabilidad de no experimentar ningún evento, ya sea enfermedad o muerte. Para hacerlo, consideremos T_{S_1} el tiempo hasta el primero de esos eventos, enfermedad o muerte, para individuos susceptibles, $T_{S_1} = \text{mín}\{T_{S_1I}, T_{S_1D}\}$. Así pues, como en el marco de riesgos competitivos, la función de supervivencia global se puede definir como

$$S_{S_1}(t) = P(T_{S_1} > t) = \exp\Big\{ -\int_0^t \Big(h_{S_1I}(u) + h_{S_1D}(u)\Big)\mathrm{d}u\Big\}, \quad t > 0. \tag{4.7}$$

Por otro lado, si por el contrario se prefiere aportar información específica de la ocurrencia de los eventos, enfermedad y muerte, se pueden definir funciones de incidencia acumulada que permitan evaluar la probabilidad de observar cada uno de los eventos competitivos. Serán pues la incidencia acumulada de enfermedad, $F_{S_1I}(t)$, y la incidencia acumulada de muerte sin enfermedad, $F_{S_1D}(t)$, que como su nombre indica se definen como la probabilidad de enfermedad y la probabilidad de muerte sin enfermedad, en el momento t, respectivamente, considerando la presencia de la otra causa competitiva de fallo. En particular, estas funciones se pueden obtener como

$$\begin{aligned} F_{S_1I}(t) &= \int_0^t h_{S_1I}(u)S_{S_1}(u)\mathrm{d}u \\ F_{S_1D}(t) &= \int_0^t h_{S_1D}(u)S_{S_1}(u)\mathrm{d}u, \quad t > 0. \end{aligned} \tag{4.8}$$

4.2.2.3. Tiempo de enfermedad hasta muerte

Por último, consideraremos la progresión tras la enfermedad. Un resultado posible es ser un individuo cero, lo que no implica ningún tiempo de transición. Por tanto, nos centraremos en los individuos susceptibles tras la enfermedad, de forma que denotamos como T_{S_2D} el tiempo de transición desde enfermedad a muerte, para aquellos pacientes susceptibles tras la enfermedad (S_2). Seguirá un modelo semi-Markov, en el que el "cronómetro" se reinicia una vez que los individuos alcanzan el estado de enfermedad, siendo este el punto inicial y tiempo 0 para esa transición, pero conservando la información relativa al tiempo que tomó la transición a la enfermedad. Bajo esta notación, su función de riesgo se define como

$$\begin{aligned} &h_{S_2D}(t - t_{S_1I} \mid T_{S_1I} = t_{S_1I}) = \\ &\qquad \lim_{\Delta t \to 0} \frac{P(t - t_{S_1I} \leq T_{S_2D} < t - t_{S_1I} + \Delta t \mid T_{S_2D} \geq t - t_{S_1I}, T_{S_1I} = t_{S_1I})}{\Delta t} \end{aligned} \tag{4.9}$$

para $t > t_{S_1I}$. Como se ha indicado anteriormente, la función de riesgo asociada al tiempo de transición desde susceptible tras la enfermedad hasta la muerte, T_{S_2D}, se ha definido condicionalmente al tiempo transcurrido en la progresión desde el estado

inicial de susceptible (S_1) hasta la enfermedad (I) , T_{S_1I}. Como resultado, la función de supervivencia que indica la proporción de pacientes susceptibles que sobreviven tras la enfermedad también se define condicionalmente mediante la expresión

$$S_{S_2D}(t - t_{S_1I} \mid T_{S_1I} = t_{S_1I}) = \exp\Big\{ - \int_{t_{S_1I}}^{t} h_{S_2D}(u - t_{S_1I} \mid T_{S_1I} = t_{S_1I})\mathrm{d}u\Big\}. \tag{4.10}$$

Por otro lado, recordemos que cuando ocurre la transición al estado de enfermedad, $T_{S_1I} = t_{S_1I}$, los individuos se dividen en ceros o susceptibles. Como consideramos una definición amplia de ceros, que tiene en cuenta eventos no específicos que ocurren en el momento $t = 0$, pueden surgir diferentes interpretaciones de una función de supervivencia general después de una enfermedad. En particular, como considerearemos en el caso práctico que mostraremos en secciones posteriores, siempre que cero sea igual a muerte en el momento $t = 0$, los ceros se considerarían no supervivientes, lo que conduciría a

$$S_{ID}(t - t_{S_1I} \mid T_{S_1I} = t_{S_1I}) = (1 - \rho_{Z_2})S_{S_2D}(t - t_{S_1I} \mid T_{S_1I} = t_{S_1I}). \tag{4.11}$$

Finalmente, cabe mencionar que también podemos hablar en términos de probabilidad de muerte, de la misma forma que para otros tiempos se puede definir la función contraria a la supervivencia. En este caso, la función F_{S_2D} sería la función de distribución asociada a este tiempo, a la que llamaremos también probabilidad de muerte tras la enfermedad o incidencia acumulada de muerte tras la enfermedad, para los pacientes susceptibles.

4.3. Inferencia bayesiana

4.3.1. Modelo

En primer lugar definiremos la parte multinomial del modelo que representa la partición del estado inicial. Sea A una variable categórica con tres posibles valores, uno para cada uno de los grupos definidos con anterioridad: cero ($A = Z_1$), curado ($A = C$) y susceptible ($A = S_1$). Los individuos partirán de uno u otro estado con distinta probabilidad. En particular, denotaremos por ρ_{Z_1} la probabilidad de cero inicial, $P(A = Z_1) = \rho_{Z_1}$, ρ_C la probabilidad de ser un paciente curado, $P(A = C) = \rho_C$, y ρ_{S_1} la probabilidad de susceptible inicial, $P(A = S_1) = \rho_{S_1}$. Esta variable seguirá una distribución categórica, es decir, una distribución multinomial en la que solo se realiza una prueba por individuo, de forma que el resultado de la misma será el grupo al que este pertenezca.

$$A_k \mid \boldsymbol{\rho}_k \sim Cat(\boldsymbol{\rho}_k) \tag{4.12}$$

donde $\boldsymbol{\rho}_k = (\rho_{Z_1,k}, \rho_{C,k}, \rho_{S_1,k})$. Para modelizar estas probabilidades se empleará un modelo de regresión logística multinomial, de forma que cada individuo k tendrá unas probabilidades individuales que dependerán de un término de regresión con unos coeficientes, $\boldsymbol{\beta}_k$, y un vector de covariables $\boldsymbol{x}_k$.

$$\rho_{Z_1,k} = \frac{\exp\{\boldsymbol{\beta}_{Z_1}\boldsymbol{x}_k\}}{1+\exp\{\boldsymbol{\beta}_{Z_1}\boldsymbol{x}_k\}+\exp\{\boldsymbol{\beta}_C\boldsymbol{x}_k\}} \tag{4.13}$$

$$\rho_{C,k} = \frac{\exp\{\boldsymbol{\beta}_C\boldsymbol{x}_k\}}{1+\exp\{\boldsymbol{\beta}_{Z_1}\boldsymbol{x}_k\}+\exp\{\boldsymbol{\beta}_C\boldsymbol{x}_k\}}$$

$$\rho_{S_1,k} = \frac{1}{1+\exp\{\boldsymbol{\beta}_{Z_1}\boldsymbol{x}_k\}+\exp\{\boldsymbol{\beta}_C\boldsymbol{x}_k\}}$$

Por otro lado, sea Z_2 la variable indicadora que identifica los ceros tras la enfermedad. Se denotará por ρ_{Z_2} la probabilidad de cero, $P(Z_2 = 1)$, y $\rho_{S_2} = 1-\rho_{Z_2}$ la probabilidad de susceptible, $P(Z_2 = 0)$. Esta segunda variable, Z_2, seguirá una distribución de Bernoulli donde la probabilidad de éxito será precisamente ρ_{Z_2}, la cual se estimará mediante un modelo de regresión logística.

$$\begin{aligned} &Z_{2,k} \mid \rho_{Z_2,k} \sim Bern(\rho_{Z_2,k}) \\ &\text{logit}(\rho_{Z_2,k}) = \boldsymbol{\beta}_{Z_2}\boldsymbol{x}_k \end{aligned} \tag{4.14}$$

Finalmente, para las transiciones entre los distintos estados se emplearán modelos de riesgos proporcionales de Cox. Sea $h_{ij}(t \mid \boldsymbol{\theta})$ la intensidad de transición del estado i a j en el instante t, de modo que la transición $i \to j$ sea una de las mostradas en secciones anteriores. Bajo esta modelización, las funciones de riesgo se definen como el producto de una intensidad de transición basal en el mismo instante t, $h_{ij,0}(t \mid \boldsymbol{\theta})$, y la exponencial de un término de regresión que incluye información de covariables, $\boldsymbol{x}$, de la forma:

$$h_{ij}(t \mid \boldsymbol{\theta}) = h_{ij,0}(t \mid \boldsymbol{\theta}) \exp\{\boldsymbol{x}'\boldsymbol{\beta_{ij}}\}. \tag{4.15}$$

En particular, consideraremos un modelo completamente paramétrico de la función de riesgo basal asumiendo distribuciones Weibull. Como ya se ha mencionado en capítulos anteriores, estas distribuciones se relacionan de forma natural con las distribuciones de tiempos y se han utilizado ampliamente en modelos de supervivencia. Alternativamente, se podrían especificar modelos semiparamétricos más flexibles para modelar las intensidades basales, como modelos por partes o splines cúbicos. Así pues, en este análisis nos centraremos en intensidades basales Weibull, que toman la forma $h_{ij,0}(t) = \alpha_{ij}\lambda_{ij}t^{\alpha_{ij}-1}$. Cabe destacar que no aparecen interceptos, $\beta_{ij,0}$, en el término de regresión de los modelos de Cox 4.15 ya que la información sobre el riesgo de referencia es capturada por el parámetro de escala λ_{ij}, el cual, de hecho, puede verse como la exponencial de ese intercepto, $\lambda_{ij} = \exp\{\beta_{ij,0}\}$.

4.3.2. Procedimiento bayesiano

Como se ha indicado en capítulos anteriores, según aumenta la complejidad de los modelos, los métodos computacionales resultan esenciales para poder llevar a cabo el proceso de inferencia bayesiana, estimando la distribución *a posteriori* de los parámetros e

hiperparámetros, $\pi(\boldsymbol{\theta} \mid \mathcal{D})$. En este sentido, INLA se postula como una buena alternativa a los métodos de Monte Carlo, y como se ha mostrado en el capítulo anterior, permite trabajar con modelos de enfermedad-muerte complejos.

No obstante, anteriormente ya se ha mencionado cómo las características de la variable indicadora de curación no permiten aplicar INLA directamente para las estimación de estas distribuciones *a posteriori*. En concreto, esto se debe a la naturaleza semiobservable de esta variable, siendo únicamente posible identificar los individuos no curados, y no teniendo absolutamente ninguna información sobre el grupo de curados. Los métodos MCMC podrían lidiar directamente con la existencia de valores faltantes en esta variable latente, sin embargo, esto tendría como resultado un proceso computacional intensivo. Es por esta razón que se plantea el uso de INLA, de forma que, si bien no es posible ajustar el modelo directamente, se puede plantear un algortimo en el que de forma iterativa se simulen valores de la variable latente y, en cada iteración, se estimen distribuciones *a posteriori* de los parámetros con INLA.

4.3.3. Distribuciones *a priori*

Dado que en general no dispondremos de conocimiento previo, consideraremos un escenario no informativo y asumiremos distribuciones *a priori*no informativas. En particular, se suponen previas gaussianas amplias para los interceptos y efectos fijos de las covariables, β, con media 0 y precisión 0.001. Esta precisión se define como la inversa de la varianza, donde una precisión baja indica una varianza alta y, por tanto, se asume poca información *a priori* acerca de estas distribuciones. A los parámetros de escala, λ_{ij}, no se les dota de una distribución *a priori*como tal, sino que la obtienen al definirse como la exponencial de los interceptos. Estas distribuciones previas normales para los efectos son estrictamente necesarias en el marco de los campos aleatorios gaussianos de Markov en el que trabaja INLA. Además, se asumen distribuciones *a priori*de complejidad penalizada (PC) para los hiperparámetros de forma, α_{ij} (consulte la documentación de INLA con la función `inla.doc("pc.alphaw")` para obtener una definición detallada). A grandes rasgos, las previas PC en el contexto de supervivencia Weibull, asumen un escenario base exponencial, $\alpha_{ij} = 1$, y se penalizan las desviaciones respecto a este modelo. Así pues, sólo la existencia de evidencia suficiente que respaldase un modelo Weibull llevaría a preferir este modelo en lugar de uno más sencillo [48].

4.3.4. INLA y muestreo de Gibbs

Si bien se ha definido un modelo y un procedimiento dentro del marco de la inferencia bayesiana que tiene sentido, cabe matizar alguna cuestión respecto al tratamiento de la curación. Como se mencionó anteriormente, en un caso hipotético en el que pudiéramos seguir a los individuos infinitamente, sería posible observar si un individuo muestra la enfermedad, y por tanto no esta curado, o si por el contrario, nunca experimenta este evento de interés, siendo así un individuo curado. Sin embargo, la censura por la derecha no permite observar en ningún caso la curación, siendo los individuos censurados candidatos a pacientes curados, en el mejor de los casos. Así pues, la curación podrá

considerarse una variable latente [46], únicamente parcialmente observada, dado que sí que se puede afirmar que aquellos individuos que desarrollan la enfermedad eran susceptibles y no curados.

A efectos prácticos, tendremos una variable indicadora de curación con valores faltantes para aquellos pacientes con tiempos hasta la enfermedad censurados. Esto no ha de suponer un problema desde la perspectiva bayesiana, dado que los valores faltantes también pueden dotarse de distribución de probabilidad. En particular, los métodos MCMC permitirían resolver esta cuestión directamente, especificando de forma simultánea el modelo multinomial y el modelo de enfermedad-muerte. No obstante, dada la exhaustividad a nivel computacional de estos métodos resulta de especial interés emplear técnicas alternativas para la inferencia.

Nuestra propuesta considera el uso de la aproximación anidada integrada de Laplace (INLA) para ajustar el modelo de curación enfermedad-muerte con cero-inflación y curación, considerando el estado de curación como una variable latente. Lázaro et al. propuso un enfoque para ajustar modelos de curación de tipo mixtura usando INLA [79], a su vez basado en trabajos previos de Gómez-Rubio y Rue [84]. En su trabajo, combinan estimaciones de INLA obtenidas tras asumir valores conocidos del indicador de curación junto con muestreo de Gibbs, proporcionando secuencialmente nuevas configuraciones de este vector indicador de curación. Para ello ajustaron dos modelos con INLA: un modelo de regresión logística para la curación (modelo de incidencia) y un modelo de supervivencia (latencia). Nuestra propuesta se basa en el mismo procedimiento, aunque resulta en una generalización de la misma ya que permite analizar datos y escenarios mucho más complejos, con un estado inicial multinomial y dentro del marco de los modelos de enfermedad-muerte.

Para poder realizar este análisis se requerirá una modificación respecto al marco metodológico propuesto en la sección anterior. Dado que INLA no permite trabajar con una distribución multinomial directamente, y a pesar de que puede aproximarse mediante una transformación multinomial-Poisson [85], se emplearán dos distribuciones de Bernoulli, definiendo dos modelos de regresión logística, que resultarán equivalentes. Este enfoque separado proporcionó estimaciones más estables por lo que se prefirió al modelo Poisson. En particular, separaremos entre individuos cero y no cero, siendo estos segundos a su vez divididos entre curados y susceptibles. En general mantendremos la notación, aunque definiremos una probabilidad de curación condicionada, $\rho'_C = P(A = C \mid A \neq Z_1)$, para aquellos individuos no cero, siendo pues el modelo empleado:

$$\begin{aligned} Z_{1,k} \mid \boldsymbol{\theta} &\sim Bern(\rho_{Z_1,k}), & \text{logit}(\rho_{Z_1,k}) &= \beta_{Z_1,0} + \boldsymbol{x}'_k\boldsymbol{\beta}_{Z_1}, \\ C_k \mid Z_{1,k} = 0, \boldsymbol{\theta} &\sim Bern(\rho'_{C,k}), & \text{logit}(\rho'_{C,k}) &= \beta_{C,0} + \boldsymbol{x}'_k\boldsymbol{\beta}_C, \end{aligned} \tag{4.16}$$

donde en lugar de tener una variable A_k que indique en cuál de los tres estados iniciales empieza el individuo k, tendremos dos variables que indicarán si el paciente tiene un tiempo cero o no, $Z_{1,k}$, y en caso de no ser cero, si el individuo está curado o no, C_k.

Para el muestreo de Gibbs, sea $\boldsymbol{c}$ el vector con los valores observados de la variable latente de curación, 1 cura y 0 susceptible. Como se sabe que los pacientes que llegan al estado de enfermedad son susceptibles, dividimos la variable latente $\boldsymbol{c} = (\boldsymbol{c}_{obs}, \boldsymbol{c}_{mis})$

, distinguiendo entre valores observados, es decir, $\boldsymbol{c}_{obs} = 0$ para aquellos individuos que muestran la enfermedad, y valores faltantes para el resto.

El procedimiento se puede resumir en el siguiente algoritmo:

- `Paso 0`. Se asignan valores iniciales para los elementos con valores faltantes de $\boldsymbol{c}$, construyendo el vector $\boldsymbol{c}^{(0)} = (\boldsymbol{c}_{mis}^{(0)}, \boldsymbol{c}_{obs})$.

  Para $n = 1, 2, \cdots$.

- `Paso 1`. Empleamos INLA para ajustar los modelos de regressión logística y el modelo de enfermedad-muerte, obteniendo las distribuciones *a posteriori* de parámetros e hiperparámetros, $\pi(\boldsymbol{\theta} \mid \boldsymbol{c}^{(n-1)})$.

- `Paso 2`. Se obtiene la media de esas distribuciones *a posteriori*, $\hat{\boldsymbol{\theta}}^{(n)}$.

- `Paso 3`. Muestreamos de nuevo $\boldsymbol{c}^{(n)} = (\boldsymbol{c}_{mis}^{(n)}, \boldsymbol{c}_{obs})$ a partir de la probabilidad

$$p_i = \delta_{I,i} + (1-\delta_{I,i}) \frac{(1 - \widehat{\rho'_{C,i}}) \cdot h_{S_1 D}(t_{D,i})^{\delta_{D,i}} \cdot S_{S_1}(t_{D,i})}{(1 - \widehat{\rho'_{C,i}}) \cdot h_{S_1 D}(t_{D,i})^{\delta_{D,i}} \cdot S_{S_1}(t_{D,i}) + \widehat{\rho'_{C,i}} \cdot h_{CD}(t_{D,i})^{\delta_{D,i}} \cdot S_C(t_{D,i})} \tag{4.17}$$

  donde $p_i = P(A_i = S_1 \mid A_i \neq Z_1, \mathcal{D}, \hat{\theta}^{(n)})$, de manera que en cada iteración obtenemos la probabilidad de susceptible (no curado); $\delta_{I,i}$ y $\delta_{D,i}$ son variables indicadoras que equivalen a 1 si el individuo i alcanza el estado de enfermedad o muerte, respectivamente; y $t_{D,i}$ es el tiempo hasta la muerte, observado o censurado, en función de lo indicado por $\delta_{D,i}$.

Mediante este procedimiento iterativo obtenemos varias configuraciones distintas de la variable latente de curación $\boldsymbol{c}$, para cada una de ellas ajustamos un modelo con INLA y estimamos distribuciones *a posteriori* para los parámetros $\boldsymbol{\theta}$. De hecho, como INLA también proporciona la verosimilitud marginal, seleccionamos la disposición cuyo ajuste lleva al valor máximo de la misma.

Después de aplicar el algoritmo basado en el muestreo de Gibbs, nos centramos en el modelo con la verosimilitud marginal más alta y calculamos la distribución *a posteriori* conjunta de parámetros e hiperparámetros, $\pi(\theta \mid \mathcal{D})$, para este escenario más probable. A partir de esta distribución es posible calcular la distribución *a posteriori* de cualquier cantidad de interés definida en términos de esos parámetros, como intensidades de transición, probabilidades de supervivencia o incidencias acumuladas.

4.4. Curación: análisis mediante datos simulados

4.4.1. Método de simulación

En esta sección se aplicará el modelo de enfermedad-muerte antes mencionado con curación y cero-inflación a un conjunto de datos simulado. Las simulaciones se han reali-

zado utilizando el paquete de `R` `simsurv` [86] que permite simular datos de supervivencia controlando covariables, censura y tiempo de seguimiento. El procedimiento de simulación se resume en los siguientes pasos:

- `Paso 1`. Se establecen valores para los parámetros del modelo de muestreo y se simulan covariables para cada individuo. En particular, se han simulado covariables de sexo y edad, a partir de una distribución binomial y normal, respectivamente.
- `Paso 2`. Se obtienen muestras a partir de dos distribuciones Bernoulli y se asignan individuos a los grupos de ceros (Z_1), curados (C) o susceptibles (S_1).
- `Paso 3`. Asignamos un tiempo de seguimiento a cada individuo y se simulan tiempos desde curación (C) hasta muerte (D), desde susceptible (S_1) hasta enfermedad (I), y desde susceptible (S_1) hasta muerte (D), censurando en función de este seguimiento.
- `Paso 4`. Para aquellos pacientes que alcanzan el estado de enfermedad, se toman muestras de una distribución Bernoulli y se les asigna a cero (Z_2) o susceptible (S_2) tras la enfermedad.
- `Paso 5`. Se simula el tiempo desde la susceptibilidad tras la enfermedad (S_2) hasta la muerte (D).

Los tiempos de transición entre estados se simulan a partir de distribuciones de Weibull, para las cuales se supone una censura no informativa. En particular, nuestro objetivo era replicar un escenario del mundo real con censura administrativa, es decir, censura por la derecha debido al final del seguimiento. Los pacientes son reclutados en diferentes puntos del período de estudio, teniendo como tiempo cero la entrada en el mismo. Por esta razón mostrarán tiempos de seguimiento distintos, de forma que algunos pacientes serán censurados cinco años después de ingresar al estudio, mientras que otros solo después de uno.

La función `simsurv` se puede utilizar para simular el tiempo hasta el evento a partir de las covariables, un tiempo de seguimiento y una distribución de probabilidad para la censura. En nuestro caso, como la censura se debe únicamente al final del estudio, no se asume ninguna otra censura al simular tiempos. Lo que hacemos es asignar aleatoriamente a cada paciente un tiempo de seguimiento, emulando ese seguimiento desigual. Posteriormente simulamos tiempos de transición, que serán censurados si el tiempo hasta el evento es mayor que el tiempo de seguimiento asignado. Los tiempos obtenidos mediante este mecanismo son muy similares a los de conjuntos de datos reales. Se espera además que las estimaciones de los parámetros sean bastante precisas, ya que este método da lugar a una censura poco informativa.

4.4.2. Escenarios simulados

En general, la precisión en la estimación de los parámetros dependerá de la adecuada identificación de la subpoblación de individuos curados. Por tanto, se espera que

la efectividad varíe entre conjuntos de datos que representan diferentes escenarios del mundo real. Nos centramos así en la capacidad de identificar pacientes curados en cinco entornos simulados diferentes, obtenidos tras controlar y modificar los valores de los parámetros durante el proceso de simulación. Los escenarios se pueden resumir de la siguiente manera.

- **Escenario simulado 1**: Riesgo decreciente de curado a muerte ($\alpha_{CD} = 0.9$) y riesgos crecientes desde susceptible a enfermedad y muerte ($\alpha_{S_1I} = \alpha_{S_1D} = 1.7$). Se espera que casi la mitad de los individuos curados muera durante el primer año (45.5 %), creciendo esta mortalidad a un ritmo prácticamente continuo durante los años posteriores. La enfermedad es un evento más frecuente que la muerte en los pacientes susceptibles: la incidencia acumulada tras 2 años alcanza el 59.6 % y el 36.2 % para enfermedad y muerte sin enfermedad, respectivamente.
- **Escenario simulado 2**: Riesgos crecientes para todas las transiciones ($\alpha_{CD} = 3.2$, $\alpha_{S_1I} = 1.6$, $\alpha_{S_1D} = 2$). Los pacientes curados mueren menos al comienzo del seguimiento, con una incidencia acumulada al año del 12.6 %, debido al valor bajo del intercepto ($\beta_{CD,0} = -2$); no obstante, el riesgo se incrementa muy rápidamente, muriendo en su mayoría (98.9 %) después de 3 años. Los pacientes susceptibles muestran incidencias acumuladas de enfermedad y muerte tras 2 años del 62.2 % y el 37.3 %, respectivamente.
- **Escenario simulado 3**: Análogamente al Escenario 1, riesgo decreciente para curados y riesgos crecientes en los susceptibles. En concreto, la transición de curación a muerte no presenta ningún cambio respecto a este otro escenario. Sin embargo, en los susceptibles el riesgo de enfermedad, comparado con el de muerte, empieza desde más abajo ($\beta_{S_1I,0} = -2$ vs $\beta_{S_1D,0} = -1$) y progresa más lentamente ($\alpha_{S_1I} = 1.7$ vs $\alpha_{S_1D} = 3$). Como resultado, los pacientes susceptibles muestran incidencias acumuladas de enfermedad notablemente menores que las de muerte, alcanzando el 18.3 % y el 78.3 % tras 2 años, respectivamente.
- **Escenario simulado 4**: Mismas condiciones que en el Escenario 3 pero intercambiando los valores relativos a las transiciones a enfermedad y muerte de la subpoblación susceptible. Así pues, la incidencia acumulada de muerte en los curados no muestra cambios. Ahora bien, los pacientes susceptibles muestran mayores incidencias acumuladas de enfermedad que de muerte, siendo estas del 78.3 % y el 18.3 % tras 2 años, respectivamente.
- **Escenario simulado 5**: Riesgo de muerte creciente en los individuos curados y riesgos decrecientes de enfermedad y muerte en la subpoblación susceptible ($\alpha_{CD} = 1.6$, $\alpha_{S_1I} = 0.8$, $\alpha_{S_1D} = 0.7$). Los pacientes curados avanzan hacia el estado de muerte a lo largo de todo el periodo de seguimiento de 5 años, con una incidencia acumulada cuyo comportamiento se asemeja a uno lineal. El 83.1 % de los curados muere después de estos 5 años. Por otro lado, las incidencias de enfermedad y muerte en los pacientes susceptibles alcanzan el 36.5 % y 55.5 % tras 2 años, respectivamente.

Cada uno de los escenarios anteriores representa distintas posibilidades en cuanto a tiempos de supervivencia y transiciones de un estado a otro, o lo que es lo mismo, un perfil de supervivencia. No obstante, hasta ahora no se ha considerado ninguna información acerca de la proporción de individuos curados en la población. Así pues, dividiremos cada escenario en dos casos, $\beta_{C,0} = -3$ y $\beta_{C,0} = -2$, el primero con menos pacientes curados que el segundo, aunque en ambos casos se sigan asumiendo probabilidades de curación bajas. En resumen, simulamos 10 conjuntos de datos, 2 para cada escenario, cada uno de ellos con un tamaño muestral de 2000 individuos.

También se simularon e incluyeron en los modelos dos covariables, sexo y edad (centrada en la media). Sin embargo, con el fin de evaluar el comportamiento del algoritmo basándose únicamente en información de supervivencia, fijamos los efectos de esas covariables en 0 y 0.05, respectivamente, en todos los escenarios. Como consecuencia, todas las incidencias acumuladas que se muestran están definidas para los valores de sexo y edad de 0, valores de referencia, que indican sexo masculino y edad media.

4.4.3. Estimación *a posteriori*

Mediante el algoritmo propuesto que combina estimaciones con INLA y muestreo de Gibbs se obtiene, en cada iteración, la distribución *a posteriori* de los parámetros e hiperparámetros del modelo. Después de 300 iteraciones del algoritmo, eliminamos las 100 primeras como *burn-in* (periodo de preparación) y seleccionamos el modelo con mayor verosimilitud marginal de las otros 200. Este modelo será el que se utilizará para estimar y analizar esas distribuciones *a posteriori*. La Tabla 4.1 incluye la media *a posteriori* e intervalos de credibilidad 0.95 para cada parámetro, así como el valor real utilizado para la simulación en cada conjunto de datos simulado. Dado que el objetivo principal es valorar cómo funciona el algoritmo a la hora de distinguir esta subpoblación de individuos curados, así como su perfil de supervivencia, no intervendrá la transición a muerte tras la enfermedad. Esto se debe a que esta última cuenta únicamente con aquellos pacientes en los que se ha observado la enfermedad, los cuales ya están identificados, y por tanto, el riesgo asociado no va a cambiar sean cuales sean los individuos curados o no curados. Es por ello que obviaremos las estimaciones para esta transición en la tabla anteriormente mencionada.

En cuanto a las estimaciones, la mayoría se acercan considerablemente a sus respectivos valores reales. De hecho, es un buen indicador de que el algoritmo está identificando una subpoblación que tiene un perfil de supervivencia similar al de los individuos curados. En general, las estimaciones de los parámetros asociados a la transición de curación a muerte son los que más difieren de su valor real y, como consecuencia, lo mismo se aplica a las incidencias acumuladas (Figura 4.3).

Estas diferencias en las estimaciones de los parámetros e incidencias acumuladas que muestra la muerte en los pacientes curados puede tener, al menos parcialmente, unas causas estructurales. Cabe destacar que, ya que en ningún caso se observa el estado de curación, las estimaciones del riesgo de muerte en los pacientes curados dependen completamente de la clasificación realizada por el algoritmo. Por el contrario, únicamente

Escenario	β_0^C	α	Media	IC 0.95	β_0	Media	IC 0.95
1	-3	0.9	1.02	(0.85, 1.21)	-0.5	-1.12	(-1.82, -0.47)
		1.7	1.63	(1.55, 1.71)	-0.5	-0.5	(-0.6, -0.4)
		1.7	1.6	(1.5, 1.7)	-1	-0.95	(-1.08, -0.84)
	-2	0.9	0.92	(0.82, 1.02)	-0.5	-0.91	(-1.26, -0.57)
		1.7	1.67	(1.58, 1.76)	-0.5	-0.47	(-0.57, -0.37)
		1.7	1.65	(1.54, 1.77)	-1	-0.98	(-1.11, -0.86)
2	-3	3.2	3.6	(3.07, 4.16)	-2	-2.62	(-3.2, -2.07)
		1.6	1.53	(1.45, 1.61)	0	-0.01	(-0.1, 0.09)
		2	2.06	(1.93, 2.18)	-0.5	-0.5	(-0.63, -0.38)
	-2	3.2	3.16	(2.88, 3.45)	-2	-2.06	(-2.38, -1.76)
		1.6	1.57	(1.49, 1.66)	0	0.1	(0.01, 0.2)
		2	2.15	(2, 2.29)	-0.5	-0.53	(-0.67, -0.4)
3	-3	0.9	0.91	(0.77, 1.04)	-0.5	-0.78	(-1.29, -0.32)
		1.7	1.73	(1.56, 1.91)	-2	-1.91	(-2.09, -1.75)
		3	2.96	(2.83, 3.09)	-1	-0.95	(-1.04, -0.85)
	-2	0.9	0.77	(0.67, 0.86)	-0.5	-0.67	(-0.98, -0.38)
		1.7	1.58	(1.41, 1.75)	-2	-1.94	(-2.13, -1.76)
		3	3.05	(2.91, 3.2)	-1	-1.05	(-1.15, -0.95)
4	-3	0.9	0.89	(0.74, 1.03)	-0.5	-0.93	(-1.5, -0.4)
		3	2.98	(2.85, 3.11)	-1	-0.96	(-1.06, -0.86)
		1.7	1.64	(1.49, 1.81)	-2	-1.92	(-2.09, -1.75)
	-2	0.9	0.91	(0.82, 1.01)	-0.5	-0.53	(-0.81, -0.27)
		3	2.92	(2.78, 3.05)	-1	-0.89	(-0.99, -0.79)
		1.7	1.74	(1.54, 1.95)	-2	-2.07	(-2.28, -1.88)
5	-3	1.6	2.53	(1.88, 3.25)	-2	-4.01	(-5.1, -3.02)
		0.8	0.79	(0.74, 0.85)	-0.5	-0.5	(-0.62, -0.38)
		0.7	0.74	(0.7, 0.78)	-0.1	-0.04	(-0.14, 0.05)
	-2	1.6	1.69	(1.46, 1.92)	-2	-2.2	(-2.61, -1.81)
		0.8	0.78	(0.72, 0.83)	-0.5	-0.44	(-0.56, -0.32)
		0.7	0.75	(0.7, 0.79)	-0.1	-0.15	(-0.26, -0.05)

Tabla 4.1: Estimaciones *a posteriori* de los paramétros del modelo con máxima verosimilitud marginal obtenido mediante muestreo de Gibbs e INLA. Valores reales para la simulación, media e intervalos de credibilidad 0.95 *a posteriori*. Cada escenario se divide en dos casos en función de los valores de $\beta_{C,0}$. Cada caso contiene 3 filas, una para cada transición desde el estado inicial, $C \to D$, $S_1 \to I$, y $S_1 \to D$, respectivamente.

considerando aquellos individuos que desarrollan la enfermedad, ya tenemos un grupo observado de pacientes susceptibles, sea de mayor o menor tamaño, lo que por sí solo proporciona información relevante sobre los riesgos de enfermedad y muerte en este grupo. Además, aunque la subpoblación de pacientes curados pudiese observarse, y asumiendo que las distribuciones propuestas por el modelo fuesen correctas, esperaríamos

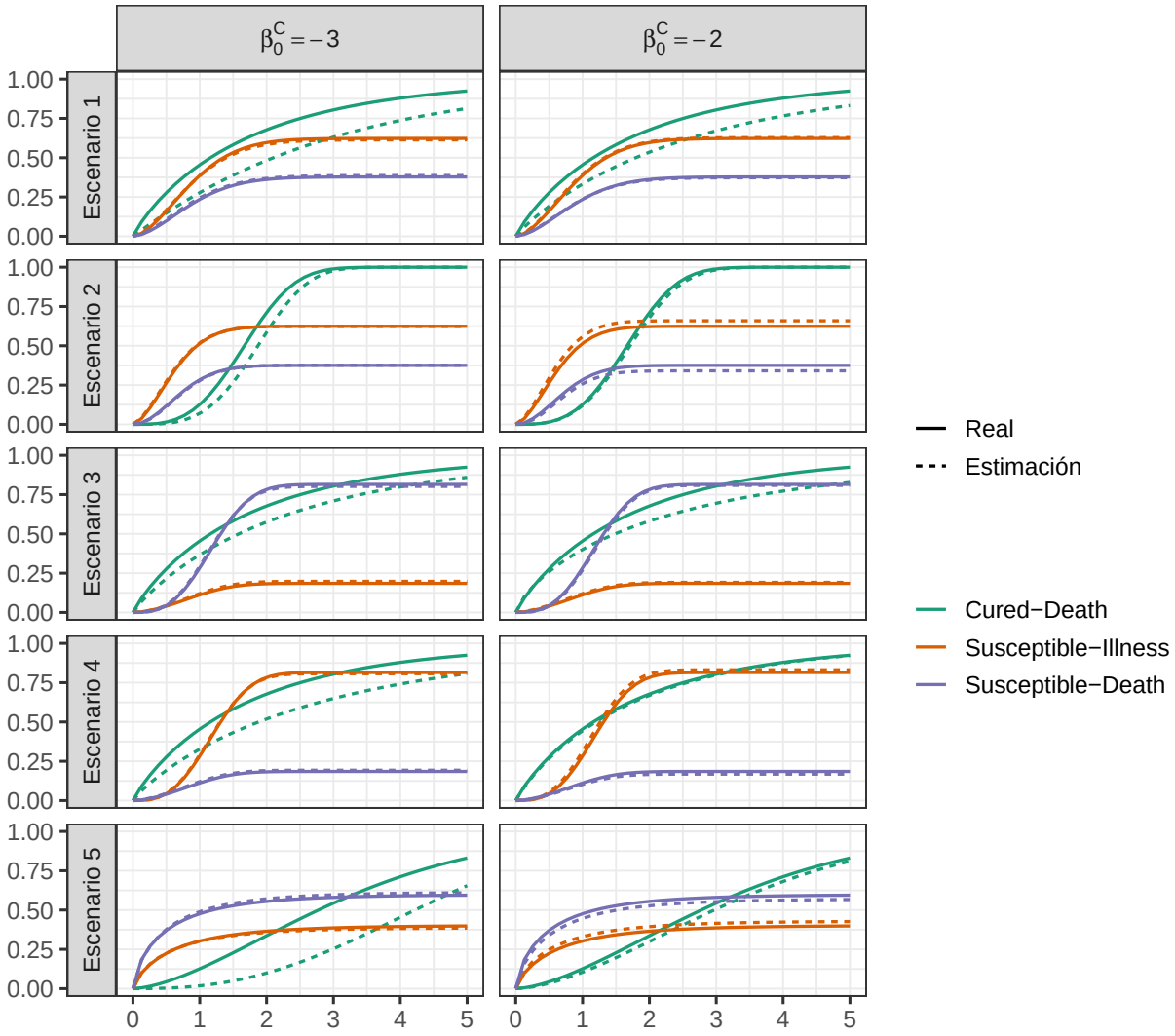

Figura 4.3: Incidencias acumuladas de las transiciones de curado a muerte, $C \to D$, de susceptible a muerte, $S_1 \to D$, y de susceptible a enfermedad, $S_1 \to I$. Las líneas continuas representan los valores reales de las incidencias acumuladas, calculadas a partir de los valores de los parámetros empleados para la simulación; las líneas discontinuas son la media *a posteriori* de las incidencias acumuladas obtenidas a partir del modelo con la máxima verosimilitud marginal.

estimaciones más precisas en la población susceptible, ya que representan un porcentaje más alto de pacientes en nuestros conjuntos de datos simulados.

Por otro lado, cuando se considera un valor del intercepto asociado a la probabilidad de curación $\beta_{C,0} = -2$, es decir, de entre los dos casos, aquel con un mayor número de individuos curados, se obtienen, en general, mejores estimaciones para los parámetros

de la transición $C \to D$, α_{CD} y $\beta_{CD,0}$. La única excepción sería el Escenario 3 en el que, a pesar de obtener una estimación más precisa del intercepto, la estimación del parámetro de forma α_{CD} presenta un mayor sesgo. Concretamente, para un valor real del intercepto $\beta_{CD,0} = -0.5$, obtenemos estimaciones de -0.78 y -0.67 en las muestras en las que $\beta_{C,0}$ es igual a -3 y -2, respectivamente; mientras que para el parámetro de forma $\alpha_{CD} = 0.9$, las estimaciones son 0.91 y 0.77, para $\beta_{C,0} = -3$ y -2, respectivamente. No obstante, a pesar de estas diferencias, se obtienen en ambos casos estimaciones similares de la incidencia acumulada de muerte en los pacientes curados.

Destaca también el Escenario 5 para el caso $\beta_{C,0} = -3$, siendo de todos el que presenta peores estimaciones. En este caso simulado, tanto el parámetro de forma como el intercepto que definen la intensidad de la transición de curado a muerte, α_{CD} y $\beta_{CD,0}$, son los que más difieren de sus valores reales, de la misma forma que lo hace la incidencia acumulada calculada a partir de los mismos. Se obtienen estimaciones sumamente mejores en el escenario con mayor proporción de individuos curados ($\beta_{C,0} = -2$).

Convergencia del método

El algoritmo de Gibbs propuesto se basa en el ajuste secuencial de los modelos mediante INLA, proporcionando cada una de esas iteraciones un ajuste y estimaciones diferentes. Así pues, de la misma forma que en los métodos MCMC se buscaba la convergencia de las cadenas de Markov, se espera en el algoritmo una especie de convergencia para poder asegurar un buen comportamiento del procedimiento y obtener estimaciones precisas.

Como se ha mencionado anteriormente, INLA proporciona estimaciones de la verosimilitud marginal, eligiéndose el modelo más probable para estimar las distribuciones *a posteriori* de parámetros y cantidades de interés. La Figura 4.4 presenta esta verosimilitud marginal en cada una de las iteraciones, para evaluar de esta forma su evolución a lo largo de la ejecución del algoritmo.

El resultado deseable sería una especie de estabilidad a partir de una determinada iteración, indicando que el algoritmo ha alcanzado un punto estacionario a partir del cual no mostrará una variación significativa. En general, se puede ver que este punto se alcanza rápidamente y que en la mayoría de los casos un *burn-in* de 100 iteraciones es suficiente, recordemos, considerando un total de 300 iteraciones. El Escenario 5 con $\beta_0^C = -3$, sin embargo, muestra una tendencia creciente de la verosimilitud marginal que se aleja de la estabilidad buscada. De hecho, este es justamente el caso simulado con peores resultados en general, cuyas estimaciones de los parámetros presentaban las mayores desviaciones con respecto a los valores reales.

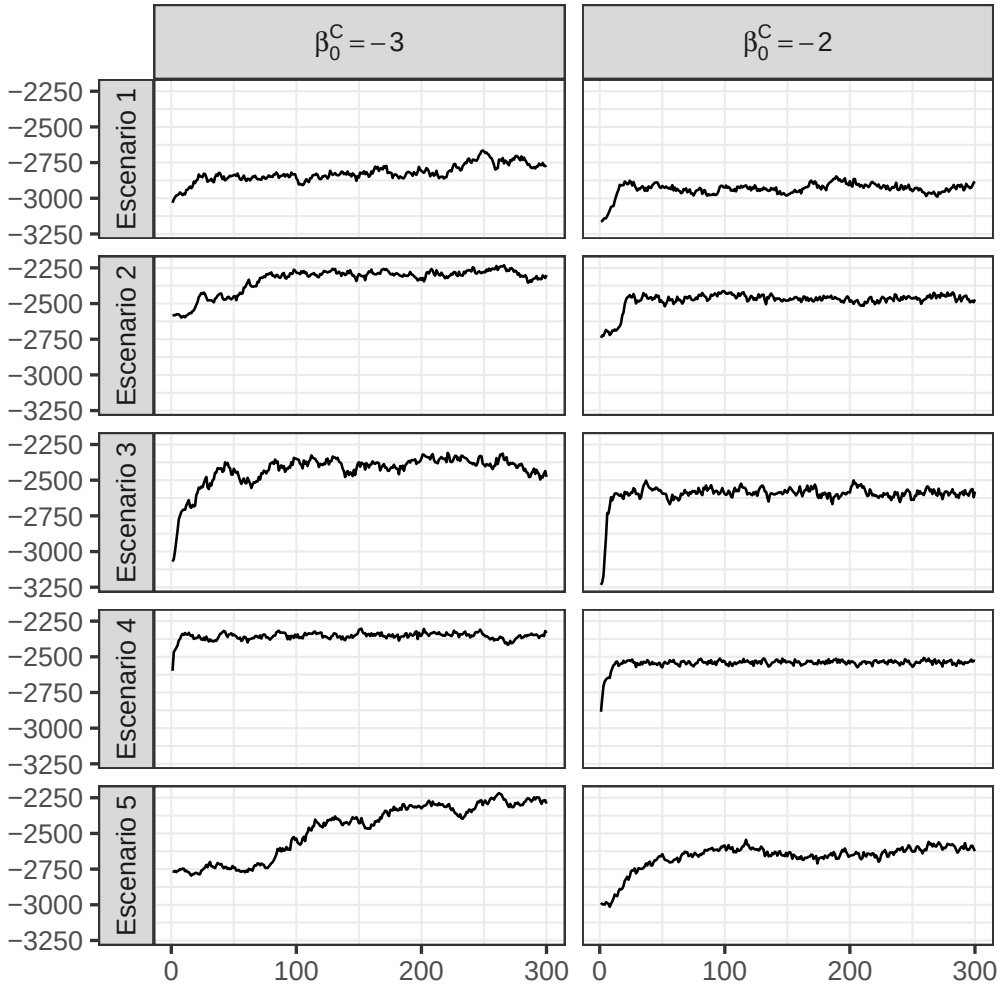

Figura 4.4: Verosimilitud marginal total (en escala logarítmica) obtenida del ajuste del modelo en cada iteración del muestreo de Gibbs, según escenario simulado y valor del intercepto $\beta_{C,0}$.

4.5. Caso práctico: muerte intrahospitalaria tras fractura

4.5.1. Cohorte PREV2FO

En esta sección se ilustrará la aplicación de los modelos planteados mediante un caso práctico con datos de la cohorte PREV2FO. En el capítulo 2 ya se ha presentado un análisis bayesiano para esta cohorte de pacientes utilizando modelos de enfermedad-muerte, el cual está basado en los artículos allí mencionados [23, 24]. El objetivo de este análisis fue explorar el potencial de un enfoque bayesiano en el marco de los modelos de enfermedad-muerte con un estudio del mundo real.

El planteamiento allí presentado puede ampliarse dentro del marco proporcionado

por un modelo de enfermedad-muerte más general, como el propuesto. En particular, dado que una pequeña proporción de pacientes no supera el ingreso e intervención, falleciendo en ese mismo ingreso, es posible incluir en el modelo un estado inicial que refleje esta muerte intrahospitalaria. Por lo general, estos pacientes que mueren en el hospital son excluidos del estudio, dada la imposibilidad de seguirlos en el tiempo. No obstante, la inclusión de estos individuos aporta información sobre la mortalidad real tras una fractura de cadera y, análogamente, sobre la mortalidad tras una refractura. Estos pacientes serán considerados individuos con tiempo $t = 0$ y analizados en consecuencia. Cabe destacar que si bien se presentarán las tasas estimadas de muerte intrahospitalaria tras la fractura inicial y tras refractura, la comparación de las mismas deberá efectuarse con cautela y en ningún caso entendiendo relaciones de causa-efecto.

	Cohorte		Refracturados	
	Total	Muertos hospital	Total	Muertos hospital
N	36237	1799 (5.0 %)	2532	104 (4.1 %)
Hombres	9389 (25.9 %)	728 (40.5 %)	516 (20.4 %)	41 (39.4 %)
Mujeres	26848 (74.1 %)	1071 (59.5 %)	2016 (79.6 %)	63 (60.6 %)
Edad, media(SD)	83.6 (6.9)	86.7 (6.7)	83.2 (6.2)	84.2 (6.5)

Tabla 4.2: Muerte intrahospitalaria en los pacientes de la cohorte PREV2FO (total y en refracturados), junto con las covariables sexo y edad.

En primer lugar, y antes de proceder a la definición del modelo que se empleará para el análisis, se realizará una breve descripción de los datos disponibles. Así pues, la Tabla 4.2 recoge información básica acerca de los pacientes que componen la cohorte, con el foco puesto en la muerte intrahospitalaria, así como en las covariables, sexo y edad, que se incluirán en el modelo. En la muestra se han observado tasas de muerte en el hospital del 5 % de los pacientes tras la fractura índice y del 4.1 % tras la refractura. La cohorte está compuesta mayormente por mujeres, 26848 (74.1 %) de un total de 36237, teniendo una presencia ligeramente mayor entre los pacientes refracturados, 2016 (79.6 %) de un total de 2532. Esta proporción se ve alterada en los individuos que mueren en el hospital, representando las mujeres el 60 % de estos. Esto se debe a que los hombres presentan mayores tasas de muerte intrahospitalaria crudas, 7.8 % y 7.9 % tras fractura y refractura, respectivamente, frente al 4.0 % y 3.1 % en las mujeres. Respecto a la edad, definida como la edad en el momento del alta tras la fractura índice, se observa que los pacientes que mueren en el hospital tienen de media edades superiores.

4.5.2. Modelización

En esta sección aplicamos el modelo propuesto al conjunto de datos de fractura de cadera. En particular, dado que se espera que todos los pacientes fracturados sean susceptibles de padecer una segunda fractura, el modelo resultante será una simplificación del caso general, pudiendo considerarse como un modelo de enfermedad-muerte cero-inflado.

La Figura 4.5) contiene un diagrama que representa el modelo empleado. Los pacientes hospitalizados debido a una fractura de cadera pueden morir en el hospital o ser dados de alta con vida. Únicamente aquellos que sobrevivan podrán ser seguidos en el tiempo hasta una fractura recurrente, hasta la muerte o hasta fin de estudio (censura por la derecha). De manera análoga, los pacientes refracturados también pueden fallecer en el hospital o ser dados de alta con vida, siendo estos últimos seguidos hasta muerte o fin de estudio.

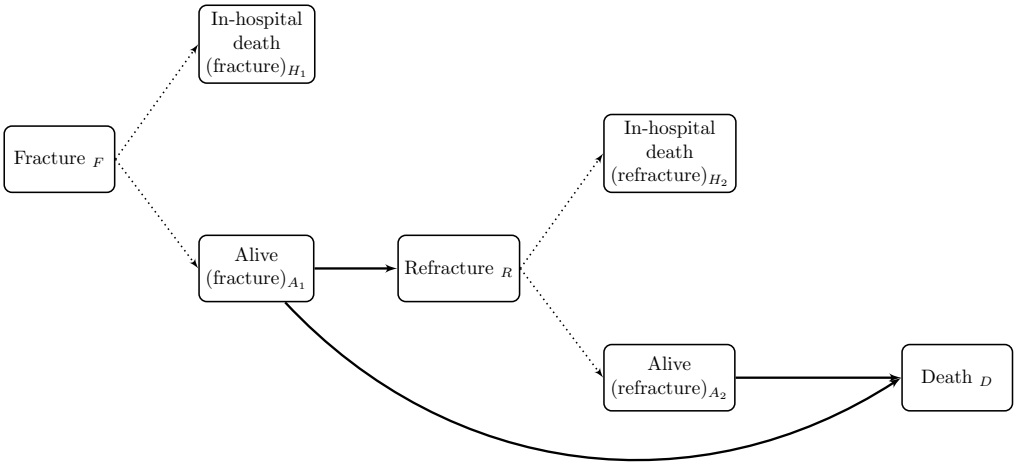

Figura 4.5: Modelo de enfermedad-muerte cero-inflado que incluye fractura, refractura y muerte, distinguiendo entre muerte intrahospitalaria y muerte durante el seguimiento.

Así pues, en términos del marco definido por los modelos multiestado, se consideran tres tiempos de transición: T_{A_1R}, el tiempo de transición desde el estado de "vivo tras la fractura" hasta el estado de refractura; T_{A_1D}, el tiempo de transición desde "vivo tras la fractura" a muerte; y T_{A_2D}, la transición tiempo desde el estado de "vivo tras la refractura" hasta la muerte. Este último tiempo de transición se define condicionado al tiempo hasta la refractura, definiendo un modelo de semi-Markov para esta transición, tal y como se ha indicado en apartados anteriores. Además, denotaremos por Z_{H_1} (Z_{H_2}) a la variable indicadora de muerte intrahospitalaria tras la fractura (refractura), cuyo valor será 1 en caso afirmativo y 0 en caso contrario. Las probabilidades de muerte intrahospitalaria asociadas a estas variables se indicarán mediante ρ_{H_1} y ρ_{H_2}, para la muerte tras la fractura índice y tras la refractura, respectivamente.

Se asumirán distribuciones de Weibull para cada tiempo de transición, así como distribuciones de Bernoulli para la variables de muerte intrahospitalaria. La información de las covariables sexo y edad se incluye a través de términos de regresión, lo que da como resultado modelos de regresión de Weibull (equivalentes a los modelos de riesgos proporcionales de Cox con una función de riesgo basal de Weibull) y modelos de regresión logística. Para los parámetros del modelo se emplearán distribuciones *a priori*no informativas, de la forma especificada en las secciones anteriores. Así pues, podemos

definir el modelo como

$$\begin{aligned} T_{A_1R} \mid \boldsymbol{\theta} &\sim We(\alpha_{A_1R}, \exp\{\eta_{A_1R}\}) \\ T_{A_1D} \mid \boldsymbol{\theta} &\sim We(\alpha_{A_1D}, \exp\{\eta_{A_1D}\}) \\ T_{A_2D} \mid T_{A_1R}, \boldsymbol{\theta} &\sim We(\alpha_{A_2D}, \exp\{\eta_{A_2D}\}) \\ Z_{H_1} \mid \boldsymbol{\theta} &\sim Bern(\rho_{H_1}), \qquad \text{logit}(\rho_{H_1}) = \eta_{H_1} \\ Z_{H_2} \mid \boldsymbol{\theta} &\sim Bern(\rho_{H_2}), \qquad \text{logit}(\rho_{H_2}) = \eta_{H_2} \end{aligned} \tag{4.18}$$

donde cada tiempo de transición o variable asociada a la muerte intrahospitalaria se define en función de un término de regresión distinto, $\eta_\star = \beta_{\star,0} + \beta_{\star,Mujer} \cdot \mathrm{I}_{Mujer} + \beta_{\star,Edad} \cdot \text{Edad}$.

4.5.3. Resultados

La Tabla 4.3 contiene un resumen de las distribuciones *a posteriori* estimadas mediante INLA para el modelo de enfermedad-muerte cero-inflado, el cual ha sido aplicado a la cohorte PREV2FO, con el fin de analizar conjuntamente la muerte en el hospital y la progresión desde la fractura índice hasta fractura recurrente o muerte.

Variable	Parámetros	Media	SD	IC 0.95
De A_1 a R	α_{A_1R}	0.921	0.016	(0.891, 0.953)
	$\beta_{A_1R,0}$	-3.581	0.048	(-3.676, -3.488)
	$\beta_{A_1R,Mujer}$	0.024	0.050	(-0.073, 0.122)
	$\beta_{A_1R,Edad}$	0.024	0.003	(0.019, 0.030)
De A_1 a D	α_{A_1D}	0.786	0.005	(0.775, 0.796)
	$\beta_{A_1D,0}$	-1.121	0.015	(-1.151, -1.092)
	$\beta_{A_1D,Mujer}$	-0.509	0.017	(-0.541, -0.476)
	$\beta_{A_1D,Edad}$	0.071	0.001	(0.068, 0.073)
De A_2 a D	α_{A_2D}	0.839	0.021	(0.799, 0.880)
	$\beta_{A_2D,0}$	-0.843	0.062	(-0.966, -0.723)
	$\beta_{A_2D,Mujer}$	-0.597	0.069	(-0.730, -0.461)
	$\beta_{A_2D,Edad}$	0.054	0.005	(0.044, 0.063)
H_1	$\beta_{H_1,0}$	-2.536	0.040	(-2.615, -2.458)
	$\beta_{H_1,Mujer}$	-0.790	0.050	(-0.889, -0.692)
	$\beta_{H_1,Edad}$	0.078	0.004	(0.070, 0.085)
H_2	$\beta_{H_2,0}$	-2.439	0.163	(-2.771, -2.131)
	$\beta_{H_2,Mujer}$	-0.995	0.207	(-1.397, -0.585)
	$\beta_{H_2,Edad}$	0.032	0.016	(0.001, 0.063)

Tabla 4.3: Estimaciones *a posteriori* para el modelo de enfermedad-muerte cero-inflado aplicado a la cohorte PREV2FO. Media, desviación estándar e intervalo de credibilidad 0.95 *a posteriori*.

Se estimaron funciones de riesgo decrecientes para todas las transiciones, refractura

(A_1R), muerte sin refractura (A_1D) y muerte tras refractura (A_2D) en los individuos dados de alta con vida, que son aquellos de los que se dispone de seguimiento. Esta información se deduce de los parámetros de forma de las distribuciones Weibull, α, con valores estimados menores que 1, siendo sus intervalos de credibilidad 0.95 claros respecto a la monotonía de estas funciones de riesgo. No se ha observado un efecto relevante de la variable sexo en el riesgo de refractura; por el contrario, en cuanto a mortalidad, se han estimado en las mujeres riesgos considerablemente menores de muerte sin y tras refractura. Finalmente, la edad se consideró un factor de riesgo para las tres transiciones, siendo el riesgo de muerte sin refractura el más relacionado con este factor.

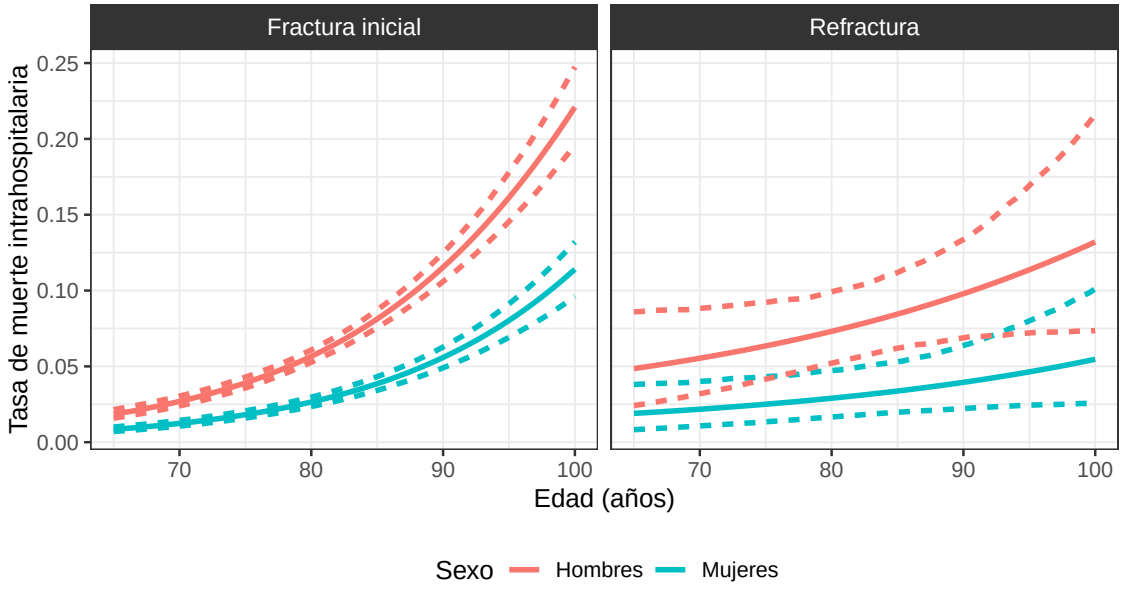

Figura 4.6: Probabilidad *a posteriori* estimada de muerte intrahospitalaria ($t = 0$) tras fractura índice, $P(Z_{H_1} = 1 \mid \mathcal{D})$, y tras refractura, $P(Z_{H_2} = 1 \mid \mathcal{D})$, según edad y sexo.

Por lo que respecta a la muerte intrahospitalaria, se estimaron probabilidades basales de muerte en el hospital similares tras la primera y la segunda fractura, con interceptos $\beta_{H_1,0} = -2.536$ y $\beta_{H_2,0} = -2.439$, respectivamente. Se estimaron efectos relevantes de ambas covariables, presentando las mujeres menores probabilidades de muerte intrahospitalaria e incrementando estas probabilidades con la edad. La Figura 4.6 muestra las probabilidades de muerte en el hospital *a posteriori* estimadas a partir de las distribuciones de los parámetros mostradas en la Tabla 4.3. Además de lo ya mencionado, se observa un mayor efecto de la edad en la probabilidad de muerte tras la fractura índice, así como intervalos de credibilidad notablemente más amplios para la muerte tras la refractura, como se podía esperar dado el menor número de pacientes refracturados. Estas comparaciones realizadas entre probabilidades de muerte tras fractura y refractura deben ser tomadas con precaución y en ningún caso debe deducirse una relación de causa-efecto.

4.6. Discusión

En este capítulo se ha propuesto una metodología estadística para evaluar la existencia de una proporción de individuos curados entre una población de estudio, dentro del marco de modelos multiestado. Creemos que esta metodología puede ser útil para aumentar el conocimiento acerca de algunas enfermedades, especialmente en aquellas en que la muerte es un riesgo competitivo que debe considerarse. Además, mediante estos modelos de mixtura no solo identificamos a una subpoblación de pacientes curados, sino que también proporcionamos información sobre su supervivencia, la cual a su vez se emplea en esta identificación.

Con respecto a la metodología propuesta, es necesario destacar que el enfoque principal del análisis realizado para los escenarios simulados es evaluar con qué precisión es capaz el algoritmo de identificar una subpoblación con un perfil de supervivencia diferente. En particular, un grupo en el que no se espera que los pacientes experimenten la enfermedad, es decir, un grupo de individuos curados. Es por eso que no se han presentado estimaciones con respecto a la progresión de la enfermedad a la muerte, ya que estas no intervienen de ninguna manera en las transiciones clave para la curación y la susceptibilidad, pudiendo estimarse los parámetros de esta transición de la misma forma que en un modelo común de enfermedad-muerte. Además, la curación después de la enfermedad no se considera en nuestro modelo, ya que por simplicidad se define que los individuos enfermos solo progresan hasta la muerte. Otros estudios podrían incluir la misma estructura de enfermedad-muerte después de la enfermedad y evaluar la curación después de esta nueva recurrencia. De hecho, esta estructura se puede definir de forma secuencial y siendo esta la razón principal por la que nuestro análisis se centra en la primera parte del modelo, ignorando la muerte después de la enfermedad, dado que sería un proceso análogo.

De manera similar, la parte de simulación del estudio no incluye resultados sobre la estimación de la probabilidad de cero, ya que es independiente de la curación y se estima en base a una variable totalmente observada. Sin embargo, sigue siendo una parte importante del modelo y su estimación se ilustra en el caso práctico basado en el estudio real sobre fractura de cadera.

Por otro lado, los análisis se han realizado asumiendo modelos paramétricos Weibull. Esta condición se ha asumido para el proceso de simulación y para el ajuste, con modelos de riesgos proporcionales de Cox y funciones de riesgo basales Weibull. Se podría utilizar otras distribuciones [87] o modelos paramétricos flexibles, para simular el tiempo hasta el evento, como por ejemplo modelos basados en splines cúbicos restringidos. Además, es posible abordar la no proporcionalidad de los riesgos, al incluir una interacción entre el tiempo y algunas de las covariables, así como incluir efectos no lineales de las mismas. Sería posible incluso introducir funciones de riesgo (o de log-riesgo) definidas por el usuario mediante el propio paquete `simsurv` [86].

Finalmente, cabe mencionar que el uso de INLA para aproximar las distribuciones *a posteriori* de los parámetros y, por ende, las incidencias acumuladas o probabilidades de transición, destaca como punto fuerte del trabajo presentado, dada su eficiencia en

comparación con los métodos MCMC dentro del marco bayesiano. Así pues, resulta relevante mencionar que a partir de la metodología propuesta, basada en un algoritmo de muestreo de Gibbs, se ha logrado extender la clase de modelos que se pueden ajustar mediante INLA, proponiendo así una metodología más rápida que permita responder simultáneamente cuestiones sobre enfermedad, muerte y curación.

Capítulo 5

Conclusiones y líneas de investigación futuras

En este último capítulo se presentarán las principales conclusiones que se derivan de nuestro trabajo, algunas ya mencionadas en los capítulos relacionados con la temática correspondiente. Además, se finalizará incluyendo algunas posibles líneas de investigación futuras en las que se pudiese profundizar.

5.1. Conclusiones

Como se ha mencionado durante toda esta memoria, nuestro trabajo se ha visto motivado por un estudio real formado por pacientes con una fractura de cadera. Es por esta razón que las metodologías planteadas buscan ofrecer respuestas a unas preguntas científicas concretas, las cuales nos han llevado a estudiar y profundizar en el marco de los modelos multiestado. Así pues, las distintas propuestas metodológicas que hemos hecho dentro de este marco multiestado bayesiano no se limitan a nuestro estudio, sino que pretenden ser útiles para otros problemas del mundo real. A lo largo de esta memoria se ha presentado los modelos multiestado como un marco metodológico útil y flexible, ideas sobre las que queremos profundizar en este apartado.

Respecto a su utilidad, en el Capítulo 2 se han presentado resultados interesantes desde una perspectiva no solo estadística sino también epidemiológica. Con su aplicación a la cohorte PREV2FO y al estudio de la fractura de cadera recurrente ha sido posible responder a cuestiones relativas a la supervivencia de los pacientes. Se han estimado *hazard ratios* para valorar la asociación entre los distintos riesgos y covariables, así como funciones de incidencia acumulada y probabilidades de transición, las cuales han aportado información relevante sobre la evolución temporal de los pacientes fracturados y su probabilidad de sufrir refracturas o fallecer. La presentación conjunta de HR e incidencias acumuladas es por sí misma reveladora, ya que pese a no encontrar diferencias entre hombres y mujeres por lo que respecta al riesgo de refractura, las mujeres se refracturan más que los hombres, como así lo indican las incidencias acumuladas. De los resultados

del modelo se deduce que esto se debe a que los hombres, al presentar mayores riesgos e incidencias de muerte, no llegan a refracturarse; pero es la muerte, y no una mayor fragilidad de las mujeres, la causante. Además, las probabilidades de transición muestran de forma dinámica la progresión de los pacientes por los distintos estados; refractura y muerte compiten inicialmente, pero posteriormente se analiza también el riesgo de muerte en los pacientes refracturados. Si bien los resultados mostrados deben entenderse como relaciones de asociación, ejemplifican claramente la potencialidad y profundidad de estos modelos en cuanto a mejorar la comprensión que se tiene de un problema o cuestión real.

Por otro lado, como se ha mencionado anteriormente, el marco multiestado es un marco flexible. No solo permite dar respuesta a cuestiones propias del ámbito de la supervivencia, sino que puede incluir elementos de otras áreas, como efectos aleatorios espaciales, que aporten una dimensión totalmente diferente y a su vez complementaria. Esto es precisamente lo que se ha hecho en el Capítulo 3. Al incluir información acerca del Área de Salud de la Comunitat Valenciana en la que los pacientes fueron ingresados y dados de alta, se amplía sustancialmente el conocimiento que se tiene sobre la epidemiología de la fractura de cadera, sus recurrencias y la mortalidad. Una información con carácter espacial que complementa los resultados relativos a la supervivencia de estos pacientes, permitiendo identificar zonas con mayores o menores riesgos de refractura, muerte sin refractura y muerte tras refractura. El punto fuerte de la metodología propuesta es que considera conjuntamente supervivencia y correlación espacial, de forma que tanto transiciones como regiones se encuentran conectadas entre sí dentro del modelo multivariante de Leroux definido para los efectos aleatorios.

El Capítulo 4 se dedica a ampliar una cuestión relevante del ámbito de la supervivencia, como lo es la curación, pero propuesta dentro del marco de los modelos multiestado. También se incluye la cero-inflación de la misma forma, ya que la inclusión de estos elementos como estados del modelo resulta natural. La cero-inflación supone la consideración de unos individuos con tiempos cero, sobre los que no se tiene seguimiento, como es el caso de la muerte intrahospitalaria tras fractura de cadera. Para este grupo de individuos solo puede estimarse la probabilidad de muerte, si bien puede compararse al inicio y tras la enfermedad. Respecto a la curación, en un estudio en el que se considerase únicamente el tiempo hasta la enfermedad, se obtendría de los individuos curados una probabilidad, la probabilidad de curación, análogamente al caso cero-inflado. No obstante, en el marco multiestado ha sido posible considerar el riesgo de muerte en pacientes curados, riesgo que incluso participa en el cálculo de la probabilidad de ser susceptible que se realiza iterativamente. En este capítulo, se ha propuesto una metodología que permite identificar un grupo de individuos con un perfil de supervivencia distinto como lo son los pacientes curados, ofreciendo además información sobre su progresión.

Finalmente, cabe mencionar que la inferencia bayesiana es la que ha ofrecido el marco metodológico estadístico que ha permitido desarrollar todas las cuestiones expuestas anteriormente. Un marco amplio y también flexible, pero sobre todo natural, en el que es posible incrementar la complejidad de los modelos para responder a nuevas preguntas científicas sin perder interpretabilidad, ya que los resultados se expresan en términos de

probabilidad. En cuanto a los métodos computacionales, INLA se postula como una buena opción para el tratamiento de los modelos multiestado bayesianos. En esta memoria se han propuesto metodologías para el análisis de estos modelos multiestado, ampliando el número de modelos posibles en INLA mediante efectos espaciales multivariantes y con un algoritmo basado en muestreo de Gibbs para analizar modelos multiestado con curación.

5.2. Líneas futuras

Los modelos multiestado bayesianos definen un marco amplio que ofrece un gran número de posibilidades. En esta última sección, se plantearán algunas líneas de investigación futuras en las que podría ser interesante profundizar.

Por lo que respecta al estudio real de fractura de cadera recurrente, la inclusión de más covariables e información sobre tratamiento con antiosteoporóticos sería relevante para profundizar en el conocimiento que se tiene acerca de esta condición médica. Asimismo, se ha empleado distribuciones Weibull para los tiempos de transición, pudiendo en estudios futuros considerarse otras distribuciones más flexibles.

En el Capítulo 3 se emplea un modelo multivariante de Leroux para los efectos aleatorios. No obstante, en el futuro se podría explorar otras estructuras de correlación que diesen lugar a modelizaciones diferentes. Por otro lado, se ha asumido un efecto fijo de la covariable sexo sobre las funciones de riesgo, dentro de un modelo de riesgos proporcionales. Sin embargo, esta asunción de proporcionalidad de los riesgos podría flexibilizarse, por ejemplo, definiendo un modelo multiestado para las mujeres y otro para los hombres. De esta forma, además, se dispondría de unos efectos aleatorios diferentes para cada grupo, lo que podría llevar a patrones espaciales distintos en mujeres y hombres, en caso de que así lo indicasen los datos. Otra posibilidad sería directamente profundizar en el desarrollo de metodologías para evaluar la proporcionalidad de los riesgos en el ámbito bayesiano.

La curación considerada en el Capítulo 4 se ha restringido a la primera parte del modelo, si bien podría incluirse también después de la enfermedad, dada la potencial estructura recursiva del modelo. Esto podría ayudar a entender mejor, dentro del ámbito del problema real al que se aplicase el modelo, el mecanismo por el cual se da la curación, comparando antes y después de la enfermedad/recurrencia.

Por otro lado, sería relevante estudiar la aplicación conjunta de modelos multiestado y datos longitudinales, lo que plantearía diversas cuestiones interesantes y probablemente motivaría el desarrollo de nuevas metodologías. La introducción de covariables dependientes del tiempo podría formar parte de esta línea, lo que a su vez se relacionaría con la no proporcionalidad de los riesgos.

Finalmente, cabe destacar que el marco metodológico planteado es un marco general y que no se limita a nuestro estudio real. Podría ser de utilidad para el estudio de otras enfermedades, o incluso fuera del ámbito epidemiológico, lo que a su vez plantearía nuevas cuestiones cuya resolución podría implicar un desarrollo metodológico estadístico.

Bibliografía

[1] Tuyl F, Gerlach R, Mengersen K. A Comparison of Bayes-Laplace, Jeffreys, and Other Priors. *Am Stat* 2008; **62**:40–44.

[2] Greenland S. Generalized Conjugate Priors for Bayesian Analysis of Risk and Survival Regressions. *Biometrics* 2003; **59**:92–99.

[3] Roos M, Martins TG, Held L, Rue H. Sensitivity Analysis for Bayesian Hierarchical Models. *Bayesian Anal* 2015; **10**:321–349.

[4] Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol* 1972; **34**: 87–220.

[5] Ibrahim JG, Chen MH and Sinha D. *Bayesian Survival Analysis*. 1th ed. New York, NY: Springer, 2001.

[6] Lau B, Cole SR and Gange SJ. Competing Risk Regression Models for Epidemiologic Data. *Am J Epidemiol* 2009; **170**: 244-256.

[7] Andersen K, Geskus RB, de Witte T, et al. Competing risks in epidemiology: possibilities and pitfalls *Int J Epidemiol* 2012; **41**: 861-870.

[8] Wei S, Tian J, Song X, et al. Causes of death and competing risk analysis of the associated factors for non-small cell lung cancer using the Surveillance, Epidemiology, and End Results database. *J Cancer Res Clin Oncol* 2018; **144**: 145-155.

[9] Schuster NA, Hoogendijk EO, Kok AAL, et al. Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. *J Clin Epidemiol* 2020; **122**: 42-48.

[10] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; **94**: 496-509.

[11] Lambert PC, Thompson JR, Weston CL, et al. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007; **8**: 576–594.

[12] Yu B, Tiwari RC, Cronin KA, et al. Cure fraction estimation from the mixture cure models for grouped survival data. *Stat Med* 2004; **23**: 1733–1747.

[13] Besag J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *J R Stat Soc Series B Stat Methodol* 1974; **36**: 192–236.

[14] Banerjee S, Carlin BP, Gelfand AE. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[15] Ver Hoef JM, Cressie N. Spatial statistics: analysis of field experiments. *Design and Analysis of Ecological Experiments* 1993; 319–341.

[16] Besag J, York J and Mollié A. Bayesian image restoration with two applications in spatial statistics. *Ann Inst Stat Math* 1991; **43**: 1–59.

[17] Riebler A, Sørbye SH, Simpson DP, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat Methods Med Res* 2016; **25**:1145–1165.

[18] Leroux BG, Lei X and Breslow N. Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. In: Halloran ME and Berry D (eds) *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York, NY: Springer, 2000, pp.179–191.

[19] Andersen PK and Keiding N. Multi-State Models for Event History Analysis. *Stat Methods Med Res* 2002; **11**: 91-115.

[20] Commenges D. Multi-state Models in Epidemiology. *Lifetime Data Anal* 1999; **5**: 315-327.

[21] Beyersmann J, Wolkewitz M, Allignol A, et al. Application of multistate models in hospital epidemiology: advances and challenges *Biom J* 2011; **53**: 332-50.

[22] Hill M, Lambert PC and Crowther MJ. Relaxing the assumption of constant transition rates in a multi-state model in hospital epidemiology. *BMC Med Res Methodol* 2021; **21**: 16.

[23] Llopis-Cardona F, Armero C, Hurtado I, et al. Incidence of Subsequent Hip Fracture and Mortality in Elderly Patients: A Multistate Population-Based Cohort Study in Eastern Spain. *J Bone Miner Res* 2020; **37**: 1200-1208.

[24] Llopis-Cardona F, Armero C, Sanfélix-Gimeno G. Estimating disease incidence rates and transition probabilities in elderly patients using multi-state models: a case study in fragility fracture using a Bayesian approach. *BMC Med Res Methodol* 2023; **23**: 40.

[25] Meira-Machado L, de Uña-Alvarez J, Cadarso-Suárez C, et al. Multi-state models for the analysis of time-to-event data. *Stat Methods Med Res* 2009; **18**: 195-222.

[26] Alvares D, Lázaro E, Gómez-Rubio V, et al. Bayesian survival analysis with BUGS. *Stat Med* 2021; 1-46.

[27] Plummer M. rjags: Bayesian Graphical Models using MCMC. R package version 4-9; 2019.

[28] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2022.

[29] Llopis-Cardona F, Armero C, Sanfélix-Gimeno G. A Bayesian multivariate spatial approach for illness-death survival models. *Stat Methods Med Res* 2023; **32**: 1633–1648.

[30] Andersen PK, Abildstrom SZ and Rosthøj S. Competing risks as a multi-state model. *Stat Methods Med Res* 2002; **11**: 203–215.

[31] Kneib T and Hennerfeind A. Bayesian semi parametric multi-state models. *Stat Model* 2008; **8**: 169–198.

[32] Le-Rademacher JG, Therneau TM and Ou FS. The Utility of Multistate Models: A Flexible Framework for Time-to-Event Data. *Curr Epidemiol Rep* 2022; **9**: 183–189.

[33] Andersen PK and Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res* 2002; **11**: 91–115.

[34] Vejakama P, Ingsathit A, McEvoy M, et al. Progression of chronic kidney disease: an illness-death model approach. *BMC Nephrol* 2017; **18**: 205.

[35] Armero C, Cabras S, Castellanos ME, et al. Bayesian analysis of a disability model for lung cancer survival. *Stat Methods Med Res* 2016; **25**: 336–351.

[36] Kuhn J, Olié V, Grave C, et al. Estimating the Future Burden of Myocardial Infarction in France Until 2035: An Illness-Death Model-Based Approach. *Clin Epidemiol* 2022; **14**: 255–264

[37] Christensen R, Johnson W, Branscum A, et al. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. 1st ed. CRC Press, 2010.

[38] Lázaro E, Armero C and Alvares D. Bayesian regularization for flexible baseline hazard functions in Cox survival models. *Biom J* 2021; **63**: 7–26.

[39] Collett D. *Modelling Survival Data in Medical Research*. 3rd ed. Chapman and Hall/CRC, 2014.

[40] Banerjee S, Wall MM and Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics* 2003; **4**: 123–142.

[41] Carlin BP and Banerjee S. Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In: Bernardo JM , Berger JO , Dawid AP , Smith AFM (eds). *Bayesian Statistics 7*. Oxford: Oxford University Press , 2003, pp.45–63.

[42] Nathoo FS and Dean CB. Spatial multistate transitional models for longitudinal event data. *Biometrics* 2008; **64**: 271–279.

[43] Eberly LE and Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Stat Med* 2000; **19**: 2279–2294.

[44] Rue H, Martino S and Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Series B Stat Methodol* 2009; **71**: 319–392.

[45] Gilks WR, Richardson S and Spiegelhalter D (eds). *Markov Chain Monte Carlo in Practice*. 1st ed. Chapman and Hall/CRC, 1995.

[46] Martino S, Akerkar R and Rue H. Approximate bayesian inference for survival models. *Scand J Stat* 2011; **38**: 514–528.

[47] Van Niekerk J, Bakka H and Rue H. Competing risks joint models using R-INLA. *Stat Model* 2021; **21**: 56–71.

[48] Van Niekerk J, Bakka H and Rue H. A principled distance-based prior for the shape of the Weibull model. *Stat Probab Lett* 2021; **174**: 109098.

[49] Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*. 3rd ed. Chapman and Hall/CRC, 2013.

[50] Schuurman NK, Grasman RPPP and Hamaker EL. A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models. *Multivar Behav Res* 2016; **51**: 185–206.

[51] O'Malley AJ and Zaslavsky AM. Domain-Level Covariance Analysis for Multilevel Survey Data with Structured Nonresponse. *J Am Stat Assoc* 2008; **103**: 1405–1418.

[52] Lewandowski D, Kurowicka D and Joe H. Generating random correlation matrices based on vines and extended onion method. *J Multivar Anal* 2009; **100**: 1989–2001.

[53] Gómez-Rubio V. *Bayesian Inference with INLA*. Boca Raton, FL: Chapman and Hall/CRC Press, 2020.

[54] Palmí-Perales F, Gómez-Rubio V and Martinez-Beneito MA. Bayesian Multivariate Spatial Models for Lattice Data with INLA. *J Stat Softw* 2020; **98**: 1–29.

[55] Meira-Machado L and Sestelo M. Estimation in the progressive illness-death model: A nonexhaustive review. *Biom J* 2019; **61**: 245–263.

[56] Touraine C, Helmer C and Joly P. Predictions in an illness-death model. *Stat Methods Med Res* 2016; **25**: 1452–1470.

[57] Chiuchiolo C, van Niekerk J and Rue H. Joint posterior inference for latent Gaussian models with R-INLA. *J Stat Comput Simul* 2022; 1–30.

[58] Berger, J O. An Overview of Robust Bayesian Analysis. *TEST* 1994; **3**: 5–124.

[59] Ogle K and Barber JJ. Ensuring identifiability in hierarchical mixed effects Bayesian models. *Ecol Appl*, 2020; **30**.

[60] Gelman A and Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research)*. Cambridge: Cambridge University Press, 2006, p. 423.

[61] Kruschke J.*Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014, p. 187.

[62] Klein J and Moeschberger M. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer, 2003, pp. 63–90.

[63] Zhang Z and Sun J. Interval censoring. *Stat Methods Med Res* 2010; **19**: 53–70.

[64] Lamarca R, Alonso J, Gómez G, et al. Left-truncated Data With Age as Time Scale: An Alternative for Survival Analysis in the Elderly Population. *J Gerontol A Biol Sci Med Sci* 1998; **53**: M337–M343.

[65] Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH, eds. *Handbook of survival analysis*. Boca Raton:CRC Press; 2016.

[66] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26:2389-2430.

[67] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992;34:1-14.

[68] Braekers R, Grouwels Y. A semi-parametric Cox's regression model for zero-inflated left-censored time to event data. *Commun Stat - Theory Methods*. 2016;45:1969–1988.

[69] Farewell VT. The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*. 1982;38:1041–46.

[70] Li CS, Taylor JMG. A semi-parametric accelerated failure time cure model. *Stat Med* 2002;21:3235-3247.

[71] Zhang J, Peng Y. Accelerated hazards mixture cure model. *Lifetime Data Anal*. 2009;15(4):455-467.

[72] de Oliveira MR, Moreira F, Louzada F. The zero-inflated promotion cure rate model applied to financial data on time-to-default. *Cogent Econ Finance*. 2017;5:1-14.

[73] Yang Q, He H, Lu B, et al. Mixture additive hazards cure model with latent variables: Application to corporate default data. *Computational Statistics & Data Analysis*. 2022;**167**.

[74] Felizzi F, Paracha N, Pöhlmann J, Ray J. Mixture Cure Models in Oncology: A Tutorial and Practical Guidance. *Pharmacoeconomics*. 2021;51:43–155.

[75] Botta L, Gatta G, Capocaccia R, et al. Long-term survival and cure fraction estimates for childhood cancer in Europe (EUROCARE-6): results from a population-based study. *The Lancet Oncology*. 2022;**23**:1525–1536.

[76] van den Berg I, Coebergh van den Braak, RR, van Vugt JL, et al. Actual survival after resection of primary colorectal cancer: results from a prospective multicenter study. *World journal of surgical oncology*. 2021;**19**:1–10.

[77] Chen MH, Shao QM, Ibrahim JG. *Monte Carlo Methods in Bayesian Computation*. New York: Springer; 2000.

[78] Rue H, Riebler A, Sørbye SH, et al. Bayesian Computing with INLA: A Review. *Annu Rev Stat Appl* 2017; **4**, 395–421.

[79] Lázaro E, Armero C, Gómez-Rubio V. Approximate Bayesian inference for mixture cure models. *TEST*. 2020;29:750–767.

[80] Yu B, Peng Y. Mixture cure models for multivariate survival data. *Comput Stat Data An*. 2008;52:1524-1532.

[81] Basu S, Tiwari, RC. Breast cancer survival, competing risks and mixture cure model: a Bayesian analysis. *J R Stat Soc A Stat*. 2010;173:307-329.

[82] Conlon ASC, Taylor JMG, Sargent DJ. Multi-state models for colon cancer recurrence and death with a cured fraction. *Stat Med*. 2014;33:1750-1766.

[83] Beesley LJ, Taylor JMG. EM algorithms for fitting multistate cure models. *Biostatistics*. 2019;20:416-432.

[84] Gómez-Rubio V, Rue H. Markov chain Monte Carlo with the Integrated Nested Laplace Approximation. *Stat Comput* 2018; 28:1033–1051.

[85] Baker, SG. The multinomial-Poisson transformation. *J R Stat Soc: Series D (The Statistician)* 1994; **43**: 495-504.

[86] Brilleman SL, Wolfe R, Moreno-Betancur M, Crowther MJ. Simulating Survival Data Using the simsurv R Package. *J Stat Softw*. 2020;97:1-27.

[87] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24:1713-1723

Código R

Capítulo 2

Modelo de enfermedad-muerte con JAGS

```
#MULTI-STATE MODEL JAGS

weib.ms.model<- "model{

#Total likelihood

  for(k in 1:n){

    #Hazard and log-survival functions
     #Transition 1,2
      base[1,2,k] <- lambda[1,2]*alpha[1,2]*pow(t[k,1],alpha[1,2]-1)
      elinpred[1,2,k] <- exp(inprod(beta[,1,2], X[k,]))
      h[1,2,k] <- base[1,2,k]*elinpred[1,2,k]
      logsurv[1,2,k]<- -lambda[1,2]*pow(t[k,1], alpha[1,2])*elinpred[1,2,k]

     #Transition 1,3
      base[1,3,k] <- lambda[1,3]*alpha[1,3]*pow(t[k,1],alpha[1,3]-1)
      elinpred[1,3,k] <- exp(inprod(beta[,1,3], X[k,]))
      h[1,3,k] <- base[1,3,k]*elinpred[1,3,k]
      logsurv[1,3,k]<- -lambda[1,3]*pow(t[k,1], alpha[1,3])*elinpred[1,3,k]


     #Definition of the log-likelihood using zeros trick
      trans12[k]<- c[k,1]*log(h[1,2,k])+logsurv[1,2,k]
      trans13[k]<- c[k,2]*log(h[1,3,k])+logsurv[1,3,k]

      phi[k] <- 100000-(trans12[k]+trans13[k]+trans23[k])
      zeros[k] ~dpois(phi[k])
  }
```

```
  #Likelihood for transition 2,3

  for(j in fracts){

    base[2,3,j] <- lambda[2,3]*alpha[2,3]*pow(t[j,2]-t[j,1],alpha[2,3]-1)
    elinpred[2,3,j] <- exp(inprod(beta[,2,3], X[j,]))
    h[2,3,j] <- base[2,3,j]*elinpred[2,3,j]
    logsurv[2,3,j]<- -lambda[2,3]*pow(t[j,2]-t[j,1], alpha[2,3])*elinpred[2,3,j]

    trans23[j] <- c[j,3]*log(h[2,3,j])+logsurv[2,3,j]
  }

  for(p in nofracts){

      trans23[p]<- 0
  }

  #Prior distributions
    for(l in 1:Nbetas){
    beta[l,1,2] ~dnorm(0,0.01)
    beta[l,1,3] ~dnorm(0,0.01)
    beta[l,2,3] ~dnorm(0,0.01)
    }

    lambda[1,2] ~ dgamma(0.01,0.01)
    lambda[1,3] ~ dgamma(0.01,0.01)
    lambda[2,3] ~ dgamma(0.01,0.01)

    alpha[1,2] ~ dgamma(0.01,0.01)
    alpha[1,3] ~ dgamma(0.01,0.01)
    alpha[2,3] ~ dgamma(0.01,0.01)

}"


#MCMC SAMPLES (Paralleled)
#Data
d.jags <- list(n=nrow(X), t=time, X=X, c=c, zeros=rep(0,nrow(X)),
               Nbetas=ncol(X), fracts=which(c[,1]==1),
               nofracts=which(c[,1]==0))
#Parameters
p.jags <- c("beta", "alpha", "lambda")
```

```
#Wrapper for parallelization
coda.samples.wrapper <- function(j)
{
  m1 = jags.model(data=d.jags, file=textConnection(weib.ms.model),
                  inits=list(.RNG.name="base::Wichmann-Hill", .RNG.seed=j),
                  n.chains=1)
  update(m1, 1000)
  coda.samples(m1, variable.names=p.jags, n.iter=10000, thin=1)
}
```

Capítulo 3

Modelo de Leroux multivariante

```
'inla.rgeneric.multiLeroux3d.model' <- function(
  cmd = c("graph", "Q", "mu", "initial", "log.norm.const",
          "log.prior", "quit"),
  theta = NULL) {

  #Internal function
  interpret.theta <- function() {

    lambda = 1 / (1 + exp(-theta[1L]))
    prec.var = exp(theta[2L:4L])
    rho=2*exp(theta[5L:7L])/(1+exp(theta[5L:7L]))-1

    prec.matrix.var        <- diag(1/prec.var)
    prec.matrix.var[2,1] <- rho[1]/sqrt(prec.var[1]*prec.var[2])
    prec.matrix.var[3,1] <- rho[2]/sqrt(prec.var[1]*prec.var[3])
    prec.matrix.var[1,2:3] <- prec.matrix.var[2:3,1]
    prec.matrix.var[2,3]   <- rho[3]/sqrt(prec.var[2]*prec.var[3])
    prec.matrix.var[3,2]   <- prec.matrix.var[2,3]
    prec.matrix.var <- solve(prec.matrix.var)

    return(list(lambda=lambda, prec.var=prec.var,
                rho=rho,prec.matrix.var=prec.matrix.var))
  }

  graph <- function(){
    require(Matrix)
```

```
  prec.matrix.var <-matrix(nrow=3,ncol=3,1)
  prec.matrix.sp <- Diagonal(nrow(W), x = 1) + W

  return(kronecker(prec.matrix.var,prec.matrix.sp))
}

Q <- function() {
  require(Matrix)

  param <- interpret.theta()
  D <- Diagonal(nrow(W), apply(W,1,sum))

  prec.matrix.sp <- (1-param$lambda)*Diagonal(nrow(W), x = 1) +
                    param$lambda*(D-W)

  return( kronecker(param$prec.matrix.var, prec.matrix.sp) )
}

mu <- function()
{
  return(numeric(0))
}

log.norm.const <- function() {
  return(numeric(0))

}

log.prior <- function() {
  param = interpret.theta()


  res <-log(1) +log(param$lambda) + log(1 - param$lambda)+
        log(MCMCpack::dwish(W=param$prec.matrix.var,v,S=diag(rep(d.elem,3))))+
        sum(log(param$prec.var))+
        sum(log(2) + theta[5:7]-2*log(1 + exp(theta[5:7])))


  return(res)
}

initial <- function() {
```

```
    return(c(0,0,0,0,0,0,0))
  }

  quit <- function() {
    return(invisible())
  }

  res <- do.call(match.arg(cmd), args = list())
  return(res)
}
```

Estimación con INLA

```
library(INLA)
library(rgdal)
library(rgeos)
library(dplyr)
library(rlang)
library(stringr)
library(spdep)

#Adjacency matrix
load("map_depart.R")
Valencia.nb <- poly2nb(mapa_depart)
adj <- Valencia.nb
W <- as(nb2mat(adj, style = "B"), "sparseMatrix")

#Data preparation
joint.data <- data_preparation()

#Define latent effects model
source("multiLeroux3d function.R")
multiLeroux3d.model <- inla.rgeneric.define(inla.rgeneric.multiLeroux3d.model,
                                            W = W, v=7,d.elem=1)
#Formula
formula.model.re.leroux=Y~-1+beta0+sex1+sex2+age1+age2+sex23+age23+
                            f(dep.aux, model=multiLeroux3d.model)
#INLA call
multistate.multiLeroux.wishart1.full <- inla(formula.model.re.leroux,
                     family=c("weibullsurv", "weibullsurv", "weibullsurv"),
                     data=joint.data, control.compute = list(dic=TRUE, config=TRUE),
                     control.predictor = list(compute = TRUE), verbose=T)
```

```
#Summary and plot of random effects
summary(multistate.multiLeroux.wishart1.full)

mapa_depart$multiLeroux_whishart1_1 <-
        multistate.multiLeroux.wishart1.full$summary.random$h.area[1:24, "mean"]
mapa_depart$multiLeroux_whishart1_2 <-
        multistate.multiLeroux.wishart1.full$summary.random$h.area[25:48, "mean"]
mapa_depart$multiLeroux_whishart1_3 <-
        multistate.multiLeroux.wishart1.full$summary.random$h.area[49:72, "mean"]

spplot(mapa_depart, c("multiLeroux_whishart1_1","multiLeroux_whishart1_2",
       "multiLeroux_whishart1_3"), col.regions=rev(Paleta(16)))
```

Capítulo 4

Muestreo de Gibbs e INLA

```
INLAGibbs_function <- function(all.susceptible.sim, dataset_full,  niter,
                               z.last, seed,p,burn){

  set.seed(seed)
  ##################
  ##  ALGORITHM  ##
  ##################

  ##########
  # STEP 0.#
  ##########
  all.sus <- all.susceptible.sim
  #all.sus$eventtime <- all.sus$eventtime/max(all.sus$eventtime)
  all.sus$event_ill   <- as.numeric(all.sus$event==1)
  all.sus$event_death <- as.numeric(all.sus$event==2)

  #z: matrix to save the different z vector configurations
  #explored.
  #z dimension: number of individuals * number of iterations
  z <- matrix(NA, nrow = nrow(all.sus), ncol = niter + 1)
  #z[,1] <- sample(0:1, nrow(bmt), rep = TRUE)
  #
  ##Not censoring observartions always  z=0 (uncured group)
  if(missing(z.last)){
    z[, 1][which(all.sus$event==1)] <- 0
```

```
  z[, 1][which(all.sus$event != 1)] <- sample(0:1, sum(all.sus$event != 1),
                                              prob= c(1-p,p),  rep = TRUE)
}
else{
  z[,1] <- z.last
}

# List of different configurations
z.list <- list()
logistic.list <- list()
survival.list.ill <-list()
survival.list.death <-list()
survival.list.death.c <-list()

n.z <- 0 #Number of different values of z altogether

z.idx <- rep(NA, niter)
#Index to indicate the z, logistic, etc. obtained in a given iteration

##############################
## USEFUL OBJECTS FOR STEP 1.#
##############################

# LOGISTIC REGRESSION MODEL

#d.logistic: array to save each iteration database.
#Note that d.logistic databases only differ in z values
#in each iteration step.
d.logistic <- all.sus
d.logistic$z <- z[, 1] #Starting point

#logistic.inla: list to save each iteration logistic model
#posterior distribution.
logistic.list <- list()
#logistic.list.zero <- list()
#mliklogistic: matrix to save the logistic model marginal
#log-likelihood.

mliktotal <- matrix(NA, nrow = 2, ncol = niter)
mliklogistic <- matrix(NA, nrow = 2, ncol = niter)

#beta1: matrix to save the modes of the conditional
#posterior marginals of the incidence model.
```

```
beta.zero <- matrix(NA, nrow = 3, ncol = niter)
beta.cur <- matrix(NA, nrow = 3, ncol = niter)

# SURVIVAL REGRESSION MODEL

#d.survival: list to save each iteration survival database.
d.survival.u <- list()
d.survival.c <- list()

#survival.inla: list to save each iteration survival model
#posterior distribution.
survival.inla.ill <- list()
survival.inla.death <- list()
survival.inla.death.c <- list()

#mliklogistic: matrix to save the survival model marginal
#log-likelihood
mliksurvival.ill     <- matrix(NA, nrow = 2, ncol = niter)
mliksurvival.death   <- matrix(NA, nrow = 2, ncol = niter)
mliksurvival.death.c <- matrix(NA, nrow = 2, ncol = niter)

#beta2 and alpha: matrix to save each iteration survival model
#posterior conditional models.
beta2.ill     <- matrix(NA, nrow = 3, ncol = niter)
beta2.death   <- matrix(NA, nrow = 3, ncol = niter)
beta2.death.c <- matrix(NA, nrow = 3, ncol = niter)

alpha.ill     <- matrix(NA, nrow = 1, ncol = niter)
alpha.death   <- matrix(NA, nrow = 1, ncol = niter)
alpha.death.c <- matrix(NA, nrow = 1, ncol = niter)



##############################
## USEFUL OBJECTS FOR STEP 2.#
##############################

x1 <- as.matrix(cbind(rep(1, nrow(all.sus)), all.sus$sex, all.sus$age))
x2 <- x1

##############################
## USEFUL OBJECTS FOR STEP 3.#
##############################
```

```
eta <- matrix(NA, nrow = nrow(all.sus), ncol = niter)
p_zero <- numeric(niter)
su  <- matrix(NA, nrow = nrow(all.sus), ncol = niter)
sc  <- matrix(NA, nrow = nrow(all.sus), ncol = niter)
hu  <- matrix(NA, nrow = nrow(all.sus), ncol = niter)
hc  <- matrix(NA, nrow = nrow(all.sus), ncol = niter)
pz  <- matrix(NA, nrow = nrow(all.sus), ncol = niter)



#Setup using index
n.z <- 0

# Just create a name for the configuration z (a better algorithm can be used)
z.id <- function(z) {
  return(paste(z, collapse = ""))
}


# Logistic proportion of zeros
logistic.inla.zero <- inla(zero ~ 1 + sex + age, family = "binomial",
                           data = dataset_full, Ntrials = 1,
                           control.predictor = list(link = 1),
                           control.fixed = list(mean.intercept = 0,
                           prec.intercept = 0.001,
                           mean = 0, prec = 0.001))

#We take intercept and coefficients for zeros
beta.zero <- logistic.inla.zero$summary.fixed[, "mean"]
# Compute zero proportion for
p_zero <- zeroprop(x1, beta.zero)



Sys.time()->start;
i <- 1
while(i <= niter) {

  print(paste0("**ITERATION: ", i))

  # CHeck wether has already been sampled
  zz <- z.id(z[, i])
```

```
    aux.idx <- which(zz == names(z.list))

    if(length(aux.idx) == 0) { #Fit models

      #########
      #STEP 1.#
      #########


      #FIT LOGISTIC REGRESSION MODEL WITH INLA
      d.logistic$z <- z[, i]

      require(INLA)

      logistic.inla <- inla(z ~ 1 + sex + age, family = "binomial",
                            data = as.data.frame(d.logistic), Ntrials = 1,
                            control.predictor = list(link = 1),
                            control.fixed = list(mean.intercept = 0,
                                                 prec.intercept = 0.001,
                                                 mean = 0, prec = 0.001))

#FIT SURVIVAL MODEL WITH INLA

d.survival.u <- subset(d.logistic, z == 0)
d.survival.c <- subset(d.logistic, z == 1)

survival.inla.ill <- inla(inla.surv(eventtime, event_ill) ~ 1 + sex + age,
                          data = d.survival.u, family = "weibullsurv",
                          control.predictor = list(link = 1),
                          control.fixed = list(mean.intercept = 0,
                                               prec.intercept = 0.001,
                                               mean = 0, prec = 0.001),
                          control.mode =  list(theta = 0.1, restart = TRUE))


survival.inla.death <- inla(inla.surv(eventtime, event_death) ~ 1 + sex + age,
                            data = d.survival.u, family = "weibullsurv",
                            control.predictor = list(link = 1),
                            control.fixed = list(mean.intercept = 0,
                                                 prec.intercept = 0.001,
                                                 mean = 0, prec = 0.001),
                            control.mode =  list(theta = 0.1, restart = TRUE))
```

```
survival.inla.death.c <- inla(inla.surv(eventtime, event_death) ~ 1 + sex + age,
                              data = d.survival.c, family = "weibullsurv",
                              control.predictor = list(link = 1),
                              control.fixed = list(mean.intercept = 0,
                                                   prec.intercept = 0.001,
                                                   mean = 0, prec = 0.001),
                                    control.mode =  list(theta = 0.1, restart = TRUE))



      #Add configuration and results
      n.z <- n.z + 1
      aux.idx <- n.z

      z.list[[n.z]] <- z[, i]
      z.idx[i] <- aux.idx
      names(z.list)[n.z] <- z.id(z[, i])
      #logistic.list.zero[[n.z]] <- logistic.inla.zero
      logistic.list[[n.z]] <- logistic.inla
      survival.list.ill[[n.z]] <- survival.inla.ill
      survival.list.death[[n.z]] <- survival.inla.death
      survival.list.death.c[[n.z]] <- survival.inla.death.c

    } else { #z has already appeared
      z.idx[i] <- aux.idx
    }

    ############
    #STEP 2.###
    ############

    #We take intercept and coefficients for cure
    beta.cur[,i]  <- logistic.list[[aux.idx]]$summary.fixed[, "mean"]

    #beta1[, i] <- logistic.list[[aux.idx]]$summary.fixed[, "mean"]
    beta2.ill[, i]   <- survival.list.ill[[aux.idx]]$summary.fixed[, "mean"]
    beta2.death[, i] <- survival.list.death[[aux.idx]]$summary.fixed[, "mean"]
    beta2.death.c[, i] <- survival.list.death.c[[aux.idx]]$summary.fixed[, "mean"]
    alpha.ill[, i]   <- survival.list.ill[[aux.idx]]$summary.hyperpar[, "mean"]
    alpha.death[, i] <- survival.list.death[[aux.idx]]$summary.hyperpar[, "mean"]
    alpha.death.c[, i] <- survival.list.death.c[[aux.idx]]$summary.hyperpar[, "mean"]
    if(is.na(alpha.death.c[,i])){
        alpha.death.c[, i] <-
```

```
        survival.list.death.c[[aux.idx]]$summary.hyperpar$`0.5quant`}
if(alpha.death.c[,i]<0){
    alpha.death.c[, i] <- mean(inla.hyperpar.sample(10000,
                              survival.list.death.c[[aux.idx]]))}


#marginal log-likelihoods to check convergence
mliklogistic[, i]       <- logistic.list[[aux.idx]]$mlik
mliksurvival.ill[, i]   <- survival.list.ill[[aux.idx]]$mlik
mliksurvival.death[, i] <- survival.list.death[[aux.idx]]$mlik
mliksurvival.death.c[, i] <- survival.list.death.c[[aux.idx]]$mlik

mliktotal[,i] <- mliklogistic[, i]+mliksurvival.ill[, i]+
                 mliksurvival.death[, i]+ mliksurvival.death.c[, i]
############
#STEP 3.###
############


# Compute cure proportion
eta[, i] <- cureprop(x1, beta.cur[, i])
# Compute survival
su[, i] <- survunc(all.sus$eventtime, x2, alpha.ill[, i],
                   alpha.death[, i], beta2.ill[, i], beta2.death[, i])
sc[, i] <- survcur(all.sus$eventtime, x2, alpha.death.c[, i],
                   beta2.death.c[, i])

# Compute hazards

hu[, i] <- hazunc(all.sus$eventtime, x2, alpha.death[, i], beta2.death[, i])
hc[, i] <- hazcur(all.sus$eventtime, x2, alpha.death.c[, i], beta2.death.c[, i])

# Sample z
#Fixed values of z
pz[, i] <- piz(all.sus$event_ill==1 , all.sus$event_death, eta[, i],
               p_zero[,i], su[, i], sc[,i], hu[,i], hc[,i])
if(sum(is.na(pz[,i]))>0){
  print("Prob. inf")

  return(list(data.frame(hu=hu[,i],hc=hc[,i],pz=pz[,i],eta=eta[,i],
                         su=su[,i],sc=sc[,i]), survival.inla.death.c))

  i<-niter+1
```

```
    }
    else{
      z[, i + 1] <- upz(pz[, i])
      i <- i + 1
    }




    print(paste0("Cured:",sum(z[,i]), " Censored-cured:",
                 sum(all.sus$event!=1)-sum(z[,i]),
                 " Prob zero:", round(mean(p_zero),3),
                 " Prob cured:", round(mean(eta[,i-1]),3),
                 " pz:", round(mean(pz[,i-1]),3)))
    if(sum(z[,i])==0){i<-niter+1}
    }
  c.time <- Sys.time() - start
  print(c.time)

  max_likely <-  burn+which.max(mliktotal[1,-c(1:burn)])

  return(list(z=z, c.time=c.time,  mliklogistic=mliklogistic,
              mliksurvival.ill=mliksurvival.ill,
              mliksurvival.death=mliksurvival.death,
              mliksurvival.death.c=mliksurvival.death.c,
              logistic.c=logistic.list[[max_likely]], mliktotal=mliktotal,
              logistic.inla.zero=logistic.inla.zero,
              survival.ill=survival.list.ill[[max_likely]],
              survival.death=survival.list.death[[max_likely]],
              survival.death.c=survival.list.death.c[[max_likely]],
              max_likely=max_likely
  ))


}
```