Trabajo Fin de Máster

Efectividad vacunal frente al SARS-CoV-2 en Galicia

Carla Guerra Tort

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Efectividade vacinal fronte ao SARS-CoV-2 en Galicia
Título en español: Efectividad vacunal frente al SARS-CoV-2 en Galicia
English title: Vaccine effectiveness against SARS-CoV-2 in Galicia
Modalidad: Modalidad B
Autora: Carla Guerra Tort, Universidade de Santiago de Compostela
Directora: Rosa María Crujeiras Casais, Universidade de Santiago de Compostela
Tutora: María Isolina Santiago Pérez, Dirección Xeral de Saúde Pública
Breve resumen del trabajo: Este trabajo tiene por objetivo estimar la efectividad de las vacunas de COVID-19 y la pérdida de inmunidad en la población gallega. Para ello, se consideran los siguientes apartados: **1.** El Coronavirus en Galicia: la pandemia de COVID-19 y la estrategia de vacunación. **2.** Introducción al Análisis de Supervivencia y a la regresión con datos censurados. **3.** Metodología: descripción de las fuentes de datos, diseño y población de estudio y técnicas analíticas. **4.** Resultados. Para finalizar, se presentan las conclusiones, fortalezas y limitaciones de los análisis y se comentan posibles líneas de trabajo futuras.

Doña Rosa María Crujeiras Casais, Profesora Titular del área de Estadística e Investigación Operativa de la Universidade de Santiago de Compostela y doña María Isolina Santiago Pérez, Técnica Estadística de la Dirección Xeral de Saúde Pública, informan que el Trabajo Fin de Máster titulado

Efectividad vacunal frente al SARS-CoV-2 en Galicia

fue realizado bajo su dirección por doña Carla Guerra Tort para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 13 de junio de 2022.

La directora:

CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Firmado digitalmente por CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-52931780V, givenName=ROSA MARIA, sn=CRUJEIRAS CASAIS, cn=CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Fecha: 2022.06.13 13:29:20 +02'00'

Doña Rosa María Crujeiras Casais

La tutora:

SANTIAGO PEREZ MARIA ISOLINA - 52496158S
Firmado digitalmente por SANTIAGO PEREZ MARIA ISOLINA - 52496158S
Fecha: 2022.06.13 12:41:48 +02'00'

Doña María Isolina Santiago Pérez

La autora:

GUERRA TORT CARLA - 45849275V
Firmado digitalmente por GUERRA TORT CARLA - 45849275V
Fecha: 2022.06.13 12:44:38 +02'00'

Doña Carla Guerra Tort

Agradecimientos

A mi directora, Rosa, por sus consejos y recomendaciones y por enseñarme a “escribir matemáticas”.

Al Servizo de Epidemioloxía de la Dirección Xeral de Saúde Pública, por su acogida y por hacerme sentir una más del grupo (nunca olvidaré esos pinchos). Le agradezco de manera especial a mi tutora, Soly, su inestimable ayuda y todo lo que me ha enseñado.

A mis padres, María Rosa y José Ramón, que siempre me han animado a estudiar y a formarme.

Y, por último, a Fernando, por su enorme paciencia, por haberme cogido fuerte de la mano y no habérmela soltado en todo este tiempo.

Índice general

Resumen

Resumen en español

El 11 de marzo de 2020 la Organización Mundial de la Salud declaró como pandemia la enfermedad por SARS-CoV-2 (COVID-19). Desde entonces y hasta el 27 de marzo de 2022, se han registrado 587.675 infecciones y más de 3.000 defunciones en Galicia. La vacunación masiva ha demostrado ser una de las principales medidas para el control de la pandemia. A medida que avanza la vacunación, es importante estimar su efectividad en el entorno real para garantizar la ganancia y mantenimiento de la inmunidad en la población. En este trabajo se presenta un estudio de cohortes para evaluar la efectividad de la vacunación frente a infección e ingreso en UCI por COVID-19 en Galicia. La población de estudio, constituida por un total de 2.129.598 individuos, se siguió del 27 de diciembre de 2020 al 27 de marzo de 2022. Se estimó la efectividad de las vacunas en individuos primovacunados frente a no vacunados y en individuos con dosis de recuerdo frente a primovacunados aplicando un modelo de Cox. También se estimó la pérdida de efectividad de la primovacunación. La efectividad ajustada de la primovacunación frente a la no inmunización fue del 27 % (IC95 %: 26,4-27,7) para infección y del 76 % (IC95 %: 75,8-76,3) para ingreso en UCI seis meses después de completar la pauta. Estos resultados sugieren que, en general, las vacunas de COVID-19 fueron efectivas frente a ambos desenlaces, observándose una pérdida de efectividad tras seis meses de un 53 % para infección y de un 17 % para ingreso en UCI.

English abstract

On March 11, 2020, the World Health Organization declared SARS-CoV-2 disease (COVID-19) a pandemic. Since then and until March, 27, 2020, 587.675 infections and more than 3.000 deaths have been reported in Galicia. Mass vaccination has proven to be one of the main measures for pandemic control. As vaccination progresses, it is important to estimate its effectiveness in the real environment to ensure the gain and maintenance of immunity in the population. In this work we present a cohort study to evaluate the effectiveness of vaccination against infection and ICU admission by COVID-19 in Galicia. The study population, consisting of 2.129.598 individuals, was followed from December 27, 2020 to March 27, 2022. Vaccine effectiveness was estimated in vaccinated versus unvaccinated individuals and in individuals with booster doses versus vaccinated individuals by applying a Cox model. The waning of effectiveness of vaccination was also estimated. The adjusted effectiveness of vaccination versus non-immunization was 27 % (95 %CI: 26,4-27,7) for infection and 76 % (95 %CI: 75,8-76,3) for ICU admission six months after completing the vaccination schedule. These results suggest that, in general, COVID-19 vaccines were effective against both outcomes, with a waning of effectiveness after six months of 53 % for infection and 17 % for ICU admission.

Prefacio

Este trabajo corresponde a una memoria sobre las prácticas realizadas en el Servizo de Epidemioloxía de la Dirección Xeral de Saúde Pública (DXSP) de la Consellería de Sanidade con el objetivo de que se considere como Trabajo Fin de Máster del Máster Universitario en Técnicas Estadísticas. El tema a tratar es la estimación de la efectividad vacunal frente a la infección por SARS-CoV-2 en Galicia.

El 11 de marzo de 2020 la Organización Mundial de la Salud (OMS) declara como pandemia la enfermedad por SARS-CoV-2 (COVID-19). En Galicia, el primer caso de COVID-19 se detecta unos días antes, el 3 de marzo. Desde entonces y hasta el 27 de marzo de 2022 se han registrado 587.675 infecciones y 3.206 fallecimientos en nuestra comunidad de acuerdo con los datos de la DXSP.

Numerosos artículos, como Wiersinga et al. (2020), Zuckerman et al. (2020) o Hu et al. (2020), describen las principales características del SARS-CoV-2, su transmisibilidad y la sintomatología de la enfermedad que provoca.

La COVID-19 es una enfermedad infecciosa causada por el SARS-CoV-2 o coronavirus del síndrome respiratorio agudo severo 2, un virus altamente transmisible y patógeno que afecta a diversos animales y al ser humano. La sintomatología y gravedad de la enfermedad son variadas, estando estrechamente relacionadas con la edad y el estado de salud del huésped. En general, las personas mayores de 60 años con comorbilidades son más propensas a desarrollar una enfermedad respiratoria grave que requiere hospitalización, o incluso puede provocar la muerte, mientras que la mayoría de jóvenes y niños suelen presentar síntomas leves o son asintomáticos. Además, la sintomatología también varía en función de la variante del virus. Algunos de los síntomas más comunes son fiebre, tos seca y fatiga; síntomas menos frecuentes son dolor de cabeza, dolor de garganta, escalofríos, diarrea, náuseas y vómitos. También se ha observado pérdida del gusto y del olfato. Los síntomas suelen aparecer tras un período de incubación de entre 1 y 14 días.

La trasmisión se produce, principalmente, a través de las gotitas líquidas que se diseminan durante el habla, al toser o al estornudar y también por aerosoles. Esta transmisión ha llevado a adoptar como principales medidas de prevención de la infección el uso de mascarilla, la distancia interpersonal y la ventilación en interiores. Con todo, la COVID-19 ha provocado un aumento de la presión hospitalaria sin precedentes y un fuerte impacto sobre la mortalidad, además de un deterioro de la situación económica y una gran repercusión sobre la vida y el comportamiento de las personas.

Con el objetivo principal de reducir la morbilidad y mortalidad causada por la COVID-19, protegiendo especialmente a aquellos grupos de la población más vulnerables, la DXSP desarrolló el "Plan galego de vacinación fronte ao SARS-CoV-2". De manera simultánea, este plan pretende reducir la presión sobre el sistema sanitario, así como permitir la recuperación social y económica.

A medida que avanza la vacunación, es importante estimar su efectividad en el entorno real para garantizar la ganancia y mantenimiento de la inmunidad en la población. La efectividad de las vacunas es una propiedad dinámica que se ve afectada por la evolución del virus y la aparición de nuevas variantes. Además, las diferencias sociodemográficas, estrategias de vacunación y tipo y características de las vacunas hacen que la efectividad vacunal sea específica para cada población.

Dado el impacto sanitario y socioeconómico de la pandemia de COVID-19 y la importancia de la vacunación como principal medida para su control, este trabajo se centra en la estimación de la efectividad vacunal (EV) y la pérdida de inmunidad en la población gallega.

El presente documento se organiza en cinco Capítulos. En el Capítulo 1 se describe la evolución de la pandemia de COVID-19 en Galicia y la estrategia de vacunación desarrollada para su control. Se definen, además, los objetivos del estudio. Seguidamente, en el Capítulo 2, se presenta una introducción al Análisis de Supervivencia y se describen en detalle las principales técnicas empleadas en los análisis. En el Capítulo 3 se detallan los métodos para cada uno de los análisis realizados y en el Capítulo 4 los resultados obtenidos. Finalmente, en el Capítulo 5 se exponen las conclusiones, fortalezas y limitaciones del estudio y se plantean posibles líneas de trabajo que esperamos poder abordar en el futuro.

Notación

- Y: variable de interés, tiempo observado hasta la ocurrencia del evento.
- $F,\ f,\ \lambda$: función de distribución, función de densidad y función de riesgo de Y, respectivamente.
- $\mathbb{P}(Y \geq y) = 1 - F(y^-)$; $\mathbb{P}(Y > y) = 1 - F(y)$.
- C: tiempo de censura.
- $G,\ g$: función de distribución y función de densidad de C, respectivamente.
- $Z = \text{mín}\{Y, C\}$.
- H: función de probabilidad acumulada de Z.
- δ: variable indicadora de censura; $\delta = \mathbb{I}\{Y \leq C\}$.
- (Z, δ): dupla de la población; (Z_i, δ_i), $1 \leq i \leq n$: duplas independientes observadas.
- $X^i = (X_{i1}, X_{i2}, \cdot, X_{iP})'$: vector traspuesto de covariables del $i - ésimo$ individuo de la muestra, con $i = 1, \cdots, n$ y $p = 1, \cdots, P$.
- $X^{(i)}$: vector de covariables del $i - ésimo$ individuo de la muestra ordenada.
- t_k, $k = 1, \cdots, K$: muestra aleatoria simple de tiempos de ocurrencia del evento de interés.
- d_k, $k = 1, \cdots, K$: número de eventos que ocurren en t_k.
- n_k, $k = 1, \cdots, K$: número de individuos a riesgo en t_k.
- $R(t_k)$, $k = 1, \cdots, K$: conjunto de individuos a riesgo en t_k.
- m_k, $k = 1, \cdots, K$: tiempos censurados en el intervalo $[t_k, t_{k+1}]$.
- ζ_k, $k = 1, \cdots, K$: cada uno de los eventos empatados en t_k.
- Q_k, $k = 1, \cdots, K$: conjunto de todos los posibles subconjuntos de ζ_k individuos que se pueden seleccionar del conjunto de riesgo $R(t_k)$.

Índice de Tablas y Figuras

Tablas

Figuras

Capítulo 1

El Coronavirus en Galicia

Este primer capítulo está centrado en la descripción de la pandemia de COVID-19 en Galicia, prestando especial atención a la estrategia de vacunación desarrollada para su control. La información indicada puede consultarse en detalle en la versión 6.3 y posteriores del "Plan galego de vacinación fronte ao SARS-CoV-2" de la Dirección Xeral de Saúde Pública (2021). Por otro lado, la información numérica proporcionada procede de las bases de datos sobre vigilancia y vacunación frente al SARS-CoV-2, descritas en la Sección 3.1 del Capítulo 3. Al final del presente capítulo, se define también el objetivo de este trabajo y los análisis realizados para su consecución.

1.1. La pandemia de COVID-19

La enfermedad por COVID-19 comenzó a extenderse por nuestro país en marzo de 2020, obligando a imponer el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el virus. El confinamiento estricto, la reducción de la movilidad y, sobre todo la vacunación, promovieron el descenso de la curva de contagios, aunque no a niveles lo suficientemente bajos como para dar por controlada la pandemia.

Hasta el 27 de marzo de 2022, los casos de COVID-19 fueron notificados a las autoridades sanitarias siguiendo la adaptación a Galicia de los criterios de la Estrategia de detección precoz, vigilancia y control de COVID-19 del Ministerio de Sanidad (2021). Para tratar de frenar la transmisión del virus y reducir la saturación de los servicios sanitarios, los esfuerzos se centraron en la detección temprana de los casos con infección activa, el establecimiento precoz de las medidas de control necesarias para evitar nuevas infecciones y la recolección de datos para la vigilancia epidemiológica y la toma de decisiones. Bajo esta estrategia se llevó a cabo una búsqueda activa y un control estricto de los casos, realizando pruebas diagnósticas a los casos sospechosos y cuarentenando a los contactos estrechos para lograr un diagnóstico temprano de la enfermedad y controlar así la transmisión del virus. Todos los casos confirmados con infección activa eran notificados con carácter obligatorio y urgente a las autoridades sanitarias pertinentes, incluyendo tanto nuevas infecciones como reinfecciones.

El 28 de marzo entró en vigor una nueva estrategia, la llamada Estrategia de vigilancia y control frente a COVID-19 tras la fase aguda de la pandemia del Ministerio de Sanidad (2022). Los altos niveles de inmunidad alcanzados en la población española determinaron un cambio en la epidemiología de la COVID-19, y las actuaciones de prevención y control pasaron a centrarse en las personas y ámbitos de mayor vulnerabilidad. Así, desde esa fecha se estableció la vigilancia solo de los casos graves y de los casos confirmados en personas de 60 años o más, inmunodeprimidos y embarazadas, o en personas asociadas a ámbitos vulnerables. Los casos confirmados fuera de estos grupos dejaron de ser de declaración obligatoria, si bien debían extremar las precauciones y reducir todo lo posible las interacciones sociales, especialmente en los días posteriores al inicio de los síntomas o al diagnóstico en el caso de las personas asintomáticas.

A lo largo de la pandemia, la tasa de incidencia acumulada a 14 días por cada 100.000 habitantes ha sido el indicador principal para evaluar la transmisión del virus. Dependiendo de su valor, se han establecido una serie de umbrales para determinar un riesgo de transmisión bajo, medio, alto o muy alto. No obstante, ha sido necesario poner en contexto el valor de esta incidencia en función del territorio y las características de la población para poder interpretar de manera adecuada las dinámicas de propagación del virus. Además, las características de los casos positivos han ido variando con el tiempo, especialmente desde el inicio de la vacunación en diciembre de 2020, a partir de la cual se ha producido un descenso de la proporción de casos graves y de la mortalidad. Por ello, los umbrales de riesgo empleados para la vigilancia de la pandemia han estado sometidos a continua revisión.

Hasta el 27 de marzo de 2022, la evolución de la pandemia ha estado marcada por un total de cinco períodos y cinco olas. Podemos definir una ola como un espacio de tiempo caracterizado por un número creciente de casos, que alcanza un máximo para, a continuación, descender más lentamente. En todos los períodos de la pandemia se ha observado una ola de contagios. En algunos casos, el inicio de una ola se ha asociado a la aparición de una nueva variante del virus. En otros, sin embargo, la principal causa del ascenso de casos ha sido la relajación de las medidas de prevención, como en los períodos vacacionales. En la Tabla 1.1 se muestran, para cada período, su duración, el número total de casos y la media diaria de casos. Además, para el total de casos se indican cuántos fueron ingresos, ingresos en unidades de cuidados intensivos (UCI) y defunciones, con sus respectivos porcentajes. Debe tenerse en cuenta que en el primer período de pandemia el número de casos y su media diaria están infraestimados, dado que apenas se realizaban pruebas diagnósticas. Por el contrario, el porcentaje de casos que ingresaron, tanto en hospital como en UCI, así como el porcentaje de defunciones, se encuentran sobreestimados.

Período	**Inicio**	**Fin**	**Días**	**Casos**		**Casos ingreso**		**Casos UCI**		**Casos defunción**	
				Total	**Media diaria**	**N**	**%**	**N**	**%**	**N**	**%**
1	03-03-2020	12-06-2020	101	10.890	107,8	2.824	26,3	318	2,9	623	5,7
2	13-06-2020	27-12-2020	197	51.184	259,8	4.521	8,8	559	1,1	843	1,6
3	28-12-2020	21-06-2021	175	68.668	392,4	6.330	9,2	996	1,4	970	1,4
4	22-06-2021	22-11-2021	153	61.669	403,1	2.341	3,8	314	0,5	241	0,4
5	23-11-2021	27-03-2022	124	395.240	3.187,4	6.379	1,6	519	0,1	559	0,1

Tabla 1.1: Períodos que han marcado la evolución de la pandemia de COVID-19 en Galicia hasta el 27 de marzo de 2022. Fuente: DXSP. Elaboración propia.

Siguiendo la Tabla 1.1, vemos que el primer período se extiende desde principios de marzo de 2020 hasta mediados de junio, momento en el que finalizó el confinamiento estricto y el país pasó a la denominada "nueva normalidad". El segundo período se define desde ese momento hasta el inicio de la vacunación, el 27 de diciembre de ese mismo año. A continuación, y hasta finales de junio de 2021, tiene lugar la tercera etapa de la pandemia, coincidiendo con la llegada y expansión de Alpha. Seguidamente, la entrada de Delta supuso el comienzo de un nuevo período, con una nueva ola de casos similar a la anterior. Aproximadamente a mediados de noviembre, la aparición de Ómicron motivó el inicio del quinto y último período considerado. Para nuestro estudio, como fecha de fin de este período hemos tomado el último día en el que todos los casos de infección activa de SARS-CoV-2 seguían siendo notificados a las autoridades sanitarias (27 de marzo de 2022).

Las variantes del SARS-CoV-2 responsables de las olas de mayor incidencia han sido las denominadas Alpha, Delta y Ómicron. Estas variantes se han caracterizado por su gran rapidez de propagación, posiblemente debido a la acumulación de mutaciones que les han conferido una ventaja evolutiva. Por ello, han sido capaces de desplazar al resto de variantes en un marco de competición y alta circulación del virus. Alpha llegó a nuestra comunidad en diciembre de 2020, y el 1 de marzo de 2021 era ya la responsable de más del 90 % de los casos detectados mediante PCR. Esta variante siguió dominando hasta casi principios de julio, cuando entró Delta, la principal responsable de los casos acontecidos durante los meses de verano y otoño. La transición entre Delta y la última variante de especial interés, Ómicron, se produjo hacia finales de año. Desde entonces, Ómicron ha sido la variante principal,

representando el 100 % de las PCR positivas a 27 de marzo de 2022.

En la Figura 1.1 se muestran los porcentajes de representación de las variantes entre los casos detectados mediante PCR, por semanas, desde el 18 de noviembre de 2020 hasta finales de marzo de 2022. Se distinguen claramente los períodos de dominio de cada variante (representación del 50 % o más), así como la gran rapidez en las transiciones. Alpha dominó hasta comienzos de julio de 2021, si bien nunca alcanzó una representación del 100 % sobre todas las PCR positivas. Seguidamente, Delta irrumpió en el escenario de la pandemia y en pocas semanas se convirtió en la variante principal, manteniéndose en porcentanjes superiores al 90 % hasta finales de año. Ya en el mes de diciembre, Delta se vio desplazada por Ómicron, que fue la responsable de la ola de casos acontecida en el quinto período de la pandemia. Si bien Delta dominó claramente durante la segunda mitad de 2021 y fue más transmisible que Alpha, los porcentajes de Ómicron en sus meses como variante dominante son prácticamente del 100 %. Este hecho, junto con la incidencia acumulada registrada para la quinta ola, lleva a pensar que Ómicron es la variante de SARS-CoV-2 con mayor capacidad de propagación.

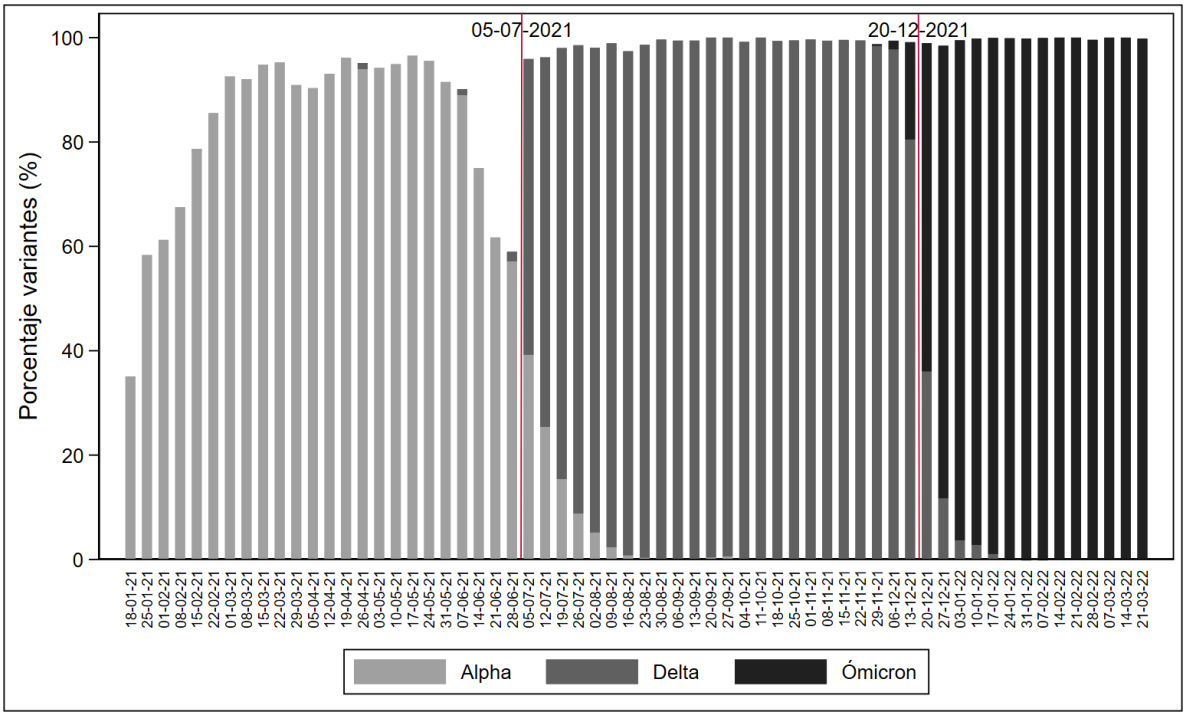

Figura 1.1: Porcentaje de representación de los casos diagnosticados por PCR, por semanas, para las principales variantes del SARS-CoV-2. Las líneas rojas verticales indican las fechas de cambio de la variante dominante. Fuente: DXSP. Elaboración propia.

A continuación, se presentan unos gráficos con las series diarias de casos y fallecimientos acumulados a 14 días por cada 100.000 habitantes en la población gallega.

La Figura 1.2 muestra la incidencia acumulada. Para cada período se observa una subida más o menos acusada del número de casos y un posterior descenso. Durante los dos primeros períodos de la pandemia, la incidencia acumulada se mantuvo por debajo de los 500 casos, si bien en el primer período sus valores están infraestimados, como ya se indicó. En la tercera fase, justo después del inicio de la vacunación y cuando ya se hacían pruebas a todos los casos sospechosos y a sus contactos estrechos, se produjo un repunte de los contagios, con valores de incidencia rozando los 1.000 casos a 14 días por cada 100.000 habitantes. Esta tercera ola comenzó tras el período navideño, en el que se relajaron las restricciones, y justó después de la llegada de Alpha a nuestra comunidad, lo que explica esta subida de la incidencia aún con el inicio de la inmunización. Alpha produjo un rápido aumento de los

casos en las primeras semanas, dando lugar a una ola que alcanzó su máximo a comienzos de febrero. A partir de marzo y hasta finales del período, la incidencia se mantuvo más o menos estable y por debajo de los 250 casos. En el mes de julio comenzó otra ola con un aumento de los casos similar al observado en la fase anterior. Este aumento fue debido a múltiples factores, como las vacaciones de verano, la pérdida de inmunidad de los primeros colectivos vacunados y, especialmente, a la llegada de Delta y al hecho de que los más jóvenes estaban sin vacunar. Más adelante, en la Figura 1.4, se puede observar cómo las personas entre 15 y 24 años fueron las más afectadas en esta ola. Ya en el quinto período, marcado por la aparición de Ómicron, se registraron las mayores tasas de incidencia de toda la pandemia, triplicando los valores de las fases anteriores y alcanzado un máximo de 3.500 casos a 14 días por cada 100.000 habitantes.

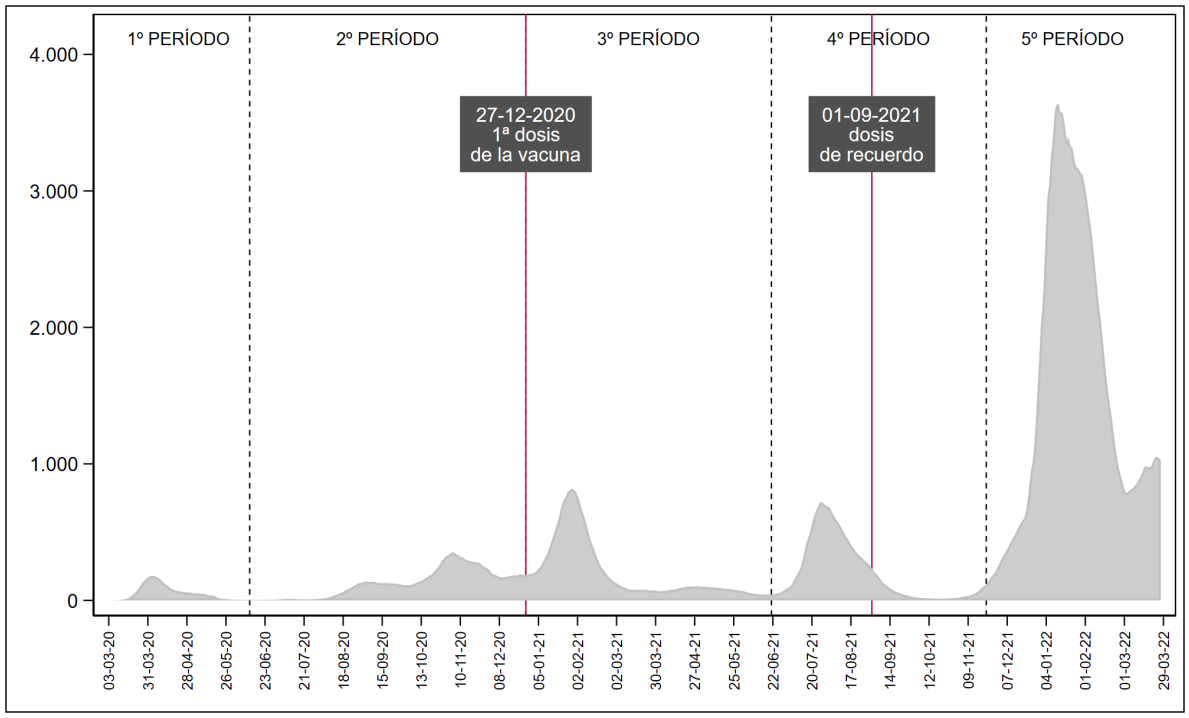

Figura 1.2: Incidencia acumulada a 14 días por cada 100.000 habitantes desde el inicio de la pandemia y hasta el 27 de marzo de 2022. Las líneas grises punteadas delimitan los cinco períodos de la pandemia y las líneas continuas en color rojo marcan el inicio de la vacunación y de la administración de la dosis de recuerdo. Fuente: DXSP. Elaboración propia.

En la Figura 1.3 se representa el número acumulado de fallecimientos por COVID-19. Los valores más altos se sitúan al comienzo de la pandemia y en el tercer período. Las medidas de prevención adoptadas para reducir el número de contagios, así como la protección de los colectivos más vulnerables motivaron, en febrero de 2021, un descenso de la mortalidad. A partir de entonces y tras el inicio de la inmunización, el número de fallecimientos a 14 días por cada 100.000 habitantes se ha mantenido por debajo de cinco.

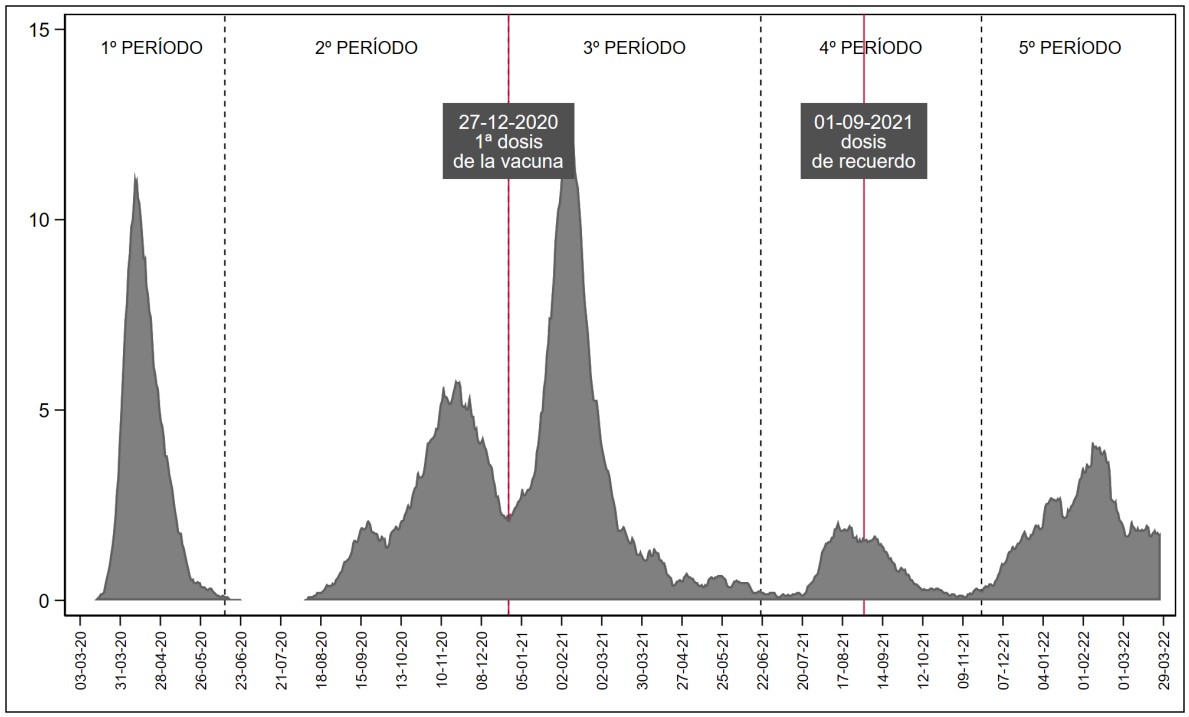

Figura 1.3: Número acumulado de fallecimientos a 14 días por cada 100.000 habitantes desde el inicio de la pandemia y hasta el 27 de marzo de 2022. Las líneas grises punteadas delimitan los cinco períodos de la pandemia y las líneas continuas en color rojo marcan el inicio de la vacunación y de la administración de la dosis de recuerdo. Fuente: DXSP. Elaboración propia.

La Figura 1.4 muestra la incidencia acumulada a 14 días por cada 100.000 habitantes por grupos de edad. En general, se observa un perfil de incidencia similar en todos los grupos, con picos en febrero y julio de 2021 y un gran repunte de los casos a comienzos del 2022. No obstante, parece que en las primeras olas se vieron más afectados los grupos de mayor edad, a partir de los 45 años, mientras que en las últimas fases de la pandemia y en la quinta ola los contagios se produjeron, principalmente, entre los más jóvenes. Las personas entre 15 y 24 años, como ya habíamos anunciado, fueron las más afectadas por la cuarta ola, acontecida en agosto de 2021. También destaca la incidencia acumulada en la última ola en este grupo y en el de 0 a 14 años, con valores próximos a los 8.000 casos. Esto se relaciona con la estrategia de vacunación por grupos de edad, en la que los más jóvenes fueron los últimos en vacunarse, estando más expuestos en los períodos tardíos de la pandemia. Con todo, a los colectivos de mayor edad también parece haberles afectado la quinta ola, pues, aunque de manera mucho menos acusada, en todos ellos se observa un repunte de los casos. En esta ocasión, la explicación podría ser la alta capacidad de Ómicron para eludir la inmunidad, así como una pérdida de esta con el paso del tiempo en los individuos que antes se vacunaron.

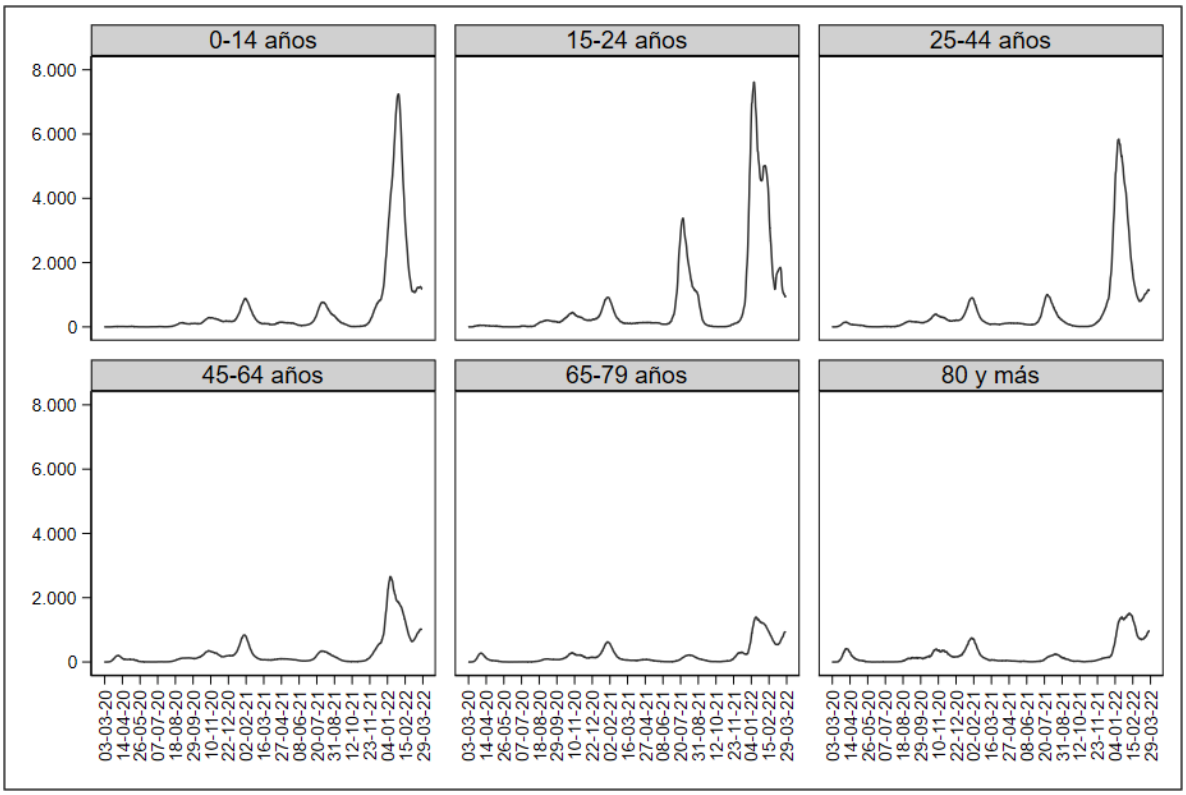

Figura 1.4: Incidencia acumulada a 14 días por cada 100.000 habitantes, por grupos de edad, desde el inicio de la pandemia y hasta el 27 de marzo de 2022. Fuente: DXSP. Elaboración propia.

1.2. Estrategia de vacunación frente al SARS-CoV-2

Desde la detección del primer caso de COVID-19 en Galicia, el 3 de marzo de 2019, el abordaje de la enfermedad por SARS-CoV-2 ha ido variando con el tiempo. La falta de información inicial sobre el virus y sus mecanismos de transmisión derivaron en pautas de tratamiento durante las primeras olas que han ido cambiando al aumentar el conocimiento sobre la enfermedad. En este sentido, lograr la inmunización de la población a través de la vacunación ha demostrado ser la medida más efectiva para el control de la pandemia.

En Galicia la vacunación se inició el 27 de diciembre de 2020 con el objetivo de reducir la morbilidad y mortalidad causada por la COVID-19. El plan de vacunación desarrollado se ha centrado en la protección de los colectivos más vulnerables de la población, así como en la reducción de la presión sobre el sistema sanitario. Es importante destacar que la inmunización frente a la COVID-19 ha sido y sigue siendo de carácter universal, accesible para toda la población sin excepción.

Debido a la limitación inicial en la disponibilidad de vacunas, se estableció un programa de vacunación en etapas progresivas, priorizando a los grupos de población en función de su vulnerabilidad, riesgo de padecer una enfermedad por COVID-19 grave y riesgo de transmisión. Así, la inmunización comenzó con los usuarios de residencias de mayores y de personas con discapacidad, personal sanitario, grandes dependientes y cuidadores profesionales. A continuación, se vacunó a los mayores de 65 años no residentes en centros de mayores, a los colectivos en activo con una función esencial para la sociedad (fuerzas y cuerpos de seguridad del estado, docentes y personal de educación infantil, primaria y secundaria) y a las personas con condiciones de alto riesgo (inmunodeprimidos, trasplantados, enfermos oncológicos, pacientes de diálisis, ...). Finalmente, se inmunizó al resto de la población en función de su año de nacimiento, estableciendo grupos de edad decenales. Esta priorización se fue adaptando de

manera periódica a la logística y necesidades específicas de la situación epidemiológica, utilizando de manera simultánea las vacunas disponibles y permitiendo el solapamiento de los grupos.

Son cuatro las vacunas aprobadas por la Agencia Europea del Medicamento que han sido administradas a la población gallega, y todas ellas se aplican por vía intramuscular. Las vacunas *Comirnaty* y *Spikevax* están compuestas por ARN mensajero, que codifica una proteína de superficie específica del SARS-CoV-2 llamada proteína S. Las vacunas *Vaxveria* y *Janssen*, por el contrario, contienen un vector viral con el gen que codifica la proteína S. En cualquier caso, es la maquinaria celular del huésped la encargada de sintetizar el antígeno.

- *Comirnaty*® de laboratorios Pfizer y BioNTech (aprobada el 21 de diciembre de 2020). La pauta recomendada es de dos dosis separadas por un intervalo de 21 días (mínimo 19). El período de inducción [1] es de siete días. La vacuna dispone de dos tipos de presentación, una para mayores de 12 años y otra para niños entre cinco y 11 años.

- *Spikevax COVID-19 Vaccine*® de Moderna (aprobada el 6 de enero de 2021). La pauta recomendada es de dos dosis separadas por un período de 28 días (mínimo 25). El período de inducción es de 14 días. Está recomendada para mayores de 12 años.

- *Vaxzevria COVID-19 Vaccine AstraZeneca*® de AstraZeneca (aprobada el 29 de enero de 2021). La pauta recomendada es de dos dosis con un intervalo entre dosis de cuatro a 12 semanas (mínimo 21 días). El período de inducción es de 14 días y se recomienda su administración en personas a partir de 18 años (de acuerdo con la última actualización del plan de vacunación, ya que la edad recomendada fue variando con el tiempo).

- *COVID-19 Vaccine Janssen*® de Johnson & Johnson (aprobada el 11 de marzo de 2021). La pauta recomendada es de una única dosis y se estima que la protección no es efectiva hasta al menos siete días después de la administración. Está indicada en personas de 18 años o más.

En la estrategia de vacunación seguida se diferencian dos etapas, a las que nos referiremos como primovacunación y dosis de recuerdo. La primovacunación se corresponde con la pauta inicial establecida para la población, consistente en una o dos dosis de la vacuna espaciadas un cierto intervalo de tiempo. En el caso de las vacunas de Pfizer, Moderna y AstraZeneca, si transcurrieron más días que el plazo establecido para la administración de la segunda dosis esta debió aplicarse igualmente y no se consideró necesario reiniciar la pauta. No obstante, si la segunda dosis se aplicó antes de los 19, 25 o 21 días, respectivamente, dicha dosis no se tuvo en cuenta y debió administrarse una nueva (tercera) respetando el intervalo mínimo de tiempo desde la segunda aplicación. Además, las pautas de vacunación con dos dosis pudieron ser homólogas (dos dosis de una misma vacuna) o heterólogas (dos dosis de vacunas distintas) en función de las dosis disponibles en cada momento. Las combinaciones aceptadas fueron Pfizer-Moderna, Moderna-Pfizer, AstraZeneca-Pfizer y AstraZeneca-Moderna. En todos los casos, para que la pauta fuese válida, debieron respetarse los tiempos mínimos entre dosis.

Dentro de la primovacunación se contempla también una dosis adicional, administrada en personas que, por tener el sistema inmune debilitado, no alcanzarían el nivel de protección adecuado si solo recibiesen la pauta indicada para la población general. Así, esta dosis adicional se reservó para personas de muy alto riesgo y personas sometidas a tratamiento con fármacos inmunosupresores.

En personas de 18 años o más, la administración de la dosis adicional debió realizarse considerando un mínimo de 28 días desde la última vacuna recibida. Además, esta dosis debió administrarse con vacunas de ARN mensajero (Pfizer y Moderna), de ser posible del mismo tipo que la vacuna o vacunas administradas con anterioridad. Las personas no vacunadas previamente recibieron una pauta de tres dosis con una vacuna de ARN mensajero.

Todos los individuos que en el momento de la primovacunación no estaban incluidos en el grupo de la dosis adicional, pero cumplieron los criterios con posterioridad, fueron captados para recibir esta dosis extra.

[1] Tiempo que transcurre desde la administración de la vacuna hasta que se adquiere una protección completa.

La dosis de recuerdo comenzó a administrarse el 1 de septiembre de 2021, aproximadamente en la mitad de la quinta ola de la pandemia. Esta dosis se administró a todas las personas vacunadas, de 18 años o más, con el fin de restaurar la protección frente al virus. Aunque la dosis de recuerdo estaba autorizada para personas entre 12 y 17 años, no se administró de manera sistemática en este grupo, quedando reservada para aquellos que habían recibido la dosis adicional por considerarse de riesgo o que la solicitaron explícitamente.

Para la dosis de recuerdo también se utilizaron las vacunas de ARN mensajero, independientemente de la vacuna utilizada en la primovacunación. En el caso de que la última dosis recibida fuera también una vacuna de ARN mensajero, el intervalo de tiempo hasta la dosis de recuerdo debió ser, como mínimo, de cinco meses. Si la última dosis de primovacunación fue con Janssen o AstraZeneca, el tiempo entre dosis pudo reducirse a tres meses. En las personas con primovacunación incompleta, se completó primero la pauta con una vacuna de ARN mensajero y se administró después la dosis de recuerdo, también con una vacuna de ARN mensajero, pasados cinco meses.

En la Tabla 1.2 se recogen los porcentajes de cobertura para la primovacunación completa y la dosis de recuerdo por grupos de edad. A 1 de septiembre de 2021, momento en el que comenzó a administrarse la dosis de recuerdo y pasados 248 días desde el inicio de la vacunación, el porcentaje de primovacunados de 70 años y más era del 100 %. Para el resto de grupos, a excepción de los de 19 años o menos, que comenzaron a vacunarse masivamente hacia finales de julio, la cobertura para la primovacunación era superior al 70 %. Durante los meses de septiembre, octubre y noviembre, la mayoría de dosis se administraron entre la población más joven, incrementándose la cobertura, sobre todo, entre los menores de 40 años. A finales de noviembre, todos los grupos superaban el 80 % de cobertura para la primovacunación. En cuanto a la dosis de recuerdo, tras tres meses de aplicación, el 100 % de las personas de 80 años y más estaban vacunadas y casi el 84 % de los mayores entre 70 y 79 años habían recibido también la dosis de recuerdo. Para el resto de grupos los porcentajes eran muy inferiores, por debajo del 50 %. A 27 de marzo de 2022, fecha que marca el fin de nuestro estudio, los grupos de edad a partir de 40 años presentaban coberturas de vacunación para la dosis de recuerdo por encima del 75 %. El porcentaje para las personas entre 30 y 39 años era del 60 % y para las de 20 a 29 años del 56 %. La cobertura para los menores de 20 años no se indica debido a que la dosis de recuerdo en este grupo no se administró de forma genérica, como ya se mencionó.

	Primovacunación completa (%)		**Dosis de recuerdo (%)**	
Grupo de edad	01-09-2021	30-11-2021	30-11-2021	27-03-2022
12-19	19,1	81,3	—	—
20-29	70,4	87,1	1,2	56,5
30-39	75,5	82,3	1,8	60,4
40-49	88,5	91,0	6,8	78,6
50-59	96,0	97,3	13,3	91,8
60-69	98,5	100,0	41,0	96,0
70-79	100,0	100,0	83,9	99,7
80+	100,0	100,0	100,0	100,0

Tabla 1.2: Porcentajes de cobertura vacunal, por grupo de edad, para la primovacunación completa y la dosis de recuerdo. Fuente: DXSP. Elaboración propia.

La Figura 1.5 representa el porcentaje de cobertura vacunal, por semanas, para la primovacunación completa y la dosis de recuerdo por grupos de edad a partir de los 12 años. Los gráficos reflejan la estrategia de vacunación seguida, priorizando por grupos, así como la intención de vacunación entre la población gallega. Las personas de 80 años y más fueron las primeras en alcanzar la inmunidad, con un rápido incremento de la cobertura tanto para la primovacunación como para la dosis de recuerdo. En el resto de grupos, sin embargo, se observa primero una cola para la cobertura y un ascenso cierto tiempo después. Esto se debe a que, en estos grupos, se vacunaron primero los individuos más

vulnerables y, posteriormente, se produjo la vacunación masiva del colectivo. En cuanto a la intención de vacunación, parece que solo se alcanzó un 100 % de cobertura en personas de 70 años y más, tanto para la primovacunación como para la dosis de recuerdo. A medida que descendemos en edad, se observa una bajada progresiva de la cobertura. Destaca especialmente el bajo porcentaje para la dosis de recuerdo entre los más jóvenes. Aproximadamente el 80 % de la población entre 20 y 39 años completó la primovacunación, pero solo el 60 % recibieron la dosis de recuerdo. El bajo porcentaje de cobertura para esta dosis en el grupo de 12 a 19 años se debe a que únicamente la recibieron las personas de riesgo o quienes la solicitaron de manera explícita.

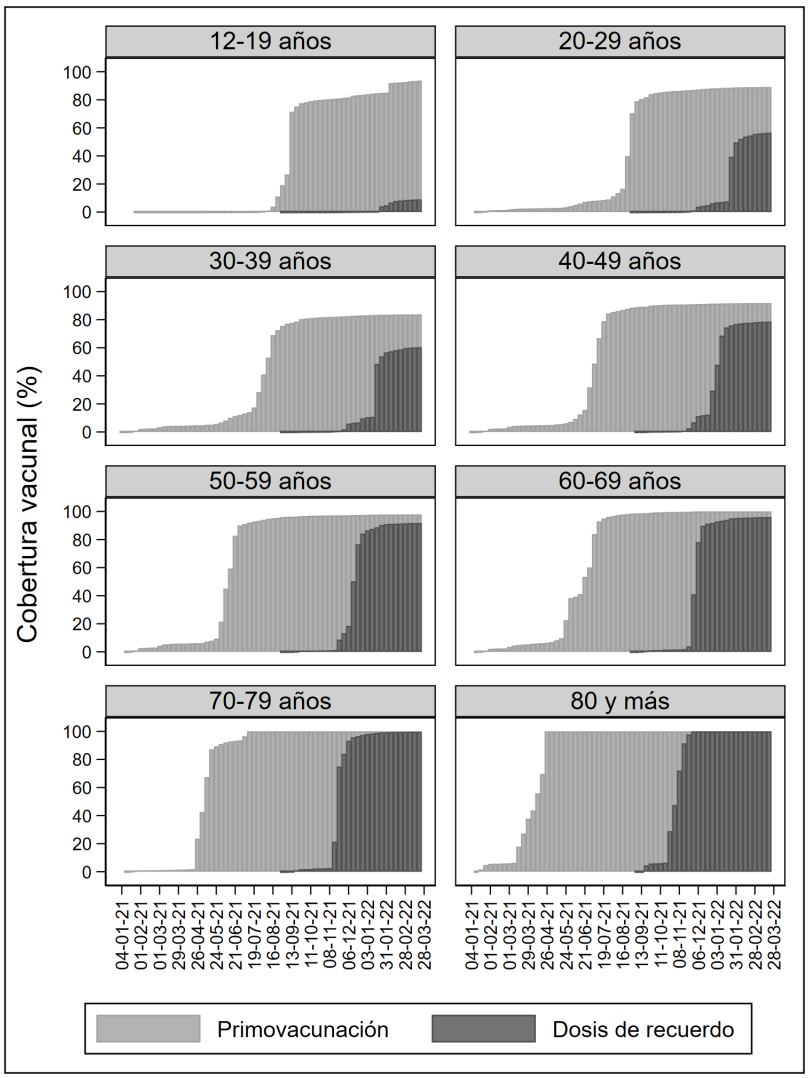

Figura 1.5: Cobertura vacunal, por semanas, para la primovacunación y la dosis de recuerdo por grupos de edad. Fuente: DXSP. Elaboración propia.

1.3. Efectividad vacunal

El objetivo principal de este estudio es la estimación de la efectividad de la vacunación frente a infección (sintomática o asintomática) e ingreso en UCI por SARS-CoV-2 en la población gallega de

12 años o más. Concretamente, en el trabajo se realizan dos análisis:

- **Análisis 1.** Estimación de la efectividad vacunal en individuos con primovacunación completa frente a no vacunados, analizando también la pérdida de inmunidad con el paso del tiempo.
- **Análisis 2.** Estimación de la efectividad de la dosis de recuerdo en individuos con primovacunación completa.

Para cumplir con el objetivo, se llevó a cabo un estudio de cohortes retrospectivo a partir de los datos de vigilancia y vacunación frente a SARS-CoV-2 de Galicia. Se calculó el riesgo relativo (RR) de infección e ingreso en UCI empleando un modelo de Cox y, a partir de su valor, se estimó la efectividad vacunal como $(1 - \text{RR})100\,\%$.

Dentro de los estudios epidemiológicos, el diseño de cohortes es ampliamente utilizado para estimar el efecto de la exposición a un determinado factor de riesgo o protector sobre la salud de una población, como puede ser la vacunación frente a la COVID-19. Su aplicación resulta adecuada para investigar cualquier exposición para la que pueda encontrarse y estudiarse un número suficiente de individuos susceptibles de experimentar el evento de interés. El primer paso del diseño es la selección de un grupo de individuos, la cohorte, que todavía no han desarrollado la enfermedad que se desea estudiar, pero están a riesgo de padecerla. Se recoge entonces información sobre la exposición o exposiciones a evaluar, así como otras posibles variables de interés. A continuación, se realiza el seguimiento de la cohorte durante un cierto período de tiempo y se detectan los casos de la enfermedad. Al finalizar el seguimiento, se compara la frecuencia de la enfermedad entre expuestos y no expuestos. Si esta frecuencia es mayor entre los expuestos, la exposición podría ser un factor de riesgo. Si, por el contrario, es menor, la exposición podría ser un factor de protección.

Una vez obtenidas las medidas de frecuencia de la enfermedad, se pueden calcular las medidas de asociación de la exposición con la enfermedad. En los estudios de cohortes, la medida básica de asociación es el riesgo relativo (RR) o *hazard ratio* (HR), esto es, el cociente entre el riesgo en el grupo de expuestos y el riesgo en el grupo de no expuestos. Uno de los métodos estadísticos más ampliamente utilizados para el cálculo del RR es el modelo de Cox de riesgos proporcionales, que permite ajustar el riesgo por posibles factores de confusión, y que pueden ser fijos o variables en el tiempo.

Los estudios de cohortes pueden ser concurrentes, lo que permite seleccionar qué variables medir al inicio del seguimiento de los individuos, o retrospectivos, cuando se estudia una cohorte reunida en el pasado para la cual la información sobre la exposición y otras variables se encuentra almacenada en registros. Este tipo de diseño ha sido el empleado en nuestros análisis, al disponer ya de los datos de vacunación frente al SARS-CoV-2 en la población gallega. El diseño retrospectivo se caracteriza por ser menos costoso, ya que la información ya ha sido recopilada y no es necesario esperar a que ocurra el evento o eventos de interés. Sin embargo, la calidad de los resultados depende en gran medida de la calidad de los registros.

El diseño de cohortes presenta como principales ventajas la posibilidad de cuantificar las variables antes de ocurrir los eventos, lo que evita sesgos en las mediciones en función del desenlace, y la capacidad de establecer una secuencia temporal entre la exposición y el evento. Sin embargo, requiere de un gran número de individuos a los que seguir durante un período de tiempo largo, lo que se traduce, en la mayor parte de los casos, en un alto riesgo de abandono y un coste elevado. Para más información sobre los estudios epidemiológicos y el diseño de cohortes, consultar Hernández-Aguado et al. (2018).

En el Apéndice A se ofrece una Tabla resumen con algunos de los estudios previos sobre efectividad de las vacunas de COVID-19. Estos estudios se centran en la estimación de la efectividad de una o varias vacunas frente a cuatro desenlaces, a saber, infección, hospitalización, ingreso en UCI y defunción por COVID-19. Todos ellos son estudios epidemiológicos de carácter observacional, en los que el investigador registra las exposiciones que han recibido los sujetos, en este caso, la vacunación. La Tabla incluye ejemplos de estudios de cohortes y también de casos y controles. Para más información sobre los métodos y resultados, consultar los artículos originales.

En el siguiente Capítulo se realiza una breve introducción al Análisis de Supervivencia y se presenta la base teórica del modelo de Cox y las técnicas empleadas en los análisis. Seguidamente, en el

Capítulo 3, se define la cohorte y sus características, así como los grupos de comparación para los dos análisis realizados.

Capítulo 2

Introducción al Análisis de Supervivencia

A menudo, el objetivo de los estudios epidemiológicos es la comparación de la supervivencia entre dos o más grupos de pacientes. Típicamente, resulta necesario ajustar la función de supervivencia en cada grupo considerando una serie de covariables (variables explicativas, independientes, predictoras o factores de riesgo) relacionadas con el evento de interés. En este contexto de regresión, uno de los métodos más utilizado es el modelo de riesgos proporcionales de Cox (1972). Esta técnica se utiliza ampliamente para investigar la asociación entre el tiempo de supervivencia de los pacientes y una o más variables explicativas, continuas o categóricas.

A continuación, ofrecemos una breve introducción al Análisis de Supervivencia, indicando las principales funciones que utiliza y las características de los datos con los que trabaja para, seguidamente, presentar en detalle el modelo de regresión desarrollado por Cox.

El Análisis de Supervivencia es la rama de la Estadística que se ocupa del estudio del tiempo transcurrido desde un instante inicial hasta que sucede un determinado evento de interés. Esta formulación general se adapta a multitud de contextos, en función de la naturaleza del evento de estudio. Así, en el campo biomédico, el evento puede ser el alta de un paciente que ha recibido un determinado tratamiento o el fallecimiento. En Ingeniería, es de especial interés analizar el tiempo de funcionamiento de las máquinas hasta que se produce un fallo. En el ámbito económico, es habitual considerar como evento el primer empleo de los individuos de una población para analizar los tiempos en el paro. Como ejemplo de la variedad de aplicaciones del Análisis del Supervivencia se proponen los artículos de Huang et al. (2019), donde desarrollan un algoritmo de aprendizaje profundo que integra Análisis de Supervivencia para predecir el pronóstico en cáncer de mama a partir de datos de expresión génica, Nabizadeh et al. (2018), donde se estudia la fiabilidad de los puentes de Wisconsin en función de la longitud y el tráfico medio diario, entre otros factores, o Dirick et al. (2017), donde se revisan modelos de supervivencia clásicos y recientes y se comparan sus rendimientos con datos reales de tiempos de impago crediticio.

En cualquier caso, en el Análisis de Supervivencia la variable respuesta Y, que representa el tiempo desde un origen bien definido hasta la ocurrencia de un evento perfectamente especificado, es una variable aleatoria no negativa y, en el caso más habitual, continua. Comúnmente, nos referimos a esta variable como tiempo de supervivencia o tiempo de fallo.

Otra característica importante en supervivencia es la presencia de problemas en la observación de la variable de interés, que denotaremos por Y. A este nivel, son habituales la censura y el truncamiento.

El fenómeno de la censura aparece cuando el evento inicial o el evento final no se observan con precisión, de modo que el tiempo de supervivencia no se conoce por completo. De acuerdo con Klein et al. (2005), podemos distinguir tres tipos principales de censura: por la derecha, por la izquierda o por intervalo.

La censura por la derecha es el tipo más frecuente y se produce cuando solo se sabe que el evento

final excede al tiempo observado. Por ejemplo, en un ensayo clínico en el que interesa estudiar el fallecimiento por una determinada enfermedad, presentarán censura por la derecha aquellos pacientes que fallezcan después del término del estudio, ya que su seguimiento finalizará antes de que se produzca el fallo. Lo mismo ocurrirá con los individuos que abandonen el ensayo o los que fallezcan por otras causas.

Menos habitual es la censura por la izquierda, que se da cuando el individuo experimenta el fallo antes del tiempo observado. En los centros de aprendizaje de la primera infancia, se hacen pruebas a los niños para saber cuándo empiezan a realizar ciertas tareas por sí solos. Es posible que algunos niños ya realicen la tarea antes de entrar en el centro. En este caso, sus tiempos se considerarían censurados por la izquierda.

Finalmente, en la censura por intervalo solo se sabe que el fallo ocurre dentro de un lapso de tiempo específico. Esta censura es típica en los estudios longitudinales, en los que se sigue a los pacientes periódicamente y solo se sabe que el evento de interés se sitúa entre dos visitas consecutivas.

A su vez, la censura por la derecha puede ser de Tipo I, de Tipo II o aleatoria. En la censura Tipo I, los tiempos de censura están preespecificados. Así, si no se experimenta el evento de interés antes del fin del estudio y no hay abandonos, todas las observaciones censuradas tienen tiempos iguales a la duración del período de seguimiento. Una variante de la censura Tipo I es la censura generalizada, donde se permite que los participantes entren en el estudio en diferentes momentos. Dado que el término del estudio está fijado por el investigador, los tiempos de censura ya son conocidos cuando los individuos inician el seguimiento.

La censura de Tipo II se produce cuando se siguen los objetos experimentales hasta el fallo de un porcentaje o número previamente especificado. Los experimentos que implican este subtipo de censura son habituales en entornos industriales, donde las máquinas se someten a pruebas de fiabilidad. En este caso, todos los dispositivos (n en total) se ponen en marcha al mismo tiempo, y el experimento finaliza cuando los r primeros dispositivos fallan ($r < n$). Así, se observan los r tiempos de fallo más pequeños (ordenados) y los n-r dispositivos restantes tienen tiempos censurados, que son iguales al mayor tiempo de fallo observado.

La última categoría de censura por la derecha es la censura aleatoria. En el ámbito biomédico, es debida principalmente a abandonos, pérdidas de seguimiento por otras causas o terminación del estudio. En los dos primeros casos es especialmente importante prestar atención a la causa de la censura para evitar estimaciones sesgadas. Si el abandono o pérdida de seguimiento del paciente ocurre realmente al azar, la censura no causará ningún problema de sesgo en el análisis. Sin embargo, si los pacientes que están próximos al fallo tienen más probabilidades de abandonar que otros individuos, pueden surgir sesgos graves. Otra causa de censura aleatoria son los eventos competitivos, que surgen cuando se quiere estimar la distribución marginal de un evento pero algunos individuos en seguimiento experimentan un suceso competitivo que los elimina del estudio. De este modo, el evento de interés no es observable para estos participantes, que son censurados aleatoriamente por la derecha en el momento en el que experimentan el suceso competidor. Por ejemplo, si estamos interesados en estudiar el tiempo de fallo por cáncer y uno de los participantes muere por otra causa, entonces ese participante estará censurado, ya que desconocemos el momento en el que habría muerto de cáncer. Para poder identificar la distribución marginal a partir de los datos de riesgos competitivos, es necesario que el tiempo del evento y los tiempos de censura sean independientes.

Por otro lado, se habla de truncamiento cuando solo se observan los tiempos de los individuos en los que se cumple una determinada condición. En caso contrario, el sujeto no aporta ningún dato al investigador. Este hecho contrasta con la censura, en la que siempre existe información sobre los individuos, aunque sea parcial. Lo más frecuente es el truncamiento por la izquierda. Aquí, para que un individuo sea observado, debe darse la condición antes del evento de interés. La condición de truncamiento puede ser el diagnóstico de una enfermedad o la entrada en una residencia de ancianos, por ejemplo. Nótese que, a diferencia de la censura por la izquierda, en la que se tiene información parcial sobre los individuos que experimentan el evento antes del inicio del estudio, en el truncamiento por la izquierda estos individuos nunca serán considerados para su inclusión en el ensayo.

En el ámbito biomédico, también podemos encontrar datos doblemente truncados. De acuerdo

con de Uña-Álvarez et al. (2022), una variable Y está doblemente truncada por un par de variables aleatorias (U, V) si la observación de Y solo es posible cuando ocurre $U \leq Y \leq V$. En tal caso, U y V se denominan variables de truncamiento a la izquierda y a la derecha, respectivamente. Un escenario que lleva al doble truncamiento es el del muestreo por intervalos, donde la muestra se restringe a los individuos con evento entre dos fechas específicas, d_0 y d_1. El tiempo de truncamiento a la derecha vendrá dado por $V = d_1 - d_I$, donde d_I denota la fecha de inicio del tiempo hasta el evento, y el tiempo de truncamiento a la izquierda será $U = d_0 - d_I = V - \varsigma$, donde $\varsigma = d_1 - d_0$ es la anchura del intervalo. Podemos considerar como ejemplo el conjunto de datos de cáncer infantil del *Instituto Português de Oncologia*, disponible en la última actualización del paquete `DTDA` de `R` de Moreira et al. (2021) . La información corresponde a todos los niños diagnosticados de cáncer entre el 1 de enero de 1999 (d_0) y el 31 de diciembre de 2003 (d_1) en la región del Norte de Portugal. La variable de interés es la edad al diagnóstico en años que, por definición del cáncer infantil, se apoya en el intervalo [0, 15]. Debido al muestreo por intervalos, la edad en el momento del diagnóstico está doblemente truncada por el par (U, V) , donde la variable de truncamiento a la derecha V es el tiempo en años desde el nacimiento hasta el 31 de diciembre de 2003, y $U = V - 5$.

En otras situaciones, las variables de truncamiento no están vinculadas mediante la ecuación lineal $V = U + \varsigma$. Por ejemplo, U y V podrían representar algunos límites de observación aleatorios más allá de los cuales la variable de interés Y no puede ser muestreada. Situaciones de este tipo aparecen, entre otros, en el campo de la Astronomía.

Una diferencia importante del truncamiento doble en comparación con el truncamiento unilateral es que, con datos doblemente truncados, el estimador no paramétrico de máxima verosimilitud de la distribución de probabilidad de la variable de interés no tiene una forma explícita. De hecho, puede ser no único e incluso inexistente. Alternativas semiparamétricas y paramétricas pueden ayudar a evitar estos inconvenientes, reduciendo también la varianza a costa de introducir algún sesgo en la estimación.

El principal impacto del truncamiento en el análisis es que el investigador debe utilizar una distribución condicional para construir la probabilidad o emplear un método estadístico que utilice un conjunto de riesgo selectivo.

El Análisis de Supervivencia tiene como objetivos principales estimar la distribución del tiempo de fallo, comparar dos o más distribuciones entre grupos (por ejemplo, pacientes sometidos a distintos tratamientos) y evaluar el efecto de covariables sobre el tiempo de fallo. Para ello, es necesario recurrir a una metodología específica que considere las características de la variable respuesta, que siempre es positiva y, frecuentemente, está censurada.

En nuestros análisis consideraremos la variable respuesta Y "tiempo hasta un diagnóstico positivo en COVID-19", que presenta las siguientes características:

i. $Y \geq 0$, es decir, Y es siempre una variable positiva.

ii. Y presenta censura por la derecha de tipo aleatorio. Para los individuos que experimentan el evento de interés durante el período de seguimiento, es decir, que son diagnosticados de COVID-19, el tiempo hasta el evento es un tiempo observado. Para los individuos que no experimentan el evento solo se sabe que no fueron positivos en COVID-19 durante el seguimiento, pero se desconoce lo que ocurrió después. Sus tiempos son, por tanto, censurados. Además, dado que el fin del seguimiento se puede dar por más causas que la ocurrencia del evento o el fin del estudio, los tiempos de censura no pueden ser conocidos de antemano.

Las covariables de ajuste empleadas serán variables de tipo cualitativo, nominales u ordinales. Dichas covariables se describen en detalle en la Sección 3.3 del Capítulo 3.

2.1. Funciones fundamentales en Análisis de Supervivencia

Denotaremos la variable de interés "tiempo hasta un diagnóstico positivo en COVID-19" por Y.

Como para cualquier variable aleatoria, la distribución de Y viene dada por su función de probabilidad acumulada o función de distribución F,

$$F(t) = \mathbb{P}(Y \leq t),\ t \geq 0, \tag{2.1}$$

que representa la probabilidad de que el diagnóstico positivo en COVID-19 ocurra en t o antes del tiempo t. Sin embargo, en Análisis de Supervivencia tienen mayor importancia la función de supervivencia, la función de riesgo o razón de fallo y la función de riesgo acumulada. Conociendo cualquiera de estas funciones, las restantes pueden ser determinadas de forma única. Veamos cómo se definen y relacionan entre ellas.

La función de supervivencia representa la probabilidad de sobrevivir al tiempo t (el evento se produce después de t):

$$S(t) = \mathbb{P}(Y > t) = 1 - F(t). \tag{2.2}$$

La función de supervivencia tiene como propiedades básicas ser continua, monótona no creciente e igual a 1 en cero ($S(0) = 1$) e igual a 0 a medida que el tiempo se acerca a infinito ($S(+\infty) = 0$).

Las funciones (2.1) y (2.2) son válidas tanto para tiempos de fallo continuos como discretos. Para las siguientes definiciones, sin embargo, consideraremos la respuesta con distribución absolutamente continua, dado que en nuestros análisis trabajaremos con tiempos continuos. Para más información sobre el caso discreto, consultar Klein et al. (2005).

La función de supervivencia es también la integral de la función de densidad, esto es,

$$S(t) = \mathbb{P}(Y > t) = \int_t^\infty f(u)du,$$

con

$$\begin{aligned} f(t) = \frac{dF(t)}{dt} &= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbb{P}(Y \leq t + \Delta t) - \mathbb{P}(Y \leq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y \leq t + \Delta t)}{\Delta t}. \end{aligned}$$

Así,

$$f(t) = -\frac{dS(t)}{dt}. \tag{2.3}$$

Nótese que la función de densidad representa la probabilidad de que Y pertenezca al intervalo infinitesimal de extremos t y $t + \Delta t$, es decir, la probabilidad instantánea de que el evento ocurra en t.

Por otro lado, la función de riesgo o razón de fallo determina la probabilidad de que un individuo que haya sobrevivido en el tiempo t experimente el evento en el instante de tiempo inmediato $[t; \Delta t)$:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y \leq t + \Delta t | Y \geq t)}{\Delta t}.$$

Esta función resulta particularmente útil para determinar la distribución de Y empleando información cualitativa sobre el mecanismo de ocurrencia del evento y para describir la forma en que la probabilidad de experimentar el evento cambia con el tiempo. La única restricción para $\lambda(t)$ es que sea no negativa. Además, es fácil ver que

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y \leq t + \Delta t | Y \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y \leq t + \Delta t, Y \geq t)}{\Delta t \cdot \mathbb{P}(Y \geq t)} \\ &= \frac{1}{\mathbb{P}(Y \geq t)} \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y \leq t + \Delta t)}{\Delta t} = \frac{f(t)}{S(t)} \end{aligned} \tag{2.4}$$

sin más que aplicar la definición de probabilidad condicionada a la definición de función de riesgo.

Una cantidad relacionada con la anterior es la función de riesgo acumulado o razón de fallo acumulada,

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

que representa el área bajo la función de razón de fallo hasta un tiempo de interés t.

Como hemos ido viendo, existen importantes relaciones entre las distintas funciones de Análisis de Supervivencia. A mayores, a partir de (2.4) se pueden obtener las siguientes expresiones aplicando la relación dada en (2.3):

$$\Lambda(t) = -\ln\left(S(t)\right)$$

$$S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(s)ds}.$$

Para finalizar, es importante considerar dos características de la variable tiempo hasta el evento, como son la vida media y la vida mediana. La primera se deriva de la vida residual media, que mide la vida restante esperada de un individuo que ha sobrevivido al instante t:

$$\text{mrl}(t) = \mathbb{E}(Y - t|Y > t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}.$$

Como se puede ver, la vida residual media es el área bajo la curva de supervivencia a la derecha de t dividida por $S(t)$. La vida media, $\mathbb{E}(Y) = \text{mrl}(0) = \int_0^\infty S(t)$, se corresponde con el área total bajo la curva de supervivencia.

El cuantil p de la distribución de Y se define como el menor t_p tal que

$$S(x_p) \leq 1 - p,\ i.e.,\ x_p = \text{ínf}\left\{t : S(t) \leq 1 - p\right\}.$$

Entonces, la vida mediana es el valor $t_{0,5}$ tal que $S(t_{0,5}) = 0{,}5$.

2.2. Estimación no paramétrica de la función de distribución

En este apartado mostramos los principales métodos no paramétricos de estimación de la distribución de Y. Estos métodos se caracterizan por trabajar con hipótesis poco restrictivas, lo que los hace muy generales y flexibles. No obstante, al basarse fundamentalmente en los datos observados, sus resultados dependen ampliamente de la muestra disponible y de su tamaño.

En el modelo general de censura aleatoria se dispone de una muestra observada de la forma $\{(Z_i, \delta_i)\}_{i=1}^n$, que es una $m.a.s$ de (Z, δ) donde:

$$Z_i = \text{mín}\left\{Y_i, C_i\right\}, \delta_i = \mathbb{I}\{Y_i \leq C_i\},$$

siendo Y_i el tiempo en el que se produce el fallo o evento de interés para el $i - ésimo$ individuo de la muestra, C_i su tiempo de censura y δ_i una variable indicadora, que vale 0 cuando Z_i es un tiempo censurado y 1 cuando Z_i es un tiempo de fallo observado. Y y C son variables no negativas.

Asociadas al mecanismo probabilístico de este modelo de censura se definen una serie de funciones:

1) La función de distribución de la variable tiempo de fallo Y, introducida previamente en 2.1, $F(y) = \mathbb{P}(Y \leq y)$, su función de supervivencia, $S(y) = 1 - F(y) = \mathbb{P}(Y > y)$, y su función de densidad, $f(\cdot)$.

2) La función de distribución de la variable de censura C, $G(y) = \mathbb{P}(C \leq y)$, su función de supervivencia, $1 - G(y) = \mathbb{P}(C > y)$, y su función de densidad, $g(\cdot)$.

3) La función de distribución de la variable observada Z, $H(y) = \mathbb{P}(Z \leq y)$, y su función de supervivencia, $1 - H(y) = \mathbb{P}(Z > y) = \mathbb{P}(\text{mín}\{Y, C\} > y) = \mathbb{P}(Y > y, C > y)$.

A partir de la muestra observada, la estimación no paramétrica de F puede realizarse por diferentes métodos, entre ellos por máxima verosimilitud. Para poder derivar la función de verosimilitud, es necesario asumir la siguiente hipótesis:

$$1 - H(y) = \mathbb{P}(Y > y, C > y) = (1 - F(y))(1 - G(y)).$$

Es decir, Y y C son independientes.

Para definir la función de verosimilitud necesitamos conocer primero la probabilidad de los datos observados.

- La contribución de una observación no censurada $(Z_i, 1)$ a la verosimilitud es:

$$\mathbb{P}(Z_i, 1) = \mathbb{P}(Y = Z_i, Y \leq C) = \mathbb{P}(Y = Z_i, Z_i \leq C) = \mathbb{P}(Y = Z_i)\mathbb{P}(C \geq Z_i) = f(Z_i)(1 - G(Z_i^-))$$

- La contribución de una observación censurada $(Z_i, 0)$ a la verosimilitud es:

$$\mathbb{P}(Z_i, 0) = \mathbb{P}(C = Z_i, Y > C) = \mathbb{P}(C = Z_i, Y > Z_i) = \mathbb{P}(C = Z_i)\mathbb{P}(Y > Z_i) = g(Z_i)(1 - F(Z_i))$$

Entonces, la función de verosimilitud viene dada por:

$$L_n = \prod_{i=1}^{n} f(Z_i)^{\delta_i}(1 - F(Z_i))^{1-\delta_i} \prod_{i=1}^{n} g(Z_i)^{1-\delta_i}(1 - G(Z_i^-)^{\delta_i}.$$

Bajo independencia, Y y C no están relacionadas y, como nuestra variable de interés es Y, tenemos que

$$L_n \propto \prod_{i=1}^{n} f(Z_i)^{\delta_i}(1 - F(Z_i))^{1-\delta_i}.$$

Por tanto, en un contexto de estimación no paramétrica, el estimador de máxima verosimilitud de F se obtiene maximizando

$$L_n(F) = \prod_{i=1}^{n} f(Z_i)^{\delta_i}(1 - F(Z_i))^{1-\delta_i} \tag{2.5}$$

sobre la clase de todas las funciones de distribución F.

En ausencia de censura, el estimador no paramétrico por excelencia de F es la función de distribución empírica, que se construye a partir de una *m.a.s* de la variable de interés:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \leq t\}.$$

$\hat{F}$ estima la probabilidad de que la variable sea menor o igual que un determinado valor como la proporción de los datos de la muestra menores o iguales a dicho valor. Sin embargo, y como hemos visto, en Análisis de Supervivencia interesa más la estimación de la función de supervivencia. Dado que $S = 1 - F$, podemos construir un estimador empírico de la función de supervivencia como

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i > t\}.$$

Este estimador aproxima $S(t)$ por la proporción muestral de observaciones con tiempos de fallo mayores que t. En presencia de censura deja de ser consistente, ya que ignora el hecho de que algunas observaciones no son completas.

2.2.1. Estimador de Kaplan-Meier

Kaplan et al. (1958) propusieron un estimador de $S(\cdot)$ que es una generalización del estimador empírico en presencia de datos censurados. El estimador de Kaplan-Meier o límite-producto se puede derivar de distintas formas. A continuación, exponemos dos de los más relevantes, el método del producto y el método de máxima verosimilitud.

Sea $t_1 < \ldots < t_K$, con $K \leq n$, una *m.a.s* de tiempos de ocurrencia del evento de interés. Sea d_k el número de eventos en t_k y sea n_k el número de individuos a riesgo en t_k, es decir, individuos que no han muerto ni han sido censurados antes de t_k. El estimador de Kaplan-Meier de la supervivencia, $S(t) = \mathbb{P}(Y > t)$, obtenido por el método del producto es

$$\hat{S}(t) = \prod_{t_k \leq t} \left(1 - \frac{d_k}{n_k}\right). \tag{2.6}$$

Obsérvese que $\frac{d_k}{n_k}$ es la proporción de eventos ocurridos en el tiempo t_k, y $1 - \frac{d_k}{n_k}$ es la proporción de individuos a riesgo en t_k que sobrevive a t_k.

Para sobrevivir al tiempo t habrá que llegar a cada $t_k \leq t$ y no fallar en t_k. Por eso se multiplican las probabilidades condicionadas de sobrevivir a t_k.

Alternativamente, podemos llegar al estimador de Kaplan-Meier de la supervivencia a partir de la expresión de la verosimilitud dada en (2.5), maximizando sobre la clase de todas las funciones de distribución F. Para ello, necesitamos que F tenga un salto positivo en los tiempos observados (si no, $L_n = 0$) y que no salte en los tiempos censurados. O, equivalentemente a lo anterior, la función de supervivencia debe estar caracterizada por los valores de la función de riesgo $\lambda_1 < \ldots < \lambda_K$ en $t_1 < \ldots < t_K$, de forma que

$$S(t_K^-) = \prod_{k=1}^{K-1} (1 - \lambda_k)$$

y

$$S(t_K) = \prod_{k=1}^{K} (1 - \lambda_k).$$

Supongamos, además, que hay d_k fallos en t_k y m_k tiempos censurados en el intervalo $[t_k, t_k + 1)$. Sea $n_k = (m_k + d_k) + \cdots + (m_K + d_K)$ el número de individuos a riesgo en t_k. Entonces, la contribución de las observaciones no censuradas será

$$\prod_{k=1}^{K} \left[S(t_k^-) - S(t_k)\right]^{d_k},$$

mientras que la contribución de las censuras será

$$\prod_{k=1}^{m_k} S(t_k).$$

Por tanto:

$$
\begin{aligned}
L_n &= \prod_{k=1}^{K} \left[S(t_k^-) - S(t_k)\right]^{d_k} \prod_{k=1}^{m_k} S(t_k) = \prod_{k=1}^{K} \left[\prod_{l=1}^{k-1}(1-\lambda_l) - \prod_{l=1}^{k}(1-\lambda_l)\right]^{d_k} S(t_k)^{m_k} \\
&= \prod_{k=1}^{K} \left[\lambda_k \prod_{l=1}^{k-1}(1-\lambda_l)\right]^{d_k} \prod_{l=1}^{k}(1-\lambda_l)^{m_k} = \prod_{k=1}^{K} \left\{\lambda_k^{d_k} \prod_{l=1}^{k-1}(1-\lambda_l)^{d_k}\right\} \prod_{l=1}^{k}(1-\lambda_l)^{m_k} \\
&= \prod_{k=1}^{K} \lambda_k^{d_k}(1-\lambda_k)^{n_k-d_k}.
\end{aligned}
$$

Tomando logaritmos:

$$\ln(L_n) = \sum_{k=1}^{K} d_k \ln(\lambda_k) + (n_k - d_k)\ln(1-\lambda_k).$$

Derivando respecto a cada λ_k e igualando a cero se obtiene $\hat{\lambda}_k = \frac{d_k}{n_k}$, y el estimador de máxima verosimilitud de la supervivencia coincide con el estimador de Kaplan-Meier.

2.2.2. Estimador de Nelson-Aalen

Un estimador no paramétrico de la supervivencia alternativo al estimador de Kaplan-Meier es el estimador de Nelson-Aalen, introducido originalmente por Nelson (1969) y extendido más allá de los datos de supervivencia por Aalen (1978). Este estimador se basa en la relación entre la supervivencia y la función de riesgo acumulada, $S(t) = e^{-\Lambda(t)}$.

El estimador de la función de riesgo acumulada o razón de fallo de Nelson-Aalen es

$$\hat{\Lambda}(t) = \sum_{t_k \le t} \frac{d_k}{n_k},$$

donde cada $\hat{\lambda}_k = \frac{d_k}{n_k}$ es la estimación condicional del riesgo en t_k. Teniendo esto en cuenta, el estimador de Nelson-Aalen de la supervivencia será

$$\hat{S}_{NA}(t) = e^{-\hat{\Lambda}(t)}.$$

2.3. Regresión con variables censuradas: el modelo de Cox

Una vez introducidas las principales funciones del Análisis de Supervivencia y sus estimaciones no paramétricas, pasaremos a centrarnos en la regresión con censura, donde el interés radica en evaluar la relación entre la distribución de la variable respuesta y una o más variables explicativas.

En un contexto de regresión con datos censurados en el que para cada individuo, $i = 1, \ldots, n$, se dispone de un vector de P covariables, la muestra observada será de la forma $\{(Z_i, \delta_i, X^i)\}_{i=1}^n$, donde:

$$
\begin{aligned}
Z_i = \text{mín}\,(Y_i, C_i), \delta_i = \mathbb{I}\{Y_i \le C_i\} \\
X^i = (X_{i1}, X_{i2}, \ldots, X_{iP})'.
\end{aligned}
$$

Como ya hemos visto, Y_i denota el tiempo de fallo del individuo i, C_i el tiempo de censura del individuo i, δ_i es la variable indicadora de censura (que vale 0 cuando Z_i es un tiempo censurado y 1 cuando Z_i es un tiempo de fallo observado) y $X_{ip},\ 1 \le p \le P$, las covariables en el individuo $i - ésimo$, que pueden ser fijas o dependientes del tiempo. Se asume que Y_i es independiente de

C_i para $i = 1, 2, \ldots, n$. Esta suposición es apropiada cuando el mecanismo que provoca la censura es independiente del mecanismo de supervivencia. Además, en la regresión con datos censurados, se considera también que las variables Y_i y C_i son condicionalmente independientes dado el vector de covariables X^i.

El modelo de riesgos proporcionales propuesto por Cox (1972) es uno de los grandes enfoques dentro de la regresión con datos censurados. Se engloba dentro de la familia de modelos de riesgo multiplicativo, en los que la tasa de riesgo de un individuo condicionada a una serie de covariables X es el resultado del producto de una tasa de referencia $\lambda_0(t)$ y una función no negativa de los predictores, $\eta(\beta' X)$. Este modelo es ampliamente utilizado para evaluar el efecto de una exposición (por ejemplo, un tratamiento) entre dos o más grupos a partir de datos de supervivencia, como puede verse en Delbarre et al. (2017) o en Katzman et al. (2018). Una de sus principales ventajas, desde el punto de vista epidemiológico, es que su interpretación resulta bastante intuitiva, ya que permite estimar el efecto directo de la exposición.

Por simplicidad, introduciremos el modelo para el caso en el que los tiempos hasta el evento de interés se distribuyen de forma continua, ignorando la posibilidad de empates, y siguiendo a Iglesias Pérez et al. (2021).

Sea $\lambda(t|x)$ la función de riesgo de un individuo con vector de covariables $X^i = x$. El modelo de Cox toma la forma

$$\lambda(t|x) = \lambda_0(t) e^{\beta' x},$$

donde $\lambda_0(t)$ es el riesgo basal y $\beta = (\beta_1, \ldots, \beta_P)'$ es un vector de parámetros desconocidos[1]. El riesgo basal es una función de riesgo de referencia, no especificada, que corresponde al riesgo para el conjunto estándar de condiciones $X = 0$. Nótese que el modelo de Cox es un modelo semiparamétrico, pues solo se asume una forma paramétrica para el efecto de las covariables.

Alternativamente, podemos escribir el modelo como

$$\ln\left[\frac{\lambda(t|x)}{\lambda_0(t|x)}\right] = \beta' x \Leftrightarrow \frac{\lambda(t|x)}{\lambda_0(t|x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_P}.$$

Así, el riesgo relativo de un individuo con covariable x respecto al riesgo de referencia, $\frac{\lambda(t|x)}{\lambda_0(t|x)}$, es igual a la exponencial del predictor lineal $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_P$.

El modelo de Cox se denomina modelo de riesgos proporcionales porque el cociente entre las funciones de riesgo de dos individuos con covariables x y x^* es una cantidad constante en el tiempo, de modo que los riesgos son proporcionales:

$$\frac{\lambda(t|x)}{\lambda(t|x^*)} = \frac{\lambda_0(t|x) e^{\beta' x}}{\lambda_0(t|x^*) e^{\beta' x^*}} = e^{\beta'(x - x^*)}. \tag{2.7}$$

El cociente anterior se conoce como razón de riesgos o *hazard ratio*, y representa el riesgo relativo de que un individuo con covariable x sufra un evento comparado con un individuo con valor de la covariable x^*. Este valor es constante en el tiempo y solo depende de los valores de las covariables.

La expresión (2.7) también nos permite una fácil interpretación de los coeficientes del modelo, que no son más que el logaritmo del *hazard ratio* o riesgo relativo. En el caso de covariables cuantitativas, el incremento en una unidad de X_p hace que el riesgo se multiplique por e^{β_p}, permaneciendo constantes el resto de predictores.

- Si $e^{\beta_p} = 1 \Rightarrow \beta_p = 0$, y el riesgo es independiente de X_p.
- Si $e^{\beta_p} > 1 \Rightarrow \beta_p > 0$, y el riesgo se incrementa al aumentar X_p.
- Si $e^{\beta_p} < 1 \Rightarrow \beta_p < 0$, y el riesgo disminuye al aumentar X_p.

[1] Consideraremos el vector de parámetros β sin intercepto para el vector de covariables $X = (X_1, \cdots, X_P)$.

Si se tiene una covariable categórica con c niveles, es necesario introducir $c-1$ variables indicadoras en el modelo. En este caso, $e^{\beta_p}, p = 1, \ldots, c-1$, es el número por el que se multiplica el riesgo al pasar de la categoría de referencia a la categoría correspondiente a la variable indicadora $p - ésima$.

En este trabajo, presentamos un estudio de seguimiento con datos individuales, en el que para cada participante se dispone de los valores de una serie de covariables que pueden determinar la tasa de incidencia de la enfermedad por COVID-19. La tasa o densidad de incidencia, se calcula como el número de casos nuevos del evento en la población en función del total de personas-tiempo a riesgo. En este contexto, $\lambda_0(t|x)$ representa la tasa de incidencia en el momento t del seguimiento para un individuo en el nivel de referencia (con todas sus covariables iguales a 0), $\lambda(t|x)$ es la tasa de incidencia en el momento t del seguimiento para un individuo con covariables X, y los coeficientes β de regresión representan, para cada covariable, el logaritmo neperiano del riesgo relativo asociado con el incremento de la misma en una unidad. A modo de ejemplo, supongamos que X_1 es una variable indicadora de tratamiento ($X_1 = 0$ si placebo y $X_1 = 1$ si tratamiento) frente a una determinada enfermedad y que el resto de covariables toman siempre el mismo valor. Entonces, $\frac{\lambda(t|x)}{\lambda(t|x^*)} = e^{\beta'(x-x^*)} = e^{\beta_1}$, representará el riesgo de enfermedad si el individuo recibió el tratamiento en relación con el riesgo de enfermedad si el individuo hubiera recibido el placebo. Por tanto, el *hazard ratio* o riesgo relativo determina el efecto del tratamiento sobre la enfermedad. Concretamente, mide cuántas veces es más frecuente la afección en el grupo de tratamiento con respecto al grupo placebo, resultando muy útil como medida de asociación entre el tratamiento y la enfermedad. Aquí, un riesgo relativo de 1 indicará que la tasa de incidencia es igual en tratados que en no tratados y, por tanto, la ausencia de asociación entre el tratamiento y la enfermedad. Un riesgo relativo por encima de 1 indicará una mayor tasa de incidencia de la enfermedad entre los tratados (el tratamiento será un factor de riesgo), mientras que un riesgo relativo por debajo de 1 indicará una menor tasa de incidencia entre los tratados (el tratamiento será un factor protector).

La condición de riesgos proporcionales afecta también a la forma de las funciones de riesgo acumulado condicional y de supervivencia condicional del siguiente modo:

$$\Lambda(t|x) = \Lambda_0(t|x)e^{\beta' x}, \text{ donde } \Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

Entonces, las funciones de riesgo acumulado para dos individuos con covariables x y x^* son proporcionales:

$$\frac{\Lambda(t|x)}{\Lambda(t|x^*)} = e^{\beta'(x-x^*)} \Rightarrow \ln(\Lambda(t|x)) - \ln(\Lambda(t|x^*)) = \beta'(x - x^*).$$

Esto implica que las curvas $\ln(\Lambda(t|x))$ y $\ln(\Lambda(t|x^*))$ respecto a t son paralelas.

La función de supervivencia condicional dada $X^i = x$ viene dada por

$$S(t|x) = S_0(t|x)^{e^{\beta' x}}, \text{ donde } S_0(t) = e^{-\Lambda_0(t)},$$

puesto que $S(t|x) = e^{-\Lambda(t|x)} = e^{-\Lambda_0(t)e^{\beta' x}} = [S_0(t)]^{e^{\beta' x}}$.

Por tanto, en un modelo de riesgos proporcionales las funciones de supervivencia de dos individuos nunca se cruzan.

2.3.1. Estimación de los parámetros de regresión: verosimilitud parcial

La estimación de los coeficientes β se obtiene maximizando la función de verosimilitud parcial, que introduciremos a continuación. Veamos primero cómo se construye cuando no hay empates entre los tiempos de ocurrencia del evento, $t_1 < t_2 < \ldots < t_K$.

Como ya hemos mencionado el modelo de Cox es semiparamétrico, porque no asume una forma paramétrica para $\lambda_0(t)$. Supongamos, entonces, que $\lambda_0(t)$ es arbitraria. No es posible hacer inferencia sobre β a partir de intervalos de tiempo en los que no se producen eventos, ya que en estos intervalos

$\lambda_0(t)$ podría tomar valor cero. Por tanto, la estimación de β debe realizarse sobre el conjunto de tiempos en los que se produce el evento.

Los datos de los que disponemos se basan en una muestra de tamaño n de la forma (Z_i, δ_i, X^i), $i = 1, \ldots, n$. Se asume que la censura es no informativa, es decir, dado Z_i, el evento y el tiempo de censura para el $i - ésimo$ individuo de la muestra son independientes. Sea $R(t_k)$ el conjunto de individuos a riesgo en el tiempo t_k, es decir, todos los individuos que no están censurados ni han experimentado el evento en t_k. La probabilidad condicional de que el $i - ésimo$ individuo de la muestra ordenada, con covariables $X^{(i)}$, experimente el evento dadas las observaciones a riesgo en ese momento es

$$\frac{\lambda(t_k|X^{(i)})}{\sum_{j\in R(t_k)}\lambda(t_k|X^j)} = \frac{\lambda_0(t_k)e^{\beta' X^{(i)}}}{\sum_{j\in R(t_k)}\lambda_0(t_k)e^{\beta' X^j}} = \frac{e^{\beta' X^{(i)}}}{\sum_{j\in R(t_k)}e^{\beta' X^j}}.$$

Cada fallo aporta un factor de este tipo, de modo que la probabilidad condicional requerida viene dada por

$$L(\beta) = \prod_{k=1}^{K}\frac{e^{\beta' X^i}}{\sum_{j\in R(t_k)}e^{\beta' X^j}}. \tag{2.8}$$

Esta función se denomina verosimilitud parcial, porque pierde los factores correspondientes a las observaciones censuradas. Es un producto de términos, uno por cada tiempo de fallo t_k. Para cada factor, el numerador es el riesgo del individuo que experimentó el evento en t_k, y el denominador es la suma de todos los riesgos en el conjunto $R(t_k)$.

En presencia de empates, existen varias propuestas para construir la verosimilitud parcial.

Sean $t_1 < t_2 < \ldots < t_K$ los K tiempos de fallo distintos ordenados y ζ_k cada uno de los eventos empatados en el tiempo t_k. En la primera aproximación, debida a Breslow (1975), se consideran los ζ_k como si fueran distintos, se calculan sus contribuciones a la función de verosimilitud y se obtiene la contribución a la verosimilitud multiplicando sobre todos los evento en t_k. De esta forma, la verosimilitud parcial toma la forma

$$L_B(\beta) = \prod_{k=1}^{K}\frac{e^{\beta' S_k}}{\left[\sum_{j\in R(t_k)}e^{\beta' X^j}\right]^{\zeta_k}},$$

donde S_k es la suma de los X^j sobre todos los individuos que experimentan el evento en t_k. Cuando hay pocos empates esta aproximación funciona bastante bien. En caso contrario, tiende a producir estimaciones de β sesgadas hacia el cero.

Otras aproximaciones son las de Efron (1977), muy similar a la de Breslow cuando hay pocos empates, y la de Cox (1972).

Efron también asume que los eventos ocurren uno en cada tiempo t, de modo que también tiende a devolver estimaciones sesgadas de los coeficientes hacia el cero cuando el número de empates es grande. Sin embargo, esta propuesta ofrece estimaciones más precisas que el método de Breslow.

$$L_E(\beta) = \prod_{k=1}^{K}\frac{e^{\beta' S_k}}{\prod_{r=1}^{\zeta_k}\left(\sum_{j\in R(t_k)}e^{\beta' X^j} - \frac{r-1}{\zeta_k}\sum_{j\in K(t_k)}e^{\beta' X^j}\right)}.$$

Por su parte, el método de Cox supone que, si hay tiempos de fallo empatados, entonces estos ocurrieron al mismo tiempo. El enfoque utiliza modelos de tiempo discreto en los que se asume un modelo logístico para la tasa de riesgo, es decir, se permite que $\lambda(t|x)$ denote la probabilidad condicional de muerte en el intervalo $(t, t+1)$ dada la supervivencia al inicio del intervalo. Para construir la verosimilitud, denotaremos por Q_k al conjunto de todos los posibles subconjuntos de ζ_k individuos que podrían ser seleccionados del conjunto de riesgo $R(t_k)$. Cada elemento de Q_k es una $\zeta_k - tupla$ de individuos que podrían haber sido uno de los ζ_K eventos en el momento t_k. Sea $q = (q_1, \ldots, q_{\zeta_k})$ uno de

estos elementos de Q_k y s_{q^*} la suma de los X^j sobre todos los individuos de q, es decir, $s_{q^*} = \sum_{j=1}^{\zeta_j} X_{qj}$. Entonces, la verosimilitud de Cox viene dada por

$$L_C(\beta) = \prod_{k=1}^{K} \left[\frac{e^{\beta' S_k}}{\sum_{j=1}^{n} \mathbb{I}\{Z_j \geq Z_i\} e^{\beta' X^j}} \right]^{\delta_i}.$$

Cuando no hay empates, las tres aproximaciones se reducen a la verosimilitud parcial definida inicialmente en (2.8).

Otra forma de escribir la verosimilitud parcial a partir de la muestra observada $\left\{(Z_i, \delta_i, X^i)\right\}_{i=1}^{n}$ es

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta' X^i}}{\sum_{j=1}^{n} \mathbb{I}\{Z_j \geq Z_i\} e^{\beta' X^j}} \right]^{\delta_i}. \tag{2.9}$$

En el caso de empates, esta expresión coincide con la verosimilitud parcial de Breslow.

A partir de (2.9) se estiman los parámetros β del modo habitual. Primero, se calcula el logaritmo de la verosimilitud:

$$\ln L(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta' X^i - \ln \left(\sum_{j=1}^{n} \mathbb{I}\{Z_j \geq Z_i\} e^{\beta' X^j} \right) \right].$$

Y, seguidamente, se deriva con respecto a cada una de las P componentes de $\beta = (\beta_1, \ldots, \beta_P)'$ para obtener la función *score*, $U(\beta) = \frac{\partial \ln L(\beta)}{\partial \beta}$, cuya componente $p-ésima$ viene dada por:

$$U_p(\beta) = \frac{\partial \ln L(\beta)}{\partial \beta_p} = \sum_{i=1}^{n} \delta_i \left[X_{ip} - \frac{\sum_{j=1}^{n} \mathbb{I}\{Z_j \geq Z_i\} X_{jp} e^{\beta' X^j}}{\sum_{j=1}^{n} \mathbb{I}\{Z_j \geq Z_i\} e^{\beta' X^j}} \right].$$

Los estimadores de máxima verosimilitud parcial de los coeficientes, $\hat{\beta}$, se obtienen resolviendo las p ecuaciones no lineales $U_p(\beta) = 0,\ p = 1, \ldots, P$. Esto puede hacerse numéricamente, por ejemplo, aplicando un método iterativo como el de Newton-Raphson, tal y como se describe en Klein et al. (2005). Nótese que las ecuaciones no dependen de $\lambda_0(t)$, de modo que se pueden estimar los parámetros de regresión sin conocer la función de riesgo basal.

La matriz de información se corresponde con el negativo de la matriz de segundas derivadas del logaritmo de la verosimilitud, $I(\beta) = \left(-\frac{\partial^2 \ln L(\beta)}{\partial \beta_p \beta_k} \right),\ 1 \leq p,\ k \leq P$. Esta matriz permite hacer inferencia sobre los coeficientes estimados.

2.3.2. Contrastes de hipótesis sobre los parámetros

Muchas de las pruebas de hipótesis utilizadas en Análisis de Supervivencia aprovechan las propiedades asintóticas de la verosimilitud y verosimilitud parcial. Estas pruebas se basan tanto en la verosimilitud maximizada, como en los estimadores estandarizados por la matriz de información o en la primera derivada del logaritmo de la verosimilitud.

Bajo condiciones bastante generales se verifica que, para muestras grandes, el estimador de máxima verosimilitud parcial de β, $\hat{\beta}$, tiene una distribución Normal P-dimensional con media β y matriz de varianzas-covarianzas estimada por $I^{-1}(\hat{\beta})$. A partir de esta normalidad asintótica y de la matriz de varianzas-covarianzas estimada, se pueden calcular regiones de confianza y hacer contrastes de hipótesis sobre los parámetros.

Son tres los test asintóticos principales para evaluar la hipótesis sobre los coeficientes de regresión $H_0 : \beta = \beta_0$, generalmente con $\beta_0 = 0$.

La primera prueba es el test de Wald, basada en la normalidad asintótica del estimador parcial de máxima verosimilitud de β:

$$X_W^2 = (\hat{\beta} - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0).$$

La segunda prueba es el test de razón de verosimilitudes, que utiliza

$$X_{LR}^2 = 2\left[\ln L(\hat{\beta}) - \ln L(\beta_0)\right].$$

La última prueba es el *score* test, que utiliza la función *score* $U(\beta)$. Para muestras grandes y bajo H_0, $U(\beta)$ se distribuye según una Normal P-dimensional con media cero y matriz de varianzas-covarianzas $I(\beta)$. El estadístico del contraste viene dado por

$$X_{SC}^2 = U(\beta_0)' I(\beta_0)^{-1} U(\beta_0).$$

X_W^2, X_{LR}^2 y X_{SC}^2 siguen una distribución χ^2 con P grados de libertad cuando la hipótesis nula es cierta.

Frecuentemente, interesa hacer contrastes sobre un subconjunto de elementos de β. Surgen así los test locales. Consideraremos ahora la hipótesis $H_0 : \beta^{(q)} = \beta_0^{(q)}$, donde $\beta^{(q)}$ es un vector de q componentes ($q < P$) de β. Para realizar el contraste, podemos aplicar los mismos tests que en el caso general, pero considerando ciertas modificaciones.

En el caso del test de Wald, se utiliza la matriz de covarianzas correspondiente a las q componentes elegidas. Sea $\hat{\beta} = (\hat{\beta}^{(q)'}, \hat{\beta}^{(P-q)'})'$ el estimador de máxima verosimilitud (parcial) de β, donde $\beta^{(q)'}$ es un vector de dimensión $q \times 1$ del subconjunto de parámetros de $\hat{\beta}$ de interés, y $\hat{\beta}^{(P-q)'}$ es el vector $P - q$ de parámetros restantes. Supongamos la siguiente partición de la matriz de información:

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

donde $I_{11}(I_{22})$ es la submatriz $q \times q[(P-q) \times (Pq)]$ de la segunda derivada parcial del menos logaritmo de la verosimilitud con respecto a $\hat{\beta}^{(q)}(\hat{\beta}^{(P-q)})$ e I_{12} e I_{21} las matrices de las segundas derivadas parciales mixtas. El estadístico del test de Wald es ahora

$$X_W^2 = (\hat{\beta^{(q)}} - \beta_0^{(q)})' \left[I^{11}(\hat{\beta})\right]^{-1} (\hat{\beta^{(q)}} - \beta_0^{(q)}),$$

donde $I^{11}(\hat{\beta})$ es la submatriz superior $q \times q$ de $I^{-1}(\hat{\beta})$.

Sea $\hat{\beta}^{(P-q)}$ la estimación de máxima verosimilitud (parcial) de $\beta^{(P-q)}$ con las primeras q componentes fijadas en un valor $\beta_0^{(q)}$. El test de razón de verosimilitudes viene dado por la expresión

$$X_{LR}^2 = 2\left[\ln L(\hat{\beta}) - \ln L(\hat{\beta}^{(P-q)})\right].$$

En el caso del *score* test, sea $U_1\left[\beta_0^{(q)}, \hat{\beta}^{(P-q)}(\beta_0^{(q)})\right]]$ el vector de *scores* $q \times 1$ para $\beta^{(q)}$, evaluado en el valor hipotético de $\beta_0^{(q)}$ y en el estimador de máxima verosimilitud parcial restringido para $\beta^{(P-q)}$. Entonces

$$X_{SC}^2 = U_1\left[\beta_0^{(q)}, \hat{\beta}^{(P-q)}(\beta_0^{(q)})\right]' I^{11}(\beta_0^{(q)}, \hat{\beta}^{(P-q)}(\beta_0^{(q)})) U_1\left[\beta_0^{(q)}, \hat{\beta}^{(P-q)}(\beta_0^{(q)})\right].$$

Para muestras grandes, los estadísticos siguen ahora una distribución χ^2 con q grados de libertad cuando la hipótesis nula es cierta.

2.3.3. Estimación de la supervivencia condicional

Una vez obtenidas las estimaciones de los coeficientes de regresión, puede ser interesante estimar la supervivencia de un nuevo paciente con un conjunto determinado de valores para las covariables. El estimador de la supervivencia se basa en el estimador del riesgo acumulado basal de Breslow. Veamos primero cómo obtener este estimador para, posteriormente, definir el estimador de la supervivencia.

La verosimilitud parcial escrita como en (2.9) puede ser obtenida a partir de la verosimilitud condicional completa de los datos censurados, dada por

$$L_n = \prod_{i=1}^{n} f(Z_i|X^i)^{\delta_i}(1-F(Z_i|X^i))^{1-\delta_i} \prod_{i=1}^{n} g(Z_i|X^i)^{1-\delta_i}(1-G(Z_i^-|X^i))^{\delta_i},$$

donde $f(\cdot|x),\ F(\cdot|x),\ g(\cdot|x)$ y $G(\cdot|x)$ denotan, respectivamente, las funciones de densidad y distribución de los tiempos de fallo Y_i y de censura C_i condicionados a $X^i = x$.

Dado que Y_i y C_i son condicionalmente independientes a X^i, $F(\cdot|x)$ y $G(\cdot|x)$ no están relacionadas y, por tanto,

$$\begin{aligned} L_n &\propto \prod_{i=1}^{n} f(Z_i|X^i)^{\delta_i}(1-F(Z_i|X^i))^{1-\delta_i} = \prod_{i=1}^{n} \lambda(Z_i|X^i)^{\delta_i} e^{(-\Lambda(Z_i|X^i))} \\ &= \prod_{i=1}^{n} \left[\lambda_0(Z_i|X^i)e^{\beta' X^i}\right] e^{-\Lambda_0(Z_i|X^i)e^{\beta' X^i}} = \left[\prod_{k=1}^{K} \lambda_0(t_k)e^{\beta' X^i}\right] \left[\exp\left\{-\sum_{i=1}^{n} \Lambda_0(Z_i)e^{\beta' X^i}\right\}\right]. \end{aligned}$$

Considerando $\lambda_0(t) = 0\ \forall t$ no observado, con $\lambda_0(t_k)$ desconocidas que han de ser estimadas, la función de riesgo acumulada será

$$\Lambda_0(Z) = \sum_{k:t_k \le Z} \lambda_0(t_k).$$

En consecuencia,

$$\begin{aligned} L_n &= \left[\prod_{k=1}^{K} \lambda_0(t_k)e^{\beta' X^i}\right] \left[\exp\left\{-\sum_{i=1}^{n} e^{\beta' X^i} \sum_{k:t_k \le Z} \lambda_0(t_k)\right\}\right] \\ &= \left[\prod_{k=1}^{K} \lambda_0(t_k)e^{\beta' X^i}\right] \left[\prod_{k=1}^{K} \exp\left\{-\lambda_0(t_k) \sum_{i:Z_i \ge t_k} e^{\beta' X^i}\right\}\right]. \end{aligned}$$

Por otra parte,

$$\{i : Z_i \ge t_k\} = \{i : i \in R(t_k)\} = R(t_k)$$

$$L_n = \prod_{k=1}^{K} \left[\lambda_0(t_k)e^{\beta' X^i} \exp\left\{-\lambda_0(t_k) \sum_{i \in R(t_k)} e^{\beta' X^i}\right\}\right].$$

Fijando primero los β y maximizando respecto a $\lambda_0(t)$, la función a maximizar es

$$L_\beta(\lambda_0(t)) = \prod_{k=1}^{K} \left[\lambda_0(t_k)e^{\beta' X^i} \exp\left\{-\lambda_0(t_k) \sum_{i \in R(t_k)} e^{\beta' X^i}\right\}\right].$$

Esta función alcanza su máximo cuando $\lambda_0(t) = 0$, excepto para los tiempos en los que ocurre el

evento. Tomando logaritmos,

$$\ln L_\beta(\lambda_0) = \sum_{k=1}^{K} \ln \lambda_0(t_k) + \beta' X^i - \lambda_0(t_k) \sum_{i \in R(t_k)} e^{\beta' X^i},$$

maximizando respecto a $\lambda_0(t)$ y sustituyendo el máximo obtenido $\hat{\lambda}_0(t)$, que depende de β fijo, llegamos al estimador de máxima verosimilitud de $\hat{\lambda}_0(t, \beta)$:

$$\frac{\partial}{\partial \lambda_0(t)} = \frac{1}{\lambda_0(t)} - \sum_{i \in R(t_k)} e^{\beta' X^i} = 0 \Rightarrow \hat{\lambda}_0(t, \beta) = \frac{1}{\sum_{i \in R(t_k)} e^{\beta' X^i}}.$$

A partir de esta expresión se obtiene el estimador de Breslow del riesgo acumulado basal, dado por

$$\hat{\Lambda}_0(t) = \sum_{t_k \leq t} \frac{d_k}{\sum_{i \in R(t_k)} e^{\hat{\beta}' X^i}}, \tag{2.10}$$

que es una función escalonada con saltos en los tiempos de fallo observados. A partir de este, llegamos al estimador de la supervivencia basal, que se define como

$$\hat{S}_0(t) = e^{\left[-\hat{\Lambda}_0(t)\right]} \tag{2.11}$$

y representa la función de supervivencia de un individuo con todas las covariables iguales a 0. Si las covariables son numéricas y están centradas, esta supervivencia equivale a la de un individuo promedio.

El estimador de la función de riesgo acumulado condicional es

$$\hat{\Lambda}(t|x) = \hat{\Lambda}_0(t) e^{\hat{\beta}' x},$$

de modo que, para estimar la supervivencia de un individuo condicionada a los valores de las covariables, aplicaremos la expresión

$$\hat{S}(t|x) = \hat{S}_0(t)^{e^{\hat{\beta}' x}}. \tag{2.12}$$

Bajo condiciones de regularidad bastante suaves, para un t fijo, este estimador de la supervivencia condicional sigue una distribución normal asintótica con media $S(t|x)$ y una varianza que puede estimarse mediante

$$\hat{V}\left[\hat{S}(t|x)\right] = \left[\hat{S}(t|x)\right]^2 \left[Q_1(t) + Q_2(t; x)\right],$$

donde:

$$Q_1(t) = \sum_{t_k \leq t} \frac{d_k}{W(t_k, \hat{\beta})^2}$$

es un estimador de la varianza de $\hat{\Lambda}_0(t)$ si $\hat{\beta}$ es el verdadero valor de β,

$$Q_2(t; x) = Q_3(t; x)' \hat{V}(\hat{\beta}) Q_3(t; x)$$

representa la incertidumbre en el proceso de estimación que se añade al estimar β, Q_3 es el P-vector cuyo elemento $p - ésimo$ viene dado por

$$Q_3(t, x)_p = \sum_{t_k \leq t} \left[\frac{W^{(p)}(t_k; \hat{\beta})}{W(t_k; \hat{\beta})} - X_p\right] \left[\frac{d_k}{W(t_k, \hat{\beta})}\right],\ p = 1, \ldots, P$$

y

$$W^{(p)}(t_k;\hat{\beta}) = \sum_{i \in R(t_k)} X_{ip} e^{\hat{\beta}' X^i}.$$

$Q_3(t,x)$ es tanto más grande cuanto más lejos estén las covariables de los valores medios en el conjunto de riesgo.

A partir de esta aproximación normal asintótica se pueden construir intervalos de confianza puntuales para la supervivencia.

2.4. Extensiones del modelo de Cox

Como hemos visto, el modelo de Cox proporciona una expresión para el riesgo de un individuo con ciertos valores de las covariables en un instante de tiempo t. Este riesgo es el producto de dos cantidades. Por un lado, $\lambda_0(t)$, la función de riesgo de referencia y, por otro, la exponencial del predictor lineal $\beta' X = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P$. Una propiedad importante de esta fórmula, relacionada con el supuesto de riesgos proporcionales, es que el riesgo de referencia es una función del tiempo, pero no implica a las covariables, mientras que el factor exponencial involucra a las covariables pero no al tiempo. Por tanto, en el modelo de Cox la razón de riesgo que compara dos especificaciones cualesquiera de covariables es constante a lo largo del tiempo. Sin embargo, hay casos en los que el supuesto de riesgos proporcionales no se cumple para alguna de las covariables.

En tales casos, existen dos alternativas principales al modelo de Cox, la estratificación y la inclusión de variables dependientes del tiempo. En los siguientes apartados se describen ambos procedimientos.

2.4.1. Estratificación

Cuando la hipótesis de riesgos proporcionales no se cumple para una o varias covariables del modelo, es posible crear estratos a partir de estas covariables y emplear el modelo de riesgos proporcionales dentro de cada estrato considerando el resto de predictores. A continuación, presentamos el modelo de Cox estratificado siguiendo a Kleinbaum et al. (2011).

El modelo de Cox estratificado es una modificación del modelo de Cox que permite controlar los predictores que incumplen el supuesto de riesgos proporcionales. La estratificación permite ajustar un modelo de riesgos proporcionales en cada estrato, con coeficientes β comunes pero funciones de riesgo basal que pueden ser distintas y no relacionadas. Las variables que satisfacen el supuesto de riesgos proporcionales se incluyen en los modelos, de modo que podemos estimar su efecto ajustado por el resto de covariables y por la variable de estratificación. Sin embargo, esta última no puede incluirse, dado que no cumple con una de las hipótesis básicas, por lo que no será posible estimar su efecto sobre el riesgo (no hay coeficiente β para esta covariable). Este es el precio a pagar por la estratificación.

Formulación del modelo de Cox estratificado

Supongamos que disponemos de K variables que no satisfacen la hipótesis de riesgos proporcionales, $W_1, W_2, \ldots, W_K$, y de P-K variables que sí cumplen el supuesto, $X_1, X_2, \cdots, X_{P-K}$. Nótese que, ahora, K denota un subconjunto de covariables en lugar del número de tiempos de ocurrencia del evento de interés en la muestra observada. El proceso de estratificación implica la definición de una nueva variable, W^*, cuyas categorías resultan de todas las posibles combinaciones de las categorías de las variables de estratificación. Cada una de las combinaciones define, por tanto, un estrato. Por ejemplo, supongamos $K = 2$, con W_1 una variable categórica binaria con clases a y b y W_2 una variable continua categorizada en tres grupos, c, d y e. La nueva variable W^* tendrá como categorías las combinaciones ac, ad, ae, bc, bd y be, que definen a cada uno de los estratos del modelo.

En general, la variable de estratificación W^* tendrá K^* categorías, donde K^* es el número total de combinaciones (estratos) formados después de categorizar cada una de las W_k variables. En el ejemplo anterior, K^* es igual a seis.

En el modelo de Cox estratificado, la función de riesgo viene dada por la expresión

$$\lambda_k(t|x) = \lambda_{0k}(t)e^{\beta' x},\ k = 1, 2, \ldots, K,$$

donde k es el indicador de estrato. Nótese que solo se incluyen explícitamente en el modelo las variables X que cumplen con el supuesto de riesgos proporcionales, pero no la variable de estratificación W_k. Obsérvese también que la función de riesgo basal, $\lambda_{0k}(t)$, puede ser diferente en cada estrato.

Dado que en el modelo estratificado las funciones de riesgo entre los estratos difieren en la medida en la que tienen distintas funciones de riesgo basal, el modelo da lugar a diferentes curvas de supervivencia. Sin embargo, los coeficientes β van a ser los mismos en todos los ajustes. Esta característica se conoce como hipótesis de no interacción. Veremos más adelante cómo evaluarla.

La estimación e inferencia sobre los coeficientes de regresión se realiza a partir de la función de log-verosimilitud parcial que resulta de sumar las log-verosimilitudes parciales en cada estrato:

$$\ln L(\beta) = \ln L_1(\beta) + \ln L_2(\beta) + \cdots + \ln L_{K^*}(\beta)$$

siendo $\ln L_k(\beta)$ la log-verosimilitud parcial definida para el modelo de Cox calculada utilizando únicamente los datos de los individuos en el estrato k. El $\ln L(\beta)$ se maximiza con respecto a β siguiendo la metodología descrita en la Sección 2.3.1.

Hipótesis de no interacción

Anteriormente señalamos que en el modelo de Cox estratificado los coeficientes de regresión no varían a lo largo de los estratos, esto es, se supone la no interacción. Cuando se permite la interacción entre la variable de estratificación y una o varias covariables, pueden obtenerse distintas estimaciones de los coeficientes para estas covariables en cada estrato. Esto ocurre si ajustamos modelos de riesgo separados con los datos de cada estrato.

Para evaluar la hipótesis de no interacción y determinar qué modelo resulta más apropiado, debemos examinar la función de riesgo del modelo con interacción. Una forma de plantear la fórmula para el riesgo cuando hay interacción, considerando $K = 2$ estratos y $P = 2$ covariables es:

$$\lambda_k(t|x) = \lambda_{0k}(t)e^{\beta_{1k}X_1 + \beta_{2k}X_2},\ k = 1, 2. \tag{2.13}$$

Cada covariable en este modelo tiene asociado un coeficiente β_{pk} diferente para cada estrato. En cambio, en un modelo sin interacción, de la forma

$$\lambda_k(t|x) = \lambda_{0k}(t)e^{\beta_1 X_1 + \beta_2 X_2},\ k = 1, 2,$$

los coeficientes β_p son los mismos para el estrato 1 y el 2.
Alternativamente, podemos escribir el modelo con interacción como

$$\lambda_k(t|x) = \lambda_{0k}(t)e^{\beta_1^* X_1 + \beta_2^* X_2 + \beta_3^*(W^* X_1) + \beta_4^*(W^* X_2)},\ k = 1, 2, \tag{2.14}$$

donde W^* es la variable de estratificación. Esta formulación contiene los efectos principales de las covariables y los términos de interacción. Los riesgos basales $\lambda_{0k}(t)$ siguen siendo diferentes para cada estrato, pero ahora los coeficientes β no incluyen el subíndice k y, por tanto, son comunes a ambos estratos. No obstante, esta formulación es equivalente a la presentada en (2.13). Para probarlo, consideremos que la variable de estratificación W^* es una variable binaria, que vale 1 en el estrato 1 y 0 en el estrato 2. Entonces:

$$\lambda_1(t|x) = \lambda_{01}(t)e^{\beta_1^* X_1 + \beta_2^* X_2 + \beta_3^*(1\times X_1) + \beta_4^*(1\times X_2)} = \lambda_{01}(t)e^{(\beta_1^* + \beta_3^*)X_1 + (\beta_2^* + \beta_4^*)X_2}$$
$$\lambda_2(t|x) = \lambda_{02}(t)e^{\beta_1^* X_1 + \beta_2^* X_2 + \beta_3^*(0\times X_1) + \beta_4^*(0\times X_2)} = \lambda_{02}(t)e^{\beta_1^* X_1 + \beta_2^* X_2}.$$

Los coeficientes de X_1 son distintos en cada estrato. Concretamente, $(\beta_1^* + \beta_3^*)$ en el estrato 1 y β_1^* en el estrato 2. Lo mismo ocurre para X_2, acompañada de $(\beta_2^* + \beta_4^*)$ en el estrato 1 y de β_2^* en el

estrato 2.

Por tanto, tenemos

$$\lambda_1(t|x) = \lambda_{01}(t)e^{\beta_{11}X_1+\beta_{21}X_2}$$
$$\lambda_1(t|x) = \lambda_{01}(t)e^{(\beta_1^*+\beta_3^*)X_1+(\beta_2^*+\beta_4^*)X_2}$$

para $k = 1$ y $W^* = 1$, y

$$\lambda_2(t|x) = \lambda_{02}(t)e^{\beta_{12}X_1+\beta_{22}X_2}$$
$$\lambda_2(t|x) = \lambda_{02}(t)e^{\beta_1^*X_1+\beta_2^*X_2}$$

para $k = 2$ y $W^* = 0$.

Entonces, para el estrato 1, β_{11} debe ser equivalente a $(\beta_1^* + \beta_3^*)$ y β_{21} debe ser equivalente a $(\beta_2^* + \beta_4^*)$. Análogamente, se deduce que β_{12} debe ser equivalente a β_1^* y β_{22} debe ser equivalente a β_2^* para el estrato 2. Luego (2.13) y (2.14) son, tal y como habíamos sugerido, modelos equivalentes.

Hemos visto que el modelo de interacción puede escribirse en una forma que contiene términos de interacción, en los que la variable de estratificación, W^*, se multiplica por cada una de las covariables no estratificadas. Emplearemos ahora este modelo para evaluar la hipótesis de no interacción.

La prueba empleada es un test de razón de verosimilitud que compara la log-verosimilitud del modelo sin interacción y del modelo con interacción. El estadístico del contraste es de la forma

$$-2\ln L_R - (-2\ln L_C),$$

donde los subíndices R y C denotan, respectivamente, el modelo sin interacción (reducido) y el modelo con interacción (completo). Bajo la hipótesis nula de no interacción, la distribución de este estadístico se aproxima a una χ^2 con tantos grados de libertad como efectos de interacción incluye el modelo completo ($g.l. = p(k^* - 1)$).

Para finalizar esta Sección, indicamos algunas consideraciones prácticas acerca del procedimiento de estratificación:

i. La estratificación funciona de forma natural para las variables categóricas, siendo los estratos los distintos niveles del factor. Las variables cuantitativas pueden ser discretizadas, por ejemplo, a partir de los cuartiles. No obstante, no existe un consenso sobre qué puntos de corte elegir para la discretización y tampoco se conoce cómo esto puede afectar al modelo.

ii. Cuando se emplea la estratificación, las pruebas de hipótesis sobre los coeficientes de regresión solo tendrán buena potencia si las desviaciones de las hipótesis nulas son las mismas en todos los estratos.

iii. Las pruebas de hipótesis sobre los coeficientes de regresión ofrecen resultados fiables cuando el tamaño de la muestra dentro de cada estrado es grande o cuando el número de estratos es grande.

iv. La estimación de la función de supervivencia y/o de la función de riesgo acumulado para cada estrato puede obtenerse utilizando los estimadores descritos en la Sección 2.3.3.

2.4.2. Variables dependientes del tiempo

Como se indicó en la Sección anterior, la segunda alternativa para lidiar con covariables que no cumplen el supuesto de riesgos proporcionales es considerar su dependencia en el tiempo.

De acuerdo con Kleinbaum et al. (2011), una variable explicativa se dice independiente del tiempo o fija cuando su valor permanece constante a lo largo del tiempo t. El ejemplo más típico de este tipo de variables es el sexo. No obstante, también pueden considerarse como variables fijas la edad, la altura o el tratamiento, en la medida en que sus valores se mantienen constantes a lo largo de la duración del

estudio. Por el contrario, una variable es dependiente del tiempo cuando su valor para un individuo determinado puede diferir a lo largo del tiempo.

Dentro de las variables dependientes del tiempo podemos distinguir variables definidas en función del tiempo, variables internas y variables auxiliares. La mayoría de las variables definidas en función del tiempo resultan del producto entre una variable fija y una función del tiempo, que aquí denotaremos con $g(t)$. Por ejemplo, $E \times \log(t-3)$, donde E es una variable indicadora de exposición a un factor en el momento de la entrada en el estudio, es una variable definida en función del tiempo que vale 0 en ausencia de exposición y $\log(t-3)$ cuando hay exposición al factor. Otro ejemplo de variable definida es $E \times g(t)$, donde $g(t)$ es una variable que toma valor 1 si t es mayor o igual a un valor específico t_0, y toma valor 0 si t es menor que t_0. Así, siempre que $t \geq t_0$, $g(t) = 1$ y $E \times g(t) = E$. Sin embargo, si $t < t_0$, $g(t) = 0$, por lo que el valor de $E \times g(t)$ es siempre 0. Las funciones tipo $g(t)$, como veremos más adelante, pueden ser utilizadas como método de análisis cuando una variable fija en el tiempo como E no satisface el supuesto de riesgos proporcionales.

El segundo tipo de variables dependientes del tiempo, llamadas variables internas, de caracterizan porque pueden cambiar con el tiempo para cualquier sujeto en estudio y, además, la razón del cambio depende de comportamientos propios del individuo. Algunos ejemplos de variables internas son el *estatus* de fumador de un individuo o su nivel de obesidad en el momento t.

Por último, las variables auxiliares son aquellas cuyo valor cambia en el tiempo debido a factores externos, que pueden afectar simultáneamente a varios individuos. El índice de contaminación atmosférica en el momento t en una zona geográfica concreta es un ejemplo de este tipo de variables.

En ocasiones, una variable dependiente del tiempo puede considerarse como interna y auxiliar. Es el caso de la situación laboral de un individuo en el momento t que, supongamos, es de demandante de empleo. La falta de trabajo puede deberse a que el individuo no busca empleo activamente (variable interna) o a una situación económica de crisis (variable auxiliar).

Sea cual sea el la tipología, en presencia de variables dependientes del tiempo el modelo de Cox sigue siendo aplicable, pero la hipótesis de riesgos proporcionales deja de cumplirse. Se habla, entonces, del modelo de Cox extendido.

Formulación del modelo de Cox con variables dependientes del tiempo

La formulación del modelo de Cox con variables tanto independientes como dependientes del tiempo es la siguiente:

$$\lambda(t|x(t)) = \lambda_0(t)e^{\beta' x(t)} = \lambda_0(t) \exp\left[\sum_{p=1}^{P} \beta_p X_p + \sum_{q=1}^{Q} \gamma_q X_q(t)\right].$$

Al igual que el modelo de riesgos proporcionales, esta extensión incluye una función de riesgo de referencia $\lambda_0(t)$ y una función exponencial. Sin embargo, esta última contiene ahora tanto predictores independientes del tiempo, denotados por X_p, como predictores dependientes del tiempo, denotados por $X_q(t)$. El conjunto completo de predictores en el tiempo t es $X(t)$.

Un supuesto importante del modelo de Cox extendido es que el efecto de una variable dependiente del tiempo sobre la probabilidad de supervivencia en el tiempo t depende del valor de la variable en t y no de su valor en un momento anterior o posterior. Aunque los valores de la variable pueden cambiar con el tiempo, el modelo solo proporciona un coeficiente para cada variable dependiente del tiempo. Por tanto, en t solo hay un valor de la variable que tiene efecto sobre el riesgo, y es el valor medido en el tiempo t. En otras palabras, la historia hasta el tiempo t, $H_{t^-} = \{x(u), 0 \leq u < t\}$, no afecta al riesgo en el instante t, $\lambda(t)$, pero sí tiene influencia sobre el riesgo acumulado

$$\Lambda(t|H_{t^-}, x(t)) = \int_0^t \lambda_0(u)e^{\beta' x(u)}du.$$

La integral anterior es difícil de calcular porque depende del proceso aleatorio $X(t)$. Por eso, estimar la supervivencia o la función de riesgo acumulada o la supervivencia en un modelo de Cox con

covariables dependientes del tiempo no resulta sencillo. No obstante, es posible modificar la definición de la variable dependiente del tiempo para permitir un efecto de "retardo". Para ilustrar esta idea consideraremos el ejemplo citado en Kleinbaum et al. (2011). Supongamos como variable dependiente del tiempo la situación laboral medida semanalmente, EMP(t). Un modelo que no considera el retardo supone que el efecto de la situación laboral sobre la probabilidad de supervivencia en la semana t depende únicamente del valor observado en esa misma semana t, y no del valor de EMP(t) en la semana anterior, por ejemplo. Sin embargo, si se desea tener en cuenta un desfase de una semana, la variable situación laboral puede modificarse para que el riesgo en t sea predicho por la situación laboral en la semana $t-1$. Así, EMP(t) se sustituye en el modelo por EMP$(t-1)$:

$$\lambda(t|x(t)) = \lambda_0(t)e^{\beta^* \text{EMP}_{(t-1)}}.$$

De forma más general, el modelo de Cox extendido puede escribirse para permitir una modificación en tiempo real de cualquier variable dependiente del tiempo. Sea L_q el tiempo de retardo especificado para la variable dependiente del tiempo q, entonces el modelo extendido en tiempo de retardo es

$$\lambda(t|x(t)) = \lambda_0(t) \exp\left[\sum_{p=1}^{P} \beta_p X_p + \sum_{q=1}^{Q} \gamma_q X_q(t - L_q)\right],$$

donde la variable $X_q(t)$ ha sido sustituida por $X_q(t - L_q)$.

Volviendo al modelo extendido general, la expresión para la razón de riesgos es

$$\frac{\lambda(t|x(t))}{\lambda(t|x^*(t))} = \frac{\lambda_0(t)e^{\beta' x(t)}}{\lambda_0(t)e^{\beta' x^*(t)}} = \exp\left[\sum_{p=1}^{P} \beta(X_p - X_p^*) + \sum_{q=1}^{Q} \gamma\left[X_q(t) + X_q^*(t)\right]\right].$$

Esta fórmula describe la razón de riesgos o *hazard ratio* en el momento t para dos individuos con conjuntos de covariables (fijas y dependientes del tiempo) $x(t)$ y $x^*(t)$.

Dado que la fórmula de la razón de riesgos implica diferencias en los valores de las covariables dependientes del tiempo en el momento t, esta relación es también una función del tiempo. Por eso, en general, el modelo de Cox con variables dependientes del tiempo no satisface la hipótesis de riesgos proporcionales si al menos un γ_q no es igual a 0. Este γ_q, que en sí mismo no depende de t y toma un valor único, representa el efecto global sobre el riesgo de la correspondiente variable dependiente del tiempo considerando todos los tiempos en los que la variable fue medida a lo largo de la duración del estudio.

Estimación de los parámetros de regresión

Los métodos de estimación e inferencia en el modelo de Cox extendido son esencialmente los mismos que para el caso de riesgos proporcionales. La función de verosimilitud parcial a maximizar es ahora

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta' X^i(Z_i)}}{\sum_{j=1}^{n} \mathbb{I}\{Z_j \geq Z_i\} e^{\beta' X^j(Z_i)}}\right]^{\delta_i}. \tag{2.15}$$

Esta expresión generaliza la verosimilitud parcial del modelo de Cox, dada en (2.9), al caso de variables dependientes del tiempo. La diferencia está en que, en cada tiempo observado no censurado, Z_i, se tiene en cuenta el valor de las covariables dependientes del tiempo en ese instante a través del término $X^j(Z_i)$. Además, la expresión anterior no incluye el riesgo basal $\lambda_0(t)$, que se cancela al igual que ocurría en la verosimilitud de riesgos proporcionales. Por tanto, tampoco es necesario conocer el riesgo de referencia para estimar los parámetros de regresión. Se pueden emplear los test de Wald y de razón de verosimilitudes para hacer contrastes de hipótesis sobre los parámetros y también se pueden construir regiones de confianza.

Para calcular la verosimilitud en (2.15), se deben escribir los datos en formato *start-stop* para

definir explícitamente las variables dependientes del tiempo. Con este formato, que permite múltiples observaciones para un mismo individuo, el tiempo total de seguimiento en riesgo de un individuo se subdivide en intervalos de tiempo más pequeños, de modo que los valores de las variables pueden cambiar de un intervalo a otro. Veamos un ejemplo.

Se dispone de datos para cuatro individuos sobre el tiempo de seguimiento, estatus (ocurrencia o no de un cierto evento de interés) y condición de fumador. La variable dependiente del tiempo es Fumador x Tiempo.

ID	**Tiempo**	**Estatus**	**Fumador**
1	2	1	1
2	3	1	0
3	5	0	0
4	8	1	1

El modelo de Cox extendido viene dado por la expresión

$$\lambda(t) = \lambda_0(t)e^{\beta_1(Fumador)+\beta_2(Fumador\times Tiempo)}.$$

Para escribir los datos en formato *start-stop* debemos definir las variables *start*, inicio del intervalo, *stop*, fin del intervalo y modificar la variable estatus para indicar si el evento de interés ocurre o no en el intervalo. Además, también se incluye la variable dependiente del tiempo Fumador x Tiempo.

ID	***Start***	***Stop***	**Estatus**	**Fumador**	**FumadorxTiempo**
1	0	2	1	1	2
2	0	3	1	0	0
3	0	5	0	0	0
4	0	2	0	1	2
4	2	3	0	1	3
4	3	8	1	1	8

Para este modelo, no solo el riesgo de referencia puede cambiar con el tiempo, sino también el valor del predictor Fumador x Tiempo. Esto puede ilustrarse examinando el riesgo del individuo 4 en cada tiempo t. Este individuo, que es fumador, experimenta el evento de interés en el tiempo 8. Sin embargo, en $t = 2,\ 3$ y 8 la covariable Fumador x Tiempo cambia de valores, lo que afecta al riesgo del individuo en cada tiempo de evento:

Tiempo	**Riesgo del individuo 4**
2	$\lambda_0(t)e^{\beta_1+2\beta_2}$
3	$\lambda_0(t)e^{\beta_1+3\beta_2}$
8	$\lambda_0(t)e^{\beta_1+8\beta_2}$

La verosimilitud parcial extendida para estos datos es

$$L = L_1 \times L_2 \times L_3 = \left[\frac{e^{\beta_1+2\beta_2}}{e^{\beta_1+2\beta_2}+e^0+e^0+e^{\beta_1+2\beta_2}}\right] \times \left[\frac{e^0}{e^0+e^0+e^{\beta_1+3\beta_2}}\right] \times \left[\frac{e^{\beta_1+8\beta_2}}{e^{\beta_1+8\beta_2}}\right] = \left[\frac{e^{\beta_1+2\beta_2}}{e^{\beta_1+2\beta_2}+2+e^{\beta_1+2\beta_2}}\right] \times \left[\frac{1}{2+e^{\beta_1+3\beta_2}}\right].$$

La expresión anterior es el resultado del producto de tres términos, L_1, L_2 y L_3, uno por cada individuo que sufre el evento. El individuo 1 lo hace en $t = 2$ y es fumador, el individuo 2 en $t = 3$ y no es fumador y el individuo 4 en $t = 8$ y es fumador. El individuo 3, no fumador y censurado en $t = 5$, todavía está en riesgo cuando los individuos 1 y 2 experimentan el fallo, por eso se tiene en cuenta en los denominadores de L_1 y L_2. Nótese que la inclusión de la covariable Fumador x Tiempo no modifica la expresión del riesgo para los no fumadores (individuos 2 y 3), ya que la variable fumador vale 0 para ellos. Sin embargo, para los fumadores el riesgo sí cambia con el tiempo.

2.5. Validación del modelo de regresión

Típicamente, la validación de los modelos de regresión se basa en el análisis de residuos. En el caso del modelo de Cox y sus extensiones, se han propuesto diferentes tipos de residuos que resultan útiles para examinar varios aspectos del modelo:

i. la forma funcional de una covariable considerando el resto de predictores;

ii. la hipótesis de riesgos proporcionales;

iii. la influencia (*leverage*) de cada individuo sobre la estimación de los coeficientes de regresión y

iv. la falta de ajuste del modelo a un determinado individuo.

A continuación, presentamos los principales residuos utilizados en la validación del modelo de Cox. Más adelante, nos centraremos en la descripción de las técnicas empleadas para chequear la hipótesis básica de este modelo, el supuesto de riesgos proporcionales.

2.5.1. Residuos de la regresión

Residuos martingala

Definiremos los residuos martingala siguiendo lo expuesto por Therneau et al. (1990).

Consideremos cada individuo de la muestra como un proceso de recuento independiente $(N_i(t), t \geq 0,\ t = 1, \ldots, n)$ con función de intensidad

$$Y_i(t)e^{\beta' X^i(t)}d\Lambda_0(t),$$

donde $Y_i(t)$ es un proceso 0-1 que indica si el $i-ésimo$ individuo está a riesgo en el tiempo t, β es el vector de parámetros de regresión, $X^i(t)$ es el vector de covariables del individuo i en el tiempo t y $d\Lambda_0(t)$ representa una función de riesgo sin especificar. Para definir los residuos martingala, tomaremos como base las diferencias entre el proceso de recuento y la integral de su función de intensidad, esto es,

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)e^{\beta' X^i(s)}d\Lambda_0(s).$$

Entonces, $M_i(.)$ será una martingala específica del $i-ésimo$ individuo. Sustituyendo en la ecuación anterior el vector de parámetros desconocidos, β, por su estimador de máxima verosimilitud parcial

$\hat{\beta}$ y la función de riesgo desconocida, Λ_0, por el estimador de Breslow definido en (2.10), los residuos martingala se definen como:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\hat{\beta}' X^i(s)} d\hat{\Lambda}_0(s).$$

En cada tiempo t, el residuo representa la diferencia sobre $[0, t]$ entre el número observado de eventos y el número esperado dado el modelo o, equivalentemente, el exceso de fallos en t. Estos residuos toman valores entre $-\infty$ y 1, con $\sum \hat{M}_i(t) = 0$ y $\mathbb{E}(\hat{M}_i) = cov(\hat{M}_i, \hat{M}_j) = 0$.

En ausencia de covariables dependientes del tiempo, donde Z_i denota el tiempo observado para el $i - ésimo$ individuo de la muestra y δ_i es el indicador de censura, los residuos martingala se reducen a la expresión

$$\hat{M}_i(t) = \delta_i - \hat{\Lambda}_0(Z_i) e^{\beta' X^i}. \tag{2.16}$$

Estos residuos son especialmente útiles para explorar la forma funcional correcta de una covariable dentro del modelo de Cox. En principio, el efecto de la covariable X_p sobre el logaritmo del riesgo relativo se recoge en el término $\beta_p X_p$ del predictor lineal. Sin embargo, puede que resulte más apropiado considerar un término del tipo $\beta_p f(X_p)$, con f una cierta transformación no lineal. Para evaluar esta posibilidad, consideremos el modelo

$$\Lambda(t, X, X_p) = h(X_p) e^{\beta' X} \Lambda_0(t) = e^{f(X_p)} e^{\beta' X} \Lambda_0(t),$$

donde la forma funcional del vector de covariables X es conocida pero la función positiva $h(X_p)$ no está especificada. La representación de los residuos martingala frente a X_p puede utilizarse para obtener estimaciones de h o f. Concretamente, si $\hat{M}_i(t)$ denota el residuo martingala bajo el modelo anterior pero ajustado ignorando X_p, entonces X_p y X son independientes y

$$\mathbb{E}\{\hat{M}(t)|X_p\} \simeq \left\{1 - \frac{\bar{h}}{h(X_p)}\right\} \mathbb{E}\left\{N(t)|X_p\right\}, \tag{2.17}$$

donde $\bar{h}$ implica promedios tanto en el tiempo como en la composición esperada del conjunto de riesgo. Aunque $\bar{h}$ también es una función de X_p, será casi constante si se compara con la variación de $h(X)$ cuando X_p y X son independientes. De acuerdo con (2.17), el número esperado de exceso de eventos es aproximadamente uno menos el *hazard ratio* multiplicado por el número medio de eventos.

Dado que $\hat{M}$ y N son conocidas, podemos invertir (2.17) para obtener

$$f(X_p) - \bar{f} \simeq -\log\left(\frac{1 - sm(\hat{M}, X_p)}{sm(N, X_p)}\right),$$

donde $\bar{f} \equiv \log(\bar{h})$, $sm(\hat{M}, X_p)$ es una estimación suavizada de $\mathbb{E}\{\hat{M}(t)|X_p\}$, obtenida de suavizar el gráfico de dispersión de $\hat{M}$ frente a X_p, y $sm(N, X_p)$ es una suavización de $\mathbb{E}\{N(t)|X_p\}$. En muchos casos, la expresión anterior para $t = \infty$ puede ser sustituida por una de estas dos aproximaciones,

$$\mathbb{E}(\hat{M}|X_p) \simeq \{f(X_p) - \bar{f}\}c \quad \text{o} \quad \mathbb{E}(\hat{M}|X_p) \simeq \{h(X_p) - \bar{h}\}c,$$

con c igual al número total de eventos dividido entre el número total de individuos. La primera aproximación es aplicable cuando la dependencia de $\mathbb{E}(N|X_p)$ en X_p es débil, como en el caso de datos completos o moderadamente censurados. La segunda se utiliza cuando $Y(.)$ es independiente de $N(.)$. Una de las ventajas de representar los residuos martingala según estas aproximaciones es la interpretabilidad: el eje Y estará en una escala directa de exceso de eventos. Por tanto, un gráfico suavizado de los $\hat{M}_i$ respecto a la covariable X_p mostrará de manera aproximada su verdadera forma funcional:

- Si X_p no está incluida en el modelo, el suavizado sugerirá la forma funcional correcta para la covariable.
- Si X_p está incluida en el modelo y bien modelada, el suavizado no deberá mostrar ninguna tendencia.

Residuos de Schoenfeld

Una de las principales asunciones del modelo de Cox es que el riesgo de los individuos permanece constante en el tiempo y solo depende de los valores de las covariables. Sin embargo, este supuesto no se mantiene cuando las covariables o coeficientes de regresión cambian con el tiempo. Por ello, siempre es necesario comprobar si el modelo se ajusta adecuadamente a los datos.

Los residuos de Schoenfeld resultan muy útiles para evaluar la hipótesis de riesgos proporcionales. Siguiendo a Xue et al. (2017), estos residuos se calculan para cada individuo que experimenta el evento como la diferencia entre el valor observado de sus covariables y el valor esperado de acuerdo con los sujetos a riesgo en el momento del fallo.

Sea $X^i(t_k),\ i = 1, \cdots, n;\ k = 1, \ldots, K$, el vector de covariables del individuo que experimenta el evento en el tiempo t_k y $R(t_k)$ el conjunto de riesgo en t_k. El residuo de Schoenfeld se define como

$$r_k(\beta) = X^i(t_k) - \mathbb{E}\left[X^i(t_k)|R(t_k)\right].$$

Cuando no hay empates entre los tiempos de evento,

$$r_k(\beta) = X^i(t_k) - \bar{x}(\beta, t_k),$$

donde

$$\bar{x}(\beta, t_k) = \frac{\sum_{i=1}^{n} Y_i(t_k)\gamma_i(t_k)X^i(t_k)}{\sum_{i=1}^{n} Y_i(t_k)\gamma_i(t_k)} \tag{2.18}$$

es una media ponderada de X con pesos $Y_i(t_k)\gamma_i(t_k)$ sobre las observaciones a riesgo en t_k. Estos pesos se corresponden con el riesgo de cada sujeto en $R(t_k)$.

En la práctica, los residuos estimados $\hat{r}_k$ se obtienen sustituyendo β por su estimador de máxima verosimilitud parcial, $\hat{\beta}$. Si la hipótesis de riesgos proporcionales se cumple, entonces $\mathbb{E}(\hat{r}_k) \simeq 0$. Así, se espera que la representación de los residuos frente a los tiempos de evento se aproxime a un gráfico de dispersión en torno al 0.

Grambsch et al. (1994) generalizaron el enfoque de Schoenfeld para probar el supuesto de riesgos proporcionales escalando cada residuo r_k por un estimador de su varianza:

$$r_k^* = r_k^*(\beta) = V^{-1}(\beta, t_k)r_k(\beta). \tag{2.19}$$

En la siguiente Sección, retomaremos los residuos de Schoenfeld escalados para describir algunos métodos de evaluación de la hipótesis de riesgos proporcionales.

Medidas de influencia

Al igual que en regresión lineal, algunos individuos pueden tener una influencia especialmente grande en las estimaciones de los coeficientes. Las observaciones influyentes pueden deberse a errores en la recogida de los datos, indicar algún problema con los mismos y/o producir distorsiones en los resultados. Por ello, su identificación es otro de los pasos importantes en el proceso de validación del modelo.

Tal y como indican Xue et al. (2017), una práctica habitual en el estudio de las observaciones influyentes es eliminar cada observación, ajustar el modelo y comparar las estimaciones de los parámetros con las obtenidas en el ajuste con los datos completos. Debido a que el modelo de Cox es semiparamétrico, una observación podrá ser influyente en términos de algo más que los coeficientes de regresión. Por

ello, es necesario medir la influencia en los coeficientes de regresión y la influencia general sobre la verosimilitud del modelo.

La influencia de las observaciones individuales en los parámetros del modelo de Cox puede ser estimada del siguiente modo. Sea $\hat{\beta}$ el estimador de máxima verosimilitud parcial del vector de coeficientes β y $\hat{\beta}_{-i}$ la estimación de β cuando la $i-ésima$ observación de la muestra ha sido eliminada. La influencia de la observación eliminada i, definida como la diferencia $\hat{\beta}-\hat{\beta}_{-i}$, puede ser aproximada asignando a dicha observación un peso w_i. Supongamos $w_j = 1$ para todo $j \neq i$. Entonces, $\hat{\beta}$ puede considerarse como una función de w_i con $\hat{\beta}(1) = \hat{\beta}$ y $\hat{\beta}(0) = \hat{\beta}_{-i}$. La expansión en serie de Taylor de primer orden sobre $w_i = 1$ da

$$\hat{\beta} - \hat{\beta}_{-i} \simeq \frac{\partial \hat{\beta}}{\partial w_i},\ i = 1, \ldots, n,$$

donde $\frac{\partial \hat{\beta}}{\partial w_i}$ es evaluado en $w_i = 1$. Si evaluamos la derivada en la función *score* U (derivada de la log-verosimilitud parcial), obtenemos:

$$\frac{\partial U}{\partial \hat{\beta}} \frac{\partial \hat{\beta}}{\partial w_i} + \frac{\partial U}{\partial w_i} = 0.$$

Nótese que $\frac{\partial U}{\partial \hat{\beta}}$ es la matriz de información observada $I(\hat{\beta})$. Por tanto, tenemos que

$$\frac{\partial \hat{\beta}}{\partial w_i} = I^{-1}(\hat{\beta}) \frac{\partial U}{\partial w_i}. \tag{2.20}$$

La derivada parcial $\frac{\partial U}{\partial w_i}$ evaluada en $w_i = 1$ se corresponde con el residuo *score* para la $i-ésima$ observación de la muestra,

$$r_{Ui}(\hat{\beta}) = \int_0^\infty \left[X^i - \bar{x}(\hat{\beta}, t) \right] d\hat{M}_i(t),$$

donde $\bar{x}(\hat{\beta}, t)$ es la media ponderada $\bar{x}(\beta, t)$ definida en (2.18) evaluada en $\beta = \hat{\beta}$ y $\hat{M}_i(t)$ es la estimación del residuo martingala. Así:

$$\left(\frac{\partial \hat{\beta}}{\partial w_i} \right)_{w_i = 1} = I^{-1}(\hat{\beta}) r_{Ui}.$$

Sea D la matriz $n \times xp$ cuyo elemento en la $i-ésima$ fila es $\hat{\beta} - \hat{\beta}_{-i}$ y r_U la matriz $n \times xp$ cuyo elemento en la $i-ésima$ fila es el vector de residuos *score* del individuo i. Entonces la aproximación anterior en forma de matriz es

$$D = r_U I^{-1}(\hat{\beta}).$$

La matriz D se conoce como matriz de residuos *dfbeta* (*difference of the beta*). Cada *dfbeta* es el cambio aproximado en el vector de coeficientes cuando se elimina la observación i. Al dividir D_{ij} por el error estándar de $\hat{\beta}_i$, que es la raíz cuadrada del $i-ésimo$ elemento de la diagonal de $I^{-1}(\hat{\beta})$, obtenemos D_S, la matriz de residuos *dfbetas*. Por convenio, una observación será influyente cuando $D_{S_{ij}} > 1$ para conjuntos de datos pequeños o medianos, mientras que para conjuntos de datos grandes el criterio a seguir es $D_{S_{ij}} > \frac{2}{\sqrt{n}}$.

Pettitt et al. (1989) apuntan que el método de eliminación de una sola observación puede hacer que algunos casos queden enmascarados, es decir, que la observación eliminada sea realmente un valor influyente y no se detecte. En su lugar, y adoptando el enfoque de Cook, sugieren emplear los pesos w para estudiar el llamado desplazamiento de la verosimilitud, definido como

$$LD(w) = 2\left[L(\hat{\beta}) - L(\hat{\beta}(w))\right],$$

donde $\hat{\beta}(w)$ maximiza la verosimilitud parcial ponderada

$$PL_w(\beta) = \prod_{i=1}^{n} \frac{e^{\beta' X^i(t)\delta_i w_i}}{\left[\sum_{j\in R(t)} w_j e^{\beta' X^j(t)}\right]^{\delta_i w_i}}. \tag{2.21}$$

Si $w_i = 0$ y $w_j = 1$, $\forall j \neq i$, entones $PL_w(\beta)$ es la verosimilitud parcial sin la $i-ésima$ observación y $\hat{\beta}(w)$ se corresponde con $\hat{\beta}_{-i}$. Una observación será tanto más influyente cuanto mayor sea el desplazamiento que provoca en la verosimilitud.

Considerando (2.21), una aproximación de segundo orden nos lleva a

$$L(\hat{\beta}) - L(\hat{\beta}(w)) \approx \frac{1}{2}\left[\hat{\beta} - \hat{\beta}(w)\right]' I^{-1}(\hat{\beta})\left[\hat{\beta} - \hat{\beta}(w)\right].$$

Sea $U_w(\beta)$ la función *score* correspondiente a la log-verosimilud parcial ponderada, $\hat{\beta} - \hat{\beta}(w)$ se puede aproximar por

$$\hat{\beta} - \hat{\beta}(w) \approx \left[\frac{\partial\hat{\beta}(w)}{\partial w'}\right]_{\hat{\beta}, w_0} (w_0 - w),$$

considerando

$$\frac{\partial\hat{\beta}(w)}{\partial w'} = I^{-1}(\hat{\beta})\frac{\partial U_w(\beta)}{\partial w'},$$

que no es más que la forma matricial de (2.20). Luego $LD(w)$ se reduce a

$$LD(w) = 2\left[L(\hat{\beta}) - L(\hat{\beta}(w))\right] \approx (w_0 - w)' r_U I^{-1}(\hat{\beta}) r_U'(w_0 - w),$$

donde w_0 es un vector de 1s y r_U es la matriz de residuos *score*.

Dada esta aproximación local de $LD(w)$ alrededor de w_0, podemos considerar la dirección del vector $l_{n\times 1}$ que maximiza $l' Bl$, donde $B = r_U I^{-1}(\hat{\beta}) r_U'$. En este caso, $I^{-1}(\hat{\beta})$ es definida positiva, de modo que la matriz simétrica B de dimensión $n \times n$ es semidefinida positiva. Si $l_{\max}$ es el eigenvector de longitud 1 de B correspondiente al mayor eigenvalor $\varsigma_{\max}$, entonces $l_{\max}' Bl_{\max}$ maximiza $l'Bl$ y tiene valor $\varsigma_{\max}$. Así, para encontrar los valores $l_{\max}$ y $\varsigma_{\max}$ es necesario calcular los eigenvectores y eigenvalores de la matriz B. Cada observación tendrá asociada una medida única de influencia, a saber, su elemento $l_{\max}$.

Puesto que $l_{\max}$ es invariante a la multiplicación de sus elementos por -1, las observaciones más influyentes a nivel local serán aquellas con valores de $l_{\max}$ grandes en términos absolutos. Una observación influyente a nivel local lo será también a nivel global, pero lo contrario no tiene por qué ser cierto.

2.5.2. Evaluación de la hipótesis de riesgos proporcionales

Son tres los métodos principales para evaluar la hipótesis de riesgos proporcionales del modelo de Cox, a saber, métodos gráficos, contrastes de bondad de ajuste y uso de variables dependientes del tiempo. Seguidamente, describimos cada una de estas técnicas siguiendo a Kleinbaum et al. (2011).

Métodos gráficos

El primer enfoque gráfico para probar la hipótesis de riesgos proporcionales consiste en representar y comparar las curvas de supervivencia logarítmica estimadas sobre las diferentes clases de una variable

categórica. Bajo la hipótesis de riesgos proporcionales, dichas curvas tienen que ser paralelas.

La curva de supervivencia log-log se obtiene tomando el logaritmo natural de la probabilidad de supervivencia estimada en cada tiempo t, esto es, $\ln(-\ln \hat{S}(t))$. Dado que $\hat{S}$ es una probabilidad, $\ln \hat{S}$ tomará siempre valores negativos. Como solo es posible calcular el logaritmo de número positivos, es necesario cambiar el signo del primer logaritmo antes de tomar el segundo. El valor $\ln(-\ln \hat{S}(t))$ puede ser positivo o negativo. Nótese que. mientras que la escala del eje Y de una curva de supervivencia estimada oscila entre 0 y 1, la escala correspondiente a una curva $\ln(-\ln)$ oscila entre $-\infty$ e ∞.

Para demostrar la utilidad de las curvas $\ln(-\ln)$ como técnica de validación de la hipótesis de riesgos proporcionales, comenzaremos escribiendo la fórmula del modelo de Cox en escala logarítmica. Recordemos que:

$$S(t|x) = S_0(t)^{e^{\beta' x}},$$

donde $S_0(t) = e^{-\Lambda_0(t)}$ es la función de supervivencia basal correspondiente a la función de riesgo acumulado basal $\Lambda_0(t)$.

La fórmula log-log requiere tomar el logaritmo de la función de supervivencia condicional dos veces, considerando un signo menos tal y como se indicó arriba:

$$\ln S(t|x) = e^{\beta' x} \times \ln S_0(t),\ 0 \leq S(t|x) \leq 1$$

$$\ln\left[-\ln S(t|x)\right] = \ln\left[-e^{\beta' x} \times \ln S_0(t)\right] = \ln\left[e^{\beta' x}\right] + \ln\left[-\ln S_0(t)\right] = \beta' x + \ln\left[-\ln S_0(t)\right]$$

La expresión resultante es la suma de dos términos, el producto de los coeficientes de regresión por las covariables y el logaritmo del logaritmo negativo de la supervivencia basal. Este segundo sumando puede ser positivo o negativo.

Consideremos dos especificaciones distintas para un conjunto de covariables, x y x^*, correspondientes a dos individuos. Las curvas de superviencia log-log para estos individuos serán

$$\ln\left[-\ln S(t|x)\right] = \beta' x + \ln\left[-\ln S_0(t)\right]$$

y

$$\ln\left[-\ln S(t|x^*)\right] = \beta' x^* + \ln\left[-\ln S_0(t)\right].$$

Si hayamos la diferencia entre ambas curvas, obtenemos

$$\ln\left[-\ln S(t|x)\right] - \left(\ln\left[-\ln S(t|x^*)\right]\right) = \beta' x + \ln\left[-\ln S_0(t)\right] - \beta' x^* - \ln\left[-\ln S_0(t)\right] = \beta'(x - x^*).$$

Obsérvese que la función de supervivencia basal $S_0(t)$ desaparece, por lo que la diferencia entre las curvas log-log no depende del tiempo t.

Alternativamente, podemos escribir la ecuación anterior expresando la curva de supervivencia log-log para uno de los individuos en función de la curva log-log del otro individuo más el término lineal que no depende de t.

$$-\ln\left[-\ln S(t|x)\right] = \ln\left[-\ln S(t|x^*)\right] + \beta'(x - x^*).$$

Por tanto, si trazamos las curvas log-log de supervivencia estimadas para dos individuos en un mismo gráfico, esperaremos que estas sean paralelas. Además, la distancia entre ellas vendrá dada por la cantidad constante $\beta' x - \beta' x^*$. Esto implica que la relación entre la supervivencia de ambos individuos se mantiene constante a lo largo del tiempo y, por tanto, que el modelo de riesgos proporcionales resulta adecuado para ese conjunto de predictores.

En la práctica, podemos estimar las curvas de supervivencia log-log a partir del estimador de Kaplan-Meier descrito en la Sección 2.2.1 del Capítulo 2 y empleando el estimador de máxima verosi-

militud parcial de β, $\hat{\beta}$. No obstante, este primer enfoque presenta algunos inconvenientes:

i. ¿Qué grado de paralelismo es aceptable?

   Esta decisión puede ser bastante subjetiva, sobre todo si el tamaño de la muestra es pequeño. En este sentido, Kleinbaum et al. (2011) recomiendan seguir una estrategia conservadora, asumiendo que se cumple la hipótesis de riesgos proporcionales a menos que la falta de paralelismo sea muy evidente.

ii. ¿Cómo discretizar una covariable continua?

   Si se definen demasiadas categorías, el número de datos en cada una de ellas puede ser demasiado pequeño, lo que dificulta la comparación de las curvas. Además, una discretización en k clases puede proporcionar un gráfico distinto dependiendo de los puntos de corte seleccionados. Así, el problema reside en seleccionar tanto el número de clases como los puntos de corte que definen cada clase. Kleinbaum et al. (2011) recomiendan, en la medida de lo posible, $k \leq 3$ y que los puntos de cortes sean lo más significativos posible, proporcionando también un equilibrio entre los tamaños de las clases.

La segunda técnica gráfica que permite evaluar la hipótesis de riesgos proporcionales se basa en comparar las curvas de supervivencia "observadas" y "predichas" por el modelo. Al igual que en el caso anterior, el método puede aplicarse para cada una de las covariables por separado o ajustadas por el resto de predictores.

En el primer caso se emplea el estimador de Kaplan-Meier de la supervivencia para obtener las curvas de supervivencia observadas. Primero, se dividen las observaciones en función de las categorías de la covariable que se desea evaluar y, a continuación, se calculan los estimadores de Kaplan-Meier de la supervivencia para cada clase.

Por otro lado, para obtener las curvas de supervivencia predichas, es necesario ajustar un modelo de Cox de riesgos proporcionales con la variable que se desea evaluar. Seguidamente, se estima la supervivencia condicional para cada categoría de la covariable sustituyendo por su valor correspondiente en la fórmula (2.12).

Al representar las curvas observadas y esperadas en un mismo gráfico, si para cada categoría del predictor estas son próximas, entonces el supuesto de riesgos proporcionales para la covariable resultará razonable. Basta con que una de las categorías no muestre una curva observada y una curva predicha similares para rechazar la hipótesis.

Este segundo enfoque comparte, en cierta medida, los dos primeros inconvenientes señalados para el método gráfico anterior:

i. ¿Qué grado de proximidad entre las curvas es aceptable?

   De nuevo, Kleinbaum et al. (2011) recomiendan aceptar la hipótesis de riesgos proporcionales a menos que las diferencias entre las curvas observadas y esperadas sean muy evidentes.

ii. ¿Como obtener las curvas esperadas para una variable continua?

   Cuando se trabaja con variables continuas, las curvas observadas se obtienen del mismo modo que en el caso de variables categóricas: se discretiza la variable para formar categorías y se calcula el estimador de Kaplan-Meier de la supervivencia en cada clase. De nuevo, el problema radica en seleccionar el número de categorías y los puntos de corte.

   En el caso de las curvas esperadas, existen dos opciones. Una vez discretizada la variable continua que se desea evaluar, la primera opción consiste en ajustar un modelo de Cox de riesgos proporcionales con $k-1$ variables *dummy* o ficticias [2], siendo k el número de categorías definidas. Para cada categoría, la curva esperada se corresponderá con la supervivencia condicional estimada considerando la variable *dummy* que define a esa categoría.

[2] Variables indicadoras que toman valor 1 cuando se cumple una condición específica y 0 en otro caso.

La segunda opción pasa también por utilizar un modelo de Cox de riesgos proporcionales pero que contiene al predictor continuo que se quiere evaluar. La curva esperada para una categoría se obtendrá a partir de la supervivencia condicional estimada considerando ahora un valor de la covariable que define a la clase, por ejemplo, el valor medio.

Para este segundo método gráfico, la estrategia que ajusta por el resto de covariables se basa en la estratificación, ajustando un modelo de Cox y calculando la supervivencia condicional en cada uno de los estratos.

Para finalizar, los residuos de Schoenfeld escalados propuestos por Grambsch et al. (1994) y definidos en (2.19) también se pueden emplear para evaluar el supuesto de riesgos proporcionales de manera gráfica. Si r^*_{pk} es el residuo de Schoenfeld escalado para la covariable p en el momento t_k y $\hat{\beta}_p$ es el coeficiente de regresión de Cox estimado y fijado en el tiempo bajo la hipótesis de riesgos proporcionales, entonces el valor esperado de r^*_{pk} es aproximadamente la desviación del valor real del coeficiente en el momento t_k, $\beta_p(t_k)$:

$$\mathbb{E}(r^*_{pk}) \approx \beta_p + \beta_p(t_k)$$

Esto permite obtener una aproximación de $\beta_p(t)$ añadiendo al residuo escalado el estimador $\hat{\beta}_p$.

A partir de estas aproximaciones, es posible chequear gráficamente si la pendiente de los coeficientes se mantiene o no constante en el tiempo y, con ello, la hipótesis de riesgos proporcionales. Los valores del eje Y para la covariable p serán las sumas de los residuos de Schoenfeld escalados con la correspondiente estimación de β_p. El resultado es un gráfico de estimación del coeficiente de regresión a lo largo del tiempo. Si resulta razonablemente plano, entonces el supuesto de riesgos proporcionales se mantendrá. Con todo, Xue et al. (2017) señalan que estos gráficos pueden ser difíciles de interpretar.

Contrastes de bondad de ajuste

Los contrastes de bondad de ajuste se basan en el uso de un estadístico de prueba y un valor p para evaluar la hipótesis de riesgos proporcionales para cada covariable del modelo. Nos centraremos en la descripción de dos pruebas que utilizan los residuos de Schoenfeld.

Como hemos visto, para cada covariable del modelo, el residuo de Schoenfeld para un individuo que experimenta el evento de interés en un tiempo t se define como la diferencia entre su valor observado de la covariable y el valor esperado de acuerdo con los individuos a riesgo en t. Si, por ejemplo, consideramos tres predictores, entonces tendremos tres residuos de Schoenfeld para cada individuo no censurado, uno por cada covariable.

La primera prueba se basa en la correlación lineal entre los residuos de Schoenfeld y el orden del tiempo de los eventos. Si la correlación es 0, entonces la covariable cumple con la hipótesis de riesgos proporcionales:

$$H_0 : \rho = 0$$

La prueba consta de los siguientes pasos:

1. Ajustar un modelo de Cox de riesgos proporcionales y obtener los residuos de Schoenfeld para la covariable que se desea evaluar (tantos como individuos no censurados hay en la muestra).

2. Crear una variable que ordene a los individuos no censurados de acuerdo con sus tiempos de evento. Es decir, para el sujeto que primero experimenta el evento esta variable debe tomar el valor 1. El individuo con el siguiente tiempo de evento más pequeño tendrá asignado el valor 2, y así sucesivamente.

3. Calcular la correlación ρ entre los residuos de Schoenfeld calculados en el paso 1 y la variable creada en el paso 2. Bajo la hipótesis nula, ρ sigue una distribución Normal estándar.

4. Calcular el valor p. Si existen evidencias suficientes en contra de H_0 se concluirá que la covariable incumple la hipótesis de riesgos proporcionales.

Es importante destacar que el valor p asociado a la prueba se calcula teniendo en cuenta el resto de predictores incluidos en el modelo.

En presencia de empates entre los tiempos de ocurrencia del evento, los residuos se dividen por el número de tiempos de fallo empatados en el conjunto de riesgo correspondiente y la estimación de la correlación se pondera por el número de tiempos empatados.

En `Stata` esta prueba utiliza los residuos de Schoenfeld escalados, propuestos por Grambsch et al. (1994), en lugar de los residuos definidos originalmente por Schoenfeld.

El otro contraste basado en los residuos de Schoenfeld escalados es el propuesto por Grambsch et al. (1994). Suponiendo que la verdadera función de riesgo es variable en el tiempo,

$$\lambda_i(t) = \lambda_0(t)e^{\beta^{'}(t)X^i(t)} = \lambda_0(t)e^{(\beta+G(t)\theta)^{'}X^i(t)},$$

donde $G(t)$ es una matriz diagonal con jj elementos $g_j(t)$, estos autores demostraron que el test estadístico

$$T(G) = \left(\sum_{k=1}^{K} G_k\hat{r}_k^*\right)^{'} D^{-1}\left(\sum_{k=1}^{K} G_k\hat{r}_k^*\right)$$

con

$$D = \sum_{k=1}^{K} G_k\hat{V}_kG_k^{'} - \left(\sum_{k=1}^{K} G_k\hat{V}_k\right)\left(\sum_{k=1}^{K}\hat{V}_k\right)^{-1}\left(\sum_{k=1}^{K} G_k\hat{V}_k\right)^{'},$$

sigue una distribución χ^2 con p grados de libertad. $\hat{V}_k$ es la varianza observada de $\hat{\beta}$ en el tiempo t_k, que puede ser aproxima por la matriz de varianza media $\bar{V} = \frac{I^{-1}(\hat{\beta})}{d}$.

Distintas elecciones de G dan lugar a distintos estadísticos. Las opciones más comunes son las transformaciones identidad, de rango, logarítmica y la versión continua a la izquierda del estimador de Kaplan-Meier de la supervivencia. La habilidad de las pruebas para detectar el incumplimiento de la hipótesis depende de la elección de G. No obstante, Grant et al. (2013) mostraron mediante simulación que el rendimiento de estas propuestas en presencia de covariables dependientes del tiempo es muy inestable y que su potencia depende en gran medida de factores desconocidos en la práctica, como cuándo cambia la razón de riesgo y en qué medida.

Variables dependientes del tiempo

Por último, discutiremos el uso de variables dependientes del tiempo como tercer método para probar la hipótesis de riesgos proporcionales. Cuando se utilizan variables dependientes del tiempo para evaluar el supuesto de riesgos proporcionales para una covariable fija, el modelo de Cox se extiende para contener términos de interacción que implican a la variable que se evalúa y a alguna función del tiempo. Por ejemplo, si se está evaluando el supuesto de riesgos proporcionales para la variable Sexo, el modelo de Cox ampliado incluiría, además de la variable Sexo, la variable creada artificialmente Sexo x Tiempo. Si el coeficiente del término de interacción resulta significativo, es decir, distinto de 0, asumiremos que la variable incumple el supuesto de riesgos proporcionales. Por el contrario, si no hay significación, solo podremos concluir que el modelo extendido no se ajusta adecuadamente a los datos.

Cuando se evalúa el supuesto para cada predictor por separado, el modelo de Cox extendido adopta la forma general

$$\lambda(t|x) = \lambda_0(t)e^{\beta^{'}x+\gamma(x\times g(t))},$$

donde $g(t)$ es una función del tiempo que puede adoptar distintas formas:

- $g(t) = t$, de modo que el término de interacción sería $x \times t$,
- $g(t) = \log t$ o
- $g(t) = \begin{cases} 1 & si \quad t \geq t_0 \\ 0 & si \quad t < t_0 \end{cases}$

La habilidad de la prueba para detectar el incumplimiento de la hipótesis de riesgos proporcionales va a depender de la elección de $g(t)$.

Aplicando el modelo anterior, la hipótesis de riesgos proporcionales se evalúa probando la significación del coeficiente de interacción γ. La hipótesis nula es, por tanto, $H_0 : \gamma = 0$. Nótese que, si H_0 es cierta, el modelo se reduce a un modelo de Cox de riesgos proporcionales con covariable $X = x$.

El contraste puede llevarse a cabo utilizando el estadístico de Wald o del test de razón de verosimilitudes. En ambos casos, bajo la hipótesis nula estos estadísticos siguen una distribución χ^2 con 1 grado de libertad, dado que estamos evaluando la hipótesis para una única covariable.

El modelo de Cox extendido también puede utilizarse para evaluar el supuesto de riesgos proporcionales para varios predictores al mismo tiempo, así como para un único predictor ajustado por el resto de covariables del modelo (que se supone que cumplen el supuesto). En el primer caso, el modelo toma la forma

$$\lambda(t|x) = \lambda_0(t) exp\left[\sum_{p=1}^{P} \beta_p X_p + \gamma_p (X_p \times g_p(t))\right].$$

Este modelo contiene las covariables que se evalúan como términos de efecto principal y también como términos de interacción. Diferentes predictores pueden requerir funciones de tiempo distintas. Por eso se utiliza la notación $g_p(t)$, que define la función del tiempo para la $p - ésima$ covariable.

La hipótesis de riesgos proporcionales se evalúa simultáneamente para distintas covariables considerando la hipótesis nula $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_P = 0$. Ahora se utiliza un estadístico de razón de verosimilitudes que calcula la diferencia entre los logaritmos de la verosimilitud del modelo de riesgos proporcionales y del modelo extendido, esto es,

$$LR = -2 \ln L_{\mathrm{RP}} - (-2 \ln L_{\mathrm{EXT}}).$$

Bajo H_0, LR sigue una distribución χ^2 con p grados de libertad, donde p es el número de predictores que se evalúan al mismo tiempo.

Si el contraste resulta significativo, podemos concluir que la hipótesis de riesgos proporcionales no se cumple para al menos uno de los predictores evaluados. Para determinar qué predictor o predictores no satisfacen el supuesto, podemos repetir la prueba sucesivamente eliminando en cada repetición uno de los términos de interacción.

Para probar la hipótesis de riesgos proporcionales para un único predictor X^* ajustado por el resto de covariables X_p, el modelo de Cox extendido es ahora

$$\lambda(t|x) = \lambda_0(t) exp\left[\sum_{p=1}^{P-1} \beta_p X_p + \beta^* X^* + \gamma^* (X^* \times g(t))\right].$$

La hipótesis nula es, por tanto, $H_0 : \gamma^* = 0$. De nuevo, podemos recurrir al estadístico de Wald o del test de razón de verosimilitudes para resolver el contraste. Bajo H_0 ambos se distribuyen según una χ^2_1.

El principal inconveniente del uso de variables dependientes del tiempo para evaluar el supuesto de riesgos proporcionales es la elección de $g_p(t)$. Esta elección no suele ser obvia. Además, distintas elecciones pueden resultar en distintas conclusiones acerca del cumplimiento de la hipótesis.

En esta Sección, hemos descrito y discutido las ventajas e inconvenientes de los métodos destinados a la evaluación de la hipótesis de riesgos proporcionales. Las pruebas de bondad de ajuste no son tan subjetivas como los métodos gráficos ni tan costosas en términos computacionales como el uso de variables dependientes del tiempo. No obstante, pueden ser poco específicas, en el sentido de que pueden no detectar desviaciones concretas de la hipótesis que sí pueden observarse con los otros métodos. Además, la potencia de las pruebas depende en gran medida del tamaño de la muestra. En consecuencia, se recomienda aplicar al menos dos de los métodos disponibles para evaluar el supuesto de riesgos proporcionales.

2.6. Comparación de la supervivencia en dos o más grupos

En Análisis de Supervivencia resulta muy interesante la comparación de las funciones de supervivencia de dos o más grupos. Con este objetivo, sin embargo, es habitual trabajar con las funciones de riesgo o razón de fallo, que caracterizan a la supervivencia.

Consideremos dos grupos de pacientes, un grupo 1 sometido a un cierto tratamiento experimental y un grupo 2 que recibió un placebo. Se desea contrastar si las funciones de riesgo en esos dos grupos son iguales y, así, analizar el efecto del tratamiento. Se plantea el siguiente contraste de hipótesis:

$$H_0 : \lambda_1(t) = \lambda_2(t),\ t \geq 0$$
$$H_a : \lambda_1(t) \neq \lambda_2(t) \text{ para algún } t$$

2.6.1. El test log-rank

El estadístico de prueba log-rank compara las funciones de riesgo estimadas de dos grupos en cada tiempo de ocurrencia del evento. Se construye calculando el número de eventos observados y esperados en uno de los grupos en cada tiempo de evento observado y sumándolos después para obtener un resumen general. Ilustraremos este procedimiento siguiendo a Iglesias Pérez et al. (2021).

Sean $t_1 < t_2 < \ldots < t_K$ los distintos tiempos de la muestra conjunta (sin separar por grupos) donde ocurre el evento. En cada tiempo t_k se observan:

- n_{k1} individuos a riesgo en t_k en el grupo 1,
- d_{k1} eventos en t_k en el grupo 1,
- n_{k2} individuos a riesgo en t_k en el grupo 2 y
- d_{k2} eventos en t_k en el grupo 2.

Además, $n_k = n_{k1} + n_{k2}$ y $d_k = d_{k1} + d_{k2}$ son, respectivamente, el número de individuos a riesgo y el número de eventos en t_k para la muestra conjunta.

Bajo la hipótesis nula, se espera que los riesgos observados en cada grupo sean parecidos y, a su vez, similares al riesgo observado en la muestra conjunta para todos los tiempos del evento. Así, para el grupo m esperamos que

$$\frac{d_{km}}{n_{km}} \approx \frac{d_k}{n_k}\ k = 1, \ldots, K.$$

Esto sugiere construir el estadístico

$$U_m = \sum_{k=1}^{K} w_m(t_k) \left[\frac{d_{mk}}{n_{mk}} - \frac{d_k}{n_k} \right],$$

que mide cuánto se separa el riesgo del grupo m del riesgo conjunto.

En la práctica, se toma una función de pesos $w_m(t_k) = n_{km}w(t_k),\ m = 1, 2$, que conduce a

$$U_m = \sum_{k=1}^{K} w(t_k) \left[d_{km} - n_{km}\frac{d_k}{n_k} \right] = \sum_{k=1}^{K} w(t_k) \left[o_{km} - e_{km} \right],$$

siendo U_m la diferencia entre el número de eventos observado y el esperado en el grupo m ponderados a lo largo del tiempo.

La función de pesos $w(t_k)$ permite dar más o menos importancia a las diferencias en momentos de tiempo específicos. Concretamente, el test log-rank toma $w(t_k) = 1$, siendo el test de mayor potencia bajo el supuesto de riesgos proporcionales del modelo de Cox. Entonces, para este test, tenemos que

$$U_m = \sum_{k=1}^{K} \left[d_{km} - n_{km}\frac{d_k}{n_k} \right] = \sum_{k=1}^{K} \left[o_{km} - e_{km} \right].$$

Bajo la hipótesis nula, d_{km}, conocidos n_{km}, n_k y d_k, sigue una distribución hipergeométrica $H(N = n_k,\ n = d_k,\ p = \frac{n_{km}}{n_k})$, de forma que

$$\mathbb{E}(d_{km}) = n_{km}\frac{d_k}{n_k} = e_{km}$$

y

$$v_{km} = Var(d_{km}) = n_{km}\frac{d_k}{n_k}\left(1 - \frac{n_{km}}{n_k}\right)\frac{n_k - d_k}{n_k - 1} = \frac{n_{km}d_k(n_k n_{km})(n_k - d_k)}{n_k^2(n_k - 1)}.$$

Se sigue que $\mathbb{E}(U_m) = 0$ y $Var(U_m) = \sum_{k=1}^{K} v_{km}$.

Teniendo esto en cuenta, se llega finalmente al test estadístico asintótico

$$\frac{U_m}{\sqrt{Var(U_m)}} \sim N(0, 1) \Leftrightarrow \frac{U_m^2}{Var(U_m)} \sim \chi_1^2.$$

Rechazaremos la hipótesis nula de igualdad de riesgos entre grupos para valores del estadístico mayores que el percentil $(1 - \alpha)100\,\%$ de una χ_1^2, siendo α del nivel de significación.

El test se puede llevar a cabo considerando indistintamente U_1 o U_2. Además, se cumple que $U_1 + U_2 = 0$.

2.6.2. El test log-rank estratificado

En ocasiones, interesa comparar el riesgo en dos grupos ajustando por una variable categórica con pocas clases, digamos, C. Esto se consigue mediante el llamado test log-rank estratificado, cuya hipótesis nula para dos grupos es

$$H_0 : \lambda_{1c} = \lambda_{2c},\ \forall c = 1,\ 2,\ \ldots,\ C,\ t \geq 0.$$

Esencialmente, en cada nivel de la covariable se calcula un estadístico U_{mc} y la correspondiente varianza V_{mc}, donde m indica el grupo y c el estrato. El estadístico del contraste se construye entonces como

$$X^2 = \frac{(\sum_{c=1}^{C} U_{mc})^2}{\sum_{k=c}^{C} V_{mc}}.$$

La única diferencia entre los test log-rank estratificado y no estratificado es que en este último caso, el número observado de eventos menos el número esperado en cada tiempo de fallo se suma sobre todos los tiempos de fallo en cada grupo m. En la versión estratificada del test, sin embargo, la suma se realiza sobre todos los tiempo de fallo en cada grupo m dentro de cada estrato c, y luego se suma en

todos los estratos. En cualquier caso, si H_0 es cierta, X^2 se distribuye de manera asintótica de acuerdo con una χ^2 con $M-1$ grados de libertad, donde M representa el número de grupos que se comparan (no de estratos).

El principal inconveniente del test log-rank estratificado es el tamaño de la muestra. Si el número de observaciones es pequeño, el número de individuos en cada estrato será pequeño y el test perderá potencia.

Los test anteriores para dos grupos pueden extenderse al caso general de K muestras. Para cada muestra $k = 1,\ 2,\ \ldots,\ K$, se define el estadístico U_k, que representa la diferencia entre los eventos observados y esperados para la muestra k ponderados a lo largo del tiempo. Se verifica que $\sum_{k=1}^{K} U_k = 0$, de modo que para construir el test solo es necesario considerar $K-1$ estadísticos U_k. El test se rechaza para valores grandes del estadístico, que bajo H_0 sigue una distribución χ^2 con $K-1$ grados de libertad cuando $n \to \infty$.

Capítulo 3

Métodos

3.1. Fuentes de datos

La información utilizada en los análisis procede fundamentalmente de dos bases de datos.

Los datos relativos a los casos de infección y su gravedad se encuentran en VIXÍA, una aplicación de la DXSP para la vigilancia de enfermedades de declaración obligatoria. Esta aplicación contiene los casos de COVID-19 diagnosticados, tanto en el sistema público de salud como en el ámbito privado, mediante una prueba diagnóstica de infección activa (PDIA), que puede ser una PCR con muestra nasofaríngea o una prueba rápida de detección de antígenos, incluidos los autotest. Para cada caso, VIXÍA registra todas las pruebas (técnicas) que dieron un resultado positivo, considerando como fecha de diagnóstico la de la primera PDIA positiva. Los casos de COVID-19 notificados a VIXÍA proceden de dos fuentes, ATENEA y REXEL.

ATENEA es una herramienta corporativa para el acceso a la información del sistema público de salud. Los laboratorios gallegos de microbiología están dotados de un sistema de información en el que registran todas las pruebas microbiológicas realizadas y, en particular, las relacionadas con la COVID-19. Esta información se cruza, mediante el NASI o identificador de persona, con otras aplicaciones de ATENEA para obtener datos personales, identificar al personal sanitario del SERGAS e identificar a los usuarios y trabajadores de centros sociosanitarios, dado el impacto del SARS-CoV-2 en estos espacios.

Por otro lado, REXEL es un formulario web, de registro electrónico, en el que se recogen todas las pruebas positivas de COVID-19 realizadas en laboratorios, hospitales y clínicas de carácter privado.

VIXÍA completa la información de los casos de COVID-19 con la encuesta epidemiológica, un cuestionario realizado a los positivos por parte de los Servicios de Medicina Preventiva de los hospitales, Jefaturas territoriales y centros sociosanitarios para la definición de cada caso en términos de su sintomatología, tipo (caso fuente, huérfano o asociado a brote), riesgo y contactos.

Por otro lado, los datos relativos a la vacunación frente al SARS-CoV-2 proceden del registro gallego de vacunas que, a día 27 de marzo de 2022, tenía información de 2.381.967 personas de 12 años o más, el 97 % de la población gallega de esa edad (que comprendía 2.455.719 personas a 1 de enero de 2021 según datos del Padrón del *IGE - Instituto Galego de Estatística* (2021)). El registro de vacunas contiene información sobre la persona (NASI, sexo, edad, lugar de residencia y grupo de prioridad para recibir la vacuna), punto de vacunación y vacuna (tipo de vacuna y fecha de administración de cada dosis). La vacunación se inició el 27 de diciembre de 2020, y la dosis de recuerdo comenzó a administrarse el 1 de septiembre de 2021.

Tanto la base de VIXÍA como la del registro de vacunas COVID-19 se actualizan a diario y comparten el NASI como identificador único de persona. A partir del cruce entre las dos bases de datos se han obtenido las distintas variables utilizadas en los análisis. A mayores, la información sobre defunciones y sus fechas se ha obtenido del Registro de Mortalidad de Galicia cruzando también por NASI.

3.2. Diseño y población

Se llevó a cabo un estudio de cohortes retrospectivo a partir de los datos de vigilancia y vacunación frente al SARS-CoV-2 de Galicia. Se incluyeron en la cohorte todas las personas de 12 o más años, vacunadas con al menos una dosis entre el 27 de diciembre de 2020 y el 27 de marzo de 2022. Se excluyeron los menores de 12 años debido a la escasez de información sobre su vacunación, ya que comenzaron a recibir la primera dosis de forma masiva a finales de 2021. Por otro lado, el diseño epidemiológico planteado parte de una cohorte de individuos que todavía no han contraído la enfermedad de interés, por lo que se descartaron también todas las personas que, antes del inicio del estudio, tuviesen al menos un resultado positivo en una PDIA. Además, como el antecedente de SARS-CoV-2 disminuye la probabilidad de tener una nueva infección en los meses posteriores, y este antecedente puede influir en la decisión de recibir la vacuna, esta exclusión evitó un posible sesgo en los resultados. Se excluyeron todos los casos de COVID-19 acontecidos fuera de Galicia. Estos casos fueron identificados porque recibieron alguna dosis de la vacuna en nuestra comunidad, pero no se disponía de información sobre la fecha de su positivo, de modo que no se pudo saber si este había sido antes o durante el estudio. Se excluyeron los usuarios de residencias de mayores y otras instituciones por ser consideradas personas de riesgo. Se filtraron los datos de vacunación para eliminar registros con datos incompletos e incoherencias en la información sobre las dosis administradas. También se excluyeron los individuos para los cuales no se respetaron los tiempos mínimos entre dosis durante la primovacunación. Finalmente, se descartaron aquellas personas que recibieron la dosis adicional, por ser consideradas de riesgo, así como aquellas en las que no fue posible diferenciar si recibieron la dosis adicional o la dosis de recuerdo.

Se definieron cuatro grupos según el estatus de vacunación, considerando un período de inducción de 14 días para todas las vacunas salvo para Pfizer (siete días):

- **No vacunados (NV):** personas que no recibieron ninguna dosis de la vacuna.
- **Vacunados con una dosis (1D):** personas que recibieron una única dosis de Moderna, AstraZeneca o Pfizer.
- **Primovacunados (PV):** personas que recibieron una dosis de Janssen, o una única dosis de Moderna, AstraZeneca o Pfizer siendo menores de 65 años y siendo caso tras recibir dicha dosis, o personas que recibieron dos dosis de Moderna, AstraZeneca o Pfizer, o las combinaciones admitidas.
- **Primovacunados con dosis de recuerdo (PV+R):** personas que completaron la primovacunación y recibieron la dosis de recuerdo.

Se estimó la efectividad de las vacunas de COVID-19 frente a dos desenlaces, a saber, infección e ingreso en UCI por COVID-19. Se consideró como caso confirmado a cualquier individuo de la cohorte con una PDIA positiva entre el 27 de diciembre de 2020 y el 27 de marzo de 2022. Se consideró como ingreso en UCI por COVID-19 el ingreso de cualquier individuo de la cohorte en una UCI con una PDIA positiva, entre el 27 de diciembre de 2020 y el 27 de marzo de 2022, hasta un máximo de 14 días antes del ingreso o siete días después del ingreso.

En un primer análisis, se estimó la EV en individuos con primovacunación completa frente a no vacunados. Para ello, se consideró el seguimiento de los individuos desde el inicio del estudio hasta el 30 de noviembre de 2021. Se seleccionó esta fecha como fin del seguimiento en lugar del 27 de marzo de 2022 para excluir el período de la pandemia dominado por Ómicron. Como ya se indicó en el Capítulo 1, Ómicron fue una variante muy transmisible capaz de eludir la inmunidad. Esto supuso que muchos de los afectados por la quinta ola fuesen vacunados, por lo que no se tuvo en cuenta este período para evitar distorsiones en el cálculo de la EV de la primovacunación.

Por otro lado, se estimó la EV de la dosis de recuerdo frente a la primovacunación. En este caso, solo se siguieron los individuos que completaron la primovacunación durante el período de estudio, y que pudieron recibir o no la dosis de recuerdo. Así, el seguimiento de los participantes se realizó desde que completaron la primovacunación y hasta el 27 de marzo de 2022.

El fin de seguimiento de un individuo se produjo con el cambio de estado de vacunación, la defunción (por cualquier causa), la infección por COVID-19 o el fin del estudio, lo que antes sucediese. En el caso del desenlace ingreso en UCI, el seguimiento de los individuos finalizó en la fecha de toma de la muestra positiva y no en la fecha del ingreso. No se consideraron las reinfecciones, dado que el primer positivo suponía ya el fin del seguimiento del individuo.

Todos los participantes iniciaron el estudio en el grupo de NV. Si el seguimiento no finalizó por otra causa, la vacunación (con el período de inducción correspondiente) produjo un cambio de grupo que generó un nuevo registro en la base de datos, en el que se modificaron las fechas de inicio y de fin del seguimiento para ese individuo. Así, la base de datos resultante podía tener entre uno y cuatro registros por individuo, según el número de grupos de vacunación por los que pasara. A modo de ejemplo, nos basaremos en la siguiente tabla para mostrar algunas de las situaciones que se podían dar en el estudio.

Id	Grupo	Inicio	Fin	F.vacuna	Caso	F.caso	UCI	F.UCI	Defunción	F.defunción
16	NV	27-12-20	08-09-21	08-09-21	0	.	0	.	0	.
16	1D	08-09-21	09-10-21	09-10-21	0	.	0	.	0	.
16	PV	09-10-21	15-02-22	15-02-22	0	.	0	.	0	.
16	PV+R	15-02-22	27-03-222	.	0	.	0	.	0	.
63	NV	27-12-20	07-07-21	11-09-21	1	07-07-21	0	.	0	.
291	NV	27-12-20	15-04-21	15-04-21	0	.	0	.	0	02-08-21
291	1D	15-04-21	06-05-21	06-05-21	0	.	0	.	0	02-08-21
291	PV	06-05-21	02-08-21	.	0	.	0	.	1	02-08-21

- El individuo 16 inició su seguimiento como no vacunado el 27 de diciembre de 2020. No fue caso de COVID-19 ni tampoco falleció. Su seguimiento como no vacunado finalizó al recibir la primera dosis de la vacuna el 8 de septiembre de 2021. Esto produjo un cambio de grupo, generando un nuevo registro en la base de datos y modificando las fechas de inicio y de fin. El individuo completó la primovacunación el 9 de octubre, al recibir la segunda dosis, y su seguimiento en el grupo de primovacunados finalizó en febrero de 2022, cuando se le administró la dosis de recuerdo. En este último grupo finalizó su seguimiento coincidiendo con el fin del estudio, el 27 de marzo de 2022.

- El individuo 63 comenzó el estudio el 27 de diciembre de 2020 como no vacunado y fue caso de COVID-19 en verano de 2021, antes de recibir la primera dosis de la vacuna. Por ello, su seguimiento finalizó en el grupo de no vacunados y solo cuenta con un registro en la base de datos.

- El individuo 291 inició el estudio como no vacunado y pasó al grupo de una dosis el 15 de abril, ya que no fue caso ni falleció antes de esa fecha. Su segundo registro finalizó tras recibir la segunda dosis el 6 de mayo. Pasó entonces al grupo de primovacunados, en el que falleció por una causa distinta a la COVID-19 el 2 de agosto, fecha de fin de su seguimiento.

3.3. Variables de ajuste

Las variables a tener en cuenta para el ajuste de los modelos fueron el grupo de edad, el sexo, el ámbito de residencia (urbano, semiurbano o rural) y el número de PDIAs negativas realizadas durante el seguimiento. La última variable se construyó una vez preparada la base de datos, y se empleó como medida de exposición de los individuos frente al virus.

Se definieron grupos de edad decenales a partir de los 12 años: 12-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79 y 80 años o más.

Los ámbitos de residencia se establecieron en función de las zonas y subzonas de urbanización del año 2016 definidas por el *IGE - Instituto Galego de Estatística* (2021). En cuanto a las zonas,

se definen tres grandes categorías: zonas densamente pobladas (ZDP; densidad> 500 hab./Km^2 y población total≥ 50.000 hab.), zonas intermedias (ZIP; densidad> 100 hab./Km^2 y, o bien población total> 50.000 hab., o bien la zona es adyacente a una ZDP) y zonas poco pobladas (ZPP; grupos de áreas locales que no pertenecen a ZDP ni ZIP). A su vez, estas zonas se subdividen con el fin de diferenciar núcleos de características diferentes en ZIP altas y bajas y ZPP altas, intermedias y bajas.

De acuerdo con lo anterior, los individuos residentes en ZDP y ZIP altas se consideraron del ámbito urbano; los individuos residentes en ZIP bajas y ZPP altas e intermedias se consideraron del ámbito semiurbano, y los individuos residentes en ZPP bajas se asignaron al ámbito rural.

El número de PDIAs negativas realizadas a cada persona durante su seguimiento en los distintos grupos de vacunación se obtuvo cruzando la base de datos de la cohorte, por NASI, con la base de pruebas de ATENEA. La variable se categorizó en cuatro niveles: 0, 1, 2 y 3 o más.

3.4. Análisis estadístico

Una vez definidas las variables de interés, se analizó su distribución en la población para determinar cuáles serían incluidas finalmente en los modelos y de qué manera. Se analizó el cumplimiento de la hipótesis de riesgos proporcionales para cada una de las covariables como herramienta de diagnóstico previa al ajuste del modelo. Para ello, se obtuvieron las curvas log-log de supervivencia y las curvas observadas y esperadas. También se obtuvieron los estadísticos y el valor p asociados al contraste $H_0 : \gamma = 0$, donde γ es el coeficiente de interacción entre la covariable y el tiempo. Se consideró la función del tiempo $g(t) = t$. Recordemos que, si H_0 es cierta, se asume que la covariable cumple con el supuesto de riesgos proporcionales (ver Sección 2.5.2).

Se ajustó un modelo de Cox controlando por las covariables, susceptibles de confundir la relación entre el estatus vacunal y la COVID-19, para calcular los RR. Se estimó la EV frente a cada desenlace como $(1 - \text{RR})100\,\%$ con sus correspondientes intervalos de confianza del 95 %, en global, por grupo de edad y también por tipo de vacuna (de la primera dosis en el caso de las pautas heterólogas). Para manejar los empates en los tiempos hasta el evento, la verosimilitud parcial se calculó utilizando el método de Breslow.

Se estimó la pérdida de EV frente a infección e ingreso en UCI en individuos primovacunados calculando la EV a t días de seguimiento, considerando valores de t de cinco en cinco días desde cinco hasta un máximo de 180 (25,7 semanas). Las estimaciones se obtuvieron ajustando por grupo de edad, sexo, ámbito de residencia y número de PDIAs negativas. La pérdida de efectividad también se estimó por grupo de edad y por tipo de vacuna.

Finalmente, se analizó la influencia de las observaciones sobre las estimaciones de los coeficientes mediante el cálculo de los valores $l_{\max}$, que están relacionados con el desplazamiento a nivel local de la verosimilitud.

Todos los análisis llevados a cabo en este trabajo fueron realizados con el programa estadístico *Stata 16.0* (2019).

Capítulo 4

Resultados

En este capítulo se presentan los resultados obtenidos respecto a la estimación de la EV frente a infección por SARS-CoV-2 e ingreso en UCI, así como los relativos a la pérdida de inmunidad frente a infección. Además, se incluye la descripción y las características de la población de estudio y de las cohortes empleadas en cada uno de los análisis.

4.1. Población de estudio

Los dos análisis llevados a cabo en este trabajo parten de una misma población, si bien sus objetivos son distintos. Dicha población se compone de todas las personas de la población gallega de 12 años o más con al menos una dosis de cualquiera de las vacunas descritas administrada entre el 27 de diciembre de 2020 y el 27 de marzo de 2022. En la Figura 4.1 se muestra un diagrama de esta población base, indicando las exclusiones descritas en la Sección 3.2.

La exclusión de los menores de 12 años supuso la eliminación de 107.524 registros. Por otro lado, se excluyeron 52.289 personas que fueron caso de COVID-19 antes del inicio del estudio, además de 1.148 casos acontecidos fuera de Galicia. El descarte de la población institucionalizada redujo la base de datos en 24.334 registros. La depuración de la base de vacunas supuso la eliminación de 2.139 personas más. El número de personas que recibieron la dosis adicional o para las que no se pudo diferenciar si recibieron esta dosis o la de recuerdo fue de 91.437, que también se excluyeron de los análisis. Finalmente, la población de estudio quedó constituida por 2.129.598 individuos, el 87 % de la población total de 12 años o más residente en Galicia a 1 de enero de 2021. A 27 de marzo de 2022, de los 2.129.598 individuos de la cohorte, un 1 % había fallecido (por cualquier causa), otro 1 % había recibido una única dosis de una de las vacunas, un 18 % había completado la primovacunación y el 80 % restante había recibido la dosis de recuerdo. Además, un 16,7 % del total fue caso de COVID-19 durante el período de estudio, lo que se traduce en 354.806 individuos. De estos, un 98,2 % tuvo una única muestra positiva (348.510 individuos), mientras que el 1,8 % tuvo alguna reinfección (6.296 individuos). De los 354.806 individuos que fueron caso durante el estudio, un 0,22 % (777) ingresaron en la UCI y un 0,14 % (510) fallecieron por COVID-19.

Según los datos que figuran en VIXÍA, entre el 27 de diciembre de 2020 y el 27 de marzo de 2022 fallecieron por COVID-19 1.822 personas. De estas, 1.020 (56 %) fallecieron como no vacunadas, por lo que no llegaron a figurar en el registro de vacunas y no pudieron formar parte de la población de estudio. Por otro lado, de los 802 individuos (44 %) que sí figuraban en el registro por haber fallecido con al menos una dosis de la vacuna, se descartaron 292 tras aplicar las exclusiones. En total, se perdieron el 72 % de las defunciones por COVID-19 acontecidas durante el período de estudio, y solo se mantuvieron 510 fallecidos a causa de la enfermedad en la población. Este porcentaje tan alto de fallecimientos excluidos imposibilitó la estimación de la EV frente a defunción por COVID-19, dado que la mayoría de fallecimientos ocurrieron en no vacunados y la efectividad se estaría infraestimando.

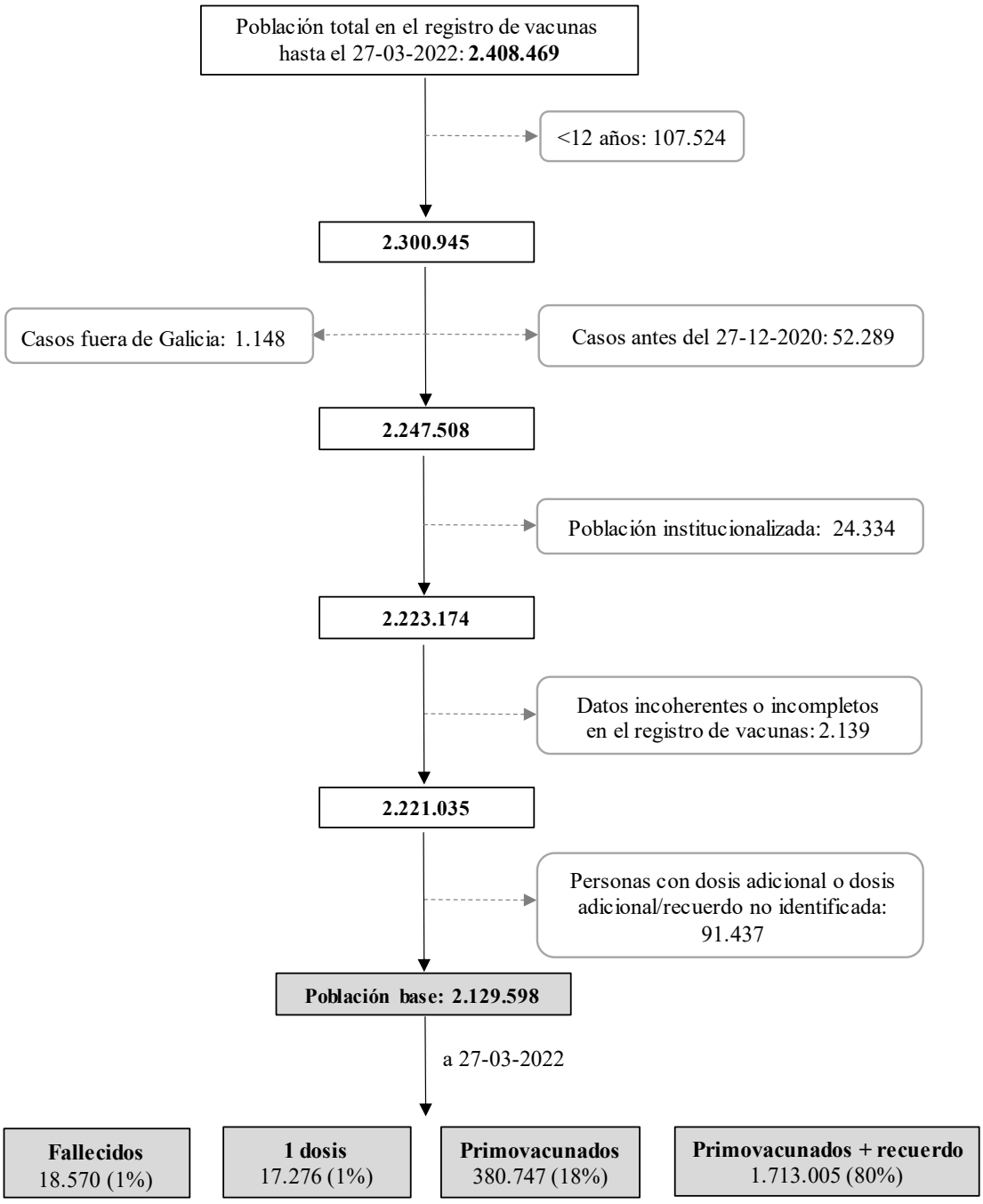

Figura 4.1: Diagrama de la población de estudio constituida, tras las exclusiones, por 2.129.598 individuos. Fuente: DXSP. Elaboración propia.

En la Tabla 4.1 se muestran las características sociodemográficas de la población de estudio y de la población de 12 años o más residente en Galicia de acuerdo con el Padrón Municipal a 1 de enero de 2021. Para el sexo y el ámbito, las distribuciones en ambas poblaciones son muy similares. En cuanto a los grupos de edad, parece que en la población de estudio las personas de 50 años y más cuentan con un poco más de representación que en el caso del total de residentes, aunque la distribución entre

poblaciones sigue siendo muy parecida. En la población de estudio, las personas de 50 años y más suponen un 56,6 % del total, mientras que en la población total este porcentaje es del 52,6 %. Esta diferencia se debe a que en nuestra población solo se consideraron individuos vacunados, y los grupos de mayor edad son los que presentan mayores coberturas vacunales.

	Población estudio		**Población total**	
Características	N	%	N	%
Sexo				
Masculino	1.006.868	47,3	1.173.588	47,8
Femenino	1.122.730	52,7	1.282.131	52,2
Grupo de edad				
12-19	162.216	7,6	182.875	7,5
20-29	152.134	7,1	229.507	9,3
30-39	237.424	11,2	309.799	12,6
40-49	378.898	17,8	443.148	18,0
50-59	366.128	17,2	408.541	16,6
60-69	321.461	15,1	353.142	14,4
70-79	274.819	12,9	291.911	11,9
80+	236.518	11,1	236.796	9,7
Ámbito				
Urbano	1.413.473	66,4	1.588.126	64,7
Semiurbano	428.748	20,1	506.793	20,6
Rural	287.377	13,5	360.800	14,7

Tabla 4.1: Características sociodemográficas de la población utilizada en los estudios y del total de la población de 12 años o más residente en Galicia a 1 de enero de 2021. Fuente: DXSP e IGE. Elaboración propia.

En la Tabla 4.2 se presenta la información relativa a los laboratorios de las dosis administradas. Concretamente, se muestran las pautas de vacunación y el número de las mismas administrado a la población gallega para la primovacunación y la dosis de recuerdo. Nótese que, a fin del estudio, el 98 % de la población de estudio había completado la primovacunación (2.093.752 personas) y un 80 % había recibido la dosis de recuerdo (1.713.005 personas). En el grupo de primovacunados, se muestran los porcentajes relativos al total de individuos que completaron la primovacunación durante el período de estudio. En el caso de la dosis de recuerdo, los porcentajes fueron calculados sobre el total de personas que recibieron esta dosis.

Para la primovacunación, la pauta más administrada fue la correspondiente a dos dosis de Pfizer, representando un 68,4 % sobre el total de primovacunados. Le siguen las pautas de dos dosis de AstraZeneca y dos dosis de Moderna, aunque en porcentajes mucho menores. Solo un 5,4 % de la población de estudio completó la primovacunación con una única dosis de Janssen. Las pautas de vacunación heterólogas fueron muy poco habituales y, de estas, solo la de AstraZeneca-Pfizer cuenta con representación en la tabla.

En cuanto a la dosis de recuerdo, el 55,1 % se administraron con Moderna (943.562) y el 44,9 % con Pfizer (769.443). Solo los que completaron la primovacunación con Janssen recibieron mayoritariamente la dosis de recuerdo con Pfizer. Para el resto de pautas de primovacunación, Moderna fue la vacuna de ARNm seleccionada en más ocasiones para la dosis de refuerzo, incluso para la pauta homóloga de Pfizer. Nótese que, para el 1 % de individuos que completaron la primovacunación con una única dosis de Moderna, AstraZeneca o Pfizer por ser menores de 65 años y haber sido caso antes de recibir su primera dosis, no se muestran datos sobre la dosis de recuerdo debido a la incapacidad para diferenciar

entra esta y la dosis adicional. De los primovacunados con Janssen, un 92,8 % recibió la dosis de recuerdo antes o el 27 de marzo de 2022; para los primovacunados con pauta homóloga de Moderna, este porcentaje fue del 68,2 %; en el caso de los primovacunados con pauta homóloga de AstraZeneca el 96,9 % recibieron la dosis de recuerdo y para los que completaron la primovacunación con dos dosis de Pfizer el porcentaje fue del 81,5 %. Estos porcentajes se relacionan con el tiempo mínimo que debía transcurrir una vez completada la primovacunación hasta poder recibir la dosis de recuerdo. Como ya se indicó, este tiempo era de tres meses si la primovacunación se había alcanzado con Janssen o AstraZeneca y de cinco meses en el caso de Moderna y Pfizer.

Pauta	**Primovacunados**		**Dosis de recuerdo**			
			Moderna		**Pfizer**	
	n	%	n	%	n	%
Janssen	112.751	5,4	14.833	0,9	89.758	5,2
Moderna/AstraZeneca/Pfizer, $\leq$ 65 años, caso	20.563	1,0	—	—	—	—
Moderna-Moderna	239.552	11,4	150.706	8,8	12.757	0,7
Moderna-Pfizer	202	0,0	25	0,0	19	0,0
AstraZeneca-AstraZeneca	282.252	13,5	167.875	9,8	105.609	6,2
AstraZeneca-Moderna	230	0,0	16	0,0	1	0,0
AstraZeneca-Pfizer	6.076	0,3	3,857	0,2	912	0,1
Pfizer-Pfizer	1.431.602	68,4	606.231	35,4	560.380	32,7
Pfizer-Moderna	524	0,0	19	0,0	7	0,0
Total	2.093.752 (98 %)		943.562 (55,1 %)		769.443 (44,9 %)	
			1.713.005 (80 %)			

Tabla 4.2: Distribución de las pautas de inmunización en la población gallega, para primovacunación y dosis de recuerdo, desde el inicio de la vacunación y hasta el 27 de marzo de 2022. Fuente: DXSP. Elaboración propia.

4.2. No vacunados *vs.* Primovacunados

4.2.1. Características de la cohorte

En el primer análisis se estimó la EV en individuos con primovacunación completa frente a no vacunados y se analizó también la pérdida de inmunidad frente a infección. Para ello, se empleó una cohorte constituida por 2.129.598 individuos y 4.159.577 registros. De estos, el 51,2 % (2.129.598) correspondían a individuos no vacunados y el 48,8 % (2.029.979) a individuos primovacunados. Nótese que los individuos con primovacunación completa contaban con dos registros en la base de datos de la cohorte, uno como no vacunados y otro como primovacunados, que se trataron de manera independiente. La cohorte se siguió desde el 27 de diciembre de 2020 hasta el 30 de noviembre de 2021; en total, 338 días. En la Figura 4.2 se muestra un diagrama de la cohorte. Para cada estado de vacunación se indica el número de individuos que pasaron por el mismo durante su seguimiento, así como los casos de COVID-19 e ingresos en UCI, defunciones y cambios de grupo. De los 2.129.598 individuos que iniciaron el seguimiento como no vacunados, 2.029.979 completaron la primovacunación. A 30 de noviembre de 2021, 36.935 personas finalizaron su seguimiento como no vacunadas, 15.735 lo hicieron con una dosis, 1.735.212 acabaron como primovacunadas y 264.060 recibieron la dosis de recuerdo. En total, en este primer análisis hubo 64.834 casos de COVID-19 (de los cuales 558 ingresaron en UCI) y 12.822 defunciones por otras causas.

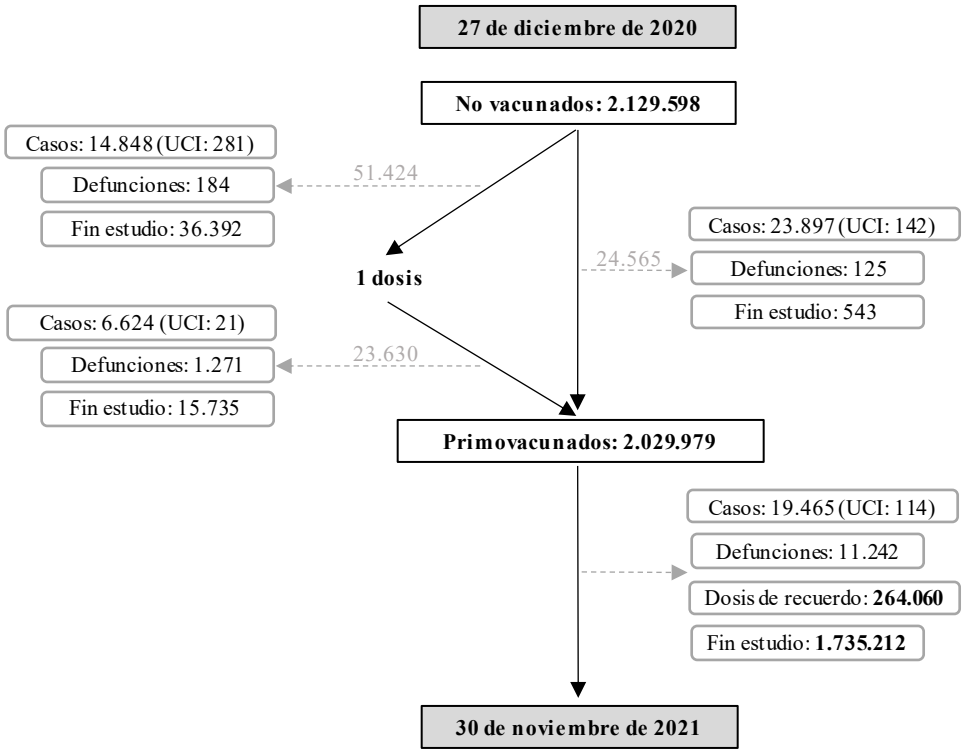

Figura 4.2: Diagrama de la cohorte para el primer análisis (personas de 12 años o más), seguida desde el 27 de diciembre de 2020 hasta el 30 de noviembre de 2021. Fuente: DXSP. Elaboración propia.

En la Tabla 4.3 se recoge la distribución de las características de la cohorte por grupos de vacunación para este primer análisis. Las variables sociodemográficas grupo de edad, sexo y ámbito de residencia presentan distribuciones muy similares. Como se vio en la Tabla 1.2 del Capítulo 1, a 30 de noviembre de 2021 la cobertura para primovacunación completa en personas de 39 años o menos rondaba el 85 %, y para personas a partir de los 39 años era superior al 90 %. Esto significa que la mayoría de individuos de la cohorte, tanto jóvenes como mayores, finalizaron su seguimiento para este primer análisis como primovacunados, de modo que ambos grupos de vacunación están constituidos, en esencia, por los mismos individuos. De ahí que apenas se observen diferencias en la distribución de estas variables.

En cuanto al número de PDIAs negativas realizadas durante el seguimiento, el 35,9 % de los individuos no vacunados y el 16,4 % de los individuos con primovacunación completa tenían al menos una prueba asociada. De estos, en más de la mitad de los casos se trataba de una única PDIA, tanto en el grupo de no vacunados como en el grupo de primovacunados.

Las características de la cohorte se encuentran estrechamente relacionadas con la estrategia de vacunación seguida, basada en grupos de edad decenales, y con la evolución de la pandemia. El primer factor hace que en los primeros compases del seguimiento la cohorte pierda individuos no vacunados en los mayores grupos de edad y, consecuentemente, gane individuos primovacunados en estos mismos grupos. A medida que avanza el seguimiento, ocurre algo similar con los grupos de edad intermedios, aunque no de manera tan acusada. Ya hacia el final, cuando nos aproximamos al 30 de noviembre, son los individuos más jóvenes de la cohorte los que pasan de estar no vacunados a estar completamente primovacunados. Este efecto se puede apreciar en los datos de la Tabla 4.4, donde se recoge el número de individuos de la cohorte en seguimiento, en global y por grupos de edad, en cada grupo de vacunación, a distintos días desde el inicio del seguimiento. A 120 días, las personas entre 30 y 79 años cuentan con representación suficiente tanto en el grupo de no vacunados como en el grupo de primovacunados.

Características	No vacunados		Primovacunados	
	n	%	n	%
Sexo				
Masculino	1.006.868	47,3	955.663	47,1
Femenino	1.122.730	52,7	1.074.316	52,9
Grupo de edad				
12-19	162.216	7,6	125.423	6,2
20-29	152.134	7,1	137.869	6,8
30-39	237.424	11,2	224.827	11,1
40-49	378.898	17,8	368.502	18,1
50-59	366.128	17,2	358.987	17,7
60-69	321.461	15,1	314.269	15,5
70-79	274.819	12,9	268.911	13,2
80+	236.518	11,1	231.191	11,4
Ámbito				
Urbano	1.413.473	66,4	1.343.360	66,2
Semiurbano	428.748	20,1	409.666	20,2
Rural	287.377	13,5	276.953	13,6
Nº PDIAs				
0	1.365.171	64,1	1.697.521	83,6
1	484.232	22,7	209.696	10,3
2	179.121	8,4	74.374	3,7
3+	101.074	4,8	48.388	2,4

Tabla 4.3: Distribución de las características de la cohorte del primer análisis por grupos de vacunación. Fuente: DXSP. Elaboración propia.

Sin embargo, en los grupos de edad más extremos, en algún momento el número de individuos en seguimiento se vuelve insuficiente, ya sean sin vacunar o como vacunados. En el caso de los más jóvenes, apenas hay representación en el grupo de primovacunados más allá de los 90 días. Esto se debe a que este colectivo comenzó a completar la primovacunación mayoritariamente en el mes de septiembre, y el seguimiento de la cohorte para este primer análisis finalizó en noviembre. Para las personas de 80 años y más, la pérdida de representación está más relacionada con la estrategia de vacunación. Este colectivo fue el primero en completar la pauta general y de manera muy rápida (en apenas cuatro meses), lo que se traduce en una falta de representación en el grupo de no vacunados a partir de los 120-150 días de seguimiento.

Analizando el número de individuos de la cohorte en seguimiento por tipo de vacuna de la primera dosis, solo se observa un número insuficiente para AstraZeneca a los 180 días de seguimiento, de 202 individuos. Este dato está relacionado con la pauta de administración. La vacuna se reservó principalmente para los colectivos en activo con una función esencial para la sociedad, que comenzaron a recibir la primera dosis al poco de iniciarse la campaña de inmunización, entre febrero y mayo de 2021. El tiempo entre dosis fue de entre cuatro y 12 semanas, esto es, entre uno y tres meses. Así, la segunda dosis de esta vacuna se administró mayoritariamente en los meses de junio y julio (96,4 % de las dosis), a 180-150 días como máximo de finalizar el seguimiento de esta cohorte, respectivamente. Por este motivo, pocos individuos primovacunados cuentan con un seguimiento de 180 días. Nótese que para los no vacunados, el número de personas en seguimiento a distintos días es siempre el mismo e igual al global.

	No vacunados				Primovacunados			
	90 días	**120 días**	**150 días**	**180 días**	**90 días**	**120 días**	**150 días**	**180 días**
Global	1.899.149	1.558.933	1.127.485	785.423	1.753.492	1.511.808	920.220	519.066
Grupo de edad								
12-19	158.591	158.194	157.707	157.460	7.269	218	138	87
20-29	141.332	138.351	137.429	135.149	40.092	14.937	10.736	5.118
30-39	211.085	206.879	205.286	202.036	200.441	45.370	25.869	12.603
40-49	337.981	331.579	329.111	244.061	355.037	324.532	41.206	19.144
50-59	328.715	308.569	250.051	23.671	350.511	342.846	257.176	21.944
60-69	302.691	238.744	31.585	16.498	307.662	296.001	122.745	17.828
70-79	269.494	159.698	11.845	4.269	265.895	263.428	240.102	224.636
Tipo de vacuna								
Janssen	1.899.149	1.558.933	1.127.485	785.423	104.743	93.500	67.294	10.238
Moderna	1.899.149	1.558.933	1.127.485	785.423	166.260	110.570	89.091	42.159
AstraZeneca	1.899.149	1.558.933	1.127.485	785.423	282.105	273.209	62.066	202
Pfizer	1.899.149	1.558.933	1.127.485	785.423	1.200.384	1.034.529	701.769	466.467

Tabla 4.4: Número total de personas en seguimiento, en global y por grupos de edad, a 90, 120, 150 y 180 días de seguimiento como no vacunadas y como completamente primovacunadas. Fuente: DXSP. Elaboración propia.

Por otro lado, las vacunas administradas a la población también guardan relación con los grupos de edad. Las vacunas de ARNm, Moderna y Pfizer, fueron administradas de manera general a todos los grupos en función de la disponibilidad. En cambio, la vacuna de Janssen se administró mayoritariamente a las personas entre 40 y 59 años, y más de la mitad de las dosis de AstraZeneca se administraron en el rango de edad de 60 a 69 años. En la Tabla 4.5 se muestra la distribución del tipo de vacuna para la primera dosis por grupos de edad decenales. En el caso de Janssen y AstraZeneca, el 82,2 % y el 83,8 % de las dosis, respectivamente, se administraron a personas entre 40 y 69 años.

	Janssen		Moderna		AstraZeneca		Pfizer	
Grupo de edad	n	%	n	%	n	%	n	%
12-19	21	0,0	43.983	17,4	39	0,0	118.173	8,0
20-29	936	0,8	17.787	7,0	7.357	2,5	126.054	8,6
30-39	1.837	1,6	56.441	22,3	16.814	5,8	162.332	11,0
40-49	28.460	24,7	24.117	9,5	26.918	9,3	299.403	20,3
50-59	50.118	43,6	36.595	14,4	34.521	11,9	244.894	16,7
60-69	16.044	13,9	32.237	12,7	181.435	62,6	91.745	6,2
70-79	8.260	7,2	14.657	5,8	22.936	7,9	228.966	15,6
80+	9.409	8,2	27.526	10,9	41	0,0	199.542	13,6

Tabla 4.5: Distribución del tipo de vacuna administrada a la población de estudio para la primera dosis de primovacunación por grupos de edad decenales. Fuente: DXSP. Elaboración propia.

En la Figura 4.3 se muestra el estimador de Nelson-Aalen del riesgo acumulado de infección (izquierda) y de ingreso en UCI (derecha) por COVID-19 en los dos grupos de vacunación. Estas gráficas sugieren que la vacunación tuvo un efecto protector frente a ambos desenlaces en esta cohorte, pues los riesgos acumulados de los primovacunados fueron inferiores a los de los no vacunados.

El valor p del estadístico del test log-rank, que compara las funciones de riesgo estimadas en los dos grupos de vacunación, fue menor de 0,001 para el desenlace infección. En el caso del ingreso en UCI, el valor p del estadístico también fue menor de 0,001. Bajo cualquier nivel de significación de los tomados

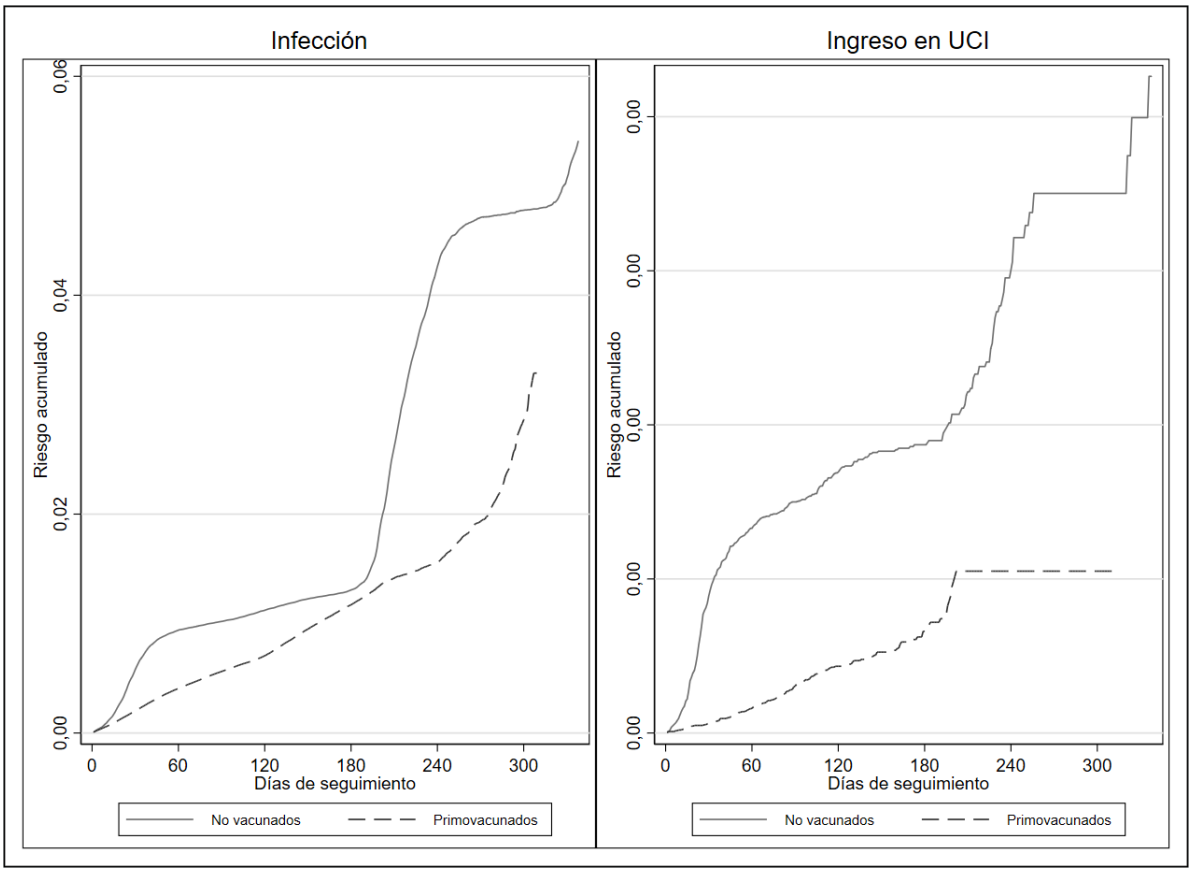

Figura 4.3: Estimador de Nelson-Aalen del riesgo acumulado frente a infección (izquierda) e ingreso en UCI (derecha) por COVID-19 en individuos no vacunados e individuos con primovacunación completa. Los valores del eje Y para el segundo gráfico guardan relación con el bajo número de individuos que ingresaron en UCI en comparación con el tamaño de la cohorte.

habitualmente ($\alpha = 0,01,\ \alpha = 0,05,$ y $\alpha = 0,001$), se concluye que las pruebas son significativas y se rechaza la igualdad de riesgos entre no vacunados y primovacunados para ambos desenlaces en esta cohorte. Este resultado coincide con lo observado a nivel gráfico.

Antes de pasar a los resultados de efectividad vacunal, en las siguientes figuras se muestra el diagnóstico previo de las covariables de ajuste a tener en cuenta en los modelos con el fin de analizar el cumplimiento de la hipótesis de riesgos proporcionales. Se presentan las curvas log-log de supervivencia, así como las curvas observadas y esperadas. Recordemos que, para el primer gráfico, se espera que las curvas sean aproximadamente paralelas mientras que, para el segundo, se busca que las curvas estén lo más próximas posible. Para mostrar el funcionamiento de ambos métodos se incluyen, por un lado, las curvas log-log para las covariables sexo, ámbito de residencia y número de PDIAs negativas, estimadas a partir del estimador Kaplan-Meier y el estimador de máxima verosimilitud parcial de β. Por otro lado, se presentan las curvas observadas y esperadas para el grupo de edad y el estado de vacunación, obtenidas a partir del estimador de Kaplan-Meier y del estimador de la supervivencia condicional definida para el modelo de Cox, respectivamente. Para cada covariable, se ofrece también el valor estimado del coeficiente de interacción de la covariable con el tiempo, $\hat{\gamma}$, así como el estadístico y el valor p asociados al contraste $H_0 : \hat{\gamma} = 0$. A mayores, se muestra el comportamiento de las curvas de supervivencia de no vacunados y de primovacunados por grupos de edad y por tipo de vacuna. Cabe destacar que estas herramientas para el diagnóstico previo de las covariables son válidas para el

desenlace infección y el desenlace ingreso en UCI por COVID-19, ya que en ambos casos el seguimiento de los individuos finalizó en la fecha de toma de la muestra positiva.

En el caso de la variable sexo se observa un buen grado de paralelismo en las curvas, lo que sugiere el cumplimiento de la hipótesis de riesgos proporcionales por parte de esta covariable. No obstante, hacia el final del seguimiento (sobre los 200 días, 5,3 en escala logarítmica) las curvas llegan a tocarse. El valor del coeficiente $\hat{\gamma}_{\text{Sexo x Tiempo}}$ fue -0,0006 (IC 95 %: -0,0010 - -0,0002), y el contraste de hipótesis $\hat{\gamma}_{\text{Sexo x Tiempo}} = 0$ resultó significativo, con un valor p asociado de 0,001.

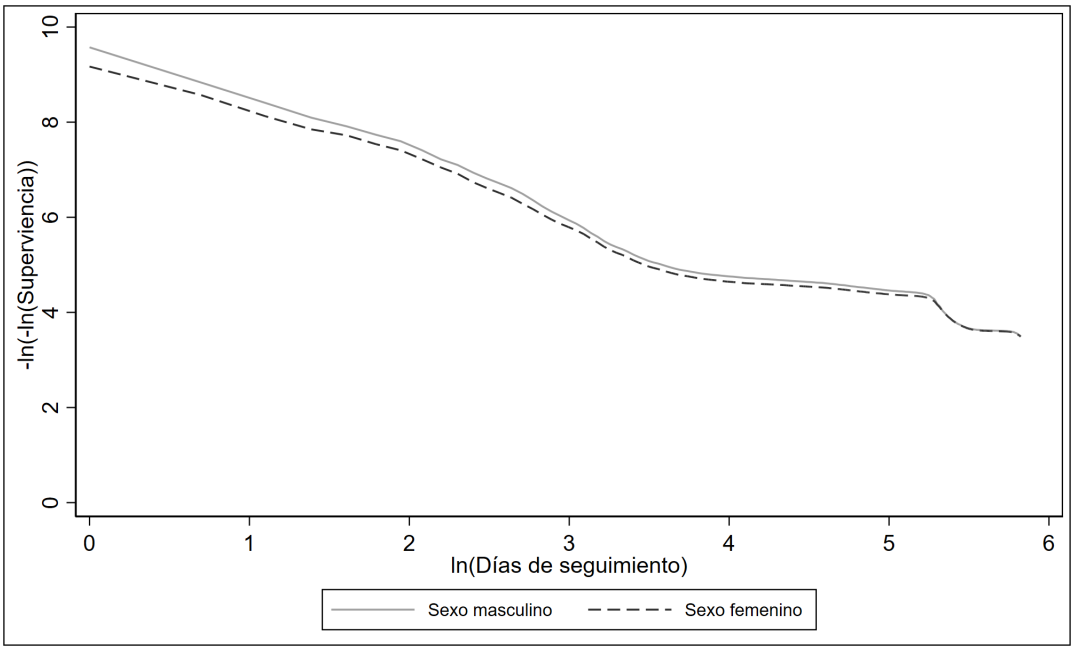

Figura 4.4: Curvas log-log de supervivencia para la variable sexo, con dos categorías.

Para el ámbito de residencia, las curvas de supervivencia de las tres categorías muestran un comportamiento muy similar, pero también se cruzan. El valor del coeficiente $\hat{\gamma}_{\text{Ámbito x Tiempo}}$ fue -0,0016 (IC 95 %: -0,0020 - -0,0013), y el contraste de hipótesis $\hat{\gamma}_{\text{Ámbito x Tiempo}} = 0$ arrojó un valor p menor de 0,001, resultando también significativo para esta covariable.

Para el número de PDIAs negativas ocurre algo parecido a lo observado en los casos anteriores. Las curvas para las distintas categorías de la variable muestran un comportamiento similar pero, en algún momento del seguimiento, se cruzan. El valor del coeficiente de interacción $\hat{\gamma}_{\text{PDIAs x Tiempo}}$ fue 0,0035 (IC 95 %: 0,0032-0,0037), y el contraste de hipótesis $\hat{\gamma}_{\text{PDIAs x Tiempo}} = 0$ arrojó un valor p inferior a 0,001 resultando, por tanto, significativo.

Las curvas observadas y esperadas para cada clase de la variable grupo de edad se muestran en la Figura 4.7. Para una mejor visibilidad, se han reducido los grupos decenales a cuatro: 12-19, 20-39, 40-59 y 70 y más años. En general, se observa proximidad entre las curvas observadas y esperadas, aunque las curvas para los grupos de 50 a 69 años y de 70 años y más, a partir de los 200 días de seguimiento, se desvían ligeramente. En este caso, el valor del coeficiente $\hat{\gamma}_{\text{Edad x Tiempo}}$ fue 0,00001 (IC 95 %: -0,00019-0,00022), y el contraste de hipótesis $\hat{\gamma}_{\text{Edad x Tiempo}} = 0$ resultó no significativo con un p valor asociado de 0,885. Considerando este mismo contraste para los grupos de edad decenales, se obtuvo un coeficiente de 0,00007 (IC95 %: -0,00003-0,00017) y un valor p de 0,190, también no significativo. Esto indica que la variable grupo de edad cumple con la hipótesis de riesgos proporcionales.

En el caso del estado de vacunación, de especial interés en nuestros análisis, para la clase de no vacunados las curvas observada y esperada se encuentran muy próximas. Para los primovacunados, sin embargo, a partir de los 200 días de seguimiento aproximadamente, la curva esperada comienza a

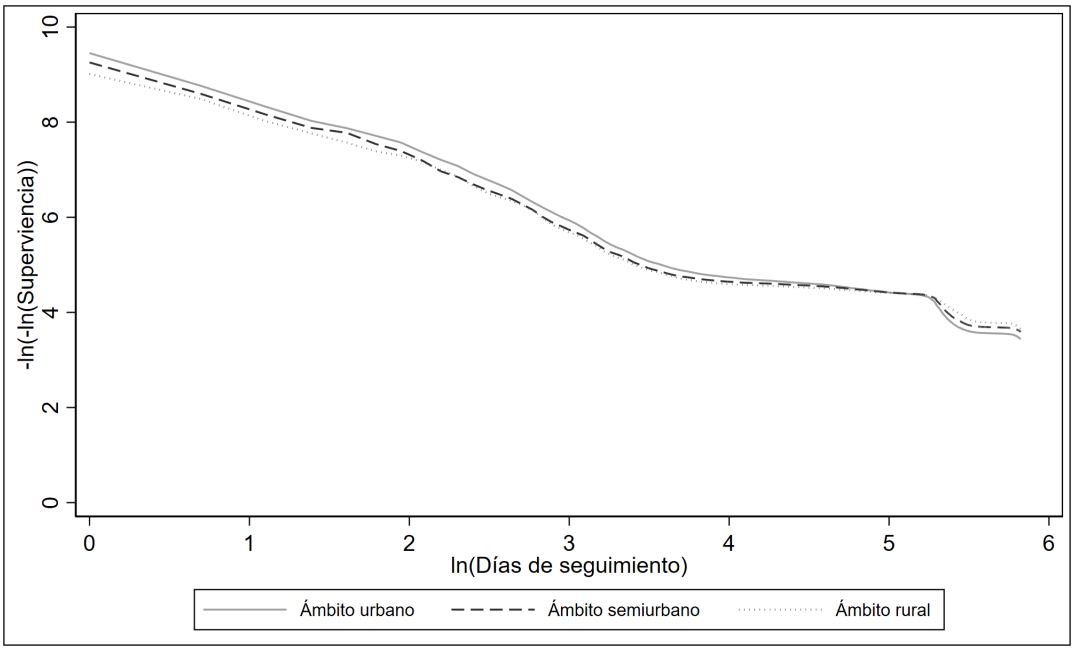

Figura 4.5: Curvas log-log de supervivencia para la variable ámbito de residencia, con tres categorías.

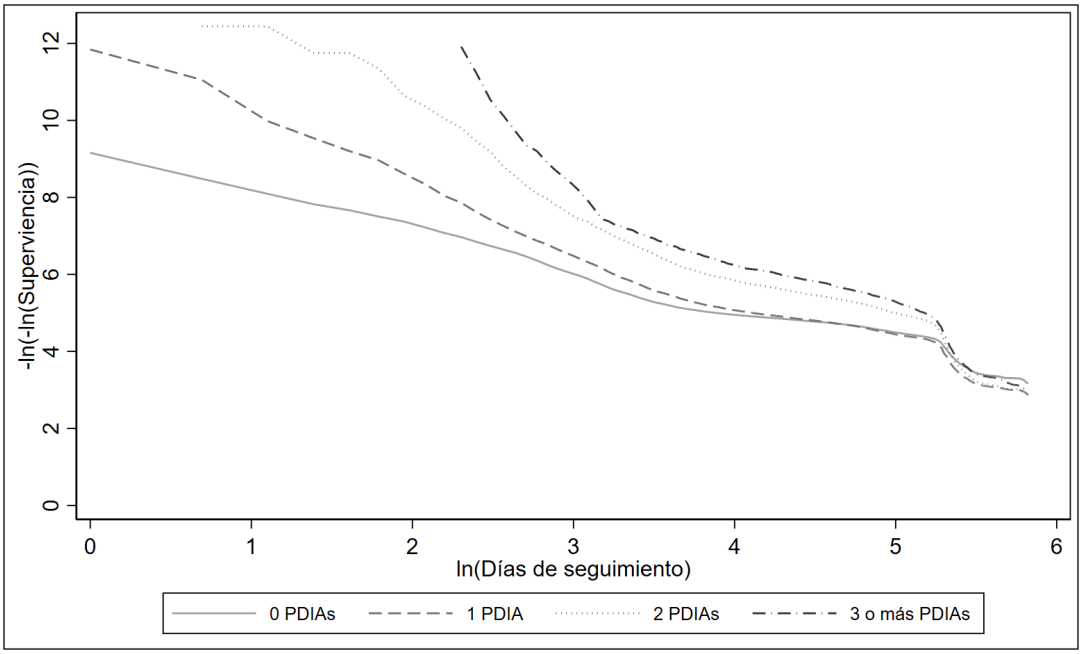

Figura 4.6: Curvas log-log de supervivencia para el número de PDIAs negativas realizadas durante el seguimiento, con cuatro categorías.

separarse de la curva observada. El coeficiente $\hat{\gamma}_{\text{Estatus x Tiempo}}$ fue 0,0077 (IC 95 %: 0,0075-0,00791), y el contraste de hipótesis $\hat{\gamma}_{\text{Estatus x Tiempo}} = 0$ resultó significativo con un p valor asociado menor de 0,001.

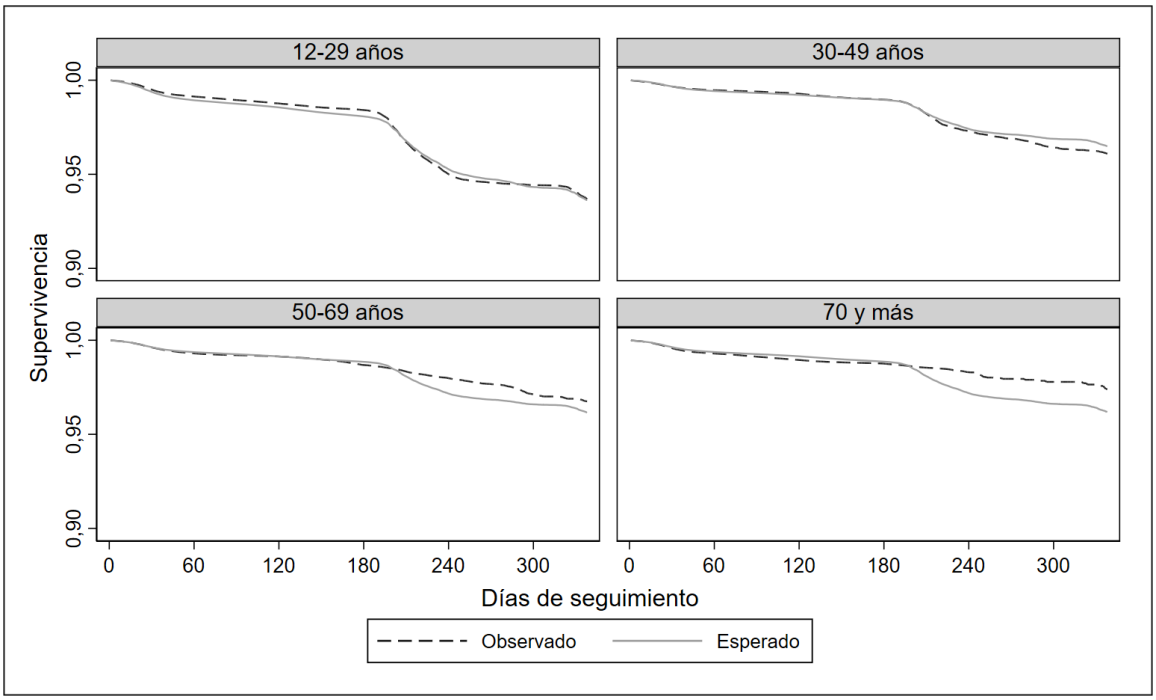

Figura 4.7: Curvas observadas y esperadas para las cuatro categorías del grupo de edad. Nótese que para el eje Y, que representa la supervivencia, se muestran valores entre 0,90 y 1,00.

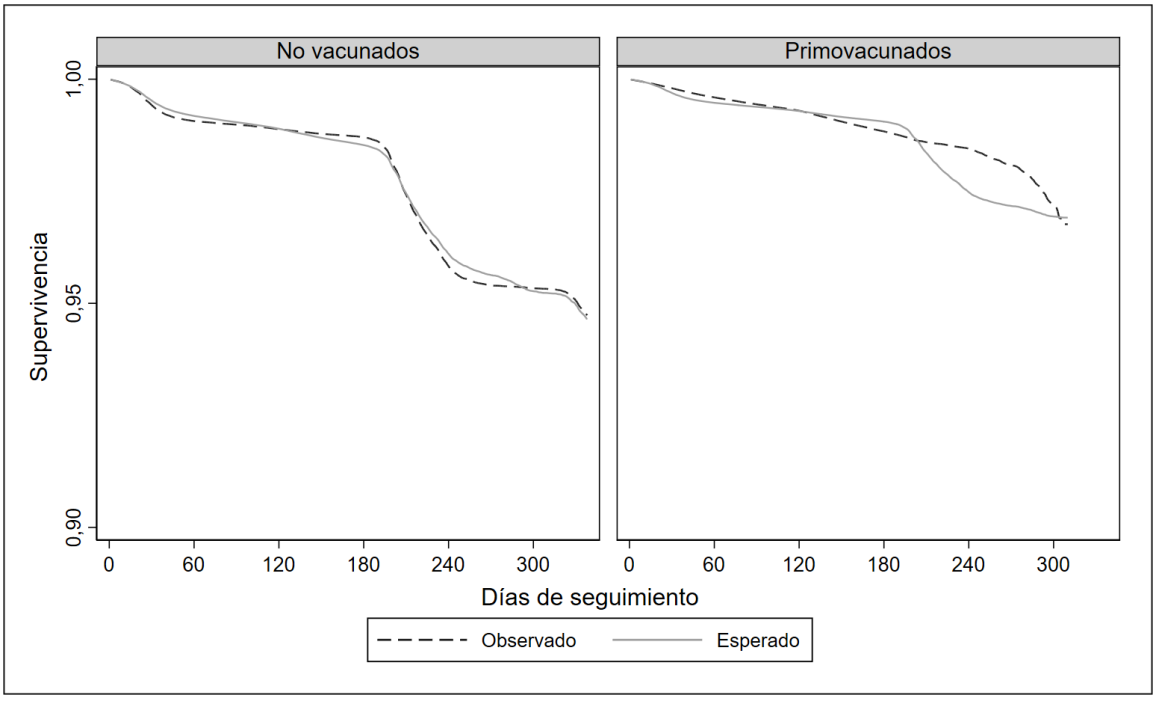

Figura 4.8: Curvas observadas y esperadas para las dos categorías de la variable estado de vacunación. Nótese que para el eje Y, que representa la supervivencia, se muestran valores entre 0,90 y 1,00.

En las Figuras 4.9 y 4.10 se muestran las estimaciones Kaplan-Meier de la supervivencia para no vacunados y primovacunados por grupos de edad y por de tipo de vacuna. En cuanto a los grupos de edad, únicamente en el caso de las personas de 70 años y más las curvas discurren más o menos paralelas y sin cruzarse. Este grupo pasó la ola de Alpha sin inmunizar y la ola de Delta con primovacunación completa, lo que igualó los riesgos en ambos grupos de vacunación. Por eso la hipótesis de riesgos proporcionales se cumple claramente para la variable estado de vacunación en este grupo de edad. Las personas más jóvenes, entre 12 y 29 años, pasaron las dos olas prácticamente como no vacunados, motivo por el cual la supervivencia de los no vacunados es menor en casi todo el tiempo de seguimiento. Los grupos de edad intermedios pasaron la ola de Alpha sin vacunar, pero en la ola de Delta algunas personas estaban sin vacunar y otras estaban ya completamente primovacunadas. Esto explica que las curvas de supervivencia estén más próximas. Además, en los primeros días de seguimiento los riesgos de ambos grupos eran similares pero, a medida que pasó el tiempo, los individuos fueron completando la primovacunación coincidiendo con la llegada de Delta, y su supervivencia fue disminuyendo con respecto a los no vacunados (que se iban reduciendo en número).

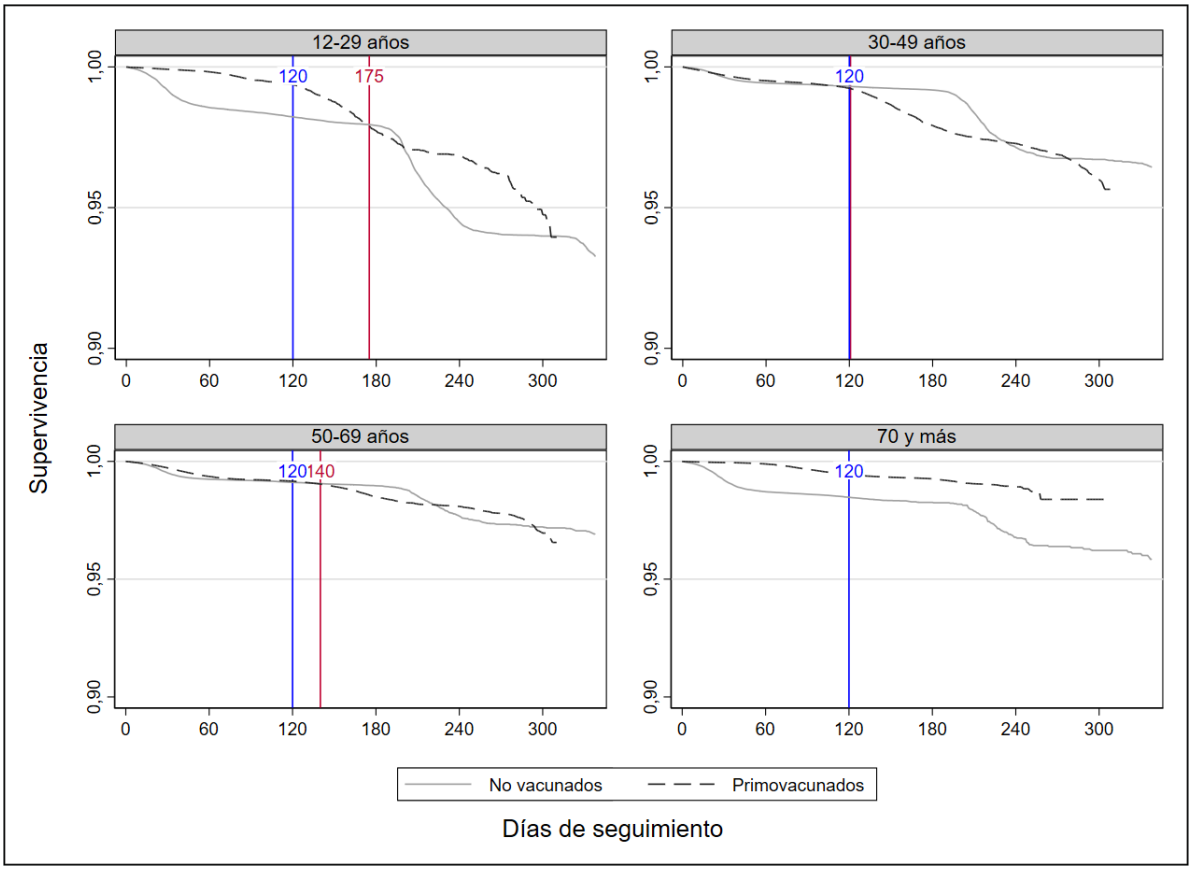

Figura 4.9: Curvas de supervivencia para cada estado de vacunación por grupos de edad. Las líneas verticales en color rojo marcan el punto del tiempo de seguimiento en el que las curvas se cruzan por primera vez. Las líneas verticales en color azul marcan el punto de corte del seguimiento. Nótese que para el eje Y, que representa la supervivencia, se muestran valores entre 0,90 y 1,00.

Por tipo de vacuna la hipótesis de riesgos proporcionales parece cumplirse para el estado de vacunación con las vacunas de ARNm, Moderna y Pfizer, y también con Janssen. En el caso de AstraZeneca, las curvas de supervivencia discurren muy próximas en ambos grupos durante los primeros tiempos de seguimiento y a los 230 días se produce el cruce. Además, se observa como para Janssen y Astra-

Zeneca el número de individuos a riesgo se reduce de manera importante a los 190 y a los 170 días de seguimiento, respectivamente.

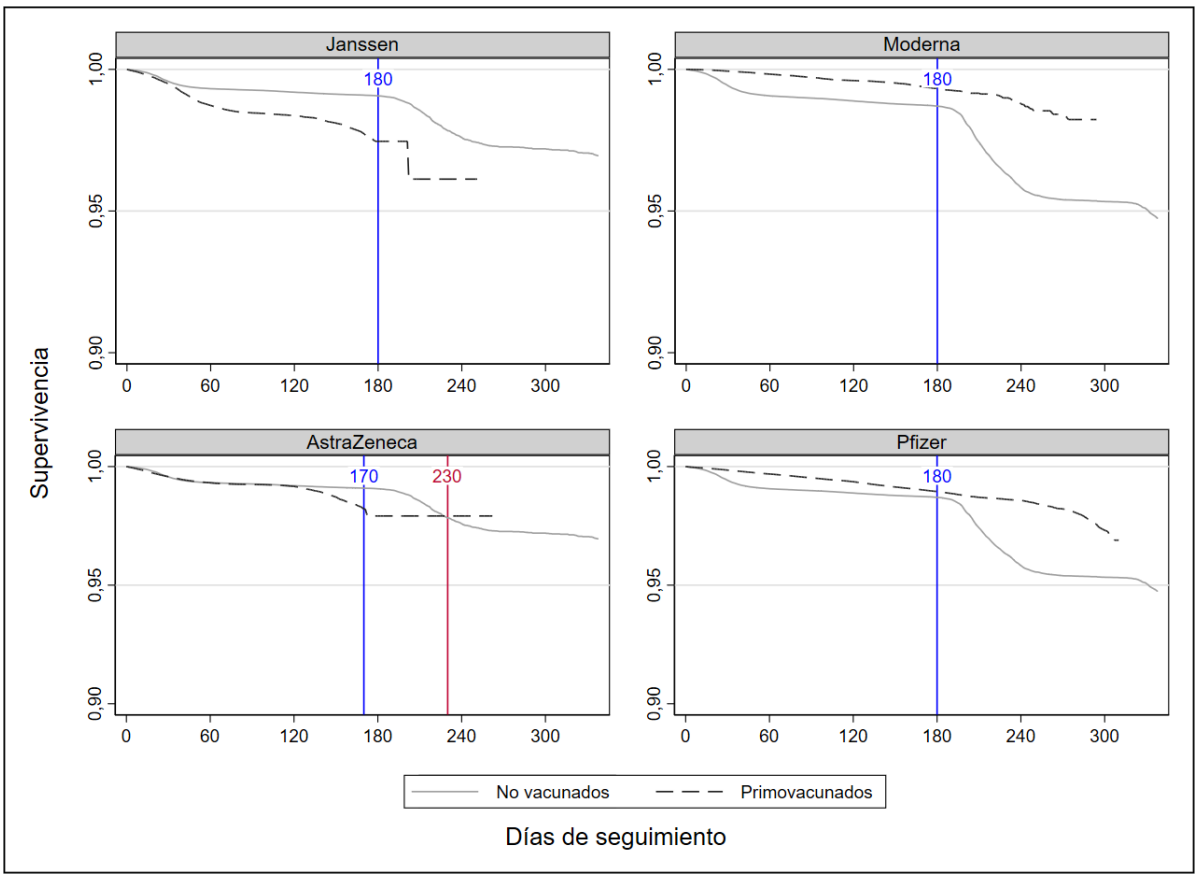

Figura 4.10: Curvas de supervivencia para cada estado de vacunación por tipo de vacuna. Las líneas verticales en color rojo marcan el punto del tiempo de seguimiento en el que las curvas se cruzan por primera vez. Las líneas verticales en color azul marcan el punto de corte del seguimiento. Nótese que para el eje Y, que representa la supervivencia, se muestran valores entre 0,90 y 1,00.

Atendiendo a lo expuesto, para evitar distorsiones en la estimación de la EV y cumplir con la hipótesis de riesgos proporcionales, se tuvieron en cuenta las siguientes consideraciones:

i. La estimación global de la EV, para ambos desenlaces, se realizó a un máximo de 180 días de seguimiento, ajustando por el grupo de edad decenal y estratificando por sexo, ámbito de residencia y número de PDIAs negativas.

ii. Para la estimación de la EV por grupos de edad:

   - Se redujeron los grupos de edad a cuatro, a saber, 12-29, 30-49, 50-69 y 70 y más con el fin de tener un número suficiente de personas en seguimiento tanto en el grupo de no vacunados como en el grupo de completamente primovacunados.
   - La estimación de la EV se realizó a un máximo de 120 días de seguimiento y estratificando por las variables sexo, ámbito de residencia y número de PDIAs negativas.

iii. Para la estimación de la EV por tipo de vacuna:

- En el caso de pautas heterólogas, se consideró la vacuna de la primera dosis.
- El cálculo de la EV para las vacunas de Janssen y AstraZeneca se restringió a las personas entre 40 y 69 años.
- La estimación de la EV se realizó a un máximo de 180 días de seguimiento salvo para AstraZeneca, para la que se consideró un seguimiento de 170 días con el fin de garantizar un número suficiente de individuos primovacunados. Como se vio en la Tabla 4.4, a los 150 días este número era de 62.066 personas y a los 180 días de 202. En tiempos intermedios, de 160 y 170 días, las personas en seguimiento primovacunadas con AstraZeneca eran, respectivamente, 21.240 y 7.170.
- Se ajustó por el grupo de edad decenal y se estratificó por sexo, ámbito de residencia y número de PDIAs negativas.

4.2.2. Efectividad vacunal global

Tasas de incidencia

En la Tabla 4.6 se recogen las tasas de incidencia globales por 1.000 personas-año y sus intervalos de confianza del 95 % para infección e ingreso en UCI por COVID-19, en el grupo de no vacunados y en el grupo de primovacunación completa. Dichas tasas representan el número de eventos nuevos en la cohorte en función del total de personas-tiempo que contribuyen al seguimiento, es decir, personas-tiempo a riesgo. Para cada desenlace, se espera que esta tasa sea menor en primovacunados que en no vacunados. En el caso de infección, sintomática o asintomática, la tasa de incidencia en personas sin vacunar fue de 41,37, mientras que en personas vacunadas fue de casi la mitad, 23,72. Para los ingresos en UCI, la tasa de incidencia en no vacunados fue de 0,45, en tanto que en personas con primovacunación completa se redujo a 0,14.

Desenlaces	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
Infección					
No vacunados	936.563,12	38.745	41,37	40,96	41,78
Primovacunados	820.626,32	19.465	23,72	23,39	24,06
Ingreso en UCI					
No vacunados	936.563,12	423	0,45	0,41	0,50
Primovacunados	820.626,32	114	0,14	0,12	0,17

Tabla 4.6: Tasas de incidencia globales de COVID-19, por 1.000 personas-año, para cada desenlace en no vacunados y completamente primovacunados.

Efectividad vacunal

En la Tabla 4.7 se presentan las estimaciones de la EV en global a un máximo de 180 días de seguimiento, para los dos desenlaces, con sus correspondientes intervalos de confianza del 95 %. Se muestran los porcentajes de efectividad obtenidos con el modelo sin ajustar y con el modelo ajustado por el grupo de edad decenal y estratificado por las variables sexo, ámbito de residencia y número de pruebas negativas realizadas durante el seguimiento.

La EV global en primovacunados respecto a no vacunados frente a infección por COVID-19 resultó ser del 27 % (IC95 %: 26,4-27,7) tras el ajuste. En el caso del ingreso en UCI, el valor estimado para la EV global ajustada fue superior, del 76 % (IC95 %: 75,8-76,3).

Desenlaces	Sin ajustar			Ajustada		
	EV(%)	IC95 %		EV(%)	IC95 %	
Infección	23,9	23,2	24,7	27,0	26,4	27,7
Ingreso en UCI	71,3	71,1	71,7	76,0	75,8	76,3

Tabla 4.7: EV frente a infección e ingreso en UCI por COVID-19 a un máximo de 180 días de seguimiento, sin ajustar y ajustada, e intervalos de confianza del 95 %.

Pérdida de efectividad

Se estimó la pérdida de efectividad de la primovacunación completa, ajustando por el grupo de edad decenal y estratificando por sexo, ámbito de residencia y número de PDIAs negativas, frente a infección e ingreso en UCI por COVID-19 (una vez transcurrido el período de inducción). La Figura 4.11 representa la pérdida de inmunidad global a partir de cinco días de seguimiento y hasta un máximo de 180 días. En el eje de abscisas se representa el tiempo de seguimiento y en el eje de ordenadas el porcentaje de EV ajustada. Se incluye la correspondiente banda de confianza del 95 % y un umbral de EV del 50 %.

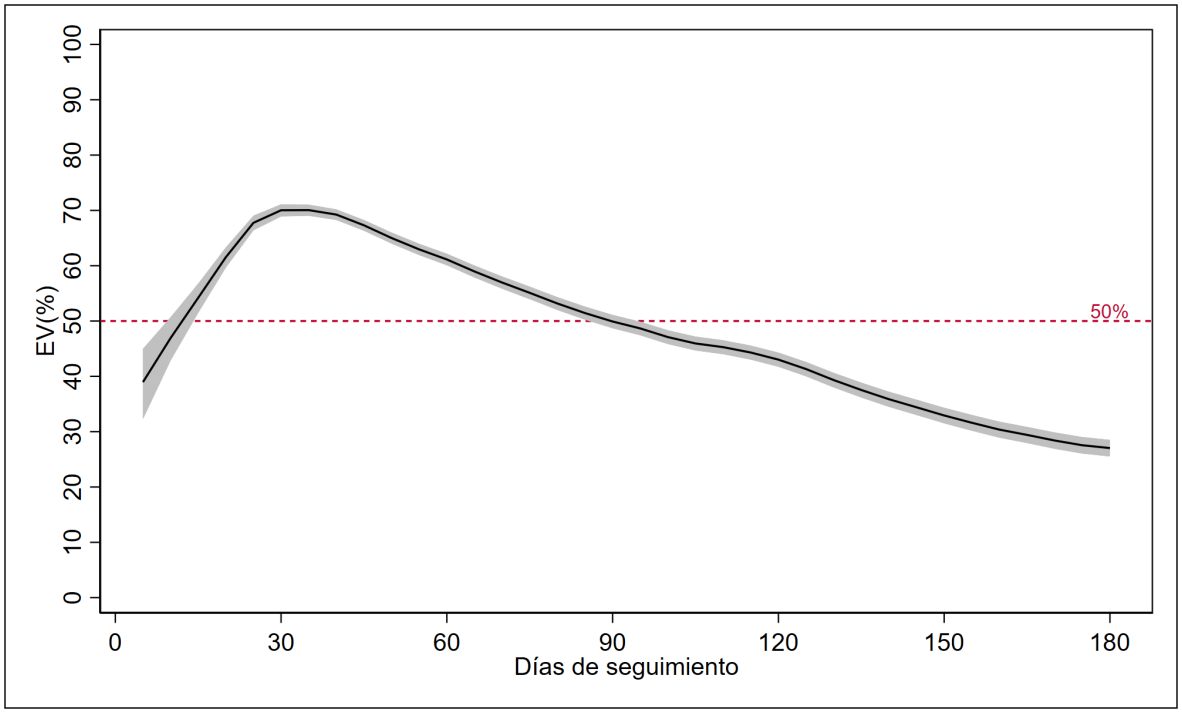

Figura 4.11: Curva de EV para primovacunación completa frente a infección por COVID-19 hasta un máximo de 180 días de seguimiento. La zona sombreada corresponde a la banda de confianza del 95 %. La línea horizontal punteada marca un umbral de EV del 50 %.

Durante los primeros 10-15 días la EV global fue inferior al 50 %. A partir de entonces y hasta aproximadamente tres meses de seguimiento se mantuvo por encima de este umbral, alcanzando un máximo del 70 % durante los 30-35 días. A partir del mes y medio comenzó a descender de manera casi proporcional al tiempo, iniciándose la pérdida de efectividad. Tras seis meses de seguimiento (180 días), la EV global frente a infección en primovacunados cayó hasta el 27 %.

La pérdida de efectividad global frente a ingreso en UCI se muestra en la Figura 4.12. La amplitud de la banda de confianza está relacionada con el bajo número de individuos que ingresaron en UCI en

relación con el tamaño de la cohorte. La EV partió de un 82 % y, en el primer mes de seguimiento, ascendió hasta el 93 %, su valor máximo. La pérdida comenzó a observarse a partir de ese momento. Hasta aproximadamente los 100 días de seguimiento, dicha pérdida fue un poco más acusada y, posteriormente, pareció mantenerse estable en un 76,5 % al menos hasta los 180 días.

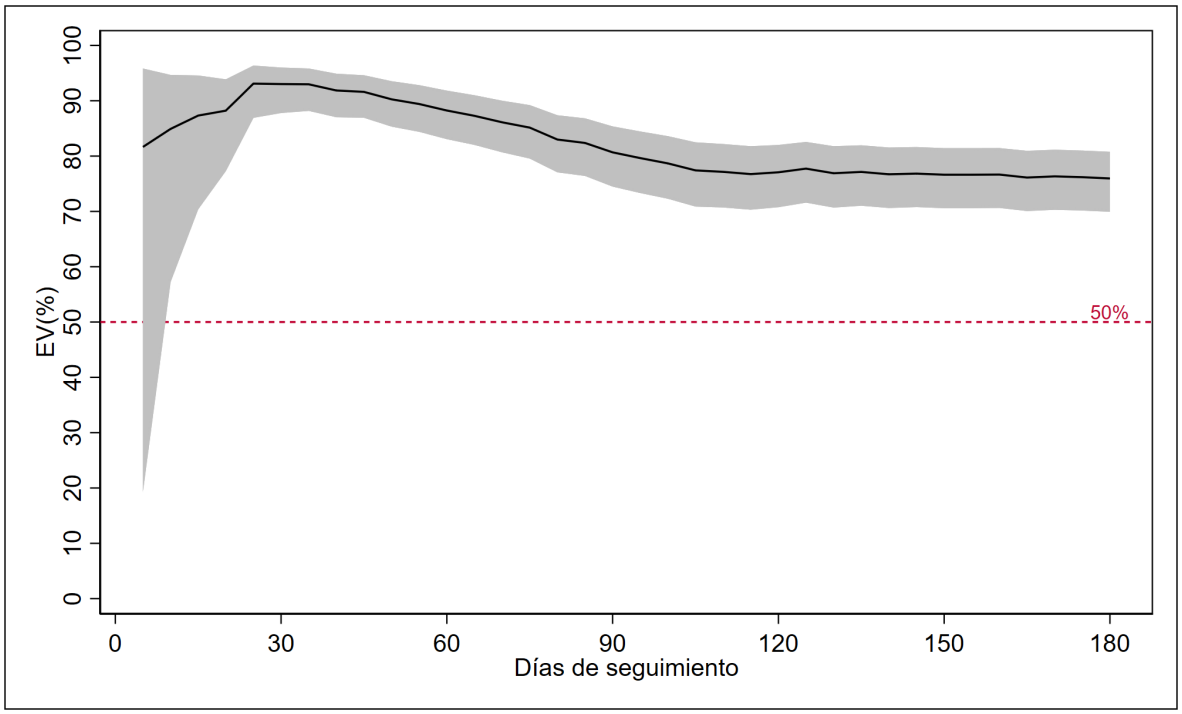

Figura 4.12: Curva de EV para primovacunación completa frente a ingreso en UCI por COVID-19 hasta un máximo de 180 días de seguimiento. La zona sombreada corresponde a la banda de confianza del 95 %. La línea horizontal punteada marca un umbral de EV del 50 %.

4.2.3. Efectividad vacunal por grupos de edad

Tasas de incidencia

A continuación, se muestran las tasas de incidencia de COVID-19 por 1.000 personas-año frente a infección en no vacunados y en completamente primovacunados por grupo de edad. Para las personas de 12 a 29 años y de 70 años y más las tasas de incidencia, como cabría esperar, son superiores en no vacunados frente a primovacunados. No obstante, estas tasas deben ser valoradas con cautela, teniendo el cuenta el momento en el que cada grupo completó la primovacunación y, por tanto, el grupo en el que los individuos estuvieron expuestos en las distintas olas de la pandemia. Los más jóvenes pasaron las olas de Alpha y Delta principalmente como no vacunados, por lo que la mayoría de casos de COVID-19 en este grupo de edad se dieron en no inmunizados (hubo 14,72 veces más casos en este grupo que en el grupo de primovacunados). Los de 70 años y más, que fueron los primeros en vacunarse, pasaron la ola de Alpha en su mayoría como no vacunados, pero todos pasaron la ola de Delta como primovacunados. Esto explica el número de personas-año en seguimiento y el número de casos en cada grupo de vacunación para este grupo de edad.

Por otro lado, en los grupos de edad intermedios se observa una tasa de incidencia superior en primovacunados frente a no vacunados, especialmente en los de 30 a 49 años. En estos grupos de edad el 7,4 % de las dosis administradas fueron de Janssen y el 20,1 % de AstraZeneca (frente a un 2,3 % y un

Infección					
Grupo de edad	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
12-29 años					
No vacunados	203788,50	16.630	81,60	80,37	82,85
Primovacunados	58.991,57	1.130	19,16	18,07	20,31
30-49 años					
No vacunados	307.942,62	7.699	25,00	24,24	25,57
Primovacunados	208.631,38	6.068	29,08	28,36	29,83
50-69 años					
No vacunados	269.726,02	6.749	25,02	24,43	25,63
Primovacunados	287.530,10	7.980	27,75	27,15	28,37
70 y más					
No vacunados	155.105,98	7.667	49,43	48,34	50,55
Primovacunados	265.473,28	4.287	16,15	15,67	16,64

Tabla 4.8: Tasas de incidencia de COVID-19 por 1.000 personas-año frente a infección por grupos de edad.

3,8 % en los otros dos grupos). La vacuna de Janssen se administró fundamentalmente de mayo a julio de 2021 (84,9 % de las dosis), y una única dosis era suficiente para completar la primovacunación. Por otro lado, la primera dosis de AstraZeneca se administró al poco de inciarse la campaña de inmunización, en febrero de 2021, y se extendió aproximadamente hasta mayo. El tiempo entre dosis fue de entre cuatro y 12 semanas, de modo que la segunda dosis de esta vacuna se administró mayoritariamente en los meses de junio y julio. Por tanto, buena parte de las personas vacunadas con Janssen y AstraZeneca completaron la primovacunación justo antes de la llegada de Delta. Además, los jóvenes entre 12 y 19 años fueron los más afectados por la ola atribuida a esta variante, con capacidad para escapar a la inmunidad, y dentro del rango de edad de 30 a 69 años se incluyen los padres de muchos de estos jóvenes. El tipo de vacuna, el momento de la vacunación y la relación con Delta podrían explicar por qué, tras completar la pauta general, las personas de 30 a 69 años se contagiaron más que cuando estaban sin vacunar.

En la Tabla 4.9 se ofrecen las tasas de incidencia de COVID-19 por 1.000 personas-año pero ahora frente a ingreso en UCI, en no vacunados y en completamente primovacunados por grupo de edad. Para este desenlace, las tasas de incidencia fueron superiores en no vacunados frente a primovacunados para todos los grupos de edad, lo que indica que la vacunación tuvo un efecto positivo sobre el agravamiento de la enfermedad asociada al virus. El mayor número de ingresos en UCI se produjo en personas de 70 años y más, para las que la tasa de incidencia de ingreso fue 5,32 veces superior en no vacunados frente a vacunados.

Ingreso en UCI					
Grupo de edad	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
12-29 años					
No vacunados	203.788,50	12	0.06	0,03	0,10
Primovacunados	58.991.57	0	0.00	—	—
30-49 años					
No vacunados	307.942,62	44	0,14	0,11	0,19
Primovacunados	208.631,38	6	0,03	0,01	0,06
50-69 años					
No vacunados	269.726,02	161	0,60	0,51	0,70
Primovacunados	287.530,10	42	0,15	0,11	0,20
70 y más					
No vacunados	155.105,98	206	1,33	1,16	1,52
Primovacunados	265.473,28	66	0,25	0,20	0,32

Tabla 4.9: Tasas de incidencia de COVID-19 por 1.000 personas-año frente a ingreso en UCI por grupos de edad.

Efectividad vacunal

En la siguiente Tabla se dan las estimaciones de la EV para cada desenlace por grupo de edad, a 120 días de seguimiento, acompañadas por su intervalo de confianza del 95 %. Los valores ausentes para la efectividad se corresponden con estimaciones negativas (lo que implica un RR > 1).

Desenlaces	**Sin ajustar**			**Ajustada**		
	EV(%)	**IC95 %**		**EV(%)**	**IC95 %**	
Infección						
12-29 años	77,7	77,5	77,9	84,0	83,8	84,2
30-49 años	—	—	—	4,9	4,0	5,9
50-69 años	6,1	5,2	7,1	9,1	8,2	10,0
70 y más	66,0	65,7	66,4	66,0	65,7	66,4
Ingreso en UCI						
12-29 años	100	100	100	100	100	100
30-49 años	63,2	63,0	64,2	70,4	70,3	71,2
50-69 años	77,0	76,8	77,3	77,0	76,9	77,4
70 y más	79,2	79,1	79,5	77,9	77,7	78,2

Tabla 4.10: EV frente a infección e ingreso en UCI por COVID-19 a 120 días de seguimiento por grupo de edad, sin ajustar y ajustada, e intervalos de confianza del 95 %.

La EV más alta frente a infección se alcanzó en el grupo de edad más joven, siendo del 77,7 % (IC95 %: 77,5-77,9) sin ajustar y del 84 % (IC95 %: 83,8-84,2) tras el ajuste. En el caso del grupo de edad más avanzada, de 70 años y más, la efectividad resultó ser un poco inferior, del 66 % (IC95 %: 65,7-66,4). Como ya se vio, para estos grupos la tasa de incidencia de COVID-19 fue superior en no vacunados frente a primovacunados, de modo que se esperaba que las vacunas resultasen efectivas en personas de estas edades. Para los de 50 a 69 años, las tasas de incidencia en ambos grupos de vacunación fueron muy similares, lo que se traduce en una EV de solo el 6,1 % (IC95 %: 5,2-7,1) sin ajustar y del 9,1 %

(IC95 %: 8,2-10) ajustada. Por otro lado, en los de 30 a 49 años se obtuvieron estimaciones negativas para la efectividad sin ajustar ni estratificar por las covariables (RR=1,1 (IC95 %: 1-1,1)). Tras el ajuste, sin embargo, se obtuvo una efectividad de casi el 5 %. Este resultado concuerda una vez más con lo ya visto en relación con las tasas de incidencia de la enfermedad en individuos de estas edades.

Con respecto al ingreso en UCI, parece que la vacunación resultó efectiva en todos los grupos de edad. Para los más jóvenes, el valor de la EV resultó ser del 100 % debido al bajo número de ingresos en UCI en personas de estas edades (tan solo 12, y todas no vacunadas, como puede observarse en la Tabla 4.9). En los grupos de edad entre 30 y 49 años y entre 50 y 69 años la efectividad frente al ingreso fue del 70,4 % (IC 95 %: 70,3-71,2) y del 77 % (IC 95 %: 76,9-77,4), respectivamente. En personas de 70 años y más, donde se dieron la mitad de los ingresos en UCI (50,6 %), la EV fue del 77,9 % (IC95 %: 77,7-78,2).

Pérdida de efectividad

En la Figura 4.13 se muestra la pérdida de EV por grupo de edad frente a infección por COVID-19 a partir de cinco días de seguimiento y hasta un máximo de 120 días. Se representa el porcentaje de EV, con su correspondiente banda de confianza del 95 %, y se incluyen dos umbrales de efectividad, del 0 % y del 50 %. Nótese que los valores de EV correspondientes a RR mayores a 1,20 se han eliminado por cuestiones de escala.

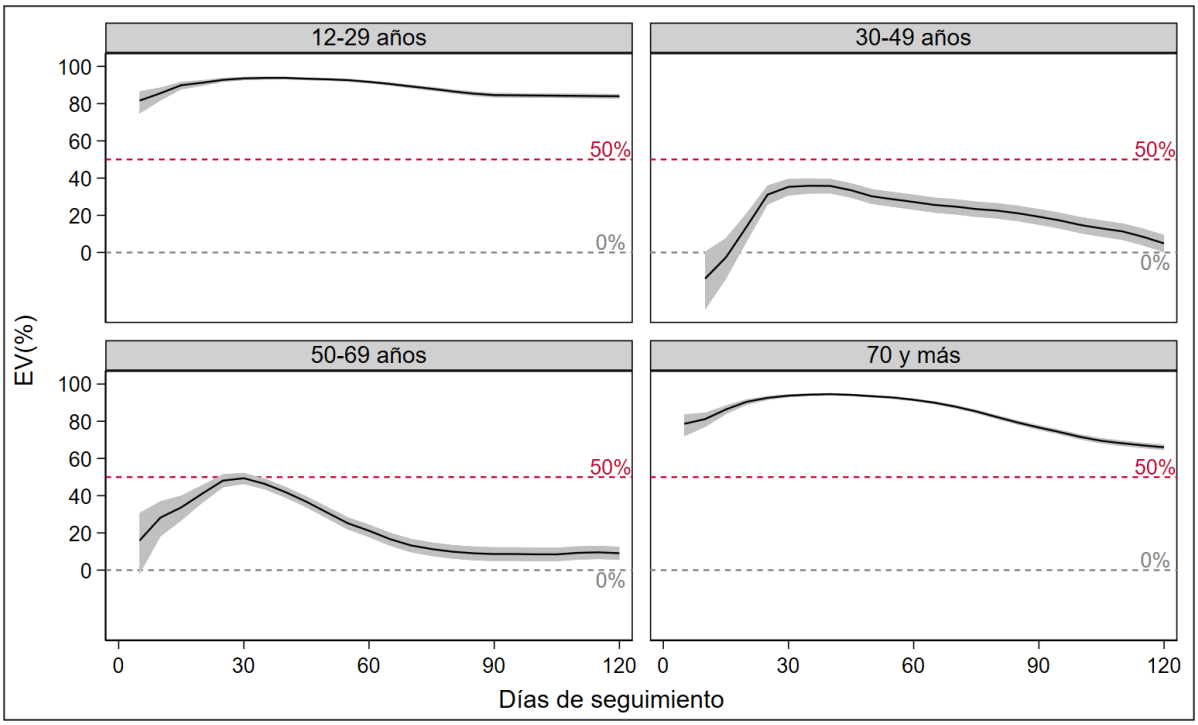

Figura 4.13: Curva de EV para primovacunación completa frente a infección por COVID-19 hasta un máximo de 120 días de seguimiento, por grupos de edad. La zona sombreada corresponde a la banda de confianza del 95 %. La línea horizontal punteada de color gris marca un umbral de EV del 0 % y la de color rojo del 50 %.

En la pérdida de EV por grupos de edad destacan, por un lado, los grupos de 12 a 29 años y de 70 años y más y, por el otro, los grupos de edad intermedios. Para los dos primeros, la EV se mantuvo por encima del 50 % hasta los 120 días de seguimiento. Además, las curvas de efectividad presentan una forma similar. Durante el primer mes la efectividad aumentó de manera suave hasta alcanzar su

máximo (del 90,8 % para los más jóvenes y del 94,6 % para los de mayor edad a 40 días de seguimiento), en el que se mantuvo hasta aproximadamente el segundo mes de seguimiento. A partir de ahí comenzó a descender, de manera un poco más acusada en el grupo de edad más avanzado, siendo de algo más del 80 % para los de 12 a 29 años a 120 días de seguimiento y del 68 % para las personas de 70 años y más en ese mismo momento.

Para los de 30 a 49 años la EV no llegó a alcanzar el 50 %, siendo incluso negativa en los primeros tiempos de seguimiento. La EV apenas alcanzó el 40 % el primer mes, momento a partir del cual comenzó a descender para, a los 120 días de seguimiento, caer al 0 %. Como ya se indicó, estos resultados guardan relación con el tipo de vacuna y el momento de administración de las dosis, lo que provocó que las personas primovacunadas de estas edades se contagiaran más que las no vacunadas.

Para el grupo de edad entre 50 y 69 años, la EV rozó el 50 % al primer mes de seguimiento. A partir de entonces comenzó la pérdida de efectividad, de forma bastante acusada hasta los dos meses y medio. Desde ahí y hasta los 120 días, se mantuvo estable en un porcentaje del 10 %.

La pérdida de efectividad frente a ingreso fue similar en los dos grupos de edad definidos a partir de los 50 años. A los 30 días de seguimiento, la EV era del 88 % para los de 50 a 69 años y del 98 % para los de 70 y más. Estos valores descendieron ligeramente y se mantuvieron más o menos estables hasta los 120 días, siendo entonces del 77 % para los primeros y del 77,9 % para las personas de mayor edad. En el grupo de edad de 30 a 49 años, la efectividad fue más variable. Durante el primer mes se produjo un primer ascenso hasta el 68,7 %, pero este porcentaje cayó rápidamente al 38,8 %. Durante el segundo mes de seguimiento (entre los 35-60 días) ascendió de nuevo para situarse en un 64,7 %, valor en torno al cual se mantuvo hasta los 120 días. Para los de 12 a 19 años la EV frente a ingreso en UCI fue del 100 % durante los 120 días de seguimiento considerados.

4.2.4. Efectividad vacunal por tipo de vacuna

Tasas de incidencia

Se muestran ahora las tasas de incidencia por 1.000 personas-año frente a infección e ingreso en UCI por COVID-19 por tipo de vacuna (Tablas 4.11 y 4.12, respectivamente). Recordemos que para Janssen y AstraZeneca únicamente se consideraron las personas en el rango de edad de 40 a 69 años.

	Infección				
Tipo de vacuna	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
Janssen					
No vacunados	450.204,23	10.114	22,47	22,03	22,91
Primovacunados	36.439,42	1.802	49,45	47,22	51,79
Moderna					
No vacunados	936.563,12	38.745	41,37	40,96	41,78
Primovacunados	77.534,38	1.008	13,00	12,22	13,83
AstraZeneca					
No vacunados	450.024,23	10.114	22,47	22,03	22,91
Primovacunados	90.753,25	2.704	29,80	28,69	30,94
Pfizer					
No vacunados	936.563,12	38.745	41,37	40,96	41,78
Primovacunados	589.626,28	12.807	21,72	21,35	22,10

Tabla 4.11: Tasas de incidencia de COVID-19 por 1.000 personas-año frente a infección por tipo de vacuna.

En el caso de Moderna y Pfizer, las tasas de incidencia frente a infección fueron superiores en no vacunados frente a primovacunados, con una diferencia de tasas superior en el caso de Moderna. Para

Ingreso en UCI					
Tipo de vacuna	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
Janssen					
No vacunados	450.204,23	191	0,42	0,37	0,49
Primovacunados	36.439,42	11	0,30	0,17	0,55
Moderna					
No vacunados	936.563,12	423	0,45	0,41	0,50
Primovacunados	77.534,38	8	0,10	0,05	0,21
AstraZeneca					
No vacunados	450.204,23	191	0,42	0,37	0,49
Primovacunados	90.753,25	19	0,21	0,13	0,33
Pfizer					
No vacunados	936.563,12	423	0,45	0,41	0,50
Primovacunados	589.626,28	65	0,11	0,09	0,19

Tabla 4.12: Tasas de incidencia de COVID-19 por 1.000 personas-año frente a ingreso en UCI por tipo de vacuna.

Janssen y AstraZeneca, sin embargo, la incidencia del virus fue superior en primovacunados frente a no inmunizados. Destaca especialmente el caso de Janssen, para la que la incidencia en primovacunados fue de más del doble que para no vacunados.

Para el desenlace ingreso en UCI, las tasas de incidencia en no vacunados se situaron en 0,45 para las vacunas de ARNm y en 0,42 para Janssen y Moderna. Para los primovacunados, las tasas fueron más heterogéneas, siendo más bajas para Moderna y Pfizer (alrededor de 0,10), de 0,21 para AstraZeneca y de 0,30 para Janssen.

Efectividad vacunal

En la Tabla 4.13 se muestran las estimaciones de la EV para cada desenlace por tipo de vacuna, a 180 días de seguimiento para todas las vacunas salvo para AstraZeneca (170 días), con su correspondiente intervalo de confianza del 95 %. Los valores ausentes para la efectividad se corresponden con estimaciones negativas (lo que implica un RR > 1).

Solo las vacunas de ARNm resultaron ser efectivas frente a infección por COVID-19. En el caso de Moderna, la EV fue del 67,7 % (IC95 %: 67,4-68,1). Para Pfizer, la vacuna administrada mayoritariamente a la población gallega, se obtuvo una EV de 34,7 % (IC95 %: 34,3-35,3). Los RR de infección en primovacunados frente a no vacunados para Janssen y AstraZeneca fueron, respectivamente, de 2,1 (IC95 %: 2-2,2) y de 1,1 (IC95 %: 1,1-1,2).

Para el desenlace ingreso en UCI, se observan de nuevo mejores resultados de efectividad para las vacunas de ARNm. Las estimaciones de la EV para Moderna y Pfizer fueron del 78,8 % (IC95 %: 78,7-79,2) y del 82,4 % (IC95 %: 82,2-82,6). Para Janssen se obtuvo una estimación de la efectividad frente a ingreso en UCI baja, del 20,1 % (IC95 %: 19,7-21,6) sin ajustar y de solo el 13,1 % (IC95 %: 12,6-14,7) tras el ajuste. Este resultado se relaciona con la proximidad entre la tasa de incidencia de ingreso en UCI en no vacunados, de 0,42 (IC95 %: 0,37-0,49), y la de vacunados con Janssen, de 0,30 (IC95 %: 0,17-0,55). Para AstraZeneca la efectividad frente a este desenlace fue del 46,7 % (IC95 %: 46,4-47,6) sin ajustar y de casi algo más del 70 % ajustada.

Desenlaces	Sin ajustar			Ajustada		
	EV(%)	IC95 %		EV(%)	IC95 %	
Infección						
Janssen	—	—	—	—	—	—
Moderna	62,9	62,5	63,3	67,7	67,4	68,1
AstraZeneca	—	—	—	—	—	—
Pfizer	31,1	30,5	31,8	34,7	34,3	35,3
Ingreso en UCI						
Janssen	20,1	19,7	21,6	13,1	12,6	14,7
Moderna	78,3	78,2	78,7	78,8	78,7	79,2
AstraZeneca	46,7	46,4	47,6	71,2	71,1	71,7
Pfizer	77,8	77,6	78,1	82,4	82,2	82,6

Tabla 4.13: EV frente a infección e ingreso en UCI por COVID-19 por tipo de vacuna a 180 días de seguimiento para todas las vacunas salvo para AstraZeneca (170 días), sin ajustar y ajustada, e intervalos de confianza del 95 %.

Pérdida de efectividad

Para la pérdida de EV frente a ingreso por tipo de vacuna, representada en la Figura 4.14, al igual que por grupos de edad se observan problemas de efectividad en algunas de las curvas. También se han eliminado los valores de EV correspondientes a RR mayores a 1,20 por motivos de escala.

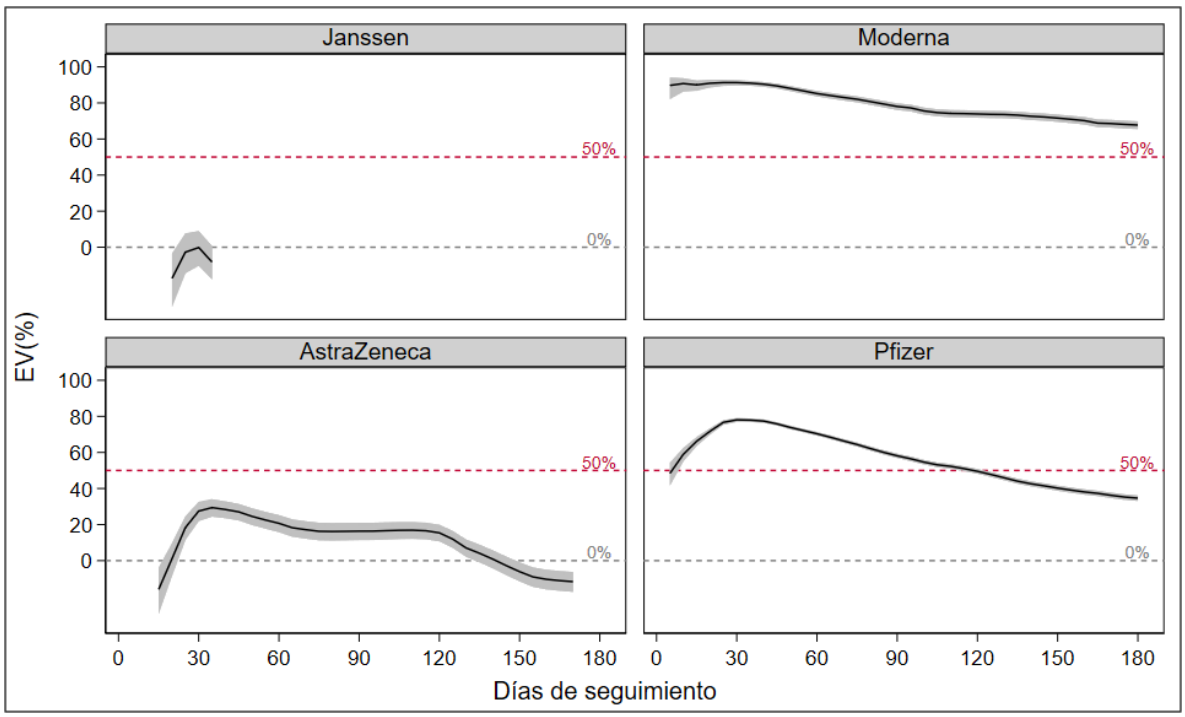

Figura 4.14: Curva de EV para primovacunación completa frente a infección por COVID-19, hasta un máximo de 180 días de seguimiento, por tipo de vacuna. La zona sombreada corresponde a la banda de confianza del 95 %. La línea horizontal punteada de color gris marca un umbral de EV del 0 % y la de color rojo del 50 %.

Moderna es la vacuna que presentó mayor efectividad y, además, menor pérdida con el paso del tiempo. El porcentaje de EV fue superior al 50 % al menos durante los primeros 180 días de seguimiento, alcanzando su máximo durante las primeras semanas tras la administración. A partir del mes y medio de seguimiento se inició la pérdida de efectividad, primero, de manera más rápida y, poco después de los tres meses, más lentamente.

Para Pfizer, la otra vacuna de ARNm, la curva de pérdida de efectividad es muy similar a la observada en global. Esto se debe a que casi el 70 % de la población recibió la primera dosis con esta vacuna. La efectividad comenzó siendo del 50 % tras la administración, llegando a un máximo del 80 % en el primer mes. A partir del mes y medio de seguimiento se produjo la pérdida de efectividad que, sobre los 120 días, cayó por debajo del 50 % para finalizar a los 180 días de seguimiento en algo más del 35 %.

La efectividad para la vacuna de AstraZeneca fue inferior al 40 % a lo largo de los 170 días de seguimiento. Comenzó siendo negativa durante los primeros 15-20 días, lo que indica que, al poco tiempo de recibir la segunda dosis, el riesgo de infección era superior en primovacunados frente a no vacunados. Al mes de seguimiento se alcanzó la efectividad máxima y, a partir del mes y medio, la inmunidad empezó a caer. Entre los 70 y los 120 días de seguimiento parece que la EV se mantuvo estable en torno al 20 %. A partir de los 120 días se produjo una nueva bajada y rápidamente la EV cayó a valores negativos.

Para la vacuna de Janssen apenas se representa la curva de pérdida de efectividad porque la mayoría de valores de RR fueron superiores a 1,20.

En relación a la pérdida de efectividad frente a ingreso en UCI, Moderna, Pfizer y AstraZeneca presentaron un buen comportamiento al menos durante los primeros 180 días tras la administración. En el primer mes de seguimiento, la efectividad de estas vacunas se situó, respectivamente, en un 88,4 %, un 99,2 % y un 82,5 %. A los 90 días, el porcentaje de EV era del 78,8 % para Moderna, del 87,1 % para Pfizer y del 80,5 % para AstraZeneca. A los 180 días estos valores apenas habían cambiado, siendo de nuevo del 78,8 % para la primera vacuna, del 71,1 % para la segunda y del 82,4 % para la tercera. En el caso de Janssen, aunque la vacuna resultó también efectiva frente a este desenlace, no llegó a alcanzar el 50 %. El valor máximo fue del 49,4 %, alcanzado al mes de seguimiento. A partir de entonces, la efectividad experimentó algunas fluctuaciones y a los 180 días era del 13,1 %.

4.3. Primovacunados *vs.* Dosis de recuerdo

4.3.1. Características de la cohorte

En el segundo análisis se estimó la EV de la dosis de recuerdo frente a primovacunación completa. Para ello, se empleó una nueva cohorte constituida por 1.919.536 individuos y 3.555.822 registros. De estos, el 54 % (1.919.536) eran primovacunados y el 46 % (1.636.286) restante primovacunados con dosis de recuerdo. Nótese que los individuos con dosis de recuerdo contaban con dos registros en la base de datos de la cohorte, uno como primovacunados y otro como primovacunados con dosis de recuerdo, que se trataron de manera independiente. Además, para este segundo análisis se excluyeron las personas de entre 12 y 19 años (467.129 registros), dado que en este grupo de edad la dosis de recuerdo no se administró de manera general, sino únicamente a aquellos que la solicitaron de forma explícita. Los individuos de la cohorte fueron seguidos desde que completaron la primovacunación hasta el 27 de marzo de 2022, siendo susceptibles de haber recibido en ese tiempo la dosis de refuerzo. La Figura 4.15 corresponde al diagrama de esta segunda cohorte. Para ambos estados de vacunación se indica el número de individuos que pasaron por el mismo durante su seguimiento, así como los casos de COVID-19 e ingresos en UCI y las defunciones. De los 1.919.536 individuos que iniciaron el seguimiento como completamente primovacunados, 1.636.286 recibieron la dosis de recuerdo. A 27 de marzo de 2022, 122.748 personas finalizaron su seguimiento como primovacunadas y 1.515.034 lo hicieron como primovacunadas con dosis de recuerdo. En total, en este segundo análisis hubo 266.026 casos de COVID-19 (de los cuales 314 ingresaron en UCI) y 15.728 defunciones por otras causas.

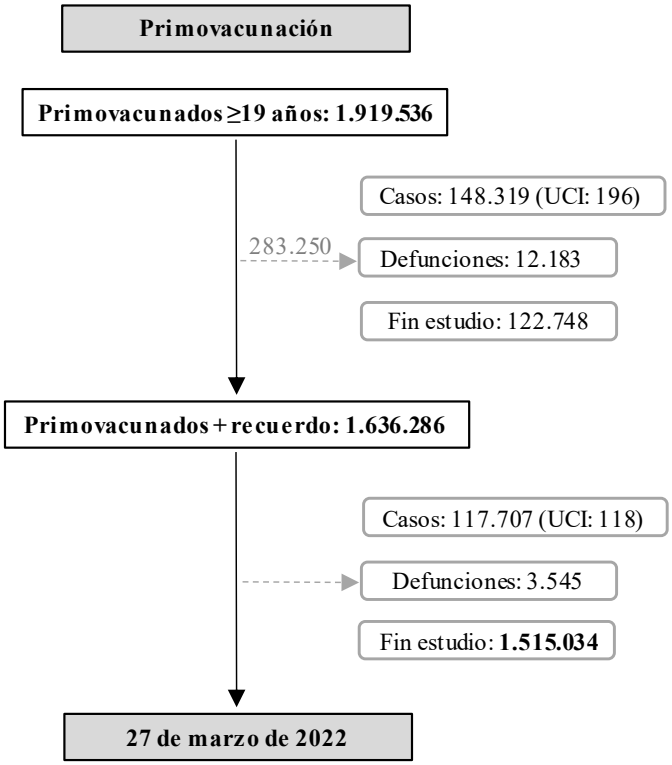

Figura 4.15: Diagrama de la cohorte para el segundo análisis (personas de 19 años o más), seguida desde la primovacunación hasta el 27 de marzo de 2022. Fuente: DXSP. Elaboración propia.

En la Tabla 4.14 se recoge la distribución de las características de la cohorte por grupos de vacunación para este segundo análisis. Las variables sexo y ámbito de residencia presentan una distribución muy similar en ambos grupos, mientras que para el grupo de edad se observan ciertas diferencias. Los más jóvenes, de entre 20 y 39 años, cuentan con menor representación en el grupo de primovacunados con dosis de recuerdo. En las personas de 60 años y más, sin embargo, el porcentaje de primovacunados con dosis de refuerzo frente a no vacunados es superior. Esto se relaciona con las coberturas de la dosis de recuerdo por grupos de edad, representadas en la Figura 1.5 del Capítulo 1. A 27 de marzo de 2022, la cobertura de la dosis de recuerdo para personas entre 20 y 39 años rondaba el 60 %. En contraposición, a partir de los 60 años, los porcentajes de cobertura se situaban por encima del 90 %.

En cuanto al número de PDIAs negativas realizadas durante el seguimiento, en ambos grupos de vacunación la mayoría de individuos no tenían ninguna prueba asociada o tenían solo una. Parece que los primovacunados, además, se realizaron más pruebas diagnósticas (negativas) durante su seguimiento que las personas con dosis de recuerdo.

Las características de esta segunda cohorte se relacionan con la estrategia de vacunación por grupos de edad y el momento en el que comenzó a administrarse la dosis de recuerdo. Los grupos de edad más jóvenes, que completaron la primovacunación más tarde, también recibieron la dosis de recuerdo más tarde. Concretamente, en el grupo de edad de 20 a 29 años, el número de personas en seguimiento para la dosis de recuerdo a los 90, 110 y 120 días de seguimiento era, respectivamente, de 4.763, 263 y 47. Por ello, la estimación de la EV en este segundo análisis se restringió a un máximo de 120 días de seguimiento, debido a la proximidad entre la administración de la dosis de recuerdo y el fin del estudio (solo el 18 % de las dosis de recuerdo se administraron antes de diciembre, de modo que, como máximo, el tiempo de seguimiento para los individuos de esta segunda cohorte como primovacunados

	Primovacunados		**Primovacunados + recuerdo**	
Características	n	%	n	%
Sexo				
Masculino	899.723	46,9	760.873	46,5
Femenino	1.019.813	53,1	875.413	53,5
Grupo de edad				
20-29	141.068	7,3	76.117	4,7
30-39	228.368	11,9	156.055	9,5
40-49	371.744	19,4	302.289	18,5
50-59	361.157	18,8	329.324	20,1
60-69	315.922	16,5	299.458	18,3
70-79	269.554	14,0	258.232	15,8
80+	231.723	12,1	214.811	13,1
Ámbito				
Urbano	1.263.959	65,8	1.063.367	65,0
Semiurbano	387.608	20,2	335.012	20,5
Rural	267.969	14,0	237.907	14,5
Nº PDIAs				
0	1.452.532	75,7	1.385.048	84,6
1	276.418	14,4	169.312	10,3
2	113.088	5,9	49.931	3,1
3+	77.498	4,0	31.9951	2,0

Tabla 4.14: Distribución de las características de la cohorte del segundo análisis por grupos de vacunación. Fuente: DXSP. Elaboración propia.

con dosis de refuerzo fue de 120 días).

En la Figura 4.16 se muestra el estimador de Nelson-Aalen del riesgo acumulado de infección (izquierda) y de ingreso en UCI (derecha) por COVID-19 en los dos grupos de vacunación. Estas gráficas sugieren que la dosis de refuerzo no tuvo efecto protector frente a ninguno de los desenlaces en esta cohorte, pues los riesgos acumulados de los primovacunados con dosis de recuerdo son superiores a los de los primovacunados. El valor p del estadístico del test log-rank para el desenlace infección fue inferior a 0,001, al igual que en el caso del ingreso en UCI. Bajo cualquier nivel de significación de los tomados habitualmente, se concluye que las pruebas son significativas y se rechaza la igualdad de riesgos entre primovacunados y primovacunados con dosis de recuerdo para ambos desenlaces en esta cohorte. Este resultado coincide con lo observado a nivel gráfico. La Figura 4.16 también sugiere, como ya se indicó, cortar el seguimiento de los primovacunados con dosis de recuerdo a los 120 días, especialmente para el desenlace ingreso en UCI. La escasez de individuos se refleja en la sección vertical de la gráfica a partir de ese tiempo de seguimiento.

Para este segundo análisis debemos tener en cuenta la irrupción de Ómicron en el escenario de la pandemia, la variante de mayor transmisibilidad y capacidad para eludir la inmunidad conocida hasta el momento. En Galicia, la presencia de Ómicron se registró por primera vez el 29 de noviembre de 2021. En aquel momento, los grupos de mayor edad (a partir de los 50 años) presentaban ya coberturas para la dosis de recuerdo superiores al 50 %, de modo que más de la mitad de los vacunados en esos grupos estuvieron expuestos al mayor riesgo de toda la pandemia con la dosis de refuerzo. Por este motivo, la dosis no parece tener un efecto protector frente a infección. En este sentido, se pensó en igualar el riesgo de exposición al virus en los dos grupos de vacunación, primovacunados y primovacunados con dosis

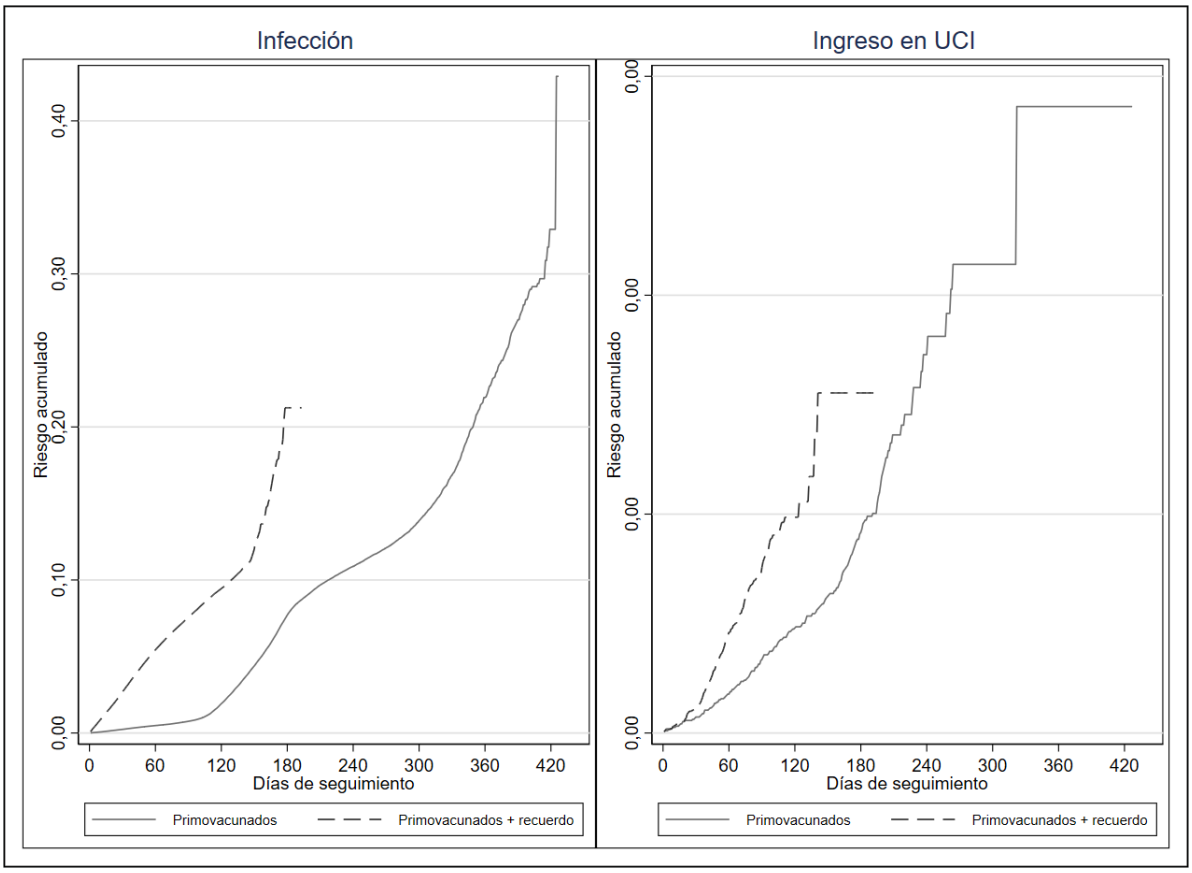

Figura 4.16: Estimador de Nelson-Aalen del riesgo acumulado frente a infección (izquierda) e ingreso en UCI (derecha) por COVID-19 en individuos primovacunados e individuos primovacunados con dosis de recuerdo. Los valores del eje Y para el segundo gráfico guardan relación con el bajo número de individuos que ingresaron en UCI en comparación con el tamaño de la cohorte.

de recuerdo. Para ello, se consideró que, durante el seguimiento, los individuos no estuvieron a riesgo hasta el 29 de noviembre de 2021. Así, todos los individuos fueron seguidos desde esta fecha hasta la recepción de la siguiente dosis de la vacuna (solo en el caso de los primovacunados), el diagnóstico positivo en COVID-19, la defunción o el fin del estudio el 27 de marzo de 2022, lo que antes sucediese.

En la Figura 4.17 se muestra de nuevo el estimador de Nelson-Aalen del riesgo acumulado para los dos desenlaces y ambos grupos de vacunación tras igualar los riesgos de exposición al virus mediante la estrategia indicada. Bajo este nuevo enfoque, la dosis de recuerdo pasa a tener un efecto protector frente a infección y también frente a ingreso en UCI, pues ahora el riesgo acumulado en personas con dosis de recuerdo es inferior al de las personas primovacunadas. El test log-rank resultó de nuevo significativo para infección, con un valor p para el estadístico del contraste menor de 0,001. En el caso del ingreso en UCI, sin embargo, el valor p asociado a la prueba fue 0,932. Esto implica que no existen evidencias suficientes para considerar que los riesgos acumulados de ingreso en UCI fueron distintos en los dos grupos de vacunación.

Con fines comparativos, para este segundo análisis se tuvieron en cuenta los dos enfoques para la estimación de la efectividad de la dosis de recuerdo frente a la primovacunación, si bien la presencia de Ómicron hace que el enfoque 1 deje de ser adecuado debido a la gran diferencia de riesgos entre los grupos de vacunación. En la Tabla 4.15 se recogen las principales características de ambos enfoques en relación al tiempo de seguimiento y al riesgo de los individuos de la cohorte.

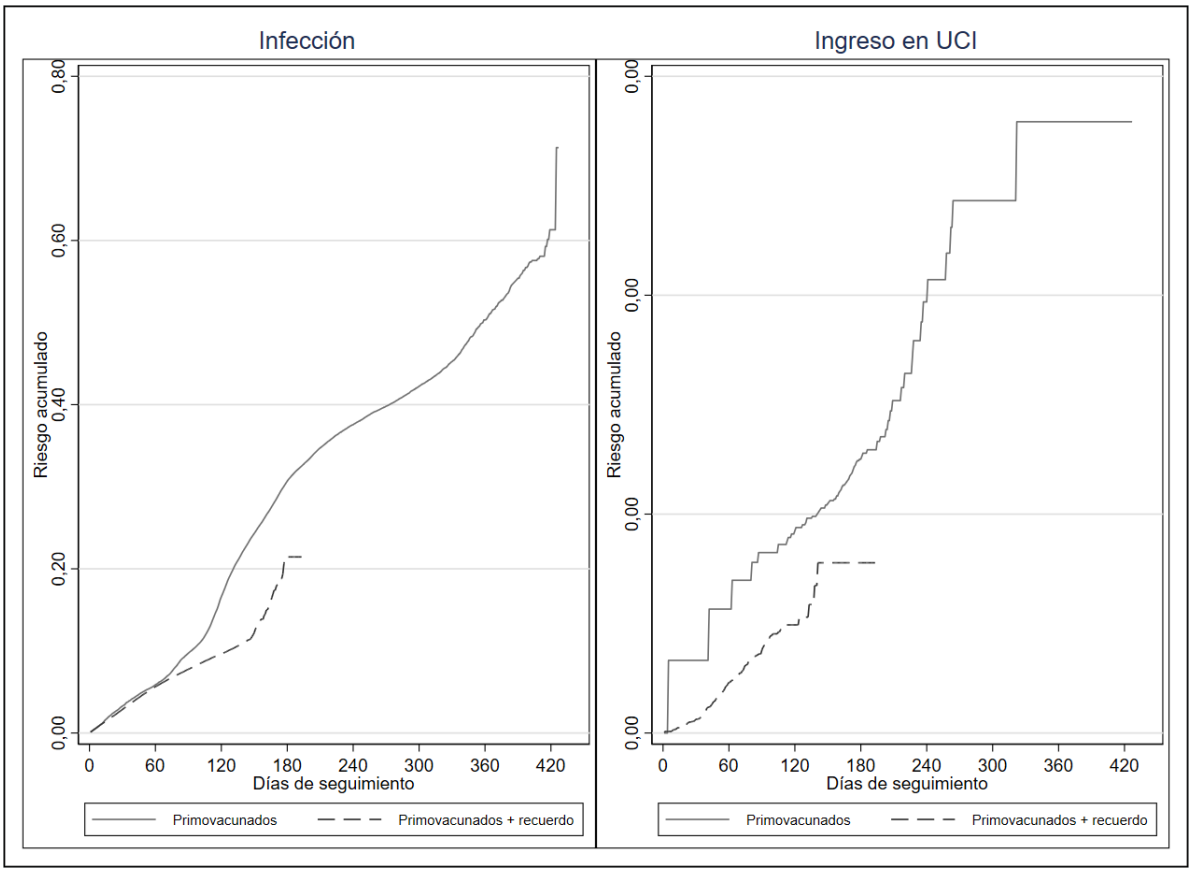

Figura 4.17: Estimador de Nelson-Aalen del riesgo acumulado frente a infección (izquierda) e ingreso en UCI (derecha) por COVID-19, en individuos primovacunados e individuos con dosis de recuerdo, desde el 29 de noviembre de 2021 hasta un máximo de 118 días de seguimiento. Los valores del eje Y para el segundo gráfico guardan relación con el bajo número de individuos que ingresaron en UCI en comparación con el tamaño de la cohorte.

El primer enfoque se corresponde con el considerado para el primer análisis. Los individuos se siguieron desde que recibieron la dosis de la vacuna (con su período de inducción correspondiente) y hasta un máximo de 120 días. Muchos primovacunados, por tanto, iniciaron su seguimiento antes de la llegada de Ómicron. Por este motivo, los individuos de ambos grupos de vacunación estuvieron en riesgo en diferentes momentos. Dado que, en general, las personas con dosis de recuerdo tuvieron un mayor riesgo de infección e ingreso en UCI por COVID-19 que los primovacunados, los resultados se ofrecen en términos del RR. Debido a esto, además, no se estimó la EV ni por grupos de edad ni por tipo de vacuna.

En el segundo enfoque se igualaron los riesgos de exposición al virus en ambos grupos de vacunación, al considerar el seguimiento de la cohorte a partir de la llegada de Ómicron el 29 de noviembre de 2021. El tiempo máximo de seguimiento fue de 118 días, hasta el 27 de marzo de 2022. En este caso, el riesgo de infección por COVID-19 es inferior en primovacunados con dosis de recuerdo, por lo que en algunos casos sí se ofrecen las estimaciones de la efectividad. Solo se considera el desenlace infección.

Las covariables de ajuste incluidas en los modelos fueron el grupo de edad decenal, el sexo, el ámbito de residencia y el número de PDIAs negativas. Para este segundo análisis ninguna de ellas cumplía con la hipótesis de riesgos proporcionales (aunque, como en el análisis anterior, los coeficientes de interacción $\hat{\gamma}$ eran muy próximos a 0), por lo que todas se incluyeron en los modelos como variables

Características	Enfoque 1	Enfoque 2
Desenlace	Infección e ingreso en UCI	Infección
Inicio del seguimiento	Vacunación	Vacunación/29-11-2021
Fin del seguimiento	Máximo 120 días	Máximo 118 días
Exposición al virus	Distinta	Igual
Estimación	RR	RR/EV

Tabla 4.15: Características de los dos enfoques planteados para la estimación de la efectividad de la dosis de recuerdo.

de estratificación.

4.3.2. Efectividad vacunal global

Tasas de incidencia

Las tasas de incidencia para el primer enfoque dependen en gran medida del momento de la pandemia en el que se sitúa el seguimiento de la segunda cohorte. En diciembre de 2021 irrumpió en el escenario de la pandemia Ómicron, la variante responsable de la ola de casos acontecida en el quinto período, en la que se alcanzaron incidencias acumuladas a 14 días de hasta 3.500 casos (ver Figura 1.2). Por este motivo, se espera que en los primovacunados con dosis de recuerdo la tasa de incidencia sea superior a la observada para los primovacunados, ya que estuvieron expuestos al mayor riesgo de toda la pandemia con la dosis de refuerzo.

En la Tabla 4.16 se recogen las tasas de incidencia globales por 1.000 personas-año y sus intervalos de confianza del 95 % para infección e ingreso en UCI por COVID-19, en el grupo de primovacunados y en el grupo de primovacunados con dosis de recuerdo. Además, a modo de referencia, se indican también las tasas de los no vacunados.

Enfoque 1					
Desenlaces	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
Infección					
No vacunados	824.804,48	29.024	35,19	34,79	35,60
Primovacunados	977.523,10	148.319	151,73	150,96	152,50
Primovacunados + recuerdo	387.920,62	117.707	303,43	301,70	305,17
Ingreso en UCI					
No vacunados	824.804,48	432	0,52	0,48	0,58
Primovacunados	977.523,10	196	0,20	0,17	0,23
Primovacunados + recuerdo	387.920,62	118	0,30	0,25	0,36

Tabla 4.16: Tasas de incidencia globales de COVID-19 por 1.000 personas-año para cada desenlace, en no vacunados, primovacunados y primovacunados con dosis de recuerdo.

Para infección, la tasa de incidencia fue de 151,73 para primovacunados y de 303,43 para primovacunados con dosis de recuerdo. Como ya se indicó, el elevado número de casos en personas con dosis de refuerzo está relacionado con la llegada de Ómicron. En el caso de los no vacunados, la tasa de incidencia es de 35,19. En cuanto al ingreso en UCI, la tasa de incidencia también es mayor en vacunados con dosis de recuerdo frente a primovacunados, siendo de 0,20 frente a 0,30, respectivamente. Los no vacunados presentan la mayor tasa de incidencia frente a este desenlace, dado que en este grupo se produjeron la mayoría de los ingresos en UCI.

Enfoque 2					
Desenlaces	**Personas-año**	**Casos**	**Tasa**	**IC95 %**	
Infección					
No vacunados	1.296,33	409	315,51	286,36	347,61
Primovacunados	176.027,38	128.475	729,86	725,88	733,86
Primovacunados + recuerdo	387.272,64	117.537	310,72	308,95	312,50
Ingreso en UCI					
No vacunados	1.296,33	11	8,49	4,70	15,32
Primovacunados	176.027,38	78	0,44	0,35	0,55
Primovacunados + recuerdo	387.272,64	117	0,31	0,26	0,37

Tabla 4.17: Tasas de incidencia globales de COVID-19 por 1.000 personas-año para cada desenlace, en no vacunados, primovacunados y primovacunados con dosis de recuerdo.

Para el segundo enfoque se observan tasas de incidencia frente a infección por COVID-19 muy distintas (ver Tabla 4.17). Para no vacunados y primovacunados con dosis de refuerzo las tasas se sitúan en torno a los 310 casos por 1.000 personas-año, mientras que para los primovacunados el valor asciende a casi 730 casos. La principal diferencia con respecto al enfoque 1 es que el número de personas-año a riesgo se ha reducido mucho en este grupo de vacunación (que ahora entra en riesgo el 29 de noviembre de 2021), pero el número de casos apenas ha disminuido. Para el ingreso en UCI, la tasa en no vacunados destaca especialmente sobre las demás. Los primovacunados y los primovacunados con dosis de refuerzo presentan tasas similares, ya que sus riesgos para este desenlace en el segundo enfoque no se pudieron considerar significativamente distintos.

Efectividad vacunal

En la Tabla 4.18 se ofrecen las estimaciones del RR bajo el primer enfoque, sin ajustar y ajustadas. En el caso del desenlace infección el ajuste dispara el riesgo a casi el doble, tomando un valor de 11,4 (IC95 %: 11,3-11,6). Esto implica que la infección por COVID-19 fue hasta 11,4 veces más frecuente en primovacunados con dosis de recuerdo que en primovacunados. Este valor nos da una idea del gran impacto de Ómicron en cuanto a infección. Para el ingreso en UCI, aunque el riesgo fue también superior en vacunados con dosis de refuerzo frente a primovacunados, la razón del riesgo de ingreso con dosis de refuerzo frente a primovacunados fue de 1,6 (IC95 %: 1,2-2,1).

Enfoque 1						
Desenlaces	**Sin ajustar**			**Ajustado**		
	RR	**IC95 %**		**RR**	**IC95 %**	
Infección	5,8	5,8	5,9	11,4	11,3	11,6
Ingreso en UCI	2,2	1,6	2,9	1,6	1,2	2,1

Tabla 4.18: RR frente a infección e ingreso en UCI por COVID-19 a un máximo de 120 días de seguimiento, sin ajustar y ajustado, e intervalos de confianza del 95 %.

Para el enfoque 2, cuyos resultados se muestran en la Tabla 4.19, se ofrece la EV sin ajustar y el RR tras el ajuste. Sin considerar ninguna covariable, la EV de la dosis de recuerdo frente a infección resultó ser del 59,2 % (IC95 %: 58,8-59,6). Sin embargo, tras el ajuste, la estimación de la efectividad se volvió negativa, obteniéndose un RR de infección del 1,2 (IC95 %: 1,2-1,3). Este valor indica que, considerando la edad, el sexo, el ámbito de residencia y el número de PDIAs negativas, en las personas con dosis de refuerzo la infección por COVID-19 fue un 20 % mayor que en primovacunados.

Enfoque 2						
Desenlace	**Sin ajustar**			**Ajustado**		
	EV(%)	**IC95 %**		**RR**	**IC95 %**	
Infección	59,2	58,8	59,6	1,2	1,2	1,3

Tabla 4.19: EV y RR frente a infección por COVID-19 a un máximo de 118 días de seguimiento, sin ajustar y ajustados, e intervalos de confianza del 95 %.

Este hecho se debe a varios motivos. Con este enfoque, en los grupos de edad más jóvenes, en los que la cobertura de la dosis de refuerzo fue más bien baja, los primovacunados tuvieron un mayor riesgo de infección que las personas con dosis de refuerzo, que eran minoritarias. Entre los 70 y los 79 años, los riesgos en ambos grupos comenzaron a igualarse. A partir de los 80 años, donde la cobertura de la dosis de recuerdo fue prácticamente del 100 % y el número de primovacunados era muy bajo en relación a las personas con dosis de refuerzo, el riesgo en primovacunados frente a infección se volvió superior. En cuanto al sexo, el riesgo fue superior en primovacunados con dosis de recuerdo frente a solo primovacunados durante los 118 días de seguimiento tanto en hombres como en mujeres. Por otro lado, en los tres ámbitos de residencia los riesgos de ambos grupos de vacunación fueron prácticamente iguales hasta aproximadamente los 70 días de seguimiento, momento a partir del cual el riesgo en primovacunados con dosis de recuerdo aumentó respecto al de los primovacunados. Por último, para el número de PDIAs negativas, para las categorías 0 y 1 los riesgos fueron similares en la mayoría del tiempo de seguimiento salvo al final, donde de nuevo el riesgo de las personas con dosis de recuerdo aumentó respecto al de las primovacunadas. Para las categorías 2 y 3 o más, sin embargo, el riesgo en primovacunados fue superior, especialmente en la última clase. Estas diferencias en los riesgos en ambos grupos de vacunación en función de las categorías de las covariables hacen que, tras la estratificación, la efectividad caiga a valores negativos y, por tanto, el RR de infección sea superior en vacunados con dosis de recuerdo frente a primovacunados.

4.4. Validación de los modelos

Observaciones influyentes

Se obtuvieron los valores $l_{\max}$ para cada uno de los análisis realizados y para cada desenlace, y se representaron frente a los días de seguimiento (resultados no mostrados). En ningún caso se detectaron observaciones influyentes, pues los valores máximos $l_{\max}$ fueron muy bajos (entre 0 y 0,1).

Hipótesis de riesgos proporcionales

Antes de ajustar los modelos, se vio que algunas de las variables de interés no cumplían con la hipótesis de riesgos proporcionales, si bien todos los coeficientes de interacción estimados $\hat{\gamma}$ eran muy próximos a 0. En el caso del grupo de edad, sexo, ámbito de residencia y número de PDIAs negativas realizadas durante el seguimiento se recurrió a la estratificación (cuando fue necesario) para permitir la variación del riesgo base $\lambda_0(t)$ en cada uno de los estratos. Para el estado de vacunación la estratificación no fue posible, ya que el proceso no permite conocer el RR relativo asociado a la variable de estratificación (y que necesitamos conocer para estimar la EV). En su lugar, se decidió cortar el seguimiento de los individuos en el momento del tiempo en el que la hipótesis parecía dejar de cumplirse para los grupos de vacunación.

Para probar la hipótesis de riesgos proporcionales se aplicaron principalmente métodos gráficos. Considerando las variables de ajuste empleadas, de tipo categórico, no se utilizaron los residuos martingala. El cálculo de estos residuos resulta más adecuado para el caso de variables continuas, con el fin de determinar la forma funcional más adecuada para su incorporación al modelo (por ejemplo, en escala logarítmica). Tampoco se recurrió a los residuos de Schoenfeld, que pueden resultar útiles en

ciertas ocasiones para chequear gráficamente si la pendiente del coeficiente estimado para una covariable varía o no con el tiempo. Con todo, ambos tipos se describen en el Capítulo 2 con la intención de introducir los principales residuos asociados a la validación del modelo Cox, así como para posibles usos futuros.

Capítulo 5

Conclusiones

Como cierre al trabajo, en este Capítulo se presentan las conclusiones, fortalezas y limitaciones de los análisis. Además, se plantean posibles líneas de trabajo futuras.

Conclusiones

1. No vacunados *vs.* Primovacunados

- En general, la primovacunación resultó efectiva frente a infección y, en mayor medida, frente a ingreso en UCI en comparación con la no inmunización. A 180 días de seguimiento, las estimaciones de la efectividad vacunal fueron del 27 % frente a infección y del 76 % frente a ingreso en UCI.
- La primovacunación resultó especialmente efectiva en los grupos de edad de 12 a 19 años y de 70 años y más frente a la no inmunización. Estos grupos fueron vacunados, principalmente, con Moderna y Pfizer.
- La primovacunación resultó efectiva en todos los grupos de edad frente a ingreso en UCI en comparación con la no inmunización, siendo superior al 70 % en todos los casos.
- Solo Moderna y Pfizer (y AstraZeneca, ligeramente) tuvieron un efecto protector frente a infección por COVID-19.
- Todas las vacunas resultaron efectivas, aunque con diferencias, frente a ingreso en UCI por COVID-19. Destacan especialmente Moderna, Pfizer y AstraZeneca, con efectividades en torno al 70-80 % frente a este desenlace.
- Parece existir una relación entre las efectividades en el grupo de 30 a 49 años y para la vacuna de Janssen y, análogamente, en personas de 50 a 69 años y para AstraZeneca.
- En general, la pérdida de efectividad de la primovacunación frente a infección comenzó a observarse a los 40-45 días tras completar la pauta (momento en el que la efectividad alcanzó su máximo, del 70 %). A partir de entonces, la efectividad comenzó a descender de manera casi proporcional al tiempo. A los 180 días, su valor era del 27 %.
- La pérdida de efectividad de la primovacunación frente a infección fue más acusada en personas entre 30 y 69 años. Además, en estos grupos de edad los valores de efectividad nunca superaron el 50 %.
- Moderna fue la vacuna que presentó una menor pérdida de efectividad (67,7 % a los 180 días), seguida de Pfizer (34,6 % a los 180 días). Para AstraZeneca la efectividad máxima fue de un 29,4 % y, a los 180 días, había caído a valores negativos. En el caso de Janssen no fue posible analizar la pérdida de efectividad, dado que los primovacunados estuvieron sometidos a un mayor riesgo de infección que los no vacunados durante los primeros 180 días de seguimiento.

- A 180 días de seguimiento, la pérdida de efectividad frente a ingreso en UCI era de un 17 %.
- Tanto en general, como por grupos de edad y tipo de vacuna, la efectividad presentó un comportamiento común: aumentó en los primeros días tras la administración de la dosis hasta alcanzar su máximo, en torno al mes o mes y medio de seguimiento y, a partir de ese momento, comenzó la pérdida de efectividad.

2. Primovacunados *vs.* Dosis de recuerdo

- Las características de Ómicron invalidan la aplicación del enfoque en el que se considera el seguimiento de los individuos en distintos momentos de la pandemia (distinta exposición al virus). Cabe destacar que el resultado se incluyó exclusivamente con fines comparativos.
- Considerando el seguimiento de los individuos en el mismo momento de la pandemia (misma exposición al virus), en general, la dosis de recuerdo no resultó efectiva frente a infección en comparación con la primovacunación. Este resultado guarda relación con las diferencias entre los tiempos desde la vacunación y con la heterogeneidad en la distribución de los grupos de vacunación comparados.

Las conclusiones obtenidas deben ser consideradas con cautela, teniendo en cuenta la relación entre los grupos de edad, el tipo de vacuna, la estrategia de vacunación y la propia evolución de la pandemia en nuestra comunidad, considerando la llegada de las variantes de principal interés y sus características en cuanto a transmisibilidad y capacidad para eludir la inmunidad. También es importante restringir las conclusiones a la población de estudio, evitando generalizaciones.

En Pardo-Seco et al. (2022) se recoge un estudio previo sobre la estimación de la efectividad de Pfizer frente a infección, hospitalización, ingreso en UCI y defunción por COVID-19 en Galicia. Los autores aplicaron un diseño de casos y controles test negativo para estimar la EV en la población de 18 años o más con al menos una prueba diagnóstica de SARS-CoV-2 entre el 27 de diciembre de 2020 y el 18 de marzo de 2021 (ola de Alpha). Se consideraron como casos a los individuos con una prueba positiva (la primera) y como controles a los individuos con una o más pruebas negativas realizadas durante el período de estudio (por tanto, un mismo individuo podía ser considerado control en más de una ocasión). La exposición se definió como el estado de vacunación el día de la prueba. El método empleado para el cálculo de la EV fue la regresión logística, considerando como covariables de ajuste el sexo, la edad, el tiempo transcurrido desde la administración de la dosis (cero en el caso de los no vacunados) y el número de semanas transcurridas desde el inicio del estudio hasta la fecha de realización de la prueba. Las estimaciones de la EV en primovacunados frente a no vacunados, para el desenlace infección, fueron del 75,4 % (IC95 %: 70,1-80,1) a los 7-13 días de seguimiento y del 90,8 % (IC95 %: 88,6-92,7) a partir de los 14 días. Asumiendo las diferencias en el diseño, estos resultados pueden compararse con los obtenidos en nuestro estudio para la vacuna de Pfizer a los 65 días de completar la primovacunación en personas de 70 años y más (la duración del estudio de Pardo-Seco et al. es de 82 días pero, considerando la estimación a 14 días o más, se reduce a un máximo de 68 días; además, en el momento del estudio las personas vacunadas eran, fundamentalmente, las de 80 años y más y los grupos de riesgo). Este dato se corresponde con una EV del 90 % (IC95 %: 89,1-90,7), un resultado muy similar al estimado por Pardo-Seco et al. Los resultados de efectividad frente a ingreso en UCI no son comparables a los mostrados en nuestro estudio, pues en su caso la EV frente a este desenlace se analizó únicamente entre los diagnosticados de COVID-19.

Fortalezas

- **Tamaño de la población de estudio**

  La principal fortaleza de este estudio es el acceso a los datos y la posibilidad de trabajar con la totalidad de la población gallega vacunada. Esto permitió disponer de información sobre un gran número de individuos, evitando los problemas analíticos derivados de trabajar con tamaños

muestrales reducidos. Por otro lado, como se indica más adelante, el hecho de que la población de estudio fuera únicamente de vacunados supuso también una limitación.

- **Duración del estudio**

  Otro de los puntos fuertes del estudio está relacionado con su duración y los tiempos de seguimiento, especialmente extensos para el primer análisis (338 días; desde el 27 de diciembre de 2020 hasta el 30 de noviembre de 2021). Esto permitió alcanzar altas coberturas de primovacunación en todos los grupos de edad, obteniendo así estimaciones de EV para todos ellos y, además, resultados representativos de toda la población. También fue posible estimar la pérdida de efectividad, a cuatro meses en el peor de los casos. Entre los análisis 1 y 2, se pudo analizar prácticamente todo el tiempo de pandemia en Galicia desde la introducción de las vacunas.

Limitaciones

- **Procedencia de la población de estudio**

  La principal limitación de este estudio está relacionada con el origen de la población empleada en los análisis. Esta población procede del registro de vacunas de COVID-19 de Galicia, que solo incluye a las personas que recibieron al menos una dosis de una de las vacunas desde el inicio de la campaña de inmunización. Por tanto, no recoge información sobre las personas que nunca se vacunaron, cuyo porcentaje aumenta a medida que disminuye la edad. Por este motivo, es posible que exista un sesgo de infraestimación de las EV, al no poder considerar a los no vacunados que fueron caso o ingresaron en UCI.

  En relación con lo anterior, se encuentra la incapacidad para estimar la EV frente a defunción. Más de la mitad de los fallecidos a consecuencia del virus fueron no vacunados que nunca entraron en el registro de vacunas (56 %), de modo que el cálculo de la EV considerando solo las defunciones en la población de estudio llevaría también a una infraestimación del valor real. Además, tampoco fue posible estimar la EV frente a hospitalización debido a que en las bases de datos utilizadas no se podían diferenciar los ingresados por COVID-19 de los ingresados por otras causas que fueron diagnosticados de COVID-19 durante el ingreso (ingresados con COVID-19).

- **Manejo del riesgo al que estuvieron sometidos los individuos**

  Otra de las limitaciones de este estudio es la imposibilidad de captar en una variable el riesgo al que estuvieron sometidos los individuos durante su seguimiento en los distintos grupos de vacunación. En este sentido, el número de PDIAs negativas resultó poco informativo.

  Para el primer análisis, se realizaron varios intentos para tratar de obtener otra medida del nivel de exposición de los individuos al virus y sus variantes. Inicialmente, se calculó para cada individuo de la cohorte la incidencia media diaria por cada 100.000 habitantes en su área sanitaria y durante su período de seguimiento. Para cada registro, se obtuvo la diferencia entre la incidencia acumulada en la fecha de fin del seguimiento y la incidencia acumulada en la fecha de inicio. El resultado se dividió por el número de días de seguimiento y se aplicó el logaritmo para tratar de normalizar la distribución de la variable. Con esto, se pretendía obtener una medida de la incidencia a la que estuvo sometido cada individuo duramente su seguimiento. No obstante, al incorporar esta variable a los modelos se obtenían resultados incoherentes y difíciles de interpretar, y tampoco era sencillo considerarla a la hora de estimar la pérdida de inmunidad. La variable presentaba el problema de que tomaba valores bajos en personas con poco tiempo de seguimiento y que habían sido caso durante una ola, especialmente al inicio, a pesar de que estaban expuestos a un riesgo muy elevado.

  Alternativamente, se trataron de controlar tanto los distintos niveles de exposición a los que pudo estar sometido un individuo, como las variantes del virus, calculando el número de días que los individuos estuvieron expuestos en las principales olas de la pandemia. Por ejemplo,

para el primer análisis, que comprende el período de tiempo del 27 de diciembre de 2020 al 30 de noviembre de 2021, se definieron dos olas principales considerando un umbral de incidencia acumulada a 14 días por 100.000 habitantes igual o superior a 150, valor que diferenciaba un riesgo medio-bajo de un riesgo medio y superiores. Las olas resultantes se correspondían con las olas tercera y cuarta de la pandemia y se relacionaban con Alpha y Delta, respectivamente. La primera ola se estableció entre el 27 de diciembre de 2020 y el 26 de febrero de 2021, con una duración de 61 días. La campaña de vacunación estaba comenzando, de modo que la mayoría de personas pasaron esta ola como no vacunadas. La segunda ola se definió entre el 8 de julio y el 7 de septiembre de 2021, lo que se traduce en otros 61 días. Salvo en los menores de 20 años, en el resto de grupos de edad la primovacunación ya estaba muy avanzada. Por ello, los principales afectados en esta ola fueron los más jóvenes. Teniendo esto en cuenta y los distintos grupos por los que podía pasar un individuo de la cohorte a lo largo del estudio, se obtuvo el número de días del seguimiento de cada individuo que caían en cada una de estas olas. Una vez obtenidos los días y sus distribuciones, se establecieron categorías. Para la primera ola se construyó una variable con dos clases, que se podían interpretar como estar o no estar expuesto en la ola debida a Alpha. Para la segunda ola se establecieron cuatro categorías según el número de días en exposición: 0, 1-30, 31-60 y 61 días. Al analizar la distribución de las variables se vio que la mayor parte de los individuos estuvieron expuestos en la ola de Alpha como no vacunados (97,9 %) y, en consecuencia, prácticamente ningún individuo pasó la ola con primovacunación completa. Esta distribución extrema en los grupos de vacunación llevó a desestimar el uso de la variable en el ajuste de los modelos, debido a su baja capacidad informativa. En el caso de la ola atribuida a Delta, la distribución de los días de exposición era un poco más heterogénea entre los dos grupos de vacunación, si bien se esperaba que pocos individuos pasaran esta ola sin vacunar. Con todo, al incluir los días de exposición en la ola de Delta en los modelos se obtenían de nuevo resultados poco coherentes. Como en el caso anterior, la variable tampoco permitía captar de manera adecuada el riesgo al que estaban sometidos los individuos, pues también tomaba valores bajos en personas con poco tiempo de seguimiento y que habían sido caso durante la ola.

Finalmente, ante las dificultades para medir la exposición al virus y dado que el tiempo de seguimiento de la cohorte fue de casi un año y las olas de Alpha y Delta tuvieron incidencias similares sobre los no vacunados y los primovacunados, respectivamente, se asumió que ambos grupos de vacunación estuvieron expuestos al mismo riesgo.

Las complicaciones encontradas a la hora de manejar el riesgo en el primer análisis condujeron al planteamiento del enfoque 2 en el segundo. Con este enfoque se buscó, precisamente, igualar el riesgo de exposición de los grupos de vacunación a comparar. Para ello, se restringió el inicio del seguimiento de la cohorte a la fecha de llegada de Ómicron. Con esto, todos los individuos fueron seguidos en el mismo momento de la pandemia asumiendo, por tanto, una exposición equivalente al virus. Sin embargo, los grupos de vacunación se comparaban ahora en distintos momentos tras la vacunación. Las personas con dosis de recuerdo comenzaban a seguirse, en su mayoría, desde la administración de la dosis, mientras que los primovacunados ya llevaban entre siete y 316 días con la pauta general completa. Esto se traducía en estados de inmunidad muy distintos para cada grupo de vacunación, lo que pudo introducir un sesgo en la estimación de la EV de la dosis de recuerdo.

- **Covariables de ajuste**

  Por otro lado, en buena parte de la bibliografía relacionada con la estimación de la efectividad de las vacunas del COVID-19 se considera como posible factor de confusión la presencia de comorbilidades, esto es, enfermedades susceptibles de incrementar el riesgo de infección por COVID-19 y/o de agravar sus síntomas. Algunos ejemplos son enfermedades renales y pulmonares crónicas, enfermedades cardiovasculares, diabetes tipo I y II u obesidad, entre otras. En nuestro caso, no fue posible acceder a este tipo de información para los individuos de la población de estudio.

  Otro factor importante que tampoco fue posible considerar en el ajuste de los modelos está

relacionado con la responsabilidad de cada individuo a la hora de cumplir con las obligaciones y recomendaciones dictadas por las autoridades sanitarias en el transcurso de la pandemia. Por ejemplo, es bien sabido que el uso de la mascarilla, especialmente en interiores, resultó ser una de las medidas más útiles para evitar el contagio. Sin embargo, no es posible diferenciar a los individuos de la cohorte que cumplieron con esta norma de los que no. Lo mismo ocurre con otros factores como el nivel de interacción social o el control del número de asistentes a reuniones entre no convivientes.

Líneas de trabajo futuras

- Analizar la efectividad de las vacunas frente a hospitalización, ahora que se empiezan a diferenciar en VIXÍA los ingresados por COVID-19 de los ingresados por otras causas con COVID-19. Otra posibilidad sería cruzar los casos de COVID-19 de VIXÍA con los datos del CMBD (Conjunto Mínimo Básico de Datos), que incluyen la información de todas las altas hospitalarias y sus diagnósticos principales y secundarios.

- Analizar la efectividad de la dosis de recuerdo frente a reinfección por COVID-19 en los casos, ahora que la cobertura de la dosis de refuerzo ha alcanzado altos porcentajes en todos los grupos de edad y que también hay más reinfecciones.

- Solicitar datos del registro poblacional de Tarxeta Sanitaria para identificar a los individuos no vacunados e incorporarlos a la población de estudio. Con esto, se completaría el colectivo de no vacunados, con lo que las estimaciones de la EV serían más ajustadas, y también se podría considerar como desenlace la defunción por COVID-19. Para este trabajo en particular, no fue posible disponer de estos datos.

- Aplicar el enfoque 2 del segundo análisis al primero por grupos de edad. Para cada grupo de edad se consideraría como fecha de inicio del seguimiento una fecha próxima a la vacunación masiva de ese colectivo. De esta forma, se podrían igualar los riesgos a los que estuvieron sometidos no vacunados y primovacunados (el seguimiento de los no vacunados se puede iniciar, digamos, en cualquier momento, ya que no presentan problemas de pérdida de inmunidad) y se obtendrían estimaciones más ajustadas de la EV y de su pérdida. No obstante, sería necesario explorar si para todos los grupos de edad existe un número suficiente de individuos no vacunados en el momento de su vacunación masiva. Para los de 80 años y más, en cierta medida, ya se aplicó este enfoque, pues comenzaron a vacunarse el 27 de diciembre de 2020, de modo que no vacunados y vacunados coincidieron en el tiempo.

Apéndice A

Estudios previos sobre efectividad vacunal frente al COVID-19

Referencia	**Diseño**	**Lugar**	**Vacunas**	**Desenlaces**	**Método estadístico**
Nasreen et al. (2022)	Casos y controles (test negativo)	Ontario	Moderna y Pfizer	Infección, ingreso y defunción	Regresión logística Regresión logística
Andrews et al. (2022)	Casos y controles (test negativo)	Reino Unido	AstraZeneca y Pfizer	Infección, ingreso y defunción	Regresión logística
Martínez-Baz et al. (2021)	Cohortes	Navarra	AstraZeneca y Pfizer	Infección e ingreso	Modelo de Cox
Nunes et al. (2021)	Cohortes	Portugal	Moderna y Pfizer	Ingreso y defunción	Modelo de Cox
Corchado-Garcia et al. (2021)	Cohortes	EE.UU.	Janssen	Infección	Modelo de Cox
Bedston et al. (2022)	Cohortes	Gales	Pfizer	Infección	Modelo de Cox
Nordström et al. (2022)	Cohortes	Suecia	Moderna, AstraZeneca y Pfizer	Infección, ingreso y defunción	Modelo de Cox

Tabla A.1: Síntesis de algunos estudios sobre efectividad vacunal y pérdida de inmunidad frente al COVID-19.

Bibliografía

Aalen, O. (1978). Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, *6*(4). https://doi.org/10.1214/aos/1176344247

Andrews, N., Tessier, E., Stowe, J., Gower, C., Kirsebom, F., Simmons, R., Gallagher, E., Thelwall, S., Groves, N., Dabrera, G., Myers, R., Campbell, C. N., Amirthalingam, G., Edmunds, M., Zambon, M., Brown, K., Hopkins, S., Chand, M., Ladhani, S. N., ... Lopez Bernal, J. (2022). Duration of Protection against Mild and Severe Disease by Covid-19 Vaccines. *New England Journal of Medicine*, *386*(4), 340–350. https://doi.org/10.1056/nejmoa2115481

Bedston, S., Akbari, A., Jarvis, C. I., Lowthian, E., Torabi, F., North, L., Lyons, J., Perry, M., Griffiths, L. J., Owen, R. K., Beggs, J., Chuter, A., Bradley, D. T., de Lusignan, S., Fry, R., Richard Hobbs, F., Hollinghurst, J., Katikireddi, S. V., Murphy, S., ... Lyons, R. A. (2022). COVID-19 vaccine uptake, effectiveness, and waning in 82,959 health care workers: A national prospective cohort study in Wales. *Vaccine*, *40*(8), 1180–1189. https://doi.org/10.1016/j.vaccine.2021.11.061

Breslow, N. E. (1975). Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review / Revue Internationale de Statistique*, *43*(1), 45. https://doi.org/10.2307/1402659

Corchado-Garcia, J., Zemmour, D., Hughes, T., Bandi, H., Cristea-Platon, T., Lenehan, P., Pawlowski, C., Bade, S., O'Horo, J. C., Gores, G. J., Williams, A. W., Badley, A. D., Halamka, J., Virk, A., Swift, M. D., Wagner, T., & Soundararajan, V. (2021). Analysis of the Effectiveness of the Ad26.COV2.S Adenoviral Vector Vaccine for Preventing COVID-19. *JAMA Network Open*, *4*(11), e2132540. https://doi.org/10.1001/jamanetworkopen.2021.32540

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Delbarre, A., Amor, B., Bardoulat, I., Tetafort, A., & Pelletier-Fleury, N. (2017). Do intra-articular hyaluronic acid injections delay total knee replacement in patients with osteoarthritis – A Cox model analysis. *PLOS ONE*, *12*(11), e0187227. https://doi.org/10.1371/journal.pone.0187227

de Uña-Álvarez, J., Moreira, C., & Crujeiras, R. M. (2022). *The Statistical Analysis of Doubly Truncated Data With Applications in R*. Willey.

Dirección Xeral de Saúde Pública. (2021). *Plan galego de vacinación fronte ao SARS-CoV-2* (tech. rep. No. 6.3). https://coronavirus.sergas.gal/Contidos/Documents/983/pgvcovid63.pdf

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, *68*(6), 652–665. https://doi.org/10.1057/s41274-016-0128-9

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, *72*(359), 557–565. https://doi.org/10.1080/01621459.1977.10480613

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526. https://doi.org/10.1093/biomet/81.3.515

Grant, S., Chen, Y. Q., & May, S. (2013). Performance of goodness-of-fit tests for the Cox proportional hazards model with time-varying covariates. *Lifetime Data Analysis*, *20*(3), 355–368. https://doi.org/10.1007/s10985-013-9277-1

Hernández-Aguado, I., & Lacarra, L. B. (2018). *Manual De Epidemiología Y Salud Pública Para Grados En Ciencias De La Salud (Incluye Acceso A Ebook)* (3rd ed.). Editorial Medica Panamericana S.A. de C.V.

Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2020). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*(3), 141–154. https://doi.org/10.1038/s41579-020-00459-7

Huang, Z., Zhan, X., Xiang, S., Johnson, T. S., Helm, B., Yu, C. Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., & Huang, K. (2019). SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00166

IGE - Instituto Galego de Estatística. (2021). https://www.ige.gal/web/index.jsp?paxina=001&idioma=gl

Iglesias Pérez, M. C., & de Uña-Álvarez, J. (2021). Análisis de Supervivencia. Apuntes del Máster en Técnicas Estadísticas.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1). https://doi.org/10.1186/s12874-018-0482-1

Klein, J., & Moeschberger, M. (2005). *Survivial Analysis: Techniques for Censored and Truncated Data* (2nd ed.). Springer.

Kleinbaum, D., & Klein, M. (2011). *Survival Analysis: A Self-Learning Text* (3rd 2012 ed.). Springer.

Martínez-Baz, I., Miqueleiz, A., Casado, I., Navascués, A., Trobajo-Sanmartín, C., Burgui, C., Guevara, M., Ezpeleta, C., & Castilla, J. (2021). Effectiveness of COVID-19 vaccines in preventing SARS-CoV-2 infection and hospitalisation, Navarre, Spain, January to April 2021. *Eurosurveillance*, *26*(21). https://doi.org/10.2807/1560-7917.es.2021.26.21.2100438

Ministerio de Sanidad. (2021). Estrategia de detección precoz, vigilancia y control de COVID-19. https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/COVID19_Estrategia_vigilancia_y_control_e_indicadores.pdf

Ministerio de Sanidad. (2022). Estrategia de vigilancia y control frente a COVID-19 tras la fase aguda de la pandemia. https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Nueva_estrategia_vigilancia_y_control.pdf

Moreira, C., de Uña-Álvarez, J., & Crujeiras, R. M. (2021). DTDA: Doubly Truncated Data Analysis. https://cran.r-project.org/web/packages/DTDA/index.html

Nabizadeh, A., Tabatabai, H., & Tabatabai, M. (2018). Survival Analysis of Bridge Superstructures in Wisconsin. *Applied Sciences*, *8*(11), 2079. https://doi.org/10.3390/app8112079

Nasreen, S., Chung, H., He, S., Brown, K. A., Gubbay, J. B., Buchan, S. A., Fell, D. B., Austin, P. C., Schwartz, K. L., Sundaram, M. E., Calzavara, A., Chen, B., Tadrous, M., Wilson, K., Wilson, S. E., & Kwong, J. C. (2022). Effectiveness of COVID-19 vaccines against symptomatic SARS-CoV-2 infection and severe outcomes with variants of concern in Ontario. *Nature Microbiology*, *7*(3), 379–385. https://doi.org/10.1038/s41564-021-01053-0

Nelson, W. (1969). Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, *1*(1), 27–52. https://doi.org/10.1080/00224065.1969.11980344

Nordström, P., Ballin, M., & Nordström, A. (2022). Risk of infection, hospitalisation, and death up to 9 months after a second dose of COVID-19 vaccine: a retrospective, total population cohort study in Sweden. *The Lancet*, *399*(10327), 814–823. https://doi.org/10.1016/s0140-6736(22)00089-7

Nunes, B., Rodrigues, A. P., Kislaya, I., Cruz, C., Peralta-Santos, A., Lima, J., Pinto Leite, P., Sequeira, D., Matias Dias, C., & Machado, A. (2021). mRNA vaccine effectiveness against COVID-19-related hospitalisations and deaths in older adults: a cohort study based on data

linkage of national health registries in Portugal, February to August 2021. *Eurosurveillance*, *26*(38). https://doi.org/10.2807/1560-7917.es.2021.26.38.2100833

Pardo-Seco, J., Mallah, N., López-Pérez, L. R., González-Pérez, J. M., Rosón, B., Otero-Barrós, M. T., Durán-Parrondo, C., Rodríguez-Tenreiro, C., Rivero-Calle, I., Gómez-Carballa, A., Salas, A., & Martinón-Torres, F. (2022). Evaluation of BNT162b2 Vaccine Effectiveness in Galicia, Northwest Spain. *International Journal of Environmental Research and Public Health*, *19*(7), 4039. https://doi.org/10.3390/ijerph19074039

Pettitt, A. N., & Daud, I. B. (1989). Case-Weighted Measures of Influence for Proportional Hazards Regression. *Applied Statistics*, *38*(1), 51. https://doi.org/10.2307/2347680

Stata 16.0. (2019). https://www.stata.com/

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160. https://doi.org/10.1093/biomet/77.1.147

Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J., & Prescott, H. C. (2020). Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19). *JAMA*, *324*(8), 782. https://doi.org/10.1001/jama.2020.12839

Xue, Y., & Schifano, E. D. (2017). Diagnostics for the Cox model. *Communications for Statistical Applications and Methods*, *24*(6), 583–604. https://doi.org/10.29220/csam.2017.24.6.583

Zuckerman, N., Pando, R., Bucris, E., Drori, Y., Lustig, Y., Erster, O., Mor, O., Mendelson, E., & Mandelboim, M. (2020). Comprehensive Analyses of SARS-CoV-2 Transmission in a Public Health Virology Laboratory. *Viruses*, *12*(8), 854. https://doi.org/10.3390/v12080854