Grado en Estadística

TRABAJO FIN DE GRADO

Investigación Epidemiológica: Estudios de caso-control y estudios de cohortes

María González Alba

Sevilla, Septiembre de 2021

Índice general

Resumen

El objetivo principal de este trabajo es ilustrar cómo tratar los estudios de cohortes y de caso-control en el software estadístico R.

En el primer capítulo se hace una introducción a la epidemiología, a la clasificación de los tipos de estudios y a las principales medidas que se aplican en ellos.

En el segundo capítulo se profundiza teóricamente en los estudios de cohortes, en las etapas que lo componen y, por último, se reproduce el código de R y se extraen las conclusiones de un estudio de cohortes retrospectivo sobre los factores de riesgo asociados a la gravedad y mortalidad de la enfermedad de COVID-19 en Kazajistán.

En el tercer capítulo, se profundiza en los estudios de caso-control y se tratan temas como la selección de los miembros del estudio, además de conocer cómo se lleva a cabo la aplicación de técnicas estadísticas para el tratamiento de estudios de este tipo reproduciendo el código R y extrayendo conclusiones de un estudio de caso-control sobre la búsqueda de factores de riesgo asociados a la infección por Escherichia Coli en un pueblo de Tasmania.

En el último capítulo se comparan ambos tipos de estudios y además se dan a conocer algunos estudios híbridos.

En el apéndice se incluye código e interpretaciones para estudios complementarios al estudio de cohortes retrospectivo hecho en R en el Capítulo 2.

Abstract

The aim of this work is to illustrate how to deal with cohort and case-control studies in R statistical software.

The first chapter introduces epidemiology, the classification of types of studies, and the main measures used on them.

In the second chapter, there is an in-depth theoretical study of cohort studies, finally, the R-code is reproduced and the conclusions of a retrospective cohort study on the risk factors associated with the severity and mortality of COVID-19 disease in Kazakhstan are drawn.

In the third chapter, the case-control studies are discussed in depth and topics such as the selection of study members, in addition to knowing how the application of statistical techniques for the treatment of studies of this type is carried out by reproducing the R code and drawing conclusions from a case-control study over the search for risk factors associated with Escherichia Coli infection in a village in Tasmania.

In the last chapter, both types of studies are compared and some hybrid studies are also released.

The appendix includes code and interpretations for studies complementary to the retrospective cohort study done in R in Chapter 2.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción a la epidemiología

1.1. Definición y características principales

La epidemiología es la ciencia encargada de estudiar la frecuencia y distribución de los fenómenos relacionados con la salud en la población humana, así como las causas que los provocan. La definición hace evidente la importancia de la estadística como medio para conseguir el propósito final de esta.

La estadística ha jugado históricamente un papel fundamental en el campo de la epidemiología, hasta el punto de que algunos de los métodos estadísticos empleados en la actualidad han surgido por la necesidad de dar respuesta a preguntas de investigación generadas en este área.

En el desarrollo de este capítulo se han utilizado como fuentes bibliográficas, fundamentalmente el trabajo de "Método Epidemiológico" (Royo-Bordonada et al., 2009) y el libro "HandBook of Epidemiology" (Ahrens & Pigeot, 2014).

La resolución de problemas en el ámbito de la epidemiología se da gracias al planteamiento del estudio adecuado, el análisis de datos y la comunicación clara y eficiente de resultados a través de técnicas estadísticas. Es por ello, que la estadística tiene un fuerte papel en el desarrollo y en los avances que se producen en la epidemiología.

La epidemiología no solo estudia enfermedades, sino todo tipo de fenómenos relacionados con la salud, entre ellos, accidentes, suicidios, hábitos de vida, etc. Los determinantes de estos fenómenos son los factores físicos, biológicos, sociales, culturales y de comportamiento que influyen sobre la salud.

En realidad, la característica que más acentúa las diferencias entre la epidemiología y otras disciplinas biológicas es el estudio de la frecuencia de fenómenos en poblaciones. Esta ciencia se

encarga de estudiar relaciones probabilísticas o estocásticas, mientras que las otras tratan las relaciones deterministas o no estocásticas. En la Tabla 1.1 se pueden ver las diferencias entre los efectos probabilísticos y deterministas (Royo-Bordonada et al., 2009).

Tabla 1.1: Diferencias entre los efectos probabilísticos y deterministas. Fuente: (Royo-Bordonada et al., 2009).

Efectos **estocásticos**	Efectos **no estocásticos**
- Probabilidad del efecto **depende** de la ***dosis de exposición***.	- Probabilidad del efecto **no depende** de la ***dosis de exposición***. La probabilidad es 0 ó 1.
- Intensidad del efecto **no depende** de la ***dosis de exposición***.	- Intensidad del efecto **depende** de la ***dosis de exposición***.
- Efecto **sin umbral** de ***dosis de exposición***.	- Efecto **con umbral** de ***dosis de exposición***.
- Tiempo de inducción **largo**.	- Tiempo de inducción **corto**.

Los factores etiológicos, es decir, aquellos que están presentes antes de que la enfermedad aparezca, aumentan la probabilidad de desarrollar una enfermedad o no (efecto probabilístico), pero no aseguran su aparición. Para poder medir el efecto se necesita medir la frecuencia con la que se produce la enfermedad en personas que presentan estos antecedentes y en personas que no los padecen (como se hace, por ejemplo, en los estudios de casos y controles). Por otra parte, el factor causal es suficiente para poder producir la enfermedad (efecto determinista), por tanto, solo hace falta estudiar a un único sujeto para ver si se manifiesta o no. Es por esto por lo que el tiempo de inducción es diferente en los experimentos en los que se estudian relaciones probabilísticas y deterministas.

Según Henquin (2013), los objetivos de la epidemiología son, entre otros:

- Identificar las causas de una enfermedad y sus factores de riesgo.
- Determinar la extensión de la enfermedad en la población.
- Estudiar la historia de la enfermedad y su pronóstico.
- Evaluar medidas preventivas y terapéuticas.

- Proporcionar fundamento para el desarrollo de políticas de salud.

Estos objetivos se resumen en lo que hoy se conoce como investigación clínica que va, también, ligada a la estadística. Es importante tener en cuenta que cuando se accede a un estudio de este área y se quiere valorar su validez hay que conocer la metodología epidemiológica empleada, así como hacer énfasis en la importancia del diseño del estudio, el método de recolección de datos y, por supuesto, el análisis de estos.

La información que se necesita para cumplir con los objetivos de esta ciencia se recoge de las investigaciones realizadas en individuos o, más frecuentemente, en grupos poblacionales. Sin embargo, nunca se podría abarcar un estudio poblacional completo ya que sería muy costoso tanto a nivel económico como temporal, por eso es importante desarrollar estrategias muestrales y de medición que permitan estudiar subgrupos de la población y extrapolar los resultados a la población completa. Reconocer la importancia que tiene la metodología en el avance del conocimiento epidemiológico ha hecho que se asuma como otro objetivo el desarrollo y estudio de nuevos métodos de aplicación en este área. Esto ha provocado una mejora de la calidad y la validez de los resultados obtenidos y de los conocimientos generados a través de esta ciencia (Hernández-Ávila et al., 2000).

Es muy probable que siempre que el diseño no sea el adecuado o el modo en el que se obtuvieron los datos no fuera el indicado, el estudio deje de tener validez ya que además de esto hay que considerar que en todos los estudios e investigaciones hay que contar de antemano con la presencia de sesgo.

Gracias a la epidemiología y a las investigaciones que promueve se conoce información de las enfermedades bajo estudio. Estos datos, que se recogen tras el proceso descrito anteriormente, pueden clasificarse en tres campos, según Royo-Bordonada et al. (2009):

- Historia natural de las enfermedades bajo estudio.
- Planificación sanitaria (pasos a seguir para conocer y controlar los problemas de salud en poblaciones).
- Mejora del proceso de toma de decisiones clínicas para mejorar la salud de los enfermos.

Dicha información, generada a posteriori, ayuda a cumplir el objetivo principal de la epidemiología, que es el control de los problemas de salud o enfermedades, del que forman parte el resto de objetivos mencionados anteriormente.

Al igual que el resto de ciencias, la epidemiología también posee un razonamiento metodológico a la hora de plantear y llevar a cabo los estudios. Este método consta de tres pasos que

mencionamos ahora pero serán desarrollados con posterioridad en el trabajo:

- Asociación entre el factor de exposición y el evento.
- Formulación de una hipótesis sobre el mecanismo de la enfermedad.
- Aplicación del diseño adecuado para confirmar la hipótesis.

1.2. Factor de exposición vs evento

Uno de los objetivos mencionados es enfocar el estudio en busca de la relación o asociación entre el factor de exposición y la enfermedad o evento (Kestenbaum, 2009). El factor de exposición es el factor de riesgo que estamos estudiando y que puede provocar el desarrollo del evento (Ecuación 1.1). El evento no tiene por qué ser siempre una enfermedad, puede ser una característica clínica como el nivel de colesterol LDL en sangre o el estado de vacunación, entre otros (Kestenbaum, 2009). Por otro lado, un factor de exposición no tiene porqué ser perjudicial en todos los casos, también puede ser beneficioso, es por esto que siempre se habla de factor de exposición y no de riesgo para no connotar un significado dañino (Ahrens & Pigeot, 2014). En el Ejemplo 1.1 (Kestenbaum, 2009) se explican a través de un caso estos nuevos conceptos.

$$\text{Exposición} \overset{\text{Asociación}}{\longrightarrow} \text{Evento} \tag{1.1}$$

■ **Ejemplo 1.1** Se realiza un estudio para ver si la insuficiencia cardíaca influye en la supervivencia después de un primer infarto de miocardio. Se estudian 1000 personas que han sobrevivido a su primer infarto de miocardio y se les somete a un examen físico y a una ecocardiografía para ver si existe o no insuficiencia cardíaca. Las personas continúan bajo estudio hasta que mueren o dejan el estudio (por diversas causas). En este caso, el objetivo del estudio es:

$$\text{Insuficiencia cardíaca} \overset{\text{Asociación}}{\longrightarrow} \text{Supervivencia}$$

donde:

$$\text{Factor de exposición} \longrightarrow \text{Insuficiencia cardíaca}$$

$$\text{Evento} \longrightarrow \text{Supervivencia}$$

En este caso, el factor de exposición es también una enfermedad. Existen otros casos en los que se ha investigado la asociación de factores de riesgo con la insuficiencia cardíaca siendo esta, ahora, el evento. ■

La asociación de un factor de exposición y un evento implica que cuando uno de los dos está presente, la probabilidad de que el otro también lo esté es muy alta. Esto ayuda a determinar

con cierta fiabilidad la existencia de una enfermedad y a cumplir con otro de los objetivos de la epidemiología, implementar políticas de salud destinadas a corregir estos factores de exposición y reducir la frecuencia de la aparición de las enfermedades asociadas.

1.3. Estudios y diseños epidemiológicos

Para seguir cumpliendo con los objetivos propuestos por la epidemiología surgen los estudios epidemiológicos. Cada estudio tiene asociado uno o más diseños diferentes en función de la hipótesis bajo estudio. La existencia de estudios y diseños epidemiológicos establecidos de antemano ayudan a prever los posibles sesgos y limitaciones que pueden aparecer en función del tipo de estudio y, como se mencionó anteriormente, asegurar la validez y fiabilidad de este.

Existen muchas clasificaciones de tipos de estudios, la mayoría de ellos se basan en criterios como la intervención sobre las características de los individuos, el modo de recolección de datos, el seguimiento de la población o el análisis de la información. Se puede hacer una primera distinción entre dos tipos de estudios que son los más importantes en este área. Los estudios experimentales y los observacionales.

En los estudios experimentales el investigador controla la asignación de la exposición a los individuos que están siendo estudiados. En este caso, la población bajo estudio se divide en dos grupos, el grupo al que se le asigna el factor de exposición y el grupo al que no se le hace ningún cambio (Henquin, 2013). Generalmente, la exposición se asigna a los individuos de manera aleatoria (Ahrens & Pigeot, 2014).

Los estudios observacionales son aquellos en los que el investigador no controla la asignación de la exposición, se limita a observarla, a conocer el comportamiento de los factores que aparecen o no en los individuos de manera independiente al estudio.

Dentro de los estudios observacionales podemos hacer otra clasificación, los estudios descriptivos y los estudios analíticos.

Los estudios observacionales descriptivos son los que se encargan de observar y describir la realidad de un factor que está presente en la población. A menudo se usan datos que se han recogido de manera rutinaria (Ahrens & Pigeot, 2014), aunque también puede plantearse la recolección de datos desde cero para este tipo de estudio (Royo-Bordonada et al., 2009). Debido a su carácter representativo de la realidad juega un papel importante en la gestión y propuestas de hipótesis epidemiológicas.

Sobre los estudios observacionales analíticos podemos decir que son aquellos en los que la

población bajo estudio está seleccionada debido a una característica que presenta y que es de interés para el investigador (Henquin, 2013), además están orientados a evaluar una hipótesis. En este caso, los datos se recopilan con el propósito de estudiarlos para esta causa (Ahrens & Pigeot, 2014).

Cabe mencionar que los estudios experimentales son todos de carácter analítico.

Para resumir y mejorar el entendimiento de lo introducido hasta ahora sobre esta sección, la Figura 1.1 muestra de forma esquemática los criterios de clasificación de los estudios epidemiológicos y los diseños asociados a estos.

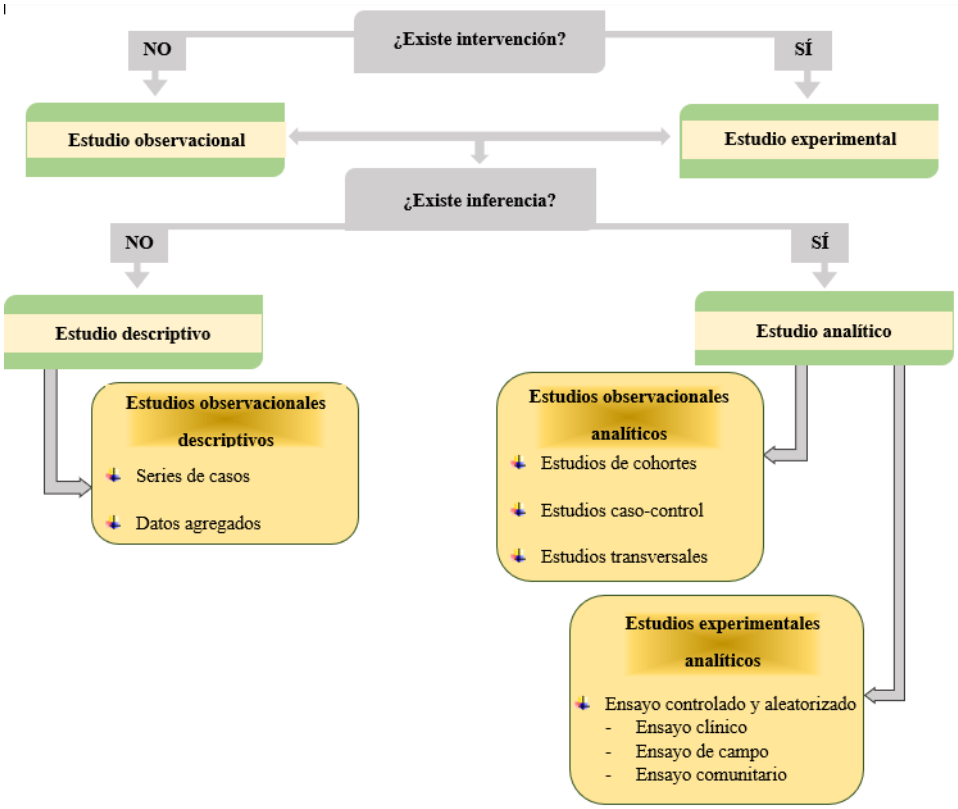


Figura 1.1: Clasificación de los estudios epidemiológicos.

A lo largo de este trabajo nos centraremos en los estudios observacionales analíticos, concretamente en los estudios de cohortes y de caso-control en los que se profundizará en los siguientes capítulos.

1.4. Medidas

Con el propósito de determinar la extensión, cuantificar la aparición y el impacto de la enfermedad en la población, nacen las medidas de frecuencia y de efecto. Tienen un papel muy importante en el ámbito epidemiológico ya que, ayudan a hacer comparaciones entre poblaciones de distinta raza, sexo, edad o localización, además de entre individuos que presentan (o están sometidos) o no el factor de exposición.

1.4.1. Medidas de frecuencia

Surgen para determinar la frecuencia de la enfermedad o evento de interés (Olsen et al., 2010). La forma más sencilla, aunque no la más útil, de medir la frecuencia de un evento o factor de exposición es contar el número de casos. Según la consideración del tiempo, existen dos tipos de medidas: las medidas de incidencia y las medidas de prevalencia.

1.4.1.1. Incidencia

La incidencia mide el número de nuevos casos que aparecen en la población (Kestenbaum, 2009). Se consideran dos tipos de medidas de incidencia: la incidencia acumulada (Ecuación 1.2) y la tasa de incidencia (Ecuación 1.3).

- Incidencia acumulada (IA) (Kestenbaum, 2009):

$$\text{IA} = \frac{\text{número de casos nuevos}}{\text{número de personas sanas al inicio } (t_0)} \tag{1.2}$$

Esta medida indica la proporción de personas que han enfermado en el período de tiempo del estudio. La incidencia acumulada se mide sobre una población fija, que no permite entradas durante el tiempo de observación.

- Tasa de incidencia (I) (Kestenbaum, 2009):

$$\text{I} = \frac{\text{número de casos nuevos}}{\text{número de personas x tiempo de observación}} \tag{1.3}$$

Se mide sobre poblaciones dinámicas, es decir, poblaciones que sí permiten la entrada de nuevos miembros al estudio. En la Ecuación 1.3 el denominador se obtiene sumando los tiempos en observación en que cada sujeto está en riesgo de que le aparezca la enfermedad o evento bajo estudio. La tasa de incidencia se mide en las unidades de tiempo inverso ($\frac{1}{\text{tiempo}}$ o tiempo^{-1}). Por tanto, el inverso de la tasa de incidencia se interpreta como el tiempo medio hasta la aparición de la enfermedad.

A continuación, se muestra cómo calcular ambas medidas de incidencia de la infección de la gripe (Ejemplo 1.2) (Kestenbaum, 2009):

■ **Ejemplo 1.2** La población bajo estudio son 500 estudiantes de medicina y se comienza en enero del año 2000 (t_0). De enero (t_0) a marzo del mismo año se desarrollan 5 nuevos casos de gripe en la población bajo estudio.

- Incidencia acumulada:

$$\text{IA} = \frac{5 \text{ nuevos casos}}{500 \text{ personas}} \text{x} 100\,\% = 1\,\%$$

1 de cada 100 personas desarrolla la enfermedad de la gripe.

- Tasa de incidencia:

$$\text{I} = \frac{5 \text{ nuevos casos}}{500 \text{ personas x } 3 \text{ meses}} = \frac{5 \text{ nuevos casos}}{1500 \text{ personas x meses}} = 0.003333333333 \text{ meses}^{-1}$$

En esta ocasión, como es una población cerrada y no se han reportado salidas del estudio, todas las personas están bajo estudio 3 meses, por eso hacemos el cálculo 500 personas x 3 meses = 1500 personas x meses.

■

En el Ejemplo 1.3 y en la Figura 1.2 (Olsen et al., 2010) se expone otro caso de cálculo de la tasa de incidencia en una población dinámica:

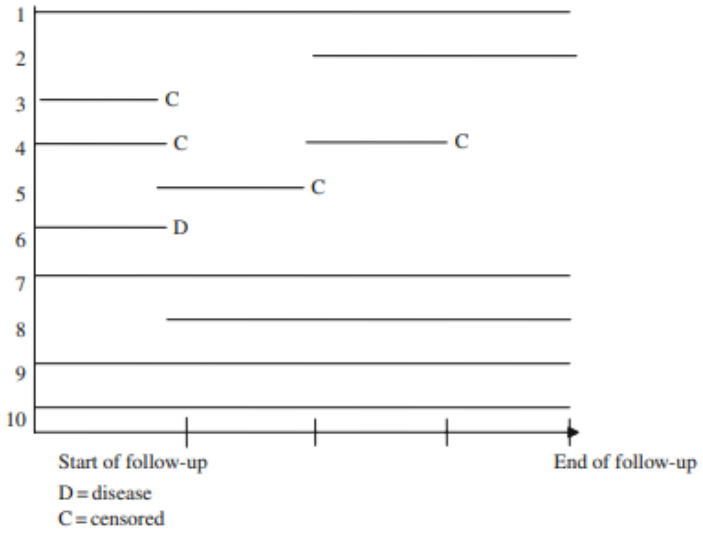


Figura 1.2: Gráfico para el cálculo de la tasa de incidencias para el Ejemplo 1.3. Fuente: (Olsen et al., 2010).

■ **Ejemplo 1.3** En el gráfico vemos representadas las entradas y salidas de los individuos bajo estudio, así como la razón de abandono del estudio siendo:

- D = aparece la enfermedad
- C = abandono censurado

En el eje de ordenadas aparecen los individuos bajo estudio que son, en total, 10 personas. El eje de abscisas de la gráfica se corresponde con el tiempo de observación, que en este caso son 2 años, este eje está dividido en 4 partes iguales, correspondientes a unos 6 meses cada una (0.5 años). La tasa de incidencia será:

$$\mathrm{I} = \frac{1 \text{ nuevo caso}}{(2 + 1 + 0.5 + 1 + 0.5 + 0.5 + 2 + 1.5 + 2 + 2) \text{ personas x años}} = \frac{1}{13} = 0.077 \text{ años}^{-1}$$

■

1.4.1.2. Prevalencia

La prevalencia (Ecuación 1.4) se encarga de medir la proporción de individuos enfermos en un período (Kestenbaum, 2009).

$$\text{Prevalencia} = \frac{\text{número de personas enfermas}}{\text{total de personas en la población}} \tag{1.4}$$

Se puede expresar como la probabilidad de estar enfermo en un determinado momento. Esta medida solo hace referencia a un momento concreto o a un período de tiempo limitado, es una medida estática y no se puede ampliar su uso más allá del período especificado.

En el Ejemplo 1.4 (Henquin, 2013) se calcula la prevalencia de la miopía en individuos de entre 50 y 80 años.

■ **Ejemplo 1.4** Se realiza una encuesta a 2700 personas de entre 50 y 80 años durante un mes. Se encuentran 310 casos de miopía entre los encuestados. Por tanto, la prevalencia es:

$$\text{Prevalencia} = \frac{310 \text{ casos}}{2700 \text{ encuestados}} = 0.1148$$

La prevalencia de la miopía es del 11.5 % en esta población durante el mes en el que se realizó la encuesta. ■

1.4.2. Medidas de efecto

Estas medidas sirven para valorar el efecto que tiene la exposición a un determinado factor de riesgo con la aparición de la enfermedad bajo estudio. Debido a su definición se aplican en estudios en los que se compara una población expuesta con otra que no lo está, ya que, si se diera el caso de que son poblaciones completamente iguales salvo en la exposición, al encontrar alguna diferencia entre ellas se va a atribuir directamente al factor de exposición.

1.4.2.1. Riesgo relativo

El riesgo relativo (RR) estima la frecuencia del efecto en el grupo de expuestos en relación al de no expuestos. Esta medida se usa en los estudios de cohortes y en los estudios experimentales. Es el cociente de la incidencia en el grupo expuesto I_1 y en el grupo que no lo está I_0 (Ecuación 1.5) (Mirón Canelo & Alonso Sardón, 2008).

$$RR = \frac{\text{Incidencia en expuestos}}{\text{Incidencia en no expuestos}} = \frac{I_1}{I_0} \tag{1.5}$$

Toma valores entre 0 y 1.

- Si RR $> 1 \longrightarrow$ Indica que el factor de exposición está asociado a la enfermedad.
- Si RR $= 1 \longrightarrow$ Indica que no existe asociación entre la enfermedad y el factor de exposición.
- Si RR $< 1 \longrightarrow$ Indica que el factor de exposición disminuye el riesgo de aparición de la enfermedad. El factor de exposición protege al individuo de contraer la enfermedad.

1.4.2.2. Odds ratio

Es la medida más usada en los estudios de casos y controles porque, como veremos más adelante, la población de casos en este diseño se selecciona a partir de individuos que ya han desarrollado la enfermedad y, por tanto, no se puede calcular la incidencia.

Odds (Ecuación 1.6) es la razón entre una proporción, p, y su complementaria, 1 - p, es decir, el cociente del número de veces que un evento pudo ocurrir y el número de veces que un evento no pudo ocurrir.

$$\text{Odds} = \frac{\text{p}}{\text{1 - p}} \tag{1.6}$$

Por consiguiente, la Odds Ratio (OR) es la razón entre la odds en los expuestos y la odds en los no expuestos.

Tabla 1.2: Tabla para la definición de la Odds Ratio. Fuente: (Moreno-Altamirano et al., 2000).

	Grupo A (casos)	Grupo B (controles)
Expuestos	a	b
No expuestos	c	d

Teniendo en cuenta la Tabla 1.2 (Moreno-Altamirano et al., 2000) la Odds Ratio se calcularía de la siguiente forma (Ecuación 1.7):

$$\mathrm{OR} = \frac{\frac{\mathrm{a}}{\mathrm{b}}}{\frac{\mathrm{c}}{\mathrm{d}}} = \frac{\mathrm{a} \cdot \mathrm{d}}{\mathrm{b} \cdot \mathrm{c}} \tag{1.7}$$

La interpretación de la Odds Ratio es similar a la que se hace del riesgo relativo.

Con el Ejemplo 1.5 (Mirón Canelo & Alonso Sardón, 2008) se hace referencia al cálculo de estas dos últimas medidas introducidas.

■ **Ejemplo 1.5** Se quiere comparar el riesgo de infección por Hepatitis-B en los sanitarios (expuestos) frente a otro tipo de personal (no expuesto). Lo haremos sobre dos grupos: Grupo A (infectados) y Grupo B (no infectados). La Tabla 1.3 con los datos es la siguiente:

Tabla 1.3: Tabla para el Ejemplo 1.5. Fuente: (Mirón Canelo & Alonso Sardón, 2008).

	Grupo A (infectados)	Grupo B (no infectados)	
Expuestos (sanitarios)	a = 40	b = 3960	$n_1 = 4000$
No expuestos (otros)	c = 2	d = 998	$n_0 = 1000$
	$m_1 = 42$	$m_0 = 4958$	$N = 5000$

Siendo:

- $n_1 \longrightarrow$ Total de individuos expuestos (sanitarios) en el estudio.
- $n_0 \longrightarrow$ Total de individuos no expuestos (otros) en el estudio.
- $m_1 \longrightarrow$ Total de individuos infectados.
- $m_0 \longrightarrow$ Total de individuos no infectados.

Vamos a calcular el riesgo relativo (RR). Por la Ecuación 1.5 sabemos que para el cálculo de esta medida, primero hay que hacer el cálculo de la incidencia en los expuestos y en los no expuestos.

- Incidencia en expuestos $= \frac{40 \text{ nuevos casos}}{4000 \text{ expuestos en total}} = \frac{40}{4000}$

- Incidencia en no expuestos $= \frac{\text{2 nuevos casos}}{\text{1000 no expuestos en total}} = \frac{2}{1000}$
- Por tanto, el riesgo relativo es:

$$\text{RR} = \frac{\frac{40}{4000}}{\frac{2}{1000}} = 5$$

Por tanto, la infección por hepatitis es 5 veces mayor en el grupo de los expuestos que en el otro. Como se vio en la definición del RR, al ser mayor que 1 indica que la exposición a la infección (ser sanitario) está asociada fuertemente a la aparición de esta misma.

Ahora vamos a calcular la Odds Ratio. Por la definición de esta medida (Ecuación 1.7) sabemos que es el cociente de la Odds de los expuestos y de la Odds de los no expuestos.

- Odds de expuestos $= \frac{\text{Proporción de contagiados}}{\text{Proporción de no contagiados}} = \frac{\frac{40}{4000}}{\frac{3960}{4000}} = \frac{\text{a} = 40}{\text{b} = 3960} = \frac{40}{3960}$
- Odds de no expuestos $= \frac{\frac{2}{1000}}{\frac{998}{1000}} = \frac{\text{c} = 2}{\text{d} = 998} = \frac{2}{998}$
- Por tanto, la Odds Ratio es:

$$\text{OR} = \frac{\frac{\text{a} = 40}{\text{b} = 3960}}{\frac{\text{c} = 2}{\text{d} = 998}} = \frac{\text{a} \cdot \text{d}}{\text{b} \cdot \text{c}} = 5.04$$

Como vemos, es muy similar al riesgo relativo que hemos obtenido.

■

A lo largo de los siguiente capítulos se irá profundizando en los estudios de cohortes y de caso-control, además se añadirán nuevas medidas específicas para estos estudios si son necesarias y se expondrán más ejemplos aplicados a los diseños ya mencionados.

Capítulo 2

Estudios de cohortes

Los estudios de cohortes son estudios observacionales analíticos que se caracterizan por la identificación y seguimiento de una o más cohortes con un determinado nivel de exposición, para detectar y cuantificar la aparición del evento o enfermedad de interés a lo largo del tiempo (Royo-Bordonada et al., 2009).

Una cohorte es un grupo de personas pertenecientes a la población bajo estudio que comparten una condición en común y cuyo resultado final es desconocido al principio del estudio (Kestenbaum, 2009).

La esencia de los estudios de cohortes es el seguimiento (Royo-Bordonada et al., 2009), esto significa que son estudios longitudinales y, además, la cohorte bajo estudio es observada varias veces en el tiempo en el que se realiza el estudio. El seguimiento sirve para medir la frecuencia (Incidencia acumulada, ecuación 1.2) y el ritmo (Tasa de incidencia, ecuación 1.3) de la enfermedad de interés. También cabe añadir que durante este período, se puede modificar el nivel de exposición en la/las cohortes.

Los estudios de cohortes ponen su foco en el evento de interés que ocurre durante el seguimiento. Para cumplir con este objetivo, los investigadores suelen excluir de la cohorte a las personas que ya tienen la enfermedad de antemano. Estos estudios ayudan fundamentalmente a buscar la relación entre el factor de exposición y el evento de interés. Son capaces de colaborar en la verificación de la asociación de la exposición y el evento de interés a través del tiempo. A veces, el factor de exposición se encuentra presente en la población bajo estudio de antemano, y en otras ocasiones no y habría que exponer a los miembros a este factor (nos encontraríamos en un estudio experimental), sin embargo, en ambos casos, se podría acabar determinando un grado de relación entre el factor y el evento bajo estudio.

Cabe señalar que si centramos el estudio en observar y obtener conclusiones de una sola cohorte, que sería la de individuos que presentan el factor de exposición, estaríamos ante un estudio de cohortes descriptivo (Henquin, 2013) ya que no habría ningún planteamiento de una hipótesis que relacione el factor y el evento bajo estudio. Al igual que en los estudios analíticos también se usaría la incidencia (Ecuación 1.2) como principal medida, puesto que cuantificaremos la frecuencia del evento entre los individuos de la cohorte que presenta el factor de exposición.

2.1. Clasificación de los estudios de cohortes y tipos de cohortes

Los estudios de cohortes se pueden clasificar según el tiempo y el momento de observación en tres tipos según Gallego Iborra et al. (2012):

- Estudios prospectivos: son los que se han comentado anteriormente. En ellos, los individuos no presentan el evento de interés al comienzo del estudio y, además, sus datos se recogen durante el período de seguimiento y por tanto se sitúan cronológicamente dentro de este. Este tipo de estudio aparece esquematizado en la figura 2.1.

Figura 2.1: Esquema de los estudios prospectivos.

- Estudios retrospectivos: también son conocidos como estudios históricos. Su principal característica es que el estudio se comienza a posteriori de la recolección de datos. En este caso, el período de seguimiento no se sitúa en el tiempo a la vez que el período en el que se hace el estudio (Figura 2.2). Se trata de reconstruir el seguimiento de los miembros de la cohorte a través de la información que se recoge de registros históricos (Royo-Bordonada et al., 2009), por tanto, si los datos no son de calidad, la validez del estudio se verá limitada.

Figura 2.2: Esquema de los estudios retrospectivos.

- Estudios ambispectivos: es la combinación de los dos anteriores. En el comienzo del estudio a algunos individuos de la cohorte habrá que reconstruirle una parte de su seguimiento, por ello, hay que añadir esta información a los datos de los individuos que se van recogiendo a la vez que avanza el estudio (Figura 2.3).

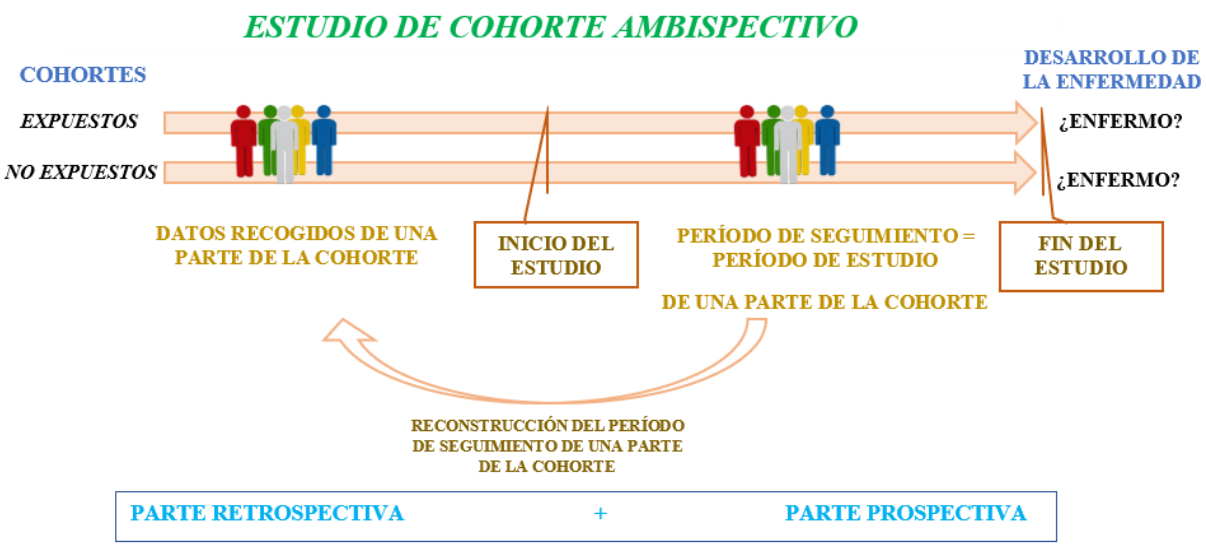


Figura 2.3: Esquema de los estudios ambispectivos.

En relación a la población, existen diferentes tipos de cohortes. En general las cohortes se clasifican en fijas o dinámicas. Las cohortes fijas son aquellas en las que no se permite la entrada de nuevos miembros al estudio más allá del período de selección de los individuos y solo admite las salidas de estos miembros ya sea por enfermedad, muerte u otras causas. Sin embargo, las

cohortes dinámicas sí permiten la entrada de nuevos miembros al estudio durante el período de seguimiento y, por supuesto, también permite su salida por las condiciones mencionadas anteriormente.

Adicionalmente y atendiendo a Gallego Iborra (2012), dependiendo del número de grupos que se estén analizando en cada estudio, podemos clasificarlos en:

- Estudios con dos cohortes: una de estas dos cohortes está expuesta y la otra no. Los tipos de estudios a los que nos hemos referido hasta ahora trabajan con estas dos cohortes, por tanto, las figuras 2.1, 2.2 y 2.3 son ejemplos de estudios con este tipo de cohorte.
- Estudios de comparación con la población general: el estudio hace el seguimiento a una cohorte expuesta y cuando finaliza los datos se comparan con los recogidos de la población general en otros registros.
- Estudios de comparaciones internas: solo hay una cohorte bajo estudio, a ella pertenecen individuos expuestos y no expuestos y, cuando finaliza el estudio, se hacen las comparaciones entre ellos.
- Estudios de cohortes múltiples: se forman varias cohortes que varían en función del grado de exposición al que están sometidas. Cuando se pone fin al estudio, se hacen comparaciones entre las cohortes, la cohorte de referencia o control es la que tenga un menor nivel de exposición o simplemente no esté expuesta. Así se irán comparando todas las cohortes con la cohorte de control.

2.2. Selección de las cohortes

Una vez que queda establecida la relación entre el factor de exposición y el evento de interés, comienza el proceso de selección de la cohorte. Esta parte del estudio es fundamental para el desarrollo y las conclusiones futuras y viene motivada por la naturaleza de la exposición, las características del evento bajo estudio y la posibilidad de seguimiento del grupo elegido. Por lo tanto, la población seleccionada para hacer el seguimiento debe estar formada por individuos que tienen riesgo de desarrollar el evento de estudio.

En el estudio de cohortes más común, estudio de dos cohortes dinámicas, se selecciona un grupo de individuos que tiene el factor de exposición y otro grupo comparable en el que los individuos no están expuestos. Cabe señalar que la cohorte expuesta debe tener un nivel de exposición suficiente como para poder desarrollar la enfermedad y poder comparar los resultados con la cohorte no expuesta.

En general, las personas que se seleccionan para el estudio de cohortes se eligen a partir de la extracción de una muestra al azar de la población de interés que está expuesta al factor de interés, esto se hace sobre todo en los estudios de cohortes descriptivos mencionados con anterioridad.

En algunos casos la selección de individuos se hace a través del muestreo de conveniencia o no probabilístico que es el que tiene lugar cuando se escogen los individuos por continuadas visitas a consultas médicas. El principal problema que da este tipo de muestreo es que normalmente el tamaño muestral resultante no es representativo de toda la población y no se podrán extrapolar los resultados a la población total. Por esto, es preferible recurrir a técnicas de muestreo probabilístico.

El método más usado es el muestreo aleatorio simple. Todos los sujetos que forman parte de la población objetivo tienen la misma probabilidad de ser seleccionados y esta es independiente de la que tienen el resto de individuos. Otro método interesante de aplicar, cuando en la población objetivo existen varias zonas geográficas en diferente proporción, es el muestreo estratificado en el que se pondera cada estrato para que tenga el peso que ocupa en la población inicial. Jugando también con las proporciones pero en cuanto a sexo, raza, edad u otra característica se aplica el muestreo por cuotas en el que se seleccionan los individuos en la muestra de manera proporcional a lo que aporta en la población objetivo un individuo con sus mismas características. Por último, si los individuos pertenecieran a grupos excluyentes se debería aplicar el muestreo por conglomerados en el que se harán conglomerados heterogéneos pero con homogeneidad dentro de cada uno de ellos, dentro de cada conglomerado se hará un muestreo aleatorio simple, el problema es que la variabilidad es mayor (inter e intragrupal) y además el tamaño muestral es mayor (Molina Arias & Ochoa Sangrador, 2013).

Antes de comenzar con el seguimiento de la/las cohortes seleccionadas se debe tener información de las variables que se van a estudiar de cada individuo para evitar confusión. En cuanto al seguimiento es importante seleccionar cohortes en las que preveamos que el porcentaje de abandonos del estudio va a ser bajo. Además, al igual que las reglas de inclusión en el estudio deben ser claras, las reglas de exclusión deben estar bien fijadas de antemano.

2.3. Seguimiento de las cohortes

El seguimiento se hace a tiempo real cuando realizamos un estudio de cohortes prospectivo, aunque en esta sección se puede adaptar todo lo que conlleva este proceso de seguimiento a estudios retrospectivos y ambispectivos en los que hay que reconstruir todo el período de

seguimiento o solo una parte de éste.

La fecha de entrada al estudio puede ser la misma para todos los individuos de una cohorte (cohorte fija) o distinta para algunos de ellos (cohorte dinámica), este momento supone el inicio de la contribución de cada individuo en el estudio. Este fecha de entrada no tiene porqué coincidir con el inicio de la exposición (de hecho, no suele hacerlo).

El período de seguimiento va a depender de la pregunta a la que se quiera responder, es decir, de la asociación entre el factor de exposición y el evento que queramos medir para responder a la hipótesis planteada por los investigadores (Kestenbaum, 2009).

Cuando hicimos referencia a los estudios retrospectivos se puntualizó que era muy importante que las fuentes de las cuales recogiésemos los datos para la reconstrucción del seguimiento de los individuos de las cohortes fueran muy buenas para que los datos que aportaran estuviesen en correcto estado. Esto mismo se ha de aplicar a los estudios prospectivos ya que también influirá en su calidad. En el período de seguimiento se deben establecer unos objetivos y una prioridades para poder recoger la información necesaria y concreta con el fin de obtener unos datos de los que sacar buenas conclusiones. Según Royo-Bordonada et al. (2009), los objetivos principales durante el período de seguimiento en un estudio de cohortes son tres:

- Detectar la aparición del evento de interés.
- Detectar cuáles son las pérdidas en el estudio (abandonos) y cuándo se han producido.
- Detectar cambios importantes en el nivel de exposición de los individuos, así como la exposición a nuevos factores de riesgo.

El seguimiento se realiza en los individuos de la cohorte con periodicidad, este se puede llevar a cabo a través de formularios, pruebas médicas u otros métodos dependiendo del factor de exposición y del evento bajo estudio. La periodicidad con la que se recogen resultados para hacer el seguimiento va de la mano con esta última idea y también con la capacidad o necesidad de mantener individuos dentro del estudio.

Es muy importante que en un estudio de dos o más cohortes, el seguimiento de cada una se haga completo y de manera similar al resto para evitar el sesgo de seguimiento (Henquin, 2013). A veces el período de seguimiento se hace muy largo, sobre todo en la investigación de enfermedades crónicas, y es por esto por lo que, en estos casos, a menudo el estudio de cohortes es retrospectivo y así se consigue evitar el largo período de seguimiento, aunque corremos el riesgo de que se ponga en duda la veracidad de los datos recogidos porque hay que reconstruir el seguimiento.

En la sección anterior se hizo hincapié en la importancia de seleccionar miembros de las cohortes en los que se prevea que no existe intención de abandonar. Aunque esta condición se cumple, es importante que se haga una campaña para mantener la tasa de participación alta como por ejemplo, crear un logo del estudio, contar con personal flexible en horarios y que faciliten la comunicación, etcétera. Aún así, es posible que varios individuos abandonen el estudio por ello es importante contar con un equipo que sea capaz de detectar los abandonos y el momento en el que se producen (Royo-Bordonada et al., 2009).

En el momento en el que se dejan de registrar datos de un individuo del estudio se guarda como la fecha de salida del estudio que puede darse por tres motivos diferentes que son:

- Abandono del estudio.
- Fin del estudio.
- Desarrollo del evento de interés (si el evento es único).

Una vez que todos los individuos han parado de contribuir al estudio (por alguno de estos motivos), se comienza con el análisis de los datos recogidos y, por último, se establece como cierta o no la hipótesis bajo estudio.

2.4. Validez del estudio y sesgos

Los estudios observacionales tienen un mayor riesgo de sesgos (Molina Arias & Ochoa Sangrador, 2014) ya que, por su naturaleza y definición, los investigadores no tienen control en varios puntos importantes, principalmente en la exposición. Además, se ven influenciados por factores de confusión y por variables modificadoras de efecto.

Los factores de confusión o variables "confundidoras" son variables externas a la relación entre el factor de exposición y el evento pero que sí están relacionadas cronológicamente con el factor de exposición y ser otro factor de riesgo para el evento y, además, que deben ser tenidas en cuenta porque son capaces de modificar o invertir la relación de interés en el estudio (Paradoja de Simpson) por ello, es importante desagregar los datos en función de estas variables y así evitar confusiones a la hora de determinar un resultado. Para valorar realmente la existencia de una variable "confundidora" debemos calcular los riesgos crudos, ya sean Odds Ratio (Ecuación 1.7) o Riesgo Relativo (Ecuación 1.5). Para reconocer posibles variables de confusión debemos establecer listas de variables relacionadas con el factor de estudio y que, además, puedan ser causa de desenlace. Una vez reconocidas tenemos dos opciones: controlarlas en la fase de diseño mediante la restricción y el emparejamiento, o controlarlas en la fase de análisis mediante la

estratificación y el ajuste (Solís Sánchez & Orejas Rodríguez-Arango, 1999).

Las variables modificadoras de efecto son capaces de modificar la relación de dos variables. La sospecha sobre la existencia de estas variables aparece cuando la incidencia de la enfermedad bajo estudio dista de la esperada. Estadísticamente, este término es conocido como interacción y puede observarse en modelos estadísticos multiplicativos y modelos estadísticos aditivos.

Los estudios de cohortes son los menos expuestos a sesgos, aún así sí que están sometidos a sesgos de selección, sesgos de clasificación y sesgos de confusión e interacción que se producen por los confundidores y las variables modificadoras de efecto explicados anteriormente.

El sesgo de selección aparece en la etapa de selección de la cohorte (Sección 2.2), afecta en gran medida a la validez del estudio y hay que tenerlo en cuenta siempre. Para intentar reducirlo las cohortes deben ser comparables en todos los aspectos menos en la exposición y además deben ser representativas de la población objetivo.

El sesgo de seguimiento se da en la etapa correspondiente al seguimiento (Sección 2.3). Para disminuir su impacto sobre el estudio las pérdidas en las diferentes cohortes deben ser similares. Además, en esta etapa hay que añadir los sesgos de clasificación e información que se deben a la mala gestión de los datos recogidos en el estudio.

2.5. Estudio de cohortes en R

En este apartado se va a analizar un estudio de cohortes retrospectivo a través del software estadístico R (R Core Team, 2020).

Se trata de un estudio de cohortes retrospectivo con pacientes hospitalizados por COVID-19 en el hospital de Kazajistán entre febrero y abril de 2020. En él se comparan datos demográficos, clínicos, de laboratorio y radiológicos de pacientes con diferentes grados de gravedad de COVID-19 al ingreso. La regresión logística se utilizó para evaluar los factores asociados con la gravedad de la enfermedad y la muerte hospitalaria (Yegorov et al., 2021b).

El estudio ha sido realizado por el Grupo de Investigación en Epidemiología Semey COVID-19 (Semey COVID-19 Epidemiology Research Group).

Los investigadores del estudio pudieron acceder a un registro de 1960 pacientes que presentaban síntomas similares al COVID-19 o que habían sido expuestos al virus entre el 20 de febrero y el 30 de abril de 2020. Todos los pacientes fueron hospitalizados en clínicas auxiliares especializadas y se les hizo un seguimiento hasta el día 5 de mayo de 2020 (si no abandonaban antes por alta hospitalaria o fallecimiento) que fue el día determinado para finalizar el estudio.

En este caso se busca encontrar factores de riesgo relacionados con la severidad de la enfermedad por COVID-19 y con la mortalidad. Estaríamos ante dos estudios de cohortes retrospectivos porque estudiamos dos variables objetivo. Los factores de exposición que se consideran en este estudio son de tipo clínico (presión arterial, náuseas, vómitos, etc.), de laboratorio (niveles de glucosa, creatinina, calcio, etc.), demográficos y radiológicos.

Se han realizado estudios complementarios a los dos principales que se encuentran detallados en el Apéndice A.

La mayoría de las órdenes que se recogen en las siguientes secciones y en el Apéndice A han sido tomadas del código disponible en GitHub del estudio de cohortes retrospectivo de COVID-19 en Kazajistán (Babenko, 2021). Parte del código ha sido modificado para obtener mejores resultados. Además, todas las interpretaciones son de elaboración propia.

2.5.1. Lectura de los datos

```
library(readr)
covidcohort <- read_delim("https://raw.githubusercontent.com/dimbage
                          /COVID-19-in-KZ/main/covid-19_db.txt",
                          delim = "\t",
                          escape_double = FALSE,
                          col_types = cols(
                            Disease_severity =
                              col_factor(levels = c("1","2", "3", "4")),
                            Sex =
                              col_factor(levels = c("0","1")),
                            Kazakh_ethnicity =
                              col_factor(levels = c("0","1")),
                            Deaths =
                              col_factor(levels = c("0","1")),
                            Any_comorbidities =
                              col_factor(levels = c("0","1")),
                            Hypertension =
                              col_factor(levels = c("0","1")),
                            Coronary_heart_disease =
                              col_factor(levels = c("0","1")),
```

```
                        COPD =
                          col_factor(levels = c("0","1")),
                        Chronic_kidney_disease =
                          col_factor(levels = c("0","1")),
                        Cancer =
                          col_factor(levels = c("0","1")),
                        Diabetes =
                          col_factor(levels = c("0","1")),
                        Liver_disease =
                          col_factor(levels = c("0","1")),
                        Othercomorb =
                          col_factor(levels = c("0","1")),
                        Cough =
                          col_factor(levels = c("0","1")),
                        Sputum_production =
                          col_factor(levels = c("0","1")),
                        Shortness_of_breath =
                          col_factor(levels = c("0","1")),
                        Dyspnoea =
                          col_factor(levels = c("0","1")),
                        Stuffy =
                          col_factor(levels = c("0","1")),
                        Sore_throat =
                          col_factor(levels = c("0","1")),
                        Oropharynx_hyperemia =
                          col_factor(levels = c("0","1")),
                        Tonsill_hypertrophy =
                          col_factor(levels = c("0","1")),
                        Chest_pain =
                          col_factor(levels = c("0","1")),
                        Chest_tightness =
                          col_factor(levels = c("0","1")),
                        Wheezing =
                          col_factor(levels = c("0","1")),
```

```
                    Diarrhoea =
                      col_factor(levels = c("0","1")),
                    Nausea_vomiting =
                      col_factor(levels = c("0","1")),
                    Headache =
                      col_factor(levels = c("0","1")),
                    Conjuctivitis =
                      col_factor(levels = c("0","1")),
                    Myalgia_fatigue =
                      col_factor(levels = c("0","1")),
                    Joint_pain =
                      col_factor(levels = c("0","1")),
                    Pneumonia =
                      col_factor(levels = c("0","1")),
                    Bronchitis =
                      col_factor(levels = c("0","1")),
                    Antiviral_medications =
                      col_factor(levels = c("0","1")),
                    Ribavirin =
                      col_factor(levels = c("0","1")),
                    Lopinavir_and_ritonavir =
                      col_factor(levels = c("0","1")),
                    Oseltamivir =
                      col_factor(levels = c("0","1")),
                    Antibiotics =
                      col_factor(levels = c("0","1")),
                    Hydroxychloroquine =
                      col_factor(levels = c("0","1")),
                    Anticoagulant =
                      col_factor(levels = c("0","1")),
                    Corticosteroids =
                      col_factor(levels = c("0","1")),
                    Oxygen_therapy =
                      col_factor(levels = c("0","1")),
```

```
                              Mechanical_ventilation =
                                col_factor(levels = c("0","1"))),
                            trim_ws = TRUE)
# Dimensiones del conjunto de datos
dim(covidcohort)
```

```
## [1] 1072   75
```

Es un conjunto de datos con 1072 observaciones y 75 variables.

2.5.2. Estudio de cohorte retrospectivo para factores de riesgo asociados a la severidad de la enfermedad por COVID-19

En este apartado vamos a realizar el análisis descriptivo e inferencial sobre la severidad de la enfermedad por SARS-CoV-2 buscando factores de riesgo.

Primero vamos a hacer el análisis descriptivo de los datos a través de una tabla resumen. Para ello hay que preparar el dataset.

```
library(dplyr)
library(finalfit)
cohort19 <- covidcohort
```

Para obtener una buena clasificación de la gravedad de la enfermedad vamos a crear una nueva variable llamada `Disease_severity_group` a partir de la variable `Disease_severity` que ya estaba en el conjunto de datos. La variable `Diasease_severity` guarda 4 niveles de menor a mayor gravedad. Los individuos de gravedad 1 serán clasificados en la nueva variable con `No_sintoma/Leve`, los de gravedad 2 formarán el grupo `Moderado` y los de niveles 3 y 4 serán los del grupo `Crítico/Severo`.

La edad es uno de los factores de exposición de interés, en el conjunto de datos original es de tipo numérico y ahora lo convertiremos a factor con 4 niveles o grupos de edad (`0-14`, `15-49`, `50-64` y `65 o mas`).

```
cohort19 %<>% mutate(Disease_severity_group = case_when(
  Disease_severity == 1 ~ "No_sintoma/Leve",
  Disease_severity == 2 ~ "Moderado",
  Disease_severity %in% c(3,4) ~ "Crítico/Severo"))

cohort19$Disease_severity_group <- factor(cohort19$Disease_severity_group,
```

```
                                        levels =
                                          c("No_sintoma/Leve",
                                            "Moderado",
                                            "Crítico/Severo"))



cohort19 %<>% mutate(Age_group = case_when(
  Age < 15 ~ "0-14",
  Age >= 15 & Age < 50 ~ "15-49",
  Age >= 50 & Age < 65 ~ "50-64",
  Age >= 65 ~ "65 o mas"))

cohort19$Age_group <- factor(cohort19$Age_group,
                             levels = c("0-14",
                                        "15-49",
                                        "50-64",
                                        "65 o mas"))
```

En el estudio se definen nuevas variables a partir de otras para tener más datos sobre la sintomatología de los pacientes. Se va a crear la variable `Fever` que guarda si un individuo tuvo fiebre o no a partir de su temperatura corporal (variable `Body_temperature`).

La variable `SP_low_B` guardará dos grupos según si la presión sistólica (`Systolic_pressure`) fue baja o no.

La variable `RR_high_B` guarda dos grupos según si la frecuencia respiratoria fue baja o no (`Respiratory_rate`).

La variable `WBC_low_B` guarda dos grupos según si la cantidad de glóbulos en sangre (`White_blood_cells`) era baja o no.

`WBC_high_B` guarda 0 y 1 si la cantidad de glóbulos blancos fue alta.

Por último, la variable `NLR_high_B` guarda información de si la cantidad de neutrófilos y leucocitos (`NLR`) fue alta.

```
# Creamos la variable fiebre basándonos en la temperatura corporal
cohort19 %<>% mutate(Fever = case_when(Body_temperature >= 37.3 ~ 1,
                                       Body_temperature < 37.3 ~ 0))
```

```r
# 1 = tuvo fiebre
# 0 = no tuvo fiebre
cohort19$Fever <- factor(cohort19$Fever,
                         levels = c(0, 1),
                         labels = c("NoFiebre",
                                    "Fiebre"))

# Creamos la variable presión sistólica baja a partir de la
# variable Systolic_pressure
cohort19 %<>% mutate(SP_low_B = case_when(
  Systolic_pressure < 90 ~ 1,
  Systolic_pressure >= 90 ~ 0))
# 1 = tensión baja
# 0 = tensión alta
cohort19$SP_low_B <- factor(
  cohort19$SP_low_B,
  levels = c(0, 1),
  labels = c("Alto", "Bajo"))

# Frecuencia respiratoria basado en Respiratory_rate
cohort19 %<>% mutate(RR_high_B =
                       case_when(Respiratory_rate > 24 ~ 1,
                                 Respiratory_rate <= 24 ~ 0))

cohort19$RR_high_B <- factor(
  cohort19$RR_high_B,
  levels = c(0, 1),
  labels = c("Bajo", "Alto"))


# Baja cantidad de glóbulos blancos basado en white_blood_cells
cohort19 %<>% mutate(WBC_low_B =
                   case_when(White_blood_cells < 4 ~ 1,
                             White_blood_cells >= 4 ~ 0))
```

```r
cohort19$WBC_low_B <- factor(
  cohort19$WBC_low_B,
  levels = c(0, 1),
  labels = c("Alto", "Bajo"))

# Alta cantidad de glóbulos blancos basado en white_blood_cells
cohort19 %<>% mutate(WBC_high_B = case_when(
  White_blood_cells > 10 ~ 1,
  White_blood_cells <= 10 ~ 0))

cohort19$WBC_high_B <- factor(
  cohort19$WBC_high_B,
  levels = c(0, 1),
  labels = c("Bajo", "Alto"))



# Alta cantidad de neutrófilos y leucocitos basado en NLR
cohort19 %<>% mutate(NLR_high_B = case_when(
  NLR > 4 ~ 1,
  NLR <= 4 ~ 0))

cohort19$NLR_high_B <- factor(
  cohort19$NLR_high_B,
  levels = c(0, 1),
  labels = c("Bajo", "Alto"))

# Los nombres de algunas de las variables son muy largos y pueden causarnos
# problemas a la hora de visualizar las tablas o gráficos que vamos a ir
# haciendo así que vamos a eliminar el caracter _ para acortarlos un poco
colnames(cohort19) <- gsub("_", replacement = "", colnames(cohort19))
colnames(cohort19)[7] <- "Daysym2add"
colnames(cohort19)[54] <- "Aspmintran"
colnames(cohort19)[60] <- "Creactprot"
```

Ahora, que ya se ha preparado el conjunto de datos, vamos a pasar a crear la tabla resumen con la función `summary_factorlist` de la librería `finalfit` (Harrison et al., 2021) que nos devuelve una tabla resumen a partir de la relación entre una variable dependiente `Diseaseseveritygroup` (creada anteriormente) y un grupo de variables independientes (características clínicas, demográficas, radiológicas y de laboratorio).

```
deps <- "Diseaseseveritygroup"


exp <- c("Age",
         "Agegroup",
         "Sex",
         "Kazakhethnicity",
         "BMI",
         "Daysym2add",
         "Daysinhospital",
         "Deaths",
         "Anycomorbidities",
         "Fever",
         "Cough",
         "Systolicpressure",
         "SPlowB",
         "Respiratoryrate",
         "RRhighB",
         "SpO2",
         "Whitebloodcells",
         "WBClowB",
         "WBChighB",
         "NLR",
         "NLRhighB",
         "Haemoglobin",
         "Prothrombintime",
         "Aspmintran",
         "Totalbilirubin",
         "Creatinine",
         "Creactprot",
```

```
          "Pneumonia",
          "Bronchitis",
          "Antiviralmedications",
          "Antibiotics",
          "Anticoagulant",
          "Corticosteroids",
          "Oxygentherapy",
          "Mechanicalventilation")



cohort19 %>%
  summary_factorlist(deps, exp,
                     p = T,
                     cont = "median",
                     total_col = TRUE,
                     column = TRUE,
                     na_include = F) -> tb1

knitr::kable(tb1,
             caption = "\\label{table04}Características demográficas y clínicas de
             los pacientes con COVID-19 confirmados por laboratorio
             clasificados al ingreso por gravedad de la enfermedad.",
             align = c("l", "r", "c", "r", "c", "c", "c"))
```

Tabla 2.1: Características demográficas y clínicas de los pacientes con COVID-19 confirmados por laboratorio clasificados al ingreso por gravedad de la enfermedad.

label	levels	No_sintoma/Leve	Moderado	Crítico/Severo	Total	p
Age	Median (IQR)	32.0 (23.0 to 47.0)	38.0 (28.0 to 53.0)	58.0 (45.5 to 71.0)	36.0 (24.0 to 50.0)	<0.001
Agegroup	0-14	74 (10.8)	18 (5.3)	1 (2.1)	93 (8.7)	<0.001
	15-49	455 (66.6)	220 (64.3)	15 (31.9)	690 (64.4)	
	50-64	117 (17.1)	77 (22.5)	13 (27.7)	207 (19.3)	
	65 o mas	37 (5.4)	27 (7.9)	18 (38.3)	82 (7.6)	

label	levels	No_sintoma/Leve	Moderado	Crítico/Severo	Total	p
Sex	Mujer	382 (55.9)	186 (54.4)	20 (42.6)	588 (54.9)	0.200
	Hombre	301 (44.1)	156 (45.6)	27 (57.4)	484 (45.1)	
Kazakhethnicity	Kazaj	132 (19.3)	70 (20.5)	18 (38.3)	220 (20.5)	0.008
	Otra	551 (80.7)	272 (79.5)	29 (61.7)	852 (79.5)	
BMI	Median (IQR)	23.0 (20.0 to 26.0)	26.0 (23.0 to 30.2)	27.5 (25.8 to 29.2)	24.0 (21.0 to 29.0)	0.007
Daysym2add	Median (IQR)	1.0 (1.0 to 2.0)	2.0 (1.0 to 6.0)	4.0 (2.0 to 7.0)	1.0 (1.0 to 3.0)	<0.001
Daysinhospital	Median (IQR)	16.0 (14.0 to 17.0)	16.0 (15.0 to 18.0)	15.0 (3.0 to 17.0)	16.0 (14.0 to 17.0)	0.058
Deaths	No muere	682 (99.9)	337 (98.5)	33 (70.2)	1052 (98.1)	<0.001
	Muere	1 (0.1)	5 (1.5)	14 (29.8)	20 (1.9)	
Anycomorbidities	No	448 (65.6)	179 (52.3)	14 (29.8)	641 (59.8)	<0.001
	Si	235 (34.4)	163 (47.7)	33 (70.2)	431 (40.2)	
Fever	NoFiebre	462 (94.7)	181 (68.8)	10 (35.7)	653 (83.8)	<0.001
	Fiebre	26 (5.3)	82 (31.2)	18 (64.3)	126 (16.2)	
Cough	No	579 (84.8)	179 (52.3)	20 (42.6)	778 (72.6)	<0.001
	Si	104 (15.2)	163 (47.7)	27 (57.4)	294 (27.4)	
Systolicpressure	Median (IQR)	110.0 (110.0 to 120.0)	120.0 (110.0 to 130.0)	117.5 (107.5 to 126.2)	110.0 (110.0 to 120.0)	<0.001
SPlowB	Alto	416 (98.1)	234 (100.0)	25 (89.3)	675 (98.4)	<0.001
	Bajo	8 (1.9)	0 (0.0)	3 (10.7)	11 (1.6)	
Respiratoryrate	Median (IQR)	18.0 (18.0 to 19.0)	18.0 (18.0 to 20.0)	22.5 (20.0 to 28.5)	18.0 (18.0 to 20.0)	<0.001
RRhighB	Bajo	431 (98.2)	236 (96.7)	19 (67.9)	686 (96.5)	<0.001
	Alto	8 (1.8)	8 (3.3)	9 (32.1)	25 (3.5)	
SpO2	Median (IQR)	98.0 (96.0 to 98.0)	97.0 (96.0 to 98.0)	91.5 (84.5 to 95.0)	97.0 (96.0 to 98.0)	<0.001
Whitebloodcells	Median (IQR)	6.5 (5.1 to 8.0)	6.1 (5.0 to 7.7)	6.5 (5.5 to 9.1)	6.3 (5.1 to 7.9)	0.096
WBClowB	Alto	593 (91.7)	293 (89.6)	42 (95.5)	928 (91.2)	0.335
	Bajo	54 (8.3)	34 (10.4)	2 (4.5)	90 (8.8)	

label	levels	No_sintoma/Leve	Moderado	Crítico/Severo	Total	p
WBChighB	Bajo	596 (92.1)	307 (93.9)	36 (81.8)	939 (92.2)	0.019
	Alto	51 (7.9)	20 (6.1)	8 (18.2)	79 (7.8)	
NLR	Median (IQR)	2.0 (1.4 to 3.1)	2.1 (1.4 to 3.0)	2.6 (0.9 to 6.5)	2.1 (1.4 to 3.2)	0.485
NLRhighB	Bajo	242 (84.3)	132 (85.7)	17 (65.4)	391 (83.7)	0.031
	Alto	45 (15.7)	22 (14.3)	9 (34.6)	76 (16.3)	
Haemoglobin	Median (IQR)	137.0 (124.0 to 149.0)	136.5 (124.0 to 152.0)	127.0 (114.5 to 142.5)	136.5 (124.0 to 150.0)	0.013
Prothrombintime	Median (IQR)	12.8 (11.4 to 16.0)	12.9 (11.9 to 15.5)	15.0 (12.9 to 19.4)	13.0 (11.7 to 16.0)	0.011
Aspmintran	Median (IQR)	17.7 (12.0 to 23.1)	20.1 (14.0 to 29.6)	37.0 (17.7 to 49.9)	18.5 (12.5 to 26.5)	<0.001
Totalbilirubin	Median (IQR)	11.5 (8.2 to 16.7)	10.5 (7.7 to 15.0)	13.1 (10.0 to 20.1)	11.2 (8.0 to 16.0)	0.003
Creatinine	Median (IQR)	70.1 (57.3 to 86.2)	75.8 (62.8 to 92.4)	85.0 (67.5 to 116.5)	73.1 (60.3 to 88.0)	<0.001
Creactprot	Median (IQR)	2.0 (0.0 to 6.0)	3.0 (0.2 to 10.0)	6.0 (0.7 to 45.1)	2.3 (0.1 to 8.1)	0.039
Pneumonia	No	373 (84.4)	131 (50.2)	3 (8.8)	507 (68.8)	<0.001
	Si	69 (15.6)	130 (49.8)	31 (91.2)	230 (31.2)	
Bronchitis	No	329 (74.4)	188 (72.0)	32 (94.1)	549 (74.5)	0.021
	Si	113 (25.6)	73 (28.0)	2 (5.9)	188 (25.5)	
Antiviralmedications	No	222 (32.5)	81 (23.7)	20 (42.6)	323 (30.1)	0.002
	Si	461 (67.5)	261 (76.3)	27 (57.4)	749 (69.9)	
Antibiotics	No	569 (83.3)	174 (50.9)	21 (44.7)	764 (71.3)	<0.001
	Si	114 (16.7)	168 (49.1)	26 (55.3)	308 (28.7)	
Anticoagulant	No	677 (99.1)	330 (96.5)	36 (76.6)	1043 (97.3)	<0.001
	Si	6 (0.9)	12 (3.5)	11 (23.4)	29 (2.7)	
Corticosteroids	No	676 (99.0)	335 (98.0)	45 (95.7)	1056 (98.5)	0.124
	Si	7 (1.0)	7 (2.0)	2 (4.3)	16 (1.5)	
Oxygentherapy	No	673 (98.5)	318 (93.0)	17 (36.2)	1008 (94.0)	<0.001
	Si	10 (1.5)	24 (7.0)	30 (63.8)	64 (6.0)	
Mechanicalventilation	No	681 (99.7)	335 (98.0)	29 (61.7)	1045 (97.5)	<0.001

label	levels	No_sintoma/Leve	Moderado	Crítico/Severo	Total	p
	Si	2 (0.3)	7 (2.0)	18 (38.3)	27 (2.5)	

En la Tabla 2.1 aparecen representadas la mediana y el rango intercuartil (IQR) que es una medida encargada de cuantificar la dispersión de la muestra en base a la variable que representa en cada caso.

Estas medidas son calculadas para la población total y para cada una de los grupos de la variable dependiente `Diseaseseveritygroup`.

Además aparece calculado el p-valor resultante de aplicar la Prueba de Wilcoxon-Mann-Whitney, la Prueba de Chi-Cuadrado (χ^2) o la Prueba Exacta de Fisher en las que se comprueban si hay diferencias significativas entre los grupos de las variables. Mirando los p-valores podemos decir que existen diferencias significativas entre los grupos de las variables `Age`, `Agegroup`, `Kazakhethnicity` entre otras. Sin embargo parece que no existen diferencias significativas entre los sexos (variable `Sex`), entre los grupos de niveles bajos de glóbulos blancos (`WBClowB`) o entre los grupos de la variable `NLR`.

La mortalidad observada es del 1.9 % en el total de casos. La columna más importante es la del grupo `Severo/Critico` porque ahí es donde veremos cuáles son los factores asociados a enfermar gravemente de COVID-19 aunque aún falta realizar el análisis inferencial que determine la asociación. Mirando este grupo de enfermos, vemos que tener más de 65 años (`Agegroup`), ser hombre (`Sex`), tener patologías previas (`Anycomoridities`) o pertenecer a la etnia kazaja (`Kazakhethnicity`) podrían ser posibles factores de riesgo asociados a enfermar gravemente por COVID-19.

Para comprobar las asociaciones vamos a realizar una regresión logística simple y múltiple.

Vamos a guardar dos grupos de severidad en la variable `SeverityGroups` para poder usarla como variable dependiente en la regresión logística.

```
# No_Severo = grupos 1 y 2
# Severo = grupos 3 y 4
cohort19 %<>% mutate(SeverityGroups = case_when(
  Diseaseseverity %in% c(1,2) ~ "No_Severo",
  Diseaseseverity %in% c(3,4) ~ "Severo"))

cohort19$SeverityGroups <- factor(
```

```
  cohort19$SeverityGroups,
  levels = c("No_Severo","Severo"))
```

Hay que hacer algunos cambios en las variables que vamos a usar como explicativas para que estén bien definidas como factor.

```
cohort19 %<>% mutate(AspmintranB = case_when(
  Aspmintran > 40 ~ 1,
  TRUE ~ 0))

cohort19$AspmintranB <- factor(
  cohort19$AspmintranB)

cohort19 %<>% mutate(TotalbilirubinB = case_when(
  Totalbilirubin > 17 ~ 1,
  TRUE ~ 0))

cohort19$TotalbilirubinB <- factor(cohort19$TotalbilirubinB)

cohort19 %<>% mutate(DirectbilirubinB = case_when(
  Directbilirubin > 5.1 ~ 1,
  TRUE ~ 0))

cohort19$DirectbilirubinB <- factor(
  cohort19$DirectbilirubinB)

cohort19 %<>% mutate(CreatinineB = case_when(
  Creatinine > 118 ~ 1,
  TRUE ~ 0));
cohort19$CreatinineB = factor(cohort19$CreatinineB)
```

Ya podemos hacer la regresión logística simple y múltiple. La haremos con la función `finalfit` de la librería que lleva el mismo nombre.

```
deps <- "SeverityGroups"

exp = c(
```

```
  "Age",
  "Sex",
  "Kazakhethnicity",
  "Anycomorbidities",
  "Whitebloodcells",
  "NLR",
  "Haemoglobin",
  "Prothrombintime",
  "Fibrinogen",
  "Albumin",
  "AspmintranB",
  "TotalbilirubinB",
  "Glucose",
  "Bloodureanitrogen",
  "CreatinineB",
  "Creactprot",
  "Potassium",
  "Calcium")



expmult = c(
  "Age",
  "Kazakhethnicity",
  "Anycomorbidities",
  "Whitebloodcells")

cohort19 %>% finalfit::finalfit(deps, exp, expmult) -> tb3

knitr::kable(tb3,
             align = "c",
             caption = "\\label{table05}Regresión logística bivariada de
             factores asociados con las probabilidades de
             enfermedad grave por COVID-19 en Kazajstán.")
```

Tabla 2.2: Regresión logística bivariada de factores asociados con las probabilidades de enfermedad grave por COVID-19 en Kazajstán.

	Dependent: SeverityGroups		No_Severo	Severo	OR (univariable)	OR (multivariable)
1	Age	Mean (SD)	36.9 (17.8)	57.5 (20.3)	1.06 (1.04-1.08, p<0.001)	1.05 (1.03-1.07, p<0.001)
21	Sex	Mujer	568 (96.6)	20 (3.4)	-	-
20		Hombre	457 (94.4)	27 (5.6)	1.68 (0.93-3.07, p=0.086)	-
15	Kazakhethnicity	Kazaj	202 (91.8)	18 (8.2)	-	-
16		Otra	823 (96.6)	29 (3.4)	0.40 (0.22-0.74, p=0.003)	0.73 (0.37-1.51, p=0.380)
3	Anycomorbidities	No	627 (97.8)	14 (2.2)	-	-
4		Si	398 (92.3)	33 (7.7)	3.71 (2.00-7.25, p<0.001)	2.34 (1.18-4.85, p=0.017)
24	Whitebloodcells	Mean (SD)	6.6 (2.3)	7.3 (3.0)	1.12 (0.99-1.25, p=0.052)	1.13 (1.00-1.27, p=0.044)
17	NLR	Mean (SD)	2.7 (2.4)	4.9 (6.6)	1.15 (1.05-1.25, p=0.001)	-
14	Haemoglobin	Mean (SD)	136.2 (20.1)	127.5 (18.8)	0.98 (0.97-0.99, p=0.005)	-
19	Prothrombintime	Mean (SD)	531.1 (4742.1)	15.9 (3.4)	1.00 (NA-1.00, p=0.871)	-
12	Fibrinogen	Mean (SD)	1701.2 (8227.8)	4.8 (2.2)	1.00 (NA-1.00, p=0.725)	-
2	Albumin	Mean (SD)	52.8 (250.5)	32.2 (11.1)	0.95 (0.91-0.99, p=0.004)	-
5	AspmintranB	0	975 (97.0)	30 (3.0)	-	-
6		1	50 (74.6)	17 (25.4)	11.05 (5.63-21.21, p<0.001)	-
22	TotalbilirubinB	0	882 (96.0)	37 (4.0)	-	-
23		1	143 (93.5)	10 (6.5)	1.67 (0.77-3.30, p=0.165)	-

	Dependent: SeverityGroups		No_Severo	Severo	OR (univariable)	OR (multivariable)
13	Glucose	Mean (SD)	5.6 (4.1)	7.0 (3.1)	1.04 (0.98-1.09, p=0.094)	-
7	Bloodureanitrogen	Mean (SD)	208.1 (2958.5)	8.1 (5.3)	1.00 (0.98-1.00, p=0.965)	-
10	CreatinineB	0	991 (96.2)	39 (3.8)	-	-
11		1	34 (81.0)	8 (19.0)	5.98 (2.44-13.22, p<0.001)	-
9	Creactprot	Mean (SD)	210.1 (2984.0)	39.5 (76.4)	1.00 (NA-1.00, p=0.847)	-
18	Potassium	Mean (SD)	4.2 (1.7)	3.9 (0.7)	0.85 (0.55-1.18, p=0.438)	-
8	Calcium	Mean (SD)	2.0 (0.7)	11.1 (34.6)	1.04 (1.00-NA, p=0.208)	-

La Tabla 2.2 devuelve la Odds Ratio para cada variable que se cree asociada a una mayor gravedad de COVID-19, además nos da el intervalo de confianza para esta medida y el p-valor de su significancia.

En el análisis de regresión logística simple, las variables significativas son la mayor edad (`Age`), pertenecer a otra etnia (`Kazakhethnicity`), tener patologías previas (`Anycomorbidities`), altos niveles de aspartato transaminasa (`AspmintranB`), creatinina (`CreatinineB`) y neutrófilos y linfocitos (`NLR`), bajos niveles de hemoglobina (`Haemoglobin`) y albúmina (`Albumin`) están asociados con la gravedad de la enfermedad por COVID-19.

En el análisis múltiple los factores asociados son la edad alta (`Age`), tener patologías previas (`Anycomorbidities`) y altos niveles de glóbulos blancos (`Whitebloodcells`).

En general, se puede concluir diciendo que los factores de riesgo asociados a enfermar de mayor gravedad por COVID-19 son la edad avanzada, tener patologías previas y altos niveles de glóbulos blancos en sangre.

2.5.3. Estudio de cohorte retrospectivo para factores de riesgo asociados a la mortalidad por COVID-19

En este apartado vamos a realizar el análisis descriptivo e inferencial sobre la mortalidad por SARS-CoV-2 buscando factores de riesgo.

Para la obtención de resultados se ha realizado la misma preparación de los datos que en el estudio para la asociación de factores a la severidad de la enfermedad de COVID-19.

En este caso, la diferencia más evidente es que la variable dependiente pasa a ser la que registra las muertes (`Deaths`).

```
deps = "Deaths"


exp <- c("Age",
         "Agegroup",
         "Sex",
         "Kazakhethnicity",
         "BMI",
         "Daysym2add",
         "Daysinhospital",
         "Anycomorbidities",
         "Fever",
         "Cough",
         "Systolicpressure",
         "SPlowB",
         "Respiratoryrate",
         "RRhighB",
         "SpO2",
         "Whitebloodcells",
         "WBClowB",
         "WBChighB",
         "NLR",
         "NLRhighB",
         "Haemoglobin",
         "Prothrombintime",
         "Aspmintran",
```

```
          "Directbilirubin",
          "Creatinine",
          "Creactprot",
          "Pneumonia",
          "Bronchitis",
          "Antiviralmedications",
          "Antibiotics",
          "Anticoagulant",
          "Corticosteroids",
          "Oxygentherapy",
          "Mechanicalventilation"
          )

cohort19 %>% summary_factorlist(deps,
                                exp,
                                p = T,
                                cont = "median",
                                total_col = TRUE,
                                column = TRUE,
                                na_include = F) -> tb2

knitr::kable(tb2,
             caption = "\\label{table06}Características demográficas y clínicas
             de los pacientes, con COVID-19 confirmados por
             laboratorio, que habían sobrevivido (No muere)
             o fallecido (Muere) antes del 30 de abril
             de 2020.",
             align = c("l", "r", "c", "c", "c"))
```

Tabla 2.3: Características demográficas y clínicas de los pacientes, con COVID-19 confirmados por laboratorio, que habían sobrevivido (No muere) o fallecido (Muere) antes del 30 de abril de 2020.

label	levels	No muere	Muere	Total	p
Age	Median (IQR)	35.0 (24.0 to 50.0)	65.0 (57.8 to 77.8)	36.0 (24.0 to 50.0)	<0.001
Agegroup	0-14	92 (8.7)	1 (5.0)	93 (8.7)	<0.001
	15-49	689 (65.5)	1 (5.0)	690 (64.4)	
	50-64	199 (18.9)	8 (40.0)	207 (19.3)	
	65 o mas	72 (6.8)	10 (50.0)	82 (7.6)	
Sex	Mujer	582 (55.3)	6 (30.0)	588 (54.9)	0.043
	Hombre	470 (44.7)	14 (70.0)	484 (45.1)	
Kazakhethnicity	Kazaj	210 (20.0)	10 (50.0)	220 (20.5)	0.003
	Otra	842 (80.0)	10 (50.0)	852 (79.5)	
BMI	Median (IQR)	24.0 (21.0 to 29.0)	43.5 (33.8 to 53.2)	24.0 (21.0 to 29.0)	0.227
Daysym2add	Median (IQR)	1.0 (1.0 to 3.0)	4.0 (3.0 to 8.0)	1.0 (1.0 to 3.0)	<0.001
Daysinhospital	Median (IQR)	16.0 (14.0 to 17.0)	4.0 (2.0 to 15.0)	16.0 (14.0 to 17.0)	<0.001
Anycomorbidities	No	637 (60.6)	4 (20.0)	641 (59.8)	0.001
	Si	415 (39.4)	16 (80.0)	431 (40.2)	
Fever	NoFiebre	651 (84.5)	2 (22.2)	653 (83.8)	<0.001
	Fiebre	119 (15.5)	7 (77.8)	126 (16.2)	
Cough	No	766 (72.8)	12 (60.0)	778 (72.6)	0.308
	Si	286 (27.2)	8 (40.0)	294 (27.4)	
Systolicpressure	Median (IQR)	112.0 (110.0 to 120.0)	105.0 (100.0 to 130.0)	110.0 (110.0 to 120.0)	0.252
SPlowB	Alto	664 (98.8)	11 (78.6)	675 (98.4)	<0.001
	Bajo	8 (1.2)	3 (21.4)	11 (1.6)	
Respiratoryrate	Median (IQR)	18.0 (18.0 to 20.0)	23.0 (22.0 to 30.0)	18.0 (18.0 to 20.0)	<0.001

label	levels	No muere	Muere	Total	p
RRhighB	Bajo	678 (97.1)	8 (61.5)	686 (96.5)	<0.001
	Alto	20 (2.9)	5 (38.5)	25 (3.5)	
SpO2	Median (IQR)	98.0 (96.0 to 98.0)	86.0 (74.0 to 95.0)	97.0 (96.0 to 98.0)	<0.001
Whitebloodcells	Median (IQR)	6.3 (5.1 to 7.9)	7.0 (5.7 to 9.3)	6.3 (5.1 to 7.9)	0.086
WBClowB	Alto	910 (91.0)	18 (100.0)	928 (91.2)	0.361
	Bajo	90 (9.0)	0 (0.0)	90 (8.8)	
WBChighB	Bajo	925 (92.5)	14 (77.8)	939 (92.2)	0.062
	Alto	75 (7.5)	4 (22.2)	79 (7.8)	
NLR	Median (IQR)	2.1 (1.4 to 3.1)	2.4 (1.1 to 6.1)	2.1 (1.4 to 3.2)	0.751
NLRhighB	Bajo	384 (84.2)	7 (63.6)	391 (83.7)	0.158
	Alto	72 (15.8)	4 (36.4)	76 (16.3)	
Haemoglobin	Median (IQR)	137.0 (124.0 to 150.0)	131.0 (118.8 to 142.5)	136.5 (124.0 to 150.0)	0.200
Prothrombintime	Median (IQR)	13.0 (11.7 to 16.0)	15.6 (12.6 to 18.5)	13.0 (11.7 to 16.0)	0.063
Aspmintran	Median (IQR)	18.3 (12.4 to 25.9)	42.0 (26.0 to 70.0)	18.5 (12.5 to 26.5)	<0.001
Directbilirubin	Median (IQR)	2.6 (1.8 to 4.3)	10.0 (4.0 to 10.9)	2.7 (1.8 to 4.3)	0.012
Creatinine	Median (IQR)	73.0 (60.0 to 88.0)	87.0 (75.8 to 162.7)	73.1 (60.3 to 88.0)	0.002
Creactprot	Median (IQR)	2.0 (0.0 to 7.9)	8.3 (4.1 to 42.0)	2.3 (0.1 to 8.1)	0.077
Pneumonia	No	504 (69.8)	3 (20.0)	507 (68.8)	<0.001
	Si	218 (30.2)	12 (80.0)	230 (31.2)	
Bronchitis	No	535 (74.1)	14 (93.3)	549 (74.5)	0.164
	Si	187 (25.9)	1 (6.7)	188 (25.5)	
Antiviralmedications	No	313 (29.8)	10 (50.0)	323 (30.1)	0.087
	Si	739 (70.2)	10 (50.0)	749 (69.9)	
Antibiotics	No	753 (71.6)	11 (55.0)	764 (71.3)	0.170

label	levels	No muere	Muere	Total	p
	Si	299 (28.4)	9 (45.0)	308 (28.7)	
Anticoagulant	No	1028 (97.7)	15 (75.0)	1043 (97.3)	<0.001
	Si	24 (2.3)	5 (25.0)	29 (2.7)	
Corticosteroids	No	1037 (98.6)	19 (95.0)	1056 (98.5)	0.708
	Si	15 (1.4)	1 (5.0)	16 (1.5)	
Oxygentherapy	No	1008 (95.8)	0 (0.0)	1008 (94.0)	<0.001
	Si	44 (4.2)	20 (100.0)	64 (6.0)	
Mechanicalventilation	No	1043 (99.1)	2 (10.0)	1045 (97.5)	<0.001
	Si	9 (0.9)	18 (90.0)	27 (2.5)	

Las medidas de la Tabla 2.3 son calculadas para la población total y para cada una de los grupos de la variable dependiente `Deaths`.

Además aparece calculado el p-valor resultante de aplicar la Prueba de Wilcoxon-Mann-Whitney, la Prueba de Chi-Cuadrado (χ^2) o la Prueba Exacta de Fisher en las que se comprueban si hay diferencias significativas entre los grupos de las variables. Mirando los p-valores podemos decir que existen diferencias significativas entre los grupos de las variables `Age`, `Agegroup`, `Kazakhethnicity` entre otras. Sin embargo parece que no existen diferencias significativas entre los grupos de niveles bajos de glóbulos blancos (`WBClowB`) o entre los grupos de la variable `NLR`.

Si nos fijamos en enfermedades derivadas del COVID-19 como la neumonía o bronquitis vemos que la primera ha sido diagnosticada en el 80 % de las personas fallecidas y la segunda en el 6.7 %. Mirando este grupo de fallecidos, vemos que tener más de 65 años (`Agegroup`), ser hombre (`Sex`), tener patologías previas (`Anycomoridities`) o pertenecer a la etnia kazaja (`Kazakhethnicity`) podrían ser posibles factores de riesgo asociados a fallecer por COVID-19.

Para comprobar las asociaciones vamos a realizar una regresión logística simple y múltiple.

De neuvo la variable dependiente es `Deaths`.

```
dependent4 = "Deaths"


explanatory4 = c(
  "Age",
  "Sex",
  "Kazakhethnicity",
```

```
  "Anycomorbidities",
  "Whitebloodcells",
  "Platelets",
  "Fibrinogen",
  "Albumin",
  "AspmintranB",
  "DirectbilirubinB",
  "Glucose",
  "Bloodureanitrogen",
  "CreatinineB",
  "Sodium")



explanatorymulti4 = c(
  "Age",
  "Sex")

cohort19 %>% finalfit::finalfit(dependent4,
                                explanatory4,
                                explanatorymulti4) -> tb4

knitr::kable(tb4,
             caption = "\\label{table07}Regresión logística bivariada de
             factores asociados con la mortalidad por
             COVID-19 en Kazajstán.",
             align = "c")
```

Tabla 2.4: Regresión logística bivariada de factores asociados con la mortalidad por COVID-19 en Kazajstán.

	Dependent: Deaths		No muere	Muere	OR (univariable)	OR (multivariable)
1	Age	Mean (SD)	37.2 (17.9)	65.2 (20.1)	1.08 (1.05-1.11, p<0.001)	1.09 (1.06-1.13, p<0.001)

	Dependent: Deaths		No muere	Muere	OR (univariable)	OR (multivariable)
18	Sex	Mujer	582 (99.0)	6 (1.0)	-	-
17		Hombre	470 (97.1)	14 (2.9)	2.89 (1.15-8.21, p=0.031)	5.63 (2.06-17.57, p=0.001)
14	Kazakhethnicity	Kazaj	210 (95.5)	10 (4.5)	-	-
15		Otra	842 (98.8)	10 (1.2)	0.25 (0.10-0.62, p=0.002)	-
3	Anycomorbidities	No	637 (99.4)	4 (0.6)	-	-
4		Si	415 (96.3)	16 (3.7)	6.14 (2.23-21.55, p=0.001)	-
20	Whitebloodcells	Mean (SD)	6.6 (2.3)	8.0 (3.4)	1.22 (1.03-1.42, p=0.014)	-
16	Platelets	Mean (SD)	247.8 (81.0)	190.4 (77.4)	0.99 (0.98-1.00, p=0.002)	-
12	Fibrinogen	Mean (SD)	1641.8 (8088.0)	5.7 (2.6)	1.00 (NA-1.00, p=0.758)	-
2	Albumin	Mean (SD)	52.3 (246.5)	27.0 (16.6)	0.93 (0.88-0.99, p=0.004)	-
5	AspmintranB	0	996 (99.1)	9 (0.9)	-	-
6		1	56 (83.6)	11 (16.4)	21.74 (8.66-56.03, p<0.001)	-
10	DirectbilirubinB	0	990 (98.3)	17 (1.7)	-	-
11		1	62 (95.4)	3 (4.6)	2.82 (0.65-8.67, p=0.105)	-
13	Glucose	Mean (SD)	5.6 (4.1)	7.3 (3.7)	1.04 (0.97-1.09, p=0.132)	-
7	Bloodureanitrogen	Mean (SD)	203.5 (2925.6)	11.6 (10.4)	1.00 (NA-1.00, p=0.889)	-
8	CreatinineB	0	1015 (98.5)	15 (1.5)	-	-
9		1	37 (88.1)	5 (11.9)	9.14 (2.85-25.05, p<0.001)	-

	Dependent: Deaths		No muere	Muere	OR (univariable)	OR (multivariable)
19	Sodium	Mean (SD)	137.7 (21.1)	136.6 (4.2)	1.00 (0.98-1.06, p=0.884)	-

Mirando la Tabla 2.4, en el análisis de regresión logística simple, las variables significativas son la mayor edad (`Age`), ser hombre (`Sex`), pertenecer a otra etnia (`Kazakhethnicity`), tener patologías previas (`Anycomorbidities`), altos niveles de aspartato transaminasa (`AspmintranB`), creatinina (`CreatinineB`), glóbulos blancos en sangre (`Whitebloodcells`) y neutrófilos y linfocitos (`NLR`), bajos niveles de plaquetas (`Platelets`) y albúmina (`Albumin`) están asociados con mortalidad por COVID-19.

En el análisis múltiple los factores asociados son la edad alta (`Age`) y ser hombre (`Sex`).

En general, se puede concluir diciendo que los factores de riesgo asociados a morir o enfermar gravemente por COVID-19 son parecidos, entre ellos se encuentran la edad avanzada y pertenecer al sexo masculino, además de otros factores como no ser de etnia kazaja, tener patologías previas o niveles altos o bajos de proteínas y enzimas en sangre ya mencionadas en el estudio.

Capítulo 3

Estudios de caso-control

Los estudios de caso-control son estudios observacionales analíticos. Con el fin de cumplir su objetivo principal que es esclarecer la relación que hay entre el factor de exposición y la enfermedad de interés, la estrategia consiste en la selección de dos grupos de individuos, el grupo de casos (que presentan la enfermedad) y el grupo de controles (que no presentan la enfermedad).

Este estudio se hace de manera retrospectiva, por ejemplo, si lo que se quiere estudiar es la relación entre el consumo del alcohol y la aparición de cáncer de esófago, se medirá el consumo de alcohol en el pasado de ambos grupos, luego se calcularán las medidas pertinentes y, por último, se hará la comparación entre casos (enfermos de cáncer de esófago en la actualidad) y controles (no enfermos de cáncer de esófago en la actualidad) para esclarecer el grado de asociación entre el consumo de alcohol y el cáncer de esófago. Por este motivo, a estos estudios se les conocía como estudios retrospectivos, porque la información de la exposición se recaba posteriormente a la aparición de la enfermedad. Aún así, pueden existir estudios de caso-control prospectivos en los que se recoge la información de la exposición a la vez que se diagnostican los casos y se seleccionan los controles.

Una manera fácil de poder comprender el funcionamiento de este tipo de estudios es considerar la existencia de una población base, a partir de la cual obtenemos los casos y se muestrean personas con riesgo de padecer la enfermedad, que son los controles. Los casos aportan información de la exposición de personas enfermas y los controles la aportan sobre personas no enfermas (Royo-Bordonada et al., 2009).

Para poder trabajar con los datos de manera correcta, se crean tablas de contingencia en las que se compararán la proporción de individuos expuestos en casos y controles (Tabla 3.1), es decir, lo que realmente calculamos es el Odds Ratio (Ecuación 1.7) (Henquin, 2013).

Tabla 3.1: Tabla de contingencia estudios de caso control. Fuente: (Henquin, 2013).

	Grupo A (casos)	Grupo B (controles)
Expuestos	a	b
No expuestos	c	d
Total	a + c	b + d
Proporción de expuestos	$\frac{a}{a+c}$	$\frac{b}{b+d}$

Existen diferentes tipos de estudios de caso-control. Unos los abordaremos a lo largo de este capítulo y otros solo los mencionaremos ya que se estudiarán en profundidad en el capítulo siguiente porque son híbridos entre los estudios de cohortes y los estudios de caso-control.

El estudio de caso-control más básico es el que se ha ido explicando pero hay otros 3 tipos (Gallego Iborra et al., 2012):

- Estudios de caso-control emparejados: los controles se escogen en función de los casos. Un control debe ser igual a un caso en ciertas características (edad, sexo, nivel socioeconómico, etc.) que se determinan previamente, este es el proceso de emparejamiento o matching en el que se profundizará más en la Sección 3.3.1.
- Estudios de caso-cohorte: también se conocen como estudios de caso-control anidados. Nace de la indagación en la estrecha relación entre los estudios de cohorte y los de caso-control.
- Estudios caso-caso: también conocidos como estudios de caso-control alternantes o case-crossover. Cada caso es su propio control, el objetivo principal de este tipo de estudio es conocer si el sujeto observado estuvo haciendo algo inusual justo antes del inicio de la enfermedad de interés y, por ello, es necesario hacer la comparación del individuo con él mismo en el momento de exposición inmediato antes del desarrollo de la enfermedad y en momentos de exposición anteriores.

Los estudios de caso-control han sido de gran utilidad sobre todo para la investigación de la influencia de los factores de riesgo en enfermedades crónicas como son el cáncer, enfermedades cardiovasculares o algunas enfermedades pulmonares, aunque cada vez se usan más en estudios de enfermedades transmisibles por ejemplo (Olsen et al., 2010).

3.1. Diseño del estudio

Al igual que en el resto de estudios epidemiológicos se requiere una planificación clara, así como determinar los objetivos y las hipótesis que se van a estudiar, detallar la metodología del estudio o preparar diferentes opciones en el caso de que aparezcan problemas.

Dentro de estos pasos que se determinan para poder llevar a cabo un buen estudio se encuentra la definición de la población base. La población base se corresponde con la población de la que provienen los casos y, a partir de la cual han de obtenerse los controles (Gómez-Gómez et al., 2003).

Los controles deben representar a la población en la que se originan los casos. En el mecanismo del estudio si la población base es primaria se definirá primero la población a partir de la que se van a generar los casos, luego se seleccionan estos y de los individuos no enfermos pertenecientes a la población base se seleccionan los controles. Sin embargo, cuando los casos se escogen antes de definir la población base, más tarde se intentará identificar esta población de la que proceden para identificar a los controles, en este caso, la población base es secundaria.

Los estudios de caso-control de base primaria son los estudios de caso-cohorte o estudios de caso-control anidados que ya han sido definidos previamente. Aunque, en general el esquema siempre acaba siendo el mismo, tal y como aparece en la Figura 3.1.

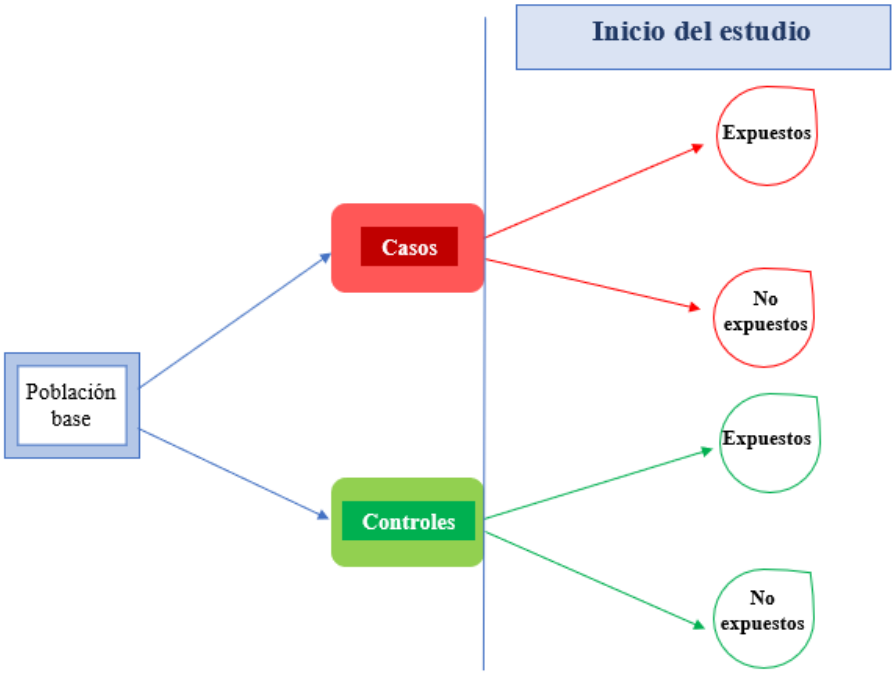


Figura 3.1: Esquema básico de los estudios de caso-control de base primaria.

Tanto en la selección de los casos como en la de los controles se deben cumplir los requisitos de representatividad, simultaneidad y homogeneidad. La representatividad indica que los casos deben representar a todos los casos que existieron en la población total en un tiempo determinado, y los controles deben representar en el tiempo a los sujetos posibles para convertirse en caso pero que no enfermaron. La simultaneidad implica que los controles deben escogerse al mismo tiempo que surgen los casos y, por último, homogeneidad significa que los controles han de pertenecer a la misma población base en la que surgen los casos e independientemente de la exposición bajo estudio (Lazcano-Ponce et al., 2001). En la Tabla 3.2 aparecen cuáles de estas características están presentes en los diferentes tipos de estudios de caso-control.

Tabla 3.2: Características de los distintos tipos de estudios caso-control. Fuente: (Lazcano-Ponce et al., 2001).

Tipo de estudio	Representatividad		Simultaneidad	Homogeneidad
	Casos	Controles		
Caso-cohorte	Sí	Sí	Asegurada	Definitiva
Caso-caso	Azar	Azar	Asegurada	Definitiva

Tanto la selección de los casos como los controles tienen su propio proceso, por ello hay que diferenciarlos, así como profundizar un poco en lo que ocurre en el tipo de estudio caso-control emparejado.

3.2. Selección de casos

Para entrar a formar parte del estudio como un caso, se deben de establecer una serie de criterios a través de pruebas diagnósticas o de laboratorio que verifiquen la existencia de la enfermedad bajo estudio en el individuo que nos interesa aunque a veces, la selección de los casos se hace mirando la historia clínica. En general, se recomienda que el grupo de casos seleccionados sea lo más homogéneo posible.

La definición de caso afecta a la validez del estudio, al tamaño de la muestra y a la precisión de los estadísticos empleados para evaluar la asociación entre el factor de exposición y la enfermedad bajo estudio. Si la sensibilidad es baja (criterios de clasificación de casos muy estrictos) se reducirá

el tamaño de muestra y una baja especificidad (una definición más abierta, poco concreta) llevará a una mala clasificación y a una mayor presencia e influencia de sesgo, ya que se incluirán falsos positivos que harán que la asociación se diluya.

Las fuentes de selección de casos más comunes son hospitales, consultorios o cualquier institución en la que puedan haber registros con base poblacional.

Según la antigüedad del diagnóstico de los casos se pueden diferenciar en dos tipos: los casos incidentes y los casos prevalentes. Los casos incidentes son los que se han diagnosticado recientemente, mientras que los casos prevalentes son aquellos individuos que alguna vez han presentado la enfermedad incluyendo ya fallecidos. Se recomienda el uso de casos incidentes para evitar algunos sesgos como la Falacia de Neyman (véase Tabla 3.3), además las pruebas y diagnósticos pueden variar con el paso de los años y se restaría homogeneidad en los casos si se mezclan los dos tipos.

3.3. Selección de controles

Los controles deben cumplir los criterios de definición de caso excepto los referentes a la enfermedad (Royo-Bordonada et al., 2009) y, además, representar la frecuencia de los individuos con riesgo de contraer la enfermedad en la población de la que provienen los casos. Por tanto, la fuente de obtención de los controles es la misma que la de los casos, registros hospitalarios o de consultorios, y, en ocasiones, personas cercanas a los casos que tengan los mismos factores de riesgo aunque a veces conlleve un sesgo de información. Además, han de tener la oportunidad de ser contados como casos si la enfermedad aparece (Kestenbaum, 2009).

Es una de las partes más difíciles de los estudios de caso-control y, como a veces no es posible contar con una buena muestra de controles, se escogen dos, una hospitalaria (base secundaria) y otra procedente de la población general (base primaria) (Henquin, 2013).

En las ocasiones en las que el número de casos es muy limitado, un aumento del número de controles puede conllevar el aumento de la potencia del estudio para detectar las asociaciones. Aún así, cuando la potencia es baja, no conviene elevar el número de controles.

Para evitar factores de confusión, se suele recurrir a técnicas como la estratificación o el emparejamiento y, como se vio anteriormente esto da lugar a diferentes tipos de estudios de caso-control.

3.3.1. Emparejamiento

El emparejamiento es un proceso de reclutamiento de controles que son iguales a los casos en una o más posibles variables de confusión. Esta técnica lo que hace es aumentar la potencia y la precisión estadística. Hay que tener en cuenta que si se empareja por una determinada variable ya no se podrá estudiar la asociación de ésta con la enfermedad de interés porque su presencia en casos y controles ha quedado en igualdad por el método del emparejamiento. Es por esta razón por la que se debe limitar el emparejamiento a variables que son factores de riesgo consolidados.

Al realizar estudios con este método se aumenta mucho su coste porque hay que hacer una intensa búsqueda de individuos con exactamente las mismas características y eso conlleva labores de recogida de información mucho más dificultosas. En ocasiones conviene no hacer este método e invertir el tiempo de recogida de información de controles pareados en la búsqueda de más controles no pareados que aumenten la potencia del estudio.

El emparejamiento puede ser individual, es decir, para cada caso se escoge un control en exclusiva o emparejamiento de frecuencia que se hace eligiendo la misma proporción de controles por categorías de la variable de emparejamiento, por grupos de edad por ejemplo.

Esta técnica requiere un análisis estadístico más complejo para que sea completamente riguroso con los datos que se manejan.

3.4. Validez del estudio y sesgos

En la Tabla 3.3 aparecen los distintos tipos de sesgos que pueden darse en un estudio de caso-control.

Tabla 3.3: Tipos de sesgos en los estudios de caso-control.

Tipo de sesgo	Características
Sesgos de selección	
Berskon	Si los casos son personas hospitalizadas no representarán a la población general.
Falacia de Neyman	Ocurre cuando se usan casos prevalentes en lugar de incidentes y la exposición implica una mayor o menor supervivencia, el nivel de exposición de los casos estará sesgado.
Efecto del voluntario	El número de personas voluntarias no suele ser representativo de la población.

Tipo de sesgo	Características
Sesgos de selección	
Efecto del obrero sano	Las personas en edad laboral suelen estar más sanas que la población general.
Sesgo de pertenencia	Ocurre cuando entre los miembros del estudio surgen subgrupos por características que tengan en común y que estén relacionadas positiva o negativamente con la variable de exposición.
Sesgos de información	
Memoria	Los casos recuerdan mejor detalles de la exposición que los controles.
Entrevistador	El entrevistador pregunta con más detalle a los casos que a los controles.
Instrumento de medida erróneo	El instrumento de medida empleado para recoger información sobre la exposición no es el adecuado.
De mala clasificación	Se produce cuando la calidad de la información recogida es distinta en el grupo de casos que en el de controles.
Sesgos de análisis	
Confusión	Se produce por la presencia de variables confundidoras.
Interacción	Se produce por la presencia de variables modificadoras de efecto.

3.5. Estudio de caso-control en R

El análisis adecuado y adaptado al tipo de datos o de estudio que se maneje es fundamental para poder obtener unos resultados buenos y fiables. Para los estudios de caso-control hay varios pasos a tener en cuenta para llevar a cabo el análisis y también hay una serie de objetivos que cumplir para poder extraer los resultados adecuados. En general, los pasos más importantes son la realización del análisis descriptivo y el cálculo de la Odds Ratio (Ecuación 1.7) y algunos riesgos atribuibles a través de tablas de contingencia. En los estudios en los que se ha realizado

la técnica del emparejamiento individual (Sección 3.3.1) es importante llevar a cabo otro tipo de análisis más adecuado para estos casos, que es la Regresión Logística Condicionada.

A lo largo de este apartado se desarrollará un análisis de datos de un estudio de caso-control en R donde se aplicarán las técnicas mencionadas en el párrafo anterior.

Con el fin de hacer una demostración del tratamiento estadístico que se realiza en este tipo de estudios a través del software R, se ha escogido el manual y ejercicio diseñado en 2013 por Martyn Kirk, Doctor en la Universidad Nacional de Australia. Toda la metodología y resultados han sido revisados por Tambri Housen y Alice Richardson en el año 2017 (Kirk, 2013).

Nos encontramos ante un estudio de caso-control para investigar la asociación entre un brote de infección por la bacteria Escherichia coli productora de la toxina Shiga con la exposición al entorno agrícola. Los datos recogidos pertenecen a habitantes de GreenCountry, un pueblo de Tasmania con unos 5500 habitantes. El brote ocurrió a principios del año 1998. Para identificar la relación entre la infección y la exposición al entorno agrícola el Departamento de Salud de Tasmania realizó un estudio de 25 casos y 25 controles. El estudio se realizó debido a las sospechas de los investigadores sobre la asociación, ya que la mayoría de los casos seleccionados vivían en zonas de granjas.

En este se pueden distinguir los dos procesos de pareamiento: por frecuencia e individual. El pareamiento por frecuencia se ha hecho por grupos de edad (3 grupos). Y el pareamiento individual se ha hecho por edad y sexo, como criterio para la edad se estableció una diferencia de $\pm$ 5 años entre el caso y el control asociado.

La mayoría de las órdenes que se recogen en las siguientes secciones han sido tomadas del código disponible en Rpubs para el ejercicio diseñado por Martyn Kirk (Kirk, 2013). Parte del código ha sido modificado para obtener mejores resultados. Además, todas las interpretaciones son de elaboración propia.

3.5.1. Lectura y preprocesamiento de los datos

Los datos se han obtenido en la plataforma GitHub, en la dirección web recogida en la orden R de lectura incluida abajo.

Primero realizamos la lectura del conjunto de datos, vemos sus dimensiones y las primeras observaciones y variables:

```
library(readr)
shigatoxin <- read_csv("https://raw.githubusercontent.com/
```

```
                        ArminderD/Case_control/master/
                        greencountry.csv",
                        col_types = cols(status = col_factor(),
                                         agematch = col_factor(),
                                         sex = col_factor(),
                                         diarrhoea = col_double(),
                                         diarrdate = col_date("%d/%m/%Y")))
# Convertirlo en data.frame (Importante para las funciones que
# usaremos más adelante)
shigatoxin <- as.data.frame(shigatoxin)
dim(shigatoxin) # 50 observaciones y 25 variables
```

```
## [1] 50 25
```

```
knitr::kable(shigatoxin[1:5, 1:10], caption = "\\label{table011}
Cabecera del conjunto de datos de caso control original.",
             align = "c")
```

Tabla 3.4: Cabecera del conjunto de datos de caso control original.

ID	status	matchid	agematch	age	sex	diarrhoea	diarrdate	vomit	stoolno
34	CASE	1	10-64	33	M	1	NA	0	20
1	CASE	5	0-9	0	F	1	1998-01-25	0	18
32	CASE	11	65+	70	M	1	1998-02-27	0	10
39	CASE	15	10-64	39	F	1	1998-01-30	0	20
26	CASE	22	65+	71	F	1	1998-04-12	1	NA

Vamos a conocer también la estructura del dataset:

```
## 'data.frame':    50 obs. of  25 variables:
##  $ ID        : num  34 1 32 39 26 48 15 20 30 28 ...
##  $ status    : Factor w/ 2 levels "CASE","CONTROL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ matchid   : num  1 5 11 15 22 24 2 3 4 6 ...
##  $ agematch  : Factor w/ 3 levels "10-64","0-9",..: 1 2 3 1 3 3 1 3 2 1 ...
##  $ age       : num  33 0 70 39 71 67 20 85 2 64 ...
##  $ sex       : Factor w/ 2 levels "M","F": 1 2 1 2 2 1 1 2 1 2 ...
##  $ diarrhoea : num  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ diarrdate : Date, format: NA "1998-01-25" ...
##  $ vomit     : num  0 0 0 0 1 0 1 0 1 0 ...
##  $ stoolno   : num  20 18 10 20 NA 5 3 8 20 10 ...
##  $ hus       : num  0 1 0 0 0 0 0 1 0 0 ...
##  $ duration  : num  4 5 3 7 5 8 2 28 5 4 ...
##  $ travel    : num  1 0 0 0 1 1 0 0 1 0 ...
##  $ animal    : num  1 0 0 1 0 0 0 0 0 0 ...
##  $ pets      : num  1 0 1 1 1 0 0 1 0 1 ...
##  $ townwat   : num  0 1 0 0 0 1 1 0 0 0 ...
##  $ watersport: num  0 1 0 1 0 1 0 0 1 0 ...
##  $ chicken   : num  0 0 0 0 1 0 1 1 1 0 ...
##  $ eggs      : num  0 1 1 1 1 1 1 1 1 1 ...
##  $ spinach   : num  0 0 1 1 1 1 1 1 0 0 ...
##  $ lettuce   : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ cucumber  : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ milkraw   : num  0 1 0 1 0 0 1 1 1 1 ...
##  $ farmlive  : num  0 1 0 0 1 0 0 1 1 1 ...
##  $ milkbottle: num  1 0 0 1 1 1 0 0 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ID = col_double(),
##   ..   status = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   ..   matchid = col_double(),
##   ..   agematch = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   ..   age = col_double(),
##   ..   sex = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   ..   diarrhoea = col_double(),
##   ..   diarrdate = col_date(format = "%d/%m/%Y"),
##   ..   vomit = col_double(),
##   ..   stoolno = col_double(),
##   ..   hus = col_double(),
##   ..   duration = col_double(),
##   ..   travel = col_double(),
##   ..   animal = col_double(),
```

```
##   ..   pets = col_double(),
##   ..   townwat = col_double(),
##   ..   watersport = col_double(),
##   ..   chicken = col_double(),
##   ..   eggs = col_double(),
##   ..   spinach = col_double(),
##   ..   lettuce = col_double(),
##   ..   cucumber = col_double(),
##   ..   milkraw = col_double(),
##   ..   farmlive = col_double(),
##   ..   milkbottle = col_double()
##   .. )
```

De este último resumen podemos sacar la Tabla 3.5 y así tener una idea de las variables con las que trabajamos para pensar en el tratamiento que hay que darles.

Tabla 3.5: Características de las variables recogidas en el estudio de caso-control de la infección por Escherichia Coli.

Nombre de la variable	Explicación	Tipo de variable
ID	Número de identificación	Numérico
status	Indica si es un CASO o un CONTROL	Factor: CASE (caso), CONTROL (control)
matchid	Número de identificación de los casos y controles pareados individualmente	Numérico
agematch	Grupos de edad para el pareado por frecuencia	Factor: 0-9, 10-64, 65+
age	Edad del sujeto observado	Numérico
sex	Sexo del sujeto observado	Factor: M (hombre), F (mujer)
diarrhoea	Registro de si los casos tuvieron diarrea	Numérico: 0 (no tuvo), 1 (sí tuvo)
diarrdate	Fecha en la que los casos tuvieron la diarrea	Fecha

Nombre de la variable	Explicación	Tipo de variable
vomit	Registro de si los casos vomitaron	Numérico: 0 (no vomitó), 1 (sí vomitó)
stoolno	Número máximo de deposiciones sueltas en 24 horas	Numérico
hus	Registro de si los casos desarrollaron síndrome urémico hemolítico	Numérico: 0 (no lo desarrolló), 1 (sí lo desarrolló)
duration	Duración de la enfermedad en los casos	Numérico
travel	Registro de si el sujeto observado viajó 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no viajó), 1 (sí viajó)
animal	Registro de si el sujeto observado ha estado en contacto con animales 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
pets	Registro de si el sujeto observado ha estado en contacto con mascotas 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
townwat	Registro de si el sujeto observado bebió agua del pueblo 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
watersport	Registro de si el sujeto observado hizo deportes de agua 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)

Nombre de la variable	Explicación	Tipo de variable
chicken	Registro de si el sujeto observado ingirió pollo 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
eggs	Registro de si el sujeto observado ingirió huevos 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
spinach	Registro de si el sujeto observado ingirió espinacas 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
lettuce	Registro de si el sujeto observado ingirió lechuga 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
cucumber	Registro de si el sujeto observado ingirió pepino 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
milkraw	Registro de si el sujeto observado ingirió leche cruda 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)
farmlive	Registro de si el sujeto observado vive en una granja	Numérico: 0 (no), 1 (sí)
milkbottle	Registro de si el sujeto observado ingirió leche embotellada 7 días antes de caer enfermo o ser entrevistado	Numérico: 0 (no), 1 (sí)

A continuación, vamos a crear una variable nueva de tipo factor cuyo nombre será `case` y guardará 0 o 1 en función de si el sujeto es un control o un caso, respectivamente. También

vamos a convertir en factor todas las variables cuyos valores numéricos sean 0 o 1 y le daremos la etiqueta No o Sí, respectivamente.

```
# Primero condicionamos 0 para control y 1 para caso
shigatoxin$case <- ifelse(shigatoxin$status == "CONTROL", 0, 1)

# La convertimos en factor con las correspondientes etiquetas (caso o control)
# según sea 1 o 0
shigatoxin$case <- factor(shigatoxin$case, levels = c(0, 1),
                          labels = c("Control", "Caso"))

# Las variables cuyos valores son 0 o 1 son las variables 7, 9 y de la 11 a la 25
shigatoxin[c(7, 9, 11:25)] <- lapply(shigatoxin[c(7, 9, 11:25)], factor,
                                     levels = c(0, 1),
                                     labels = c("No", "Si"))

# Damos las correspondientes etiquetas a la variable agematch
shigatoxin$agematch <- factor(shigatoxin$agematch,
                              levels = c("0-9", "10-64", "65+"),
                              labels = c("0-9", "10-64", "65 o más"))
```

3.5.2. Análisis descriptivo

En los estudios de caso-control es importante hacer un buen análisis descriptivo para asegurarnos que los casos y los controles son similares.

Se puede hacer la comparación entre casos y controles creando tablas de proporciones que nos dejen comprobar qué proporción de casos y controles cumplieron algunas de las características de interés como por ejemplo si tuvieron vómitos, diarrea o síndrome urémico hemolítico.

```
# Vamos a comparar las talbas de proporciones que resultan de los síntomas
# más graves y significativos de la infección

knitr::kable(prop.table(table(shigatoxin$vomit, shigatoxin$case)),
             caption = "\\label{table013}
             Tabla de frecuencias relativas
             con respecto a la variable
```

```
             vomit en casos y controles.",
             align = "c")
```

Tabla 3.6: Tabla de frecuencias relativas con respecto a la variable vomit en casos y controles.

	Control	Caso
No	0.5	0.26
Si	0.0	0.24

```
knitr::kable(prop.table(table(shigatoxin$diarrhoea, shigatoxin$case)),
             caption = "\\label{table014}
             Tabla de frecuencias relativas
             con respecto a la variable
             diarrhoea en casos y controles.",
             align = "c")
```

Tabla 3.7: Tabla de frecuencias relativas con respecto a la variable diarrhoea en casos y controles.

	Control	Caso
No	0.5	0.0
Si	0.0	0.5

```
knitr::kable(prop.table(table(shigatoxin$hus, shigatoxin$case)),
             caption = "\\label{table015}
             Tabla de frecuencias relativas
             con respecto a la variable
             hus en casos y controles.",
             align = "c")
```

Tabla 3.8: Tabla de frecuencias relativas con respecto a la variable hus en casos y controles.

	Control	Caso
No	0.5	0.42
Si	0.0	0.08

Los controles no han presentado ningún síntoma que se crea asociado con la infección, sin

embargo, en los casos si se han manifestado todos estos síntomas. La diarrea ha estado presente en todos los casos, seguida de los vómitos que los han presentado algo menos de la mitad de los casos y por último, una minoría de casos ha llegado a tener el síndrome urémico hemolítico.

```
library(ggplot2)
ggplot(shigatoxin, aes(y = sex, color = agematch, fill = agematch)) +
  geom_bar() +
  facet_grid(agematch~case)  +
  labs(title = "Comparación de sexo y grupos de edad en casos y controles")
```

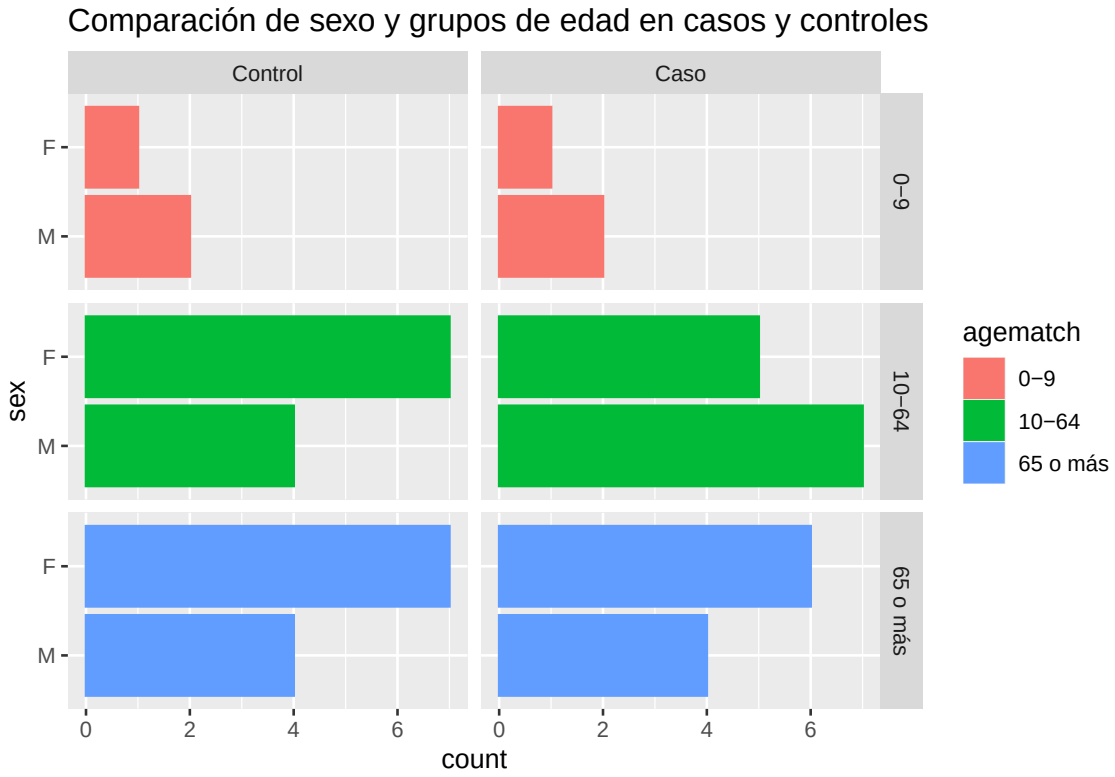


Figura 3.2: Gráfico comparador de variables de emparejamiento en casos y controles.

Como se puede apreciar en el gráfico se han incluido individuos al estudio con características muy similares en cuanto a edad y sexo, por tanto nos está dejando ver que el emparejamiento está bien hecho. Si nos fijamos este, el número de mujeres de 0 a 9 años incluidas en el estudio es parecido en casos y controles, y esto mismo ocurre en los hombres pertenecientes al mismo grupo de edad. En el grupo de 10 a 64 años hay más disparidad entre casos y controles pero aún así son muy similares, y en el grupo de 65 años o más ocurre lo mismo que en el primero, el número de hombres y mujeres seleccionados dentro de este grupo de edad es muy similar en casos y controles.

Para comprobar que realmente es cierto lo que el gráfico nos está mostrando podríamos hacer el test de Chi-Cuadrado de independencia. Para ello hay que agrupar los datos en una tabla de frecuencias absolutas por sexo y papel que tienen en el estudio y hacer el test con la función `chisq.test`. El contraste de hipótesis que se quiere resolver es:

$$\begin{cases} H_0: & No\ hay\ diferencias\ significativas\ entre\ casos\ y\ controles\ según\ el\ sexo \\ H_1: & Hay\ diferencias\ significativas\ entre\ casos\ y\ controles\ según\ el\ sexo \end{cases}$$

```
chisq.test(table(shigatoxin$sex, shigatoxin$case))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(shigatoxin$sex, shigatoxin$case)
## X-squared = 0.32206, df = 1, p-value = 0.5704
```

El p-valor obtenido es 0.57, y por tanto, podemos concluir diciendo que no existen diferencias significativas entre los casos y controles según el sexo, es decir, se encuentran en la misma proporción en el estudio, lo que nos puede asegurar que a la hora de emparejar los casos todos van a tener un control con características de sexo muy similares.

Podemos hacer lo mismo con la edad pero aplicando el Test Exacto de Fisher ya que hay algunas frecuencias menores a 5. El contraste a resolver es:

$$\begin{cases} H_0: & No\ hay\ diferencias\ significativas\ entre\ casos\ y\ controles\ según\ los\ grupos\ de\ edad \\ H_1: & Hay\ diferencias\ significativas\ entre\ casos\ y\ controles\ según\ los\ grupos\ de\ edad \end{cases}$$

```
fisher.test(table(shigatoxin$agematch, shigatoxin$case))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(shigatoxin$agematch, shigatoxin$case)
## p-value = 1
## alternative hypothesis: two.sided
```

Obtenemos un p-valor igual a 1 y concluimos al igual que en el caso del sexo, no existen diferencias significativas entre casos y controles con respecto a la edad.

Ya podemos pasar a hacer un análisis más preciso sobre la asociación entre el evento de interés (variable `case`) y algún factor de exposición (variable `travel` entre otras).

3.5.3. Análisis inferencial y cálculo de medidas de asociación

El cálculo de la odds ratio es una de las partes más importantes del análisis de los estudios de caso-control. Para su cálculo es importante crear tablas de contingencia y de ahí sacar su valor como ya se explicó en la Sección 1.4.2.2.

Dado que en este estudio hay varias variables que son posibles factores de riesgo podríamos plantear hacer un análisis univariante de algunas variables con la variable evento final y, después hacer un análisis multivariante con todas las variables que son consideradas de riesgo y así comprobar su efecto sobre la variable evento final.

3.5.3.1. Análisis univariante

Una de las variables que nos puede hacer sospechar más sobre la asociación directa con la infección es la variable `travel`. Podemos asociar el haber viajado con el contraer la infección, es por ello por lo que vamos a hacer el primer análisis para estudiar la asociación de la variable `travel` con la variable `case`.

Es importante tener en cuenta que estamos trabajando con un estudio emparejado por frecuencias de edad, por tanto las tablas de contingencia deberían de hacerse teniendo en cuenta este método.

Para realizar la tabla de contingencia a partir de tres variables, volvemos a usar la función `table`.

```
# Tabla de contingencia de 3 vías
table(shigatoxin$travel, shigatoxin$case, shigatoxin$agematch)
```

```
## , ,  = 0-9
##
##
##      Control Caso
##   No       2    2
##   Si       1    1
##
## , ,  = 10-64
##
```

```
##
##      Control Caso
##   No       5    8
##   Si       6    4
##
## , ,  = 65 o más
##
##
##      Control Caso
##   No       9    8
##   Si       2    2
```

Como vemos salen tres tablas de contingencia separadas según el grupo de edad, por tanto, estamos seguros de estar manteniendo la técnica de emparejamiento por frecuencia intacta a la hora de dar resultados.

Vamos a pasar al cálculo de la odds ratio (Ecuación 1.7), para ello usaremos la librería `epiR` (Stevenson et al., 2021) que nos proporciona algunas funciones como `epi.2by2` para realizar un buen análisis epidemiológico. Esta función hace un resumen de las medidas de asociación así como su cálculo e intervalo de confianza en crudo y por el método ajustado de Mantel-Haenszel así como el cálculo del valor de la prueba Chi-Cuadrado de Homogeneidad. Si pasamos como argumento `method = "case.control` devolverá el resumen adaptado al estudio de caso-control.

```
library(epiR)
epicc <- epi.2by2(table(shigatoxin$travel, shigatoxin$case,
                        shigatoxin$agematch),
         method = "case.control")
```

```
# Tabla con valores de prevalencia y odds ratio
knitr::kable(epicc[["tab"]], caption = "\\label{table016}
Prevalencia y odds de la asociación
entre travel y case.",
             align = "c")
```

Tabla 3.9: Prevalencia y odds de la asociación entre travel y case.

	Outcome +	Outcome -	Total	Prevalence *	Odds
Exposed +	16	18	34	47.1	0.889
Exposed -	9	7	16	56.2	1.286
Total	25	25	50	50.0	1.000

En esta tabla vemos un resumen que se ha producido por las tres tablas de contingencia generadas anteriormente con la función `table` y teniendo en cuenta a la variable `agematch`.

De aquí podemos interpretar medidas como la prevalencia a la exposición. Para calcular esta medida lo que se hace es:

$$\text{Prevalencia a la exposición} = \frac{\text{Número de casos expuestos}}{\text{Número total de expuestos}} = \frac{16}{34} = 0.4705$$

Esto significa que la prevalencia a la exposición es del 47.1 % en el momento en el que se hizo la entrevista para este estudio de casos y controles. El cálculo para los individuos no expuestos se hace de manera análoga. La prevalencia de la infección por E. Coli es del 50 % en el momento en el que se hizo la encuesta, en este caso sí se hace el cociente entre el número de casos y el número total de participantes en el estudio.

En la columna denominada Odds los dos primeros valores nos ofrecen la Odss de expuestos y la Odds de no expuestos, respectivamente. El cálculo se hace de la siguiente forma:

$$\text{Odds de expuestos} = \frac{\text{Número de casos expuestos}}{\text{Número total de expuestos}} = \frac{16}{18} = 0.889$$

$$\text{Odds de no expuestos} = \frac{\text{Número de casos no expuestos}}{\text{Número total de no expuestos}} = \frac{9}{7} = 1.286$$

La interpretación de la Odds de expuestos es que las personas que han viajado tienen un 0.889 menos de posibilidades de convertirse en casos.

La interpretación de la Odds en no expuestos es que las personas que no han viajado tienen un 0.286 más de posibilidades de convertirse en casos.

```
# Tabla resumen con valores crudos y ajustados por M-H y los IC
knitr::kable(epicc[["massoc.summary"]],
             digits = 3,
```

```
            caption = "\\label{table017}Tabla resumen
            de la asociación entre travel y case.",
            align = "c")
```

Tabla 3.10: Tabla resumen de la asociación entre travel y case.

var	est	lower	upper
Odds ratio (crude)	0.691	0.209	2.285
Odds ratio (M-H)	0.647	0.190	2.205
Odds ratio (crude:M-H)	1.068	NA	NA
Attrib fraction (est) in exposed (crude %)	-0.446	-3.659	0.552
Attrib fraction (est) in population (crude %) *	-0.286	-1.437	0.192

En la segunda tabla nos aparece el cálculo de Odds Ratio en crudo, además del ajustado por Mantel-Haesnzel y del híbrido entre ambos métodos.

La Odds Ratio en crudo se calcula a partir de la tabla de contingencia anterior y como se explicó en la Sección 1.4.2.2.

$$\text{OR} = \frac{\frac{16}{9}}{\frac{18}{7}} = 0.691$$

La interpretación que podemos darle es que si no tenemos en cuenta la variable que se ha usado para el emparejamiento `agematch`, la odds de exposición entre los casos fue 0.691 veces menor que la odds de exposición entre los controles. Este medida además se da con un intervalo de confianza al 95 % cuyos extremos superior e inferior son (0.21, 0.285).

La Odds Ratio calculada por Mantel-Haenszel devuelve el valor y el intervalo de confianza teniendo en cuenta a la variable emparejadora `agematch` y, por tanto, si existen diferencias entre el OR crudo con el OR ajustado por M-H o con el híbrido podría indicar que la edad (`agematch`) es un posible factor de confusión en este estudio, por tanto, haya sido correcto el emparejamiento por grupos edad.

Con todo esto, podemos decir que el haber viajado (variable `travel`) parece que no está directamente asociado con el haber contraido la infección ya que el valor de la Odds Ratio es menor a 1.

Las pruebas que se hacen a continuación son para contrastar el efecto de la variable `agematch` sobre el valor de la Odds Ratio cuando se tiene en cuenta esta variable y cuando no. El contraste

que va a resolverse a un nivel de confianza del 95 % es:

$$\begin{cases} H_0: & OR_{estratificado} = OR_{no\ estratificado} \\ H_1: & OR_{estratificado} \neq OR_{no\ estratificado} \end{cases}$$

```
# Prueba de M-H de Homogeneidad
knitr::kable(epicc[["massoc.detail"]][["chi2.mh"]],
             digits = 3,
             caption = "\\label{table018}Prueba de Mantel-Haenszel
             de Homogeneidad del OR",
             align = "c")
```

Tabla 3.11: Prueba de Mantel-Haenszel de Homogeneidad del OR

test.statistic	df	p.value.1s	p.value.2s
0.463	1	0.248	0.496

```
# Prueba de M-H del OR ajustado a 1
knitr::kable(epicc[["massoc.detail"]][["OR.homog.woolf"]],
             digits = 3,
             caption = "\\label{table019}Prueba de Mantel-Haenszel
             del OR ajustado a 1",
             align = "c")
```

Tabla 3.12: Prueba de Mantel-Haenszel del OR ajustado a 1

test.statistic	df	p.value
0.568	2	0.753

Por el valor de los p-valores, podemos concluir diciendo que parece que la variable `agematch` no es una variable confundidora aunque sí hay que tenerla en cuenta porque es con la que se ha hecho el emparejamiento por frecuencias.

3.5.3.2. Análisis multivariante

Como en este estudio hay varios factores de exposición, conviene hacer un análisis multivariante para poder buscar la asociación entre cada factor con la variable `case`. Podríamos crear un bucle

para aplicar la función `epi.2by2` a todas las variables que nos quedan por investigar pero existe una función llamada `cctable` de la librería `EpiStats` (Decorps, 2021) que devuelve una tabla resumen con el valor de la Odds y la Prueba exacta de Fisher para cada factor de exposición.

En este primer momento, no se va a tener en cuenta a la variable de emparejamiento `agematch`, luego cuando saquemos claramente las variables asociadas con la infección se estratificará para obtener resultados del todo fiables. Además, en el apartado anterior hemos visto que no parece ser un factor de confusión por lo que no corremos el riesgo de que pueda llegar a invertir la asociación existente entre cada factor de exposición y la variable `case`.

```
library(EpiStats)
epistat <- cctable(shigatoxin, "case",
                   exposure = colnames(shigatoxin[, c(13:25)]),
                   exact = TRUE)
knitr::kable(epistat[["df"]], caption = "\\label{table020}
Tabla análisis multivariante estudio caso-control", align = "c")
```

Tabla 3.13: Tabla análisis multivariante estudio caso-control

	Tot.Cases	Exposed	%	Tot.Ctrls	Exposed	%	OR	CI ll	CI ul	p(Fisher)
milkraw	25	21	84.00	25	5	20.00	21.00	4.16	117.47	0.000
farmlive	25	17	68.00	25	8	32.00	4.52	1.19	17.60	0.023
watersport	25	9	36.00	25	5	20.00	2.25	0.54	10.21	0.345
milkbottle	25	21	84.00	25	18	72.00	2.04	0.43	10.98	0.496
townwat	25	8	32.00	25	6	24.00	1.49	0.36	6.34	0.754
travel	25	7	28.00	25	9	36.00	0.69	0.17	2.67	0.762
pets	25	15	60.00	25	17	68.00	0.71	0.19	2.61	0.769
chicken	25	12	48.00	25	10	40.00	1.38	0.39	4.92	0.776
animal	25	6	24.00	25	7	28.00	0.81	0.19	3.46	1.000
eggs	25	19	76.00	25	18	72.00	1.23	0.29	5.37	1.000
spinach	25	12	48.00	25	11	44.00	1.17	0.34	4.12	1.000
lettuce	25	20	80.00	25	20	80.00	1.00	0.20	5.10	1.000
cucumber	25	21	84.00	25	22	88.00	0.72	0.09	4.83	1.000

Para obtener conclusiones de esta tabla es conveniente mirar el valor de la Odds Ratio (OR) y el p-valor que resulta de hacer la prueba de Fisher. La prueba de Fisher resuelve el contraste

siguiente:

$$\begin{cases} H_0: & El\ factor\ de\ exposición\ y\ enfermar\ por\ la\ infección\ son\ independientes.\ No\ hay\ relación. \\ H_1: & El\ factor\ de\ exposición\ y\ enfermar\ por\ la\ infección\ son\ dependientes.\ Hay\ relación. \end{cases}$$

Si nos fijamos en estas pruebas, solo dos variables son las que están relacionadas con infectarse por E. Coli. Las variables son `milkraw` (consumo de leche cruda) y `farmlive` (vivir en una granja).

Para la variable `milkraw` el valor de la Odds Ratio es igual a 21 lo que implica una fuerte asociación con la variable `case` y, además, al hacer la prueba de Fisher se confirma esta sospecha ya que se obtiene un p-valor 0 que implica dependencia entre las variables. Lo mismo ocurre con la variable `farmlive` cuyo OR es 4.52.

Hay otras variables que sí están asociadas a la infección por el valor de la Odds Ratio, pero sin embargo el p-valor que se obtiene por Fisher nos obliga a no tenerlas en cuenta para el paso siguiente.

A continuación, hay que estudiar más en profundidad la asociación entre estas dos variables con la variable `case`. Para este paso, vamos a comprobar si alguna de las variables `milkraw` o `farmlive` son confundidores. Lo que haremos será tomar una de las variables para el estudio de la asociación y otra como variable estratificadora.

- Asociación entre `milkraw` y `case`, estratificando por `farmlive`:

```
t1 <- xtabs(~milkraw+case+farmlive, data = shigatoxin)
ftable(t1)
```

```
##                 farmlive No Si
## milkraw case
## No      Control          14  6
##         Caso              3  1
## Si      Control           3  2
##         Caso              5 16
```

Pasamos a hacer el test Chi-Cuadrado de independencia entre las variables:

```
summary(t1)
```

```
## Call: xtabs(formula = ~milkraw + case + farmlive, data = shigatoxin)
## Number of cases in table: 50
```

```
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 35.56, df = 4, p-value = 3.557e-07
```

Según el p-valor existe independencia entre los tres factores. Podemos pasar ya a estudiar la asociación.

```
t11 <- epi.2by2(t1, method = "case.control")
knitr::kable(t11[["tab"]],
             digits = 3,
             align = "c",
             caption = "\\label{table021}
             Prevalencia y Odds Ratio de la
             asociación entre milkraw y case estratificando
             por farmlive")
```

Tabla 3.14: Prevalencia y Odds Ratio de la asociación entre milkraw y case estratificando por farmlive

	Outcome +	Outcome -	Total	Prevalence *	Odds
Exposed +	20	4	24	83.3	5.000
Exposed -	5	21	26	19.2	0.238
Total	25	25	50	50.0	1.000

```
knitr::kable(t11[["massoc.summary"]],
             digits = 3,
             align = "c",
             caption = "\\label{table022}
             Tabla resumen de la asociación
             entre milkraw y case estratificando con farmlive")
```

Tabla 3.15: Tabla resumen de la asociación entre milkraw y case estratificando con farmlive

var	est	lower	upper
Odds ratio (crude)	21.000	4.924	89.561
Odds ratio (M-H)	15.091	3.467	65.686
Odds ratio (crude:M-H)	1.392	NA	NA

var	est	lower	upper
Attrib fraction (est) in exposed (crude %)	0.952	0.803	0.988
Attrib fraction (est) in population (crude %) *	0.762	0.702	0.929

```
# Prueba de M-H de Homogeneidad
knitr::kable(t11[["massoc.detail"]][["chi2.mh"]],
             digits = 3,
             caption = "\\label{table023}Prueba de Mantel-Haenszel
             de Homogeneidad del OR",
             align = "c")
```

Tabla 3.16: Prueba de Mantel-Haenszel de Homogeneidad del OR

test.statistic	df	p.value.1s	p.value.2s
16.182	1	0	0

```
# Prueba de M-H del OR ajustado a 1
knitr::kable(t11[["massoc.detail"]][["OR.homog.woolf"]],
             digits = 3,
             caption = "\\label{table024}Prueba de Mantel-Haenszel
             del OR ajustado a 1",
             align = "c")
```

Tabla 3.17: Prueba de Mantel-Haenszel del OR ajustado a 1

test.statistic	df	p.value
1.041	1	0.307

- Asociación entre `farmlive` y `case` estratificando por `milkraw`:

```
##                 milkraw No Si
## farmlive case
## No       Control         14  3
##          Caso             3  5
## Si       Control          6  2
##          Caso             1 16
```

Test Chi-Cuadrado de independencia:

```
## Call: xtabs(formula = ~farmlive + case + milkraw, data = shigatoxin)
## Number of cases in table: 50
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 35.56, df = 4, p-value = 3.557e-07
```

Los tres factores son independientes.

Tabla 3.18: Prevalencia y Odds Ratio de laasociación entre farmlive y case estratificando por milkraw

	Outcome +	Outcome -	Total	Prevalence *	Odds
Exposed +	17	8	25	68	2.125
Exposed -	8	17	25	32	0.471
Total	25	25	50	50	1.000

```
knitr::kable(t22[["massoc.summary"]],
             digits = 3,
             align = "c",
             caption = "\\label{table026}Tabla resumen de
             la asociación entre farmlive y case
             estratificando con milkraw")
```

Tabla 3.19: Tabla resumen de la asociación entre farmlive y case estratificando con milkraw

var	est	lower	upper
Odds ratio (crude)	4.516	1.376	14.820
Odds ratio (M-H)	2.141	0.494	9.288
Odds ratio (crude:M-H)	2.109	NA	NA
Attrib fraction (est) in exposed (crude %)	0.779	0.285	0.931
Attrib fraction (est) in population (crude %) *	0.529	0.391	0.661

```
# Prueba de M-H de Homogeneidad
knitr::kable(t22[["massoc.detail"]][["chi2.mh"]],
             digits = 3,
```

```
            caption = "\\label{table027}Prueba de Mantel-Haenszel
            de Homogeneidad del OR",
            align = "c")
```

Tabla 3.20: Prueba de Mantel-Haenszel de Homogeneidad del OR

test.statistic	df	p.value.1s	p.value.2s
1.039	1	0.154	0.308

```
# Prueba de M-H del OR ajustado a 1
knitr::kable(t22[["massoc.detail"]][["OR.homog.woolf"]],
            digits = 3,
            caption = "\\label{table028}Prueba de Mantel-Haenszel
            del OR ajustado a 1",
            align = "c")
```

Tabla 3.21: Prueba de Mantel-Haenszel del OR ajustado a 1

test.statistic	df	p.value
1.041	1	0.307

Si nos fijamos, cuando estudiamos la asociación entre `milkraw` y `case` usando con variable estratificadora `farmlive`, existe diferencia entre el OR crudo y el OR combinado con Mantel-Haenszel. Esto nos indica que `farmlive` es una variable confundidora en el estudio de la asociación entre el consumo de leche cruda y caer enfermo por la infección.

La asociación entre las variables `case` y `milkraw` es fuerte, y por tanto, se podría determinar que el consumo de leche cruda es un factor de riesgo para contraer la infección por la bacteria Escherichia Coli.

La asociación entre las variables `case` y `farmlive` es fuerte por tanto, vivir en una granja es un factor de riesgo para contraer la infección por la bacteria Escherichia Coli. No hay diferencias significativas entre el OR estratificado (OR híbrido) y el no estratificado (OR crudo), por tanto, la variable `milkraw` no parece ser un factor de confusión.

3.5.4. Análisis del estudio a partir del emparejamiento individual

La variable `matchid` guarda el número de identificación de cada caso con su control, es decir, es la variable a tener en cuenta cuando vayamos a hacer un análisis de la asociación entre un factor de exposición y la enfermedad si queremos trabajar con un estudio en el que los casos y controles están pareados individualmente.

En estos casos, lo que se hace es un análisis de Regresión Logística Condicionada. Las tablas de contingencia se disponen de manera diferente, tal y como aparece en la tabla 3.22.

Tabla 3.22: Tabla de contingencia en estudios caso-control pareados individualmente.

		Controles		
		Expuestos	No expuestos	
Casos	Expuestos	e	f	e + f
	No expuestos	g	h	g + h
		e + g	f + h	e + f + g + h

Donde la Odds Ratio ahora es:

$$OR = \frac{f}{g}$$

La Regresión Logística Condicionada es una técnica que se aplica cuando disponemos de un conjunto de datos en el que los individuos se encuentran emparejados. El objetivo principal es estimar la asociación de la variable dependiente con la exposición en cada estrato y controlar los posibles factores de confusión (Pallarés Mestre, 2016). La ventaja principal es el aumento de la eficiencia predictiva, mientras que la mayor desventaja es el coste de establecer los emparejamientos y la pérdida de información por descartar algunos controles que no son válidos para emparejar con un caso (Pallarés Mestre, 2016).

En este procedimiento se hace la estimación de la máxima verosimilitud condicional, que solo estima los efectos de las variables de interés y no de las variables que definen los estratos (Pallarés Mestre, 2016).

Vamos a realizar la Regresión Logística Condicionada sobre las variables `case` y `milkraw`. En este caso, como el pareamiento se hace de manera individual, el resto de variables van a estar controladas en el caso de ser confundidoras (como ocurría con `farmlive`). Para poder aplicar esta técnica es necesario que la variable objetivo `case` sea de tipo numérico y la variable de emparejamiento individual `matchid` sea de tipo factor.

```r
shigatoxin$case <- as.numeric(shigatoxin$case)


shigatoxin$matchid <- as.factor(shigatoxin$matchid)
```

Ahora para hacer la regresión logística vamos a cargar la librería `survival` (Therneau, 2020) y a usar la función `clogit`. Le pasaremos la fórmula de la regresión condicionada que vamos a hacer dando como variable de estratificación la variable `matchid`.

```r
library(survival)
logit <- clogit(case ~ milkraw + strata(matchid), data = shigatoxin)
summary(logit)
```

```
## Call:
## coxph(formula = Surv(rep(1, 50L), case) ~ milkraw + strata(matchid),
##     data = shigatoxin, method = "exact")
##
##   n= 50, number of events= 25
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
## milkrawSi  2.833    17.000    1.029 2.753   0.0059 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## milkrawSi        17    0.05882     2.262     127.7
##
## Concordance= 0.82  (se = 0.079 )
## Likelihood ratio test= 17.23  on 1 df,   p=3e-05
## Wald test            = 7.58  on 1 df,   p=0.006
## Score (logrank) test = 14.22  on 1 df,   p=2e-04
```

La Odds Ratio en esta ocasión es 17, lo que significa que si una persona consume leche cruda la probabilidad de convertirse en un caso (enfermo) se multiplican por 17. El p-valor 0.0059 es resultado de hacer el contraste de significación de la variable `milkraw` en el estudio, al ser menor a 0.05 (nivel de significación) no podemos considerar nulo el consumir leche cruda, y por tanto, si podemos analizar su OR.

Este análisis se puede hacer con más variables. Vamos a hacerlo ahora con las variables `farmlive`, `milkbottle` y `watersport` que son las que tienen un OR más grande en la tabla del análisis multivariante.

```
logit <- clogit(case~ milkbottle + watersport + farmlive + strata(matchid),
                data = shigatoxin)
summary(logit)
```

```
## Call:
## coxph(formula = Surv(rep(1, 50L), case) ~ milkbottle + watersport +
##     farmlive + strata(matchid), data = shigatoxin, method = "exact")
##
##   n= 50, number of events= 25
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## milkbottleSi 1.3989    4.0508   1.1871 1.178   0.2386
## watersportSi 0.9107    2.4862   0.8059 1.130   0.2584
## farmliveSi   1.3255    3.7641   0.6688 1.982   0.0475 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## milkbottleSi     4.051     0.2469    0.3954     41.50
## watersportSi     2.486     0.4022    0.5124     12.06
## farmliveSi       3.764     0.2657    1.0147     13.96
##
## Concordance= 0.7  (se = 0.12 )
## Likelihood ratio test= 8.51  on 3 df,   p=0.04
## Wald test            = 5.84  on 3 df,   p=0.1
## Score (logrank) test = 7.41  on 3 df,   p=0.06
```

En este caso, las variables `milkbottle` y `watersport` no son significativas y no es necesario valorar su OR. La variable `farmlive` si es significativa aunque tenga un p-valor cercano a 0.05 (si subiéramos el nivel de significación a 0.1 dejaría de ser significativa) y su OR nos dice que si una persona vive en una granja la probabilidad de convertirse en un caso (enfermo) es 3.76 veces mayor.

Como vemos, por el análisis del estudio pareado individualmente hemos llegado a las mismas conclusiones que con el análisis pareado por frecuencias de edad y además, hemos controlado las variables confundidoras.

Capítulo 4

Estudios de cohortes frente a estudios de caso-control

Tras la introducción y profundización en los estudios de cohortes y de caso-control se pueden analizar algunas de las ventajas e inconvenientes de los estudios observacionales.

De ellos, se puede destacar su practicidad o facilidad para realizarlos ya que el sujeto solo ha de participar en las entrevistas que se le realicen y el investigador solo tiene que observar los sucesos que se van desencadenando ya que no debe alterar la exposición. También los resultados que se obtienen se pueden extrapolar a la población general, lo que es de gran ayuda para establecer metas de salud pública.

Una de las principales desventajas de estos estudios es el poco control que existe sobre los factores de confusión si no se usan técnicas para controlarlos como, por ejemplo, la estratificación o el emparejamiento en los estudios de caso-control. Además, este tipo de estudios cuentan con la debilidad de que es prácticamente imposible repetirlos como experimento, por ello resulta complicado poder hacer comparaciones de estudios observacionales entre sí.

Los estudios de cohortes y caso-control tienen algunas características en común (como las que acabamos de mencionar ya que son estudios observacionales) pero también poseen características que los diferencian y hacen que sean preferibles para según qué situaciones. A continuación vamos a estudiar las ventajas (en verde) y desventajas (en rojo) de los estudios de cohortes frente a los estudios de caso-control en la Tabla 4.1.

Tabla 4.1: Ventajas y desventajas de los estudios de cohortes frente a caso-control

Cohortes	Caso-control
Estiman la incidencia	No estiman la incidencia directamente, solo devuelven medidas de efecto
Son más costosos	Son ligeramente menos costosos y más sencillos desde el punto de vista logístico
Posibilitan la medición de la exposición de forma repetida en el tiempo	No es posible actualizar la medida de la exposición porque los datos se recogen una vez esta ha finalizado
Ayudan a establecer la relación temporal entre la exposición y el evento de interés	Es difícil establecer la relación temporal entre la exposición y el evento de interés
Tienen un riesgo más alto de perder sujetos durante el seguimiento y por ello perder validez	No tanto riesgo de perder sujetos y además no tiene limitaciones éticas
Facilitan el estudio de exposiciones infrecuentes	No es fácil incluir exposiciones infrecuentes
No son útiles para estudios sobre enfermedades raras	Son útiles para estudios de enfermedades raras y enfermedades con largos períodos de latencia
Permiten el estudio de más de un evento o enfermedad	Solo permiten el estudio de un solo evento de interés

Viendo las debilidades y fortalezas de cada tipo de estudio podemos concluir diciendo que son estudios complementarios. Los estudios de cohortes proporcionan una población bien definida en la que identificar a los casos de forma no sesgada y con información disponible de la secuencia temporal del seguimiento (Royo-Bordonada et al., 2009). Los estudios de caso-control hacen especial hincapié en los individuos que aportan información de interés sobre la exposición y además son menos costosos. Es por esto por lo que existen diseños híbridos como los estudios de caso-cohorte o los estudios de caso-control anidados en una cohorte.

4.1. Diseños híbridos

Los estudios de caso-control que pretenden maximizar la información obtenida a partir de los estudios de cohortes minimizando el coste. Los dos tipos de estudios híbridos más importantes son los estudios caso-cohorte y los estudios caso-control anidados en una cohorte. A menudo se recurre a ellos cuando los estudios de cohortes o de caso-control reúnen procedimientos muy costosos o muestras muy difíciles de obtener (Royo-Bordonada et al., 2009).

- El estudio de caso-control anidado en una cohorte parte de una cohorte inicial amplia (base primaria) y una vez definido se establece el diseño por el que se van a seleccionar casos y controles a la vez. Es común que en este tipo de estudios se recurra a la técnica del emparejamiento para controlar el efecto de las variables confundidoras debido al método de selección de los sujetos que se van a observar. A diferencia de los estudios de caso-control clásicos, los anidados suelen ser estudios prospectivos (Molina Arias, 2016).
- Los estudios caso-cohorte también parten de una cohorte inicial amplia a partir de la que se selecciona una sub-cohorte de la que saldrán los casos. En este tipo no se seleccionan casos y controles a la vez sino que, cuando ya se han seleccionado los casos, se seleccionan los controles procedentes de la cohorte inicial y que representan a la población en riesgo de enfermar pero que no han enfermado.

Apéndice A

Apéndice: Estudios complementarios al estudio de cohortes de enfermos por COVID-19 en Kazajistán

En este apartado se reflejará el código y resultados de estudios complementarios al estudio de cohortes de la Sección 2.5 (Yegorov et al., 2021a), disponible en el repositorio de GitHub "COVID-19-in-KZ" (Babenko, 2021).

El código que se recoge a continuación ha sido tomado del repositorio disponible en GitHub referente al estudio de cohortes retrospectivo de COVID-19 en Kazajistán (Babenko, 2021). Parte de las órdenes han sido modificadas para obtener mejores resultados. Todas las interpretaciones son de elaboración propia.

A.1. Correlaciones entre las características clínicas

Algunas de las variables presentes en el estudio recogen datos clínicos y otras recogen datos de laboratorio, es importante ver la correlación que hay entre ellas. Para evaluarla se usó la función `cor_mat`, perteneciente a la librería `rstatix` (Kassambara, 2021) que calcula la matriz de correlaciones con los p-valores, lo haremos por el Prueba de Rango de Spearman pasando el argumento `method = "spearman"`. Sacaremos las correlaciones entre las variables clínicas y luego entre las variables de laboratorio.

```
library(rstatix)
library(dplyr)
```

```r
# Primeros seleccionamos todos los síntomas clínicos
# que aparecen en el conjunto de datos
sintomas_c <- c(
  "Body_temperature",
  "Cough",
  "Sputum_production",
  "Shortness_of_breath",
  "Dyspnoea",
  "Stuffy",
  "Sore_throat",
  "Oropharynx_hyperemia",
  "Tonsill_hypertrophy",
  "Chest_pain",
  "Chest_tightness",
  "Wheezing",
  "Diarrhoea",
  "Nausea_vomiting",
  "Headache",
  "Conjuctivitis",
  "Myalgia_fatigue",
  "Joint_pain",
  "Pulse",
  "Systolic_pressure",
  "Diastolic_pressure",
  "Respiratory_rate",
  "SpO2"
)

# Creamos la matriz de correlaciones con los síntomas seleccionados
# con la función cor_mat y método Spearman
cor_matrix <- covidcohort %>%
  dplyr::select(dplyr::all_of(sintomas_c)) %>%
  mutate_if(is.factor, as.numeric) %>%
  data.frame() %>%
```

```
  rstatix::cor_mat(
    method = "spearman",
    alternative = "two.sided",
    conf.level = 0.95
  )
# Reordenamos la matriz de correlaciones y
# nos quedamos con el triangulo inferior
cor_inftri <- cor_matrix %>%
  rstatix::cor_reorder() %>%
  rstatix::pull_lower_triangle()
```

```
# Se pasa a un gráfico para verlo mejor con la función cor_plot
col <- ggpubr::get_palette(c("#00468BFF", "white", "#AD002AFF"), 20)


cor_inftri %>%
  rstatix::cor_plot(
    type = "lower",
    palette = col,
    font.label = list(
      size = 6,
      color = "black",
      style = "bold"
    )
  )
```

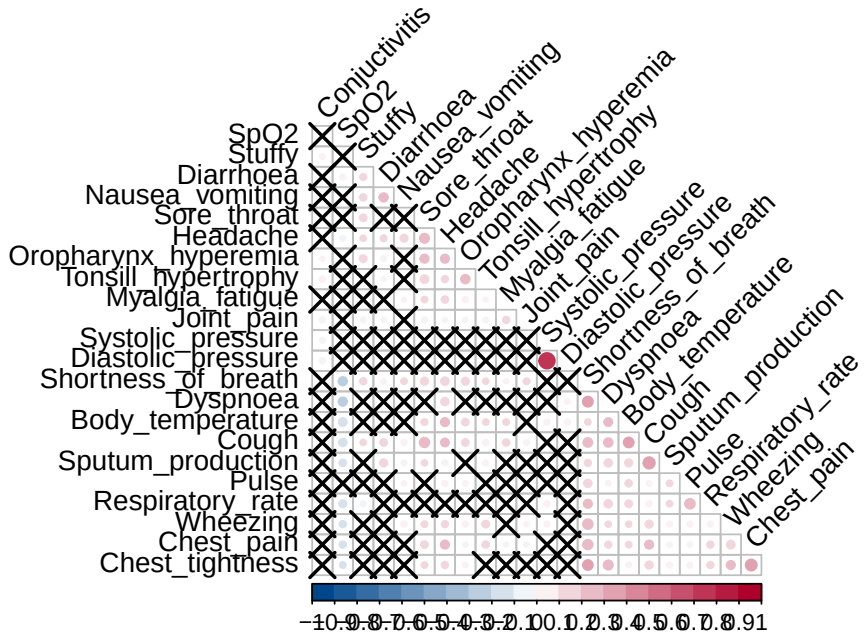


Figura A.1: Matriz de correlaciones de síntomas clínicos por COVID-19.

Las cruces implican una correlación no significativa. Hay correlaciones entre algunas variables pero sin duda la más significativa es la que existe entre la presión diastólica (`Diastolic_pressure`) y la presión sistólica (`Systolic_pressure`). Esto implica que un aumento de la presión diastólica provoca un aumento de la presión sistólica. Hay otras correlaciones positivas entre los síntomas y signos clínicos pero ninguna es tan fuerte como esta.

A.2. Correlaciones entre las características de laboratorio

```
# Seleccionamos los síntomas de laboratorio por COVID-19
sintomas_l <- c(
  "White_blood_cells",
  "Neutrophil",
  "Lymphocyte",
  "NLR",
  "Haemoglobin",
  "Monocytes",
  "Eosinophils",
  "Platelets",
```

```
  "Prothrombin_time",
  "Fibrinogen",
  "Albumin",
  "Alanine_aminotransferase",
  "Aspartate_aminotransferase",
  "Total_bilirubin",
  "Direct_bilirubin",
  "Glucose",
  "Blood_urea_nitrogen",
  "Creatinine",
  "C_reactive_protein",
  "Sodium",
  "Potassium",
  "Calcium"
)

# Creamos la matriz de correlaciones con los síntomas seleccionados
# con la función cor_mat y método Spearman
cor_matrix <- covidcohort %>%
  dplyr::select(dplyr::all_of(sintomas_l)) %>%
  data.frame() %>%
  rstatix::cor_mat(
    method = "spearman",
    alternative = "two.sided",
    conf.level = 0.95
  )

# Reordenamos la matriz de correlaciones y
# nos quedamos con el triangulo inferior
cor_inftri <- cor_matrix %>%
  rstatix::cor_reorder() %>%
  rstatix::pull_lower_triangle()
```

```r
# Se pasa a un gráfico para verlo mejor con la función cor_plot
col <- ggpubr::get_palette(c("#00468BFF", "white", "#AD002AFF"), 20)


cor_inftri %>%
  rstatix::cor_plot(
    type = "lower",
    palette = col,
    font.label = list(
      size = 6,
      color = "black",
      style = "bold"
    )
  )
```

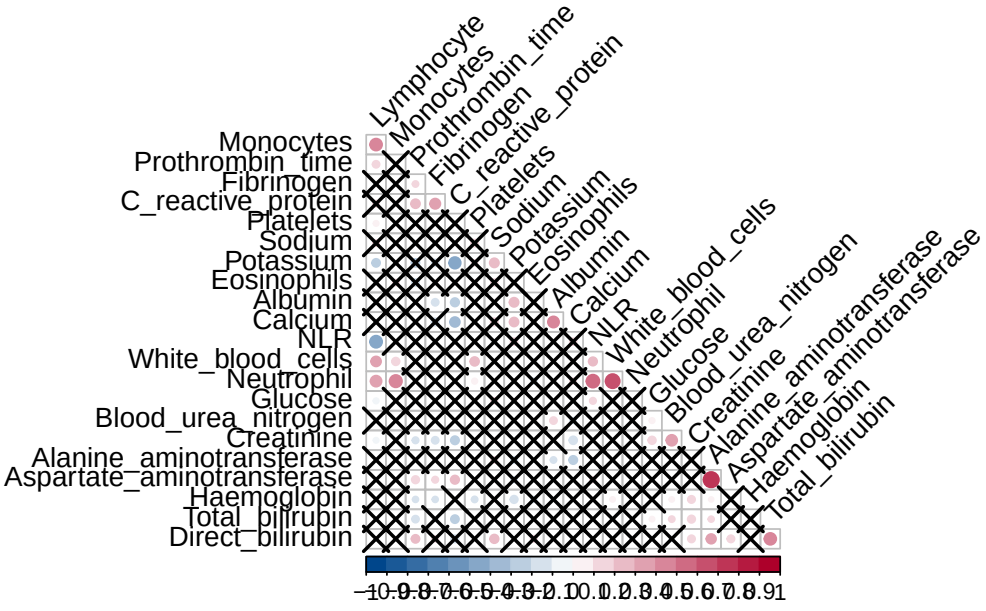


Figura A.2: Matriz de correlaciones de síntomas de laboratorio por COVID-19.

Entre los síntomas de laboratorio existen algunas correlaciones positivas como la que hay entre las enzimas alanina aminotransferasa (`Alanine_aminotransferase`) y aspartato aminotransferasa (`Aspartate_aminotransferase`). Las correlaciones negativas, como la que existe entre el potasio (`Potassium`) y la proteína C reactiva (`C_reactive_protein`), implican que cuando una

de las dos sustancias aumenta, la otra disminuye.

Para seleccionar los síntomas clínicos y de laboratorio con más representatividad en los datos se ha hecho un Análisis de Componentes Principales.

A.3. Análisis de Componentes Principales en sintomatología clínica

```
library(finalfit)
síntomas_c <- c(
  "Body_temperature",
  "Cough",
  "Sputum_production",
  "Shortness_of_breath",
  "Dyspnoea",
  "Stuffy",
  "Sore_throat",
  "Oropharynx_hyperemia",
  "Tonsill_hypertrophy",
  "Chest_pain",
  "Chest_tightness",
  "Wheezing",
  "Diarrhoea",
  "Nausea_vomiting",
  "Headache",
  "Conjuctivitis",
  "Myalgia_fatigue",
  "Joint_pain",
  "Pulse",
  "Respiratory_rate"
)



# Creamos un data.frame a partir del conjunto incial
# con los síntomas clínicos escogidos
```

```r
acp_c <- covidcohort %>%
  dplyr::select(c(
    Deaths, Disease_severity,
    dplyr::all_of(sintomas_c)
  )) %>%
  dplyr::mutate_if(is.factor, as.character) %>%
  dplyr::mutate_if(is.character, as.numeric)



# Agrupamos la severidad de la enfermedad en 2 grupos
# en una variable nueva Disease_severity_group_B
# Niveles 1 y 2 = No severo
# Niveles 3 y 4 = Severo
acp_c %<>% mutate(
  Disease_severity_group_B =
    case_when(
      Disease_severity %in%
        c(1, 2) ~ "No_Severo",
      Disease_severity %in% c(3, 4) ~ "Severo"
    )
)


acp_c %<>% dplyr::select(-Disease_severity)


acp_c$Deaths <- factor(acp_c$Deaths, levels = c(0, 1))
acp_c$Disease_severity_group_B <- factor(
  acp_c$Disease_severity_group_B,
  levels = c("No_Severo", "Severo")
)


acp_c %<>% dplyr::relocate(
  Disease_severity_group_B,
  .before = Deaths
)
```

```r
# Quitamos los outliers de las variables
# Temperatura corporal (Body_Temperature),
# Presión sistólica (Systolic_pressure), Pulso (Pulse),
# Presión diastólica (Diastolic_pressure),
# Frecuencia respiratoria (Respiratory_rate) y
# Saturación de oxígeno (SpO2)
for (i in c(3, 21:22)) {
  acp_c %>%
    dplyr::select(dplyr::all_of(i)) %>%
    rstatix::identify_outliers() %>%
    filter(is.outlier == T) %>%
    dplyr::select(1) %>%
    tibble::deframe() -> nn
  if (length(nn) != 0) {
    acp_c[acp_c[, i] %in% nn, i] <- NA
  }
  remove(nn)
}
# Escogemos los síntomas con un porcentaje de valores perdidos
# menor a 40
acp_c_nm <- acp_c %>%
  is.na() %>%
  colSums() < nrow(acp_c) * 0.4
acp_c_nm <- which(acp_c_nm == T) %>% names()

# Nos quedamos con las variables seleccionadas en el paso anterior
acp_c <- acp_c %>% dplyr::select(dplyr::all_of(acp_c_nm))

# Renombramos las dos primeras columnas
colnames(acp_c)[1:2] <- c("Odds de severidad", "Mortalidad")

# Renombramos las etiquetas de las dos primeras columnas
levels(acp_c$`Odds de severidad`)[levels(
```

```
  acp_c$`Odds de severidad`
) == "No_Severo"] <- "Leve/Moderado"
levels(acp_c$`Odds de severidad`)[levels(
  acp_c$`Odds de severidad`
) == "Severo"] <- "Severo/Critico"
levels(acp_c$Mortalidad)[levels(acp_c$Mortalidad) == "0"] <- "Supervivientes"
levels(acp_c$Mortalidad)[levels(acp_c$Mortalidad) == "1"] <- "No_Supervivientes"

# Imputamos los valores perdidos del dataset. Es un paso preliminar.
acp_c_imp <- missMDA::imputePCA(acp_c, quali.sup = 1:2, scale = T, seed = 2021)
```

Una vez que hemos preparado el conjunto de datos para poder hacer el Análisis de Componentes Principales vamos a usar la función `PCA` de la librería `factoMineR` (Lê et al., 2008) para aplicar esta técnica.

```
library(FactoMineR)
# Hacemos el análisis de componentes principales con la función PCA
# de la librería FactoMineR
res.pca <- FactoMineR::PCA(
  X = acp_c_imp$completeObs,
  scale.unit = T, quali.sup = 1:2,
  graph = FALSE
)
```

A continuación vamos a ver los resultados en gráficos.

```
knitr::kable(res.pca[["eig"]],
             caption = "\\label{table031}Varianza total
             explicada en sintomatología clínica")
```

Tabla A.1: Varianza total explicada en sintomatología clínica

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.2598629	14.817558	14.81756
comp 2	1.9099970	8.681805	23.49936
comp 3	1.4016985	6.371357	29.87072
comp 4	1.2835650	5.834386	35.70511

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 5	1.1492876	5.224035	40.92914
comp 6	1.0948321	4.976510	45.90565
comp 7	1.0244373	4.656533	50.56218
comp 8	0.9956797	4.525817	55.08800
comp 9	0.9694680	4.406673	59.49467
comp 10	0.9455871	4.298123	63.79280
comp 11	0.8890685	4.041220	67.83402
comp 12	0.8308247	3.776476	71.61049
comp 13	0.7932294	3.605588	75.21608
comp 14	0.7617058	3.462299	78.67838
comp 15	0.7513120	3.415055	82.09343
comp 16	0.6950446	3.159293	85.25273
comp 17	0.6496744	2.953066	88.20579
comp 18	0.6408365	2.912893	91.11869
comp 19	0.5907492	2.685224	93.80391
comp 20	0.5305426	2.411557	96.21547
comp 21	0.4898356	2.226525	98.44199
comp 22	0.3427614	1.558007	100.00000

```
library(factoextra)
library(ggcorrplot)
p_scree <- factoextra::fviz_eig(res.pca,
  addlabels = F,
  ylim = c(0, 23),
  ncp = 50, barfill = "#00468BFF",
  ggtheme = theme_classic()
) +
  theme(axis.title.x = element_blank())
p_scree
```

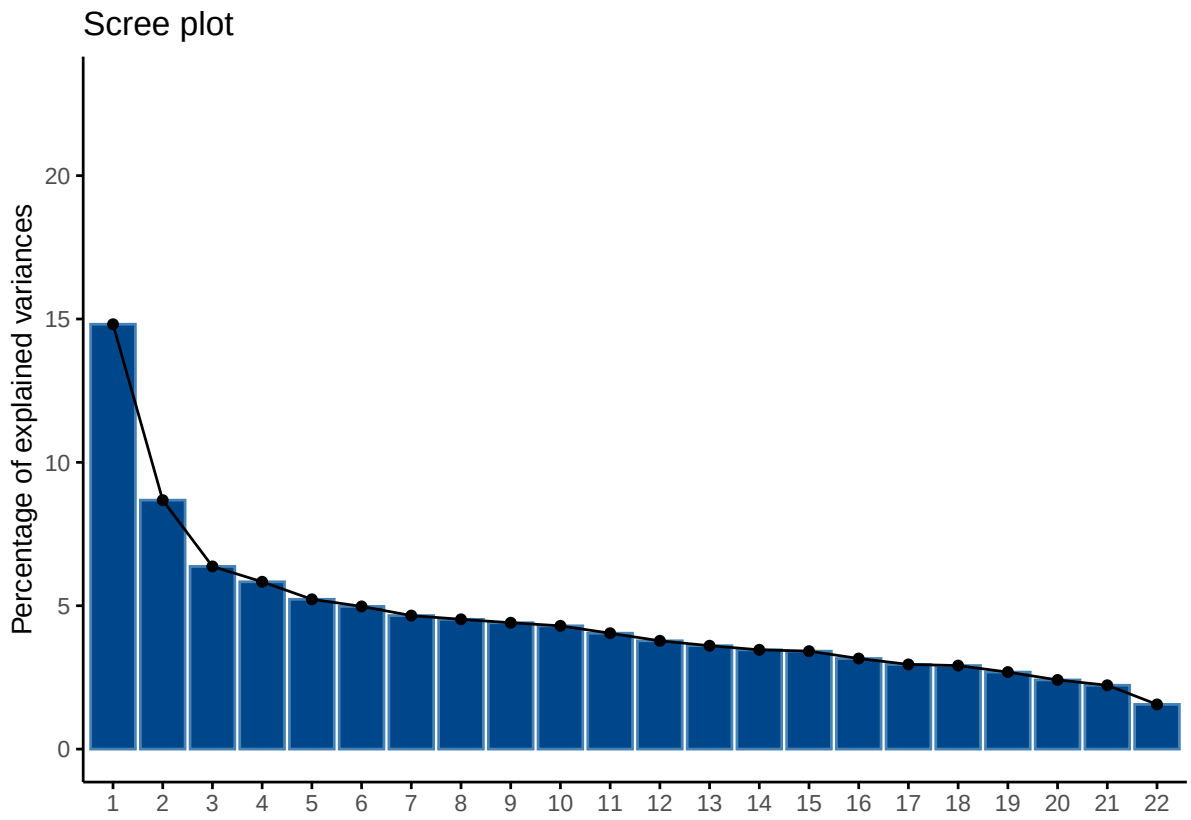


Figura A.3: Variabilidad representada por cada componente principal.

Hemos pasado de tener 23 variables de sintomatología clínica a contar con 12 componentes principales (que son combinaciones lineales de estas variables) que guardan el 71.6 % de la información que teníamos al principio (Tabla A.1 y Figura A.3). Si consideramos como máximo 5 componentes principales mantendremos el 40 % de la información inicial.

A continuación se muestra la correlación entre cada variable y la componente principal (las 5 primeras).

```
var <- factoextra::get_pca_var(res.pca)
p_corr <- ggcorrplot::ggcorrplot(var$cos2,
                                 method = "circle",
                                 ggtheme = theme_classic(),
                                 colors = c("#00468BFF",
                                            "white",
                                            "#AD002AFF"),
                                 tl.cex = 10)
p_corr
```

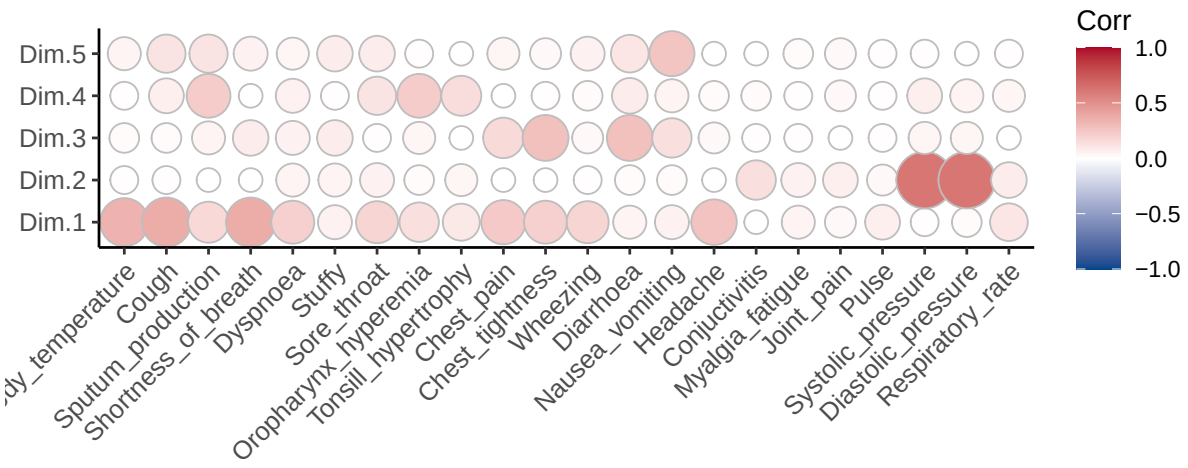


Figura A.4: Correlación entre cada variable y la componente principal.

Las variables fuertemente asociadas a la primera componente principal son `Body_temperature`, `Cough`, `Shortness_of_breath` y `Headache`. Las variables principalmente asociadas a la segunda componente son `Systolic_pressure`, `Diastolic_pressure`. En la tercera componente las variables más representativas son `Chest_tightness` y `Diarrhoea`. `Sputum_production` y `Oropharynx_hyperemia` están asociadas a la cuarta componente principal. Por último, `Nausea_vomiting` es la variable con mayor correlación con la quinta componente.

Podemos ver los gráficos que surgen entre las tres primeras componentes según la severidad o la mortalidad por COVID-19.

```
combn(1:3, 2) -> cmbn
p_sever <- list()
p_mortalidad <- list()

for (i in 1:3) {
  if (i == 3) {
    p_sever[[i]] <- factoextra::fviz_pca_biplot(res.pca,
      axes = cmbn[, i],
```

```
      geom.ind = "point",
      habillage = 1, pointsize = 1.6,
      invisible = "var", palette = "lanonc",
      addEllipses = TRUE, title = NULL
    ) +
      theme(
        panel.background = element_rect(fill = NULL),
        panel.grid.major = element_line(colour = "white"),
        panel.grid.minor = element_line(colour = "white"),
        legend.position = "right"
      ) + ggsci::scale_color_lancet()

    p_mortalidad[[i]] <- factoextra::fviz_pca_biplot(res.pca,
      axes = cmbn[, i],
      geom.ind = "point",
      habillage = 2,
      pointsize = 1.6,
      invisible = "var",
      palette = "lanonc",
      addEllipses = TRUE,
      title = NULL
    ) +
      theme(
        panel.background = element_rect(fill = NULL),
        panel.grid.major = element_line(colour = "white"),
        panel.grid.minor = element_line(colour = "white"),
        legend.position = "right"
      ) + ggsci::scale_color_lancet()
  } else {
    p_sever[[i]] <- factoextra::fviz_pca_biplot(res.pca,
      axes = cmbn[, i],
      geom.ind = "point",
      habillage = 1,
      pointsize = 1.6,
```

```
      invisible = "var",
      palette = "lanonc",
      addEllipses = TRUE,
      title = NULL
    ) +
      theme(
        panel.background = element_rect(fill = NULL),
        panel.grid.major = element_line(colour = "white"),
        panel.grid.minor = element_line(colour = "white"),
        legend.position = "none"
      ) + ggsci::scale_color_lancet()

    p_mortalidad[[i]] <- factoextra::fviz_pca_biplot(res.pca,
      axes = cmbn[, i],
      geom.ind = "point",
      habillage = 2,
      pointsize = 1.6,
      invisible = "var",
      palette = "lanonc",
      addEllipses = TRUE,
      title = NULL
    ) +
      theme(
        panel.background = element_rect(fill = NULL),
        panel.grid.major = element_line(colour = "white"),
        panel.grid.minor = element_line(colour = "white"),
        legend.position = "none"
      ) + ggsci::scale_color_lancet()
  }
}
```

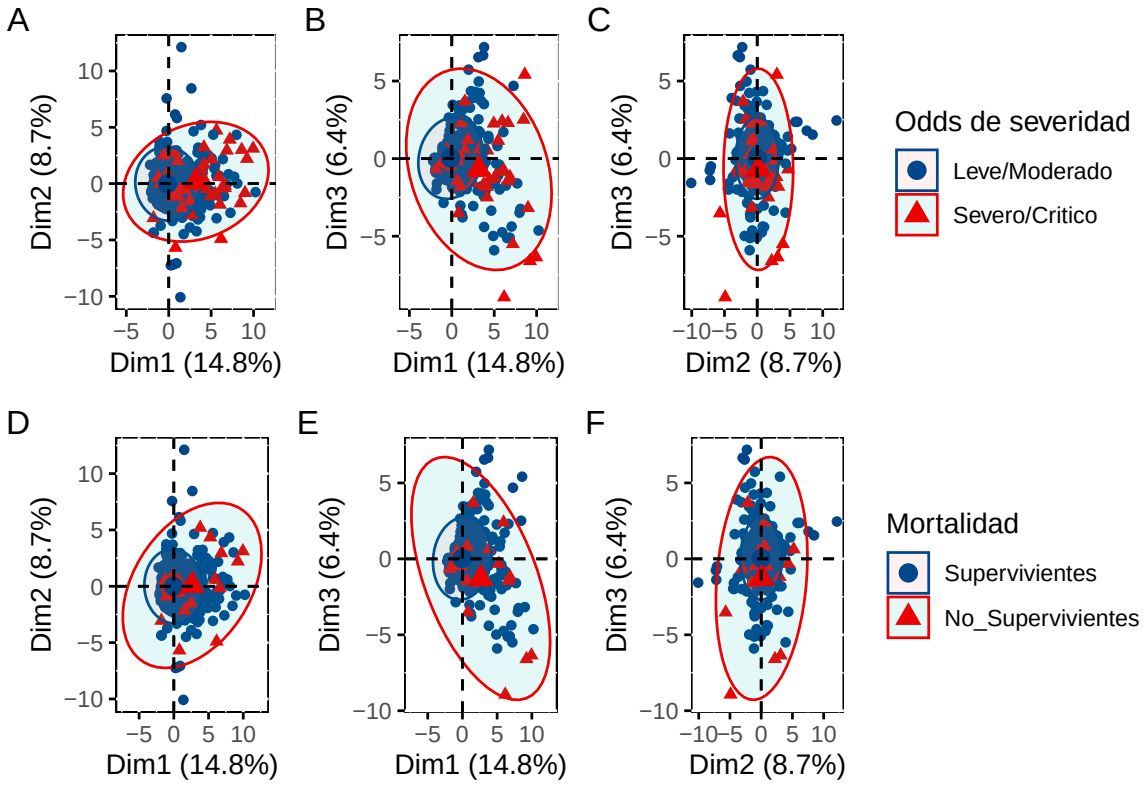


Figura A.5: Gráfico entre las 3 primeras componentes según severidad o mortalidad.

```
## null device
##           1
```

A.4. Análisis de Componentes Principales en sintomatología de laboratorio

```
sintomas_l <- c(
  "White_blood_cells",
  "Neutrophil",
  "Lymphocyte",
  "NLR",
  "Haemoglobin",
  "Monocytes",
  "Eosinophils",
  "Platelets",
  "Prothrombin_time",
```

```
  "Fibrinogen",
  "Albumin",
  "Alanine_aminotransferase",
  "Aspartate_aminotransferase",
  "Total_bilirubin",
  "Direct_bilirubin",
  "Glucose",
  "Blood_urea_nitrogen",
  "Creatinine",
  "C_reactive_protein",
  "Sodium",
  "Potassium",
  "Calcium"
)

# Creamos un data.frame a partir del conjunto incial
# con los síntomas de laboratorio escogidos
acp_l <- covidcohort %>%
  dplyr::select(c(
    Deaths, Disease_severity,
    dplyr::all_of(sintomas_l)
  ))

# Agrupamos la severidad de la enfermedad en 2 grupos
# en una variable nueva Disease_severity_group_B
# Niveles 1 y 2 = No severo
# Niveles 3 y 4 = Severo
acp_l %<>% mutate(Disease_severity_group_B = case_when(
  Disease_severity %in% c(1, 2) ~ "No_Severo",
  Disease_severity %in% c(3, 4) ~ "Severo"
))

acp_l %<>% dplyr::select(-Disease_severity)
```

```r
acp_l$Deaths <- factor(acp_l$Deaths, levels = c(0, 1))
acp_l$Disease_severity_group_B <- factor(acp_l$Disease_severity_group_B,
  levels = c("No_Severo", "Severo")
)

acp_l %<>% dplyr::relocate(Disease_severity_group_B, .before = Deaths)



# Quitamos los outliers
for (i in c(3:ncol(acp_l))) {
  acp_l %>%
    dplyr::select(i) %>%
    rstatix::identify_outliers() %>%
    dplyr::filter(is.outlier == T) %>%
    dplyr::select(1) %>%
    tibble::deframe() -> nn
  if (length(nn) != 0) {
    acp_l[acp_l[, i] %in% nn, i] <- NA
  }
  remove(nn)
}
for (i in c(3:ncol(acp_l))) {
  acp_l[, i] <- acp_l %>%
    dplyr::select(i) %>%
    tibble::deframe() %>%
    scales::rescale()
}


# Escogemos los síntomas con un porcentaje de valores perdidos
# menor a 40
acp_l_nm <- acp_l %>%
  is.na() %>%
  colSums() < nrow(acp_l) * 0.4
acp_l_nm <- which(acp_l_nm == T) %>% names()
```

```r
# Nos quedamos con las variables seleccionadas en el paso anterior
acp_l <- acp_l %>% dplyr::select(dplyr::all_of(acp_l_nm))

# Renombramos las dos primeras columnas
colnames(acp_l)[1:2] <- c("Odds de severidad", "Mortalidad")

# Renombramos las etiquetas de las dos primeras columnas
levels(acp_l$`Odds de severidad`)[levels(
  acp_l$`Odds de severidad`
) == "No_Severo"] <- "Leve/Moderado"
levels(acp_l$`Odds de severidad`)[
  levels(acp_l$`Odds de severidad`) == "Severo"
] <- "Severo/Critico"
levels(acp_l$Mortalidad)[levels(acp_l$Mortalidad) == "0"] <- "Supervivientes"
levels(acp_l$Mortalidad)[levels(acp_l$Mortalidad) == "1"] <- "No_Supervivientes"



# Imputamos los valores perdidos del dataset. Es un paso preliminar.
acp_l_imp <- missMDA::imputePCA(acp_l, quali.sup = 1:2, scale = F, seed = 2021)
```

Una vez que hemos preparado el conjunto de datos para poder hacer el Análisis de Componentes Principales vamos a usar la función `PCA` de la librería `factoMineR` (Lê et al., 2008) para aplicar esta técnica.

```r
library("FactoMineR")
# Hacemos el análisis de componentes principales con la función PCA
# de la librería FactoMineR
res.pca <- FactoMineR::PCA(
  X = acp_l_imp$completeObs,
  scale.unit = F, quali.sup = 1:2,
  graph = FALSE
)
```

A continuación vamos a ver los resultados en gráficos.

```
knitr::kable(res.pca[["eig"]],
             caption = "\\label{table032}Varianza total
             explicada en sintomatología de laboratorio")
```

Tabla A.2: Varianza total explicada en sintomatología de laboratorio

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	0.0214393	33.355183	33.35518
comp 2	0.0121796	18.948968	52.30415
comp 3	0.0076712	11.934772	64.23892
comp 4	0.0069793	10.858333	75.09726
comp 5	0.0062514	9.725958	84.82321
comp 6	0.0025282	3.933340	88.75655
comp 7	0.0024562	3.821344	92.57790
comp 8	0.0018425	2.866484	95.44438
comp 9	0.0016421	2.554803	97.99919
comp 10	0.0012860	2.000814	100.00000

```
library(factoextra)
library(ggcorrplot)


p_scree <- factoextra::fviz_eig(res.pca,
  addlabels = F,
  ylim = c(0, 35),
  ncp = 50,
  barfill = "#00468BFF",
  ggtheme = theme_classic()
) +
  theme(axis.title.x = element_blank())
p_scree
```

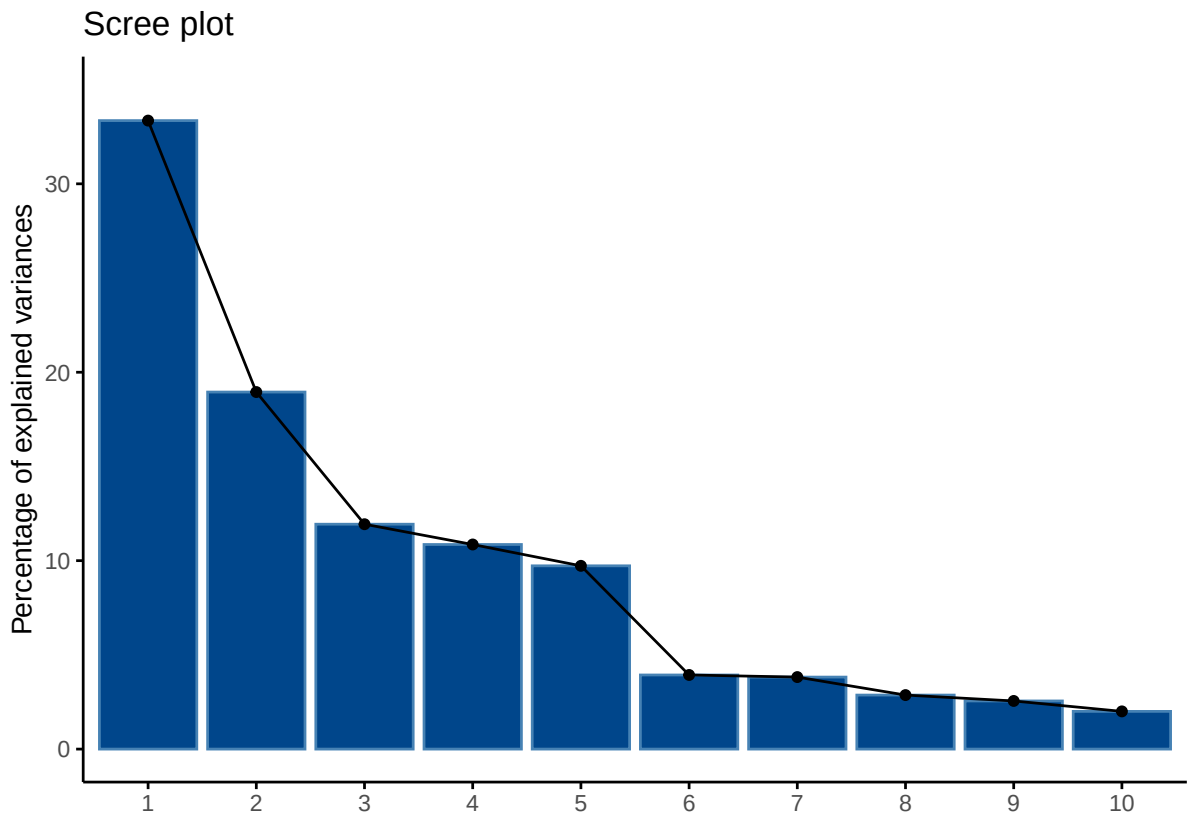


Figura A.6: Variabilidad representada por cada componente principal.

Hemos pasado de tener 22 variables de sintomatología de laboratorio a contar con 4 componentes principales (que son combinaciones lineales de estas variables) que guardan el 75.05 % de la información que teníamos al principio (Tabla A.2 y Figura A.6).

A continuación se muestra la correlación entre cada variable y la componente principal.

```
var <- factoextra::get_pca_var(res.pca)
p_corr <- ggcorrplot::ggcorrplot(var$cos2,
  method = "circle",
  ggtheme = theme_classic(),
  colors = c("#00468BFF", "white", "#AD002AFF"),
  tl.cex = 10
)
p_corr
```

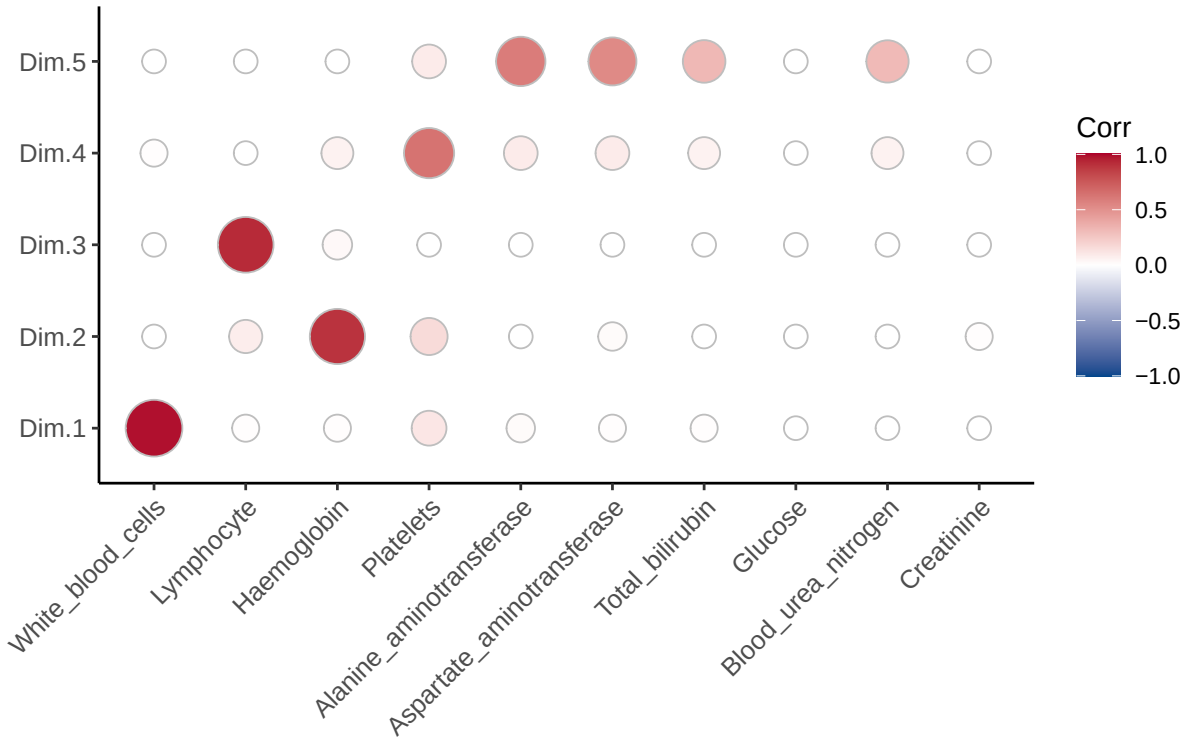


Figura A.7: Correlación entre cada variable y la componente principal.

La variable fuertemente asociada a la primera componente principal es `White_blood_cells`. La variable principalmente asociadas a la segunda componente es `Haemoglobin`. `Lymphocyte` está asociada a la tercera componente principal. Por último, `Platelets` es la variable con mayor correlación con la cuarta componente.

Podemos ver los gráficos que surgen entre las tres primeras componentes según la severidad o la mortalidad por COVID-19.

```
combn(1:3, 2) -> cmbn
p_sever <- list()
p_Mortalidad <- list()

for (i in 1:3) {
  if (i == 3) {
    p_sever[[i]] <- factoextra::fviz_pca_biplot(
      res.pca,
      axes = cmbn[, i],
      geom.ind = "point",
```

```
      habillage = 1,
      pointsize = 1.6,
      invisible = "var",
      palette = "lanonc",
      addEllipses = TRUE,
      title = NULL
    ) +
      theme(
        panel.background = element_rect(fill = NULL),
        panel.grid.major = element_line(colour = "white"),
        panel.grid.minor = element_line(colour = "white"),
        legend.position = "right"
      ) +
      ggsci::scale_color_lancet()

    p_Mortalidad[[i]] <- factoextra::fviz_pca_biplot(
      res.pca,
      axes = cmbn[, i],
      geom.ind = "point",
      habillage = 2,
      pointsize = 1.6,
      invisible = "var",
      palette = "lanonc",
      addEllipses = TRUE, title = NULL
    ) +
      theme(
        panel.background = element_rect(fill = NULL),
        panel.grid.major = element_line(colour = "white"),
        panel.grid.minor = element_line(colour = "white"),
        legend.position = "right"
      ) +
      ggsci::scale_color_lancet()
  } else {
    p_sever[[i]] <- factoextra::fviz_pca_biplot(
```

```
    res.pca,
    axes = cmbn[, i],
    geom.ind = "point",
    habillage = 1, pointsize = 1.6,
    invisible = "var",
    palette = "lanonc",
    addEllipses = TRUE,
    title = NULL
  ) +
    theme(
      panel.background = element_rect(fill = NULL),
      panel.grid.major = element_line(colour = "white"),
      panel.grid.minor = element_line(colour = "white"),
      legend.position = "none"
    ) +
    ggsci::scale_color_lancet()

  p_Mortalidad[[i]] <- factoextra::fviz_pca_biplot(
    res.pca,
    axes = cmbn[, i],
    geom.ind = "point",
    habillage = 2,
    pointsize = 1.6,
    invisible = "var",
    palette = "lanonc",
    addEllipses = TRUE,
    title = NULL
  ) +
    theme(
      panel.background = element_rect(fill = NULL),
      panel.grid.major = element_line(colour = "white"),
      panel.grid.minor = element_line(colour = "white"),
      legend.position = "none"
    ) +
```

```
      ggsci::scale_color_lancet()
  }
}
```

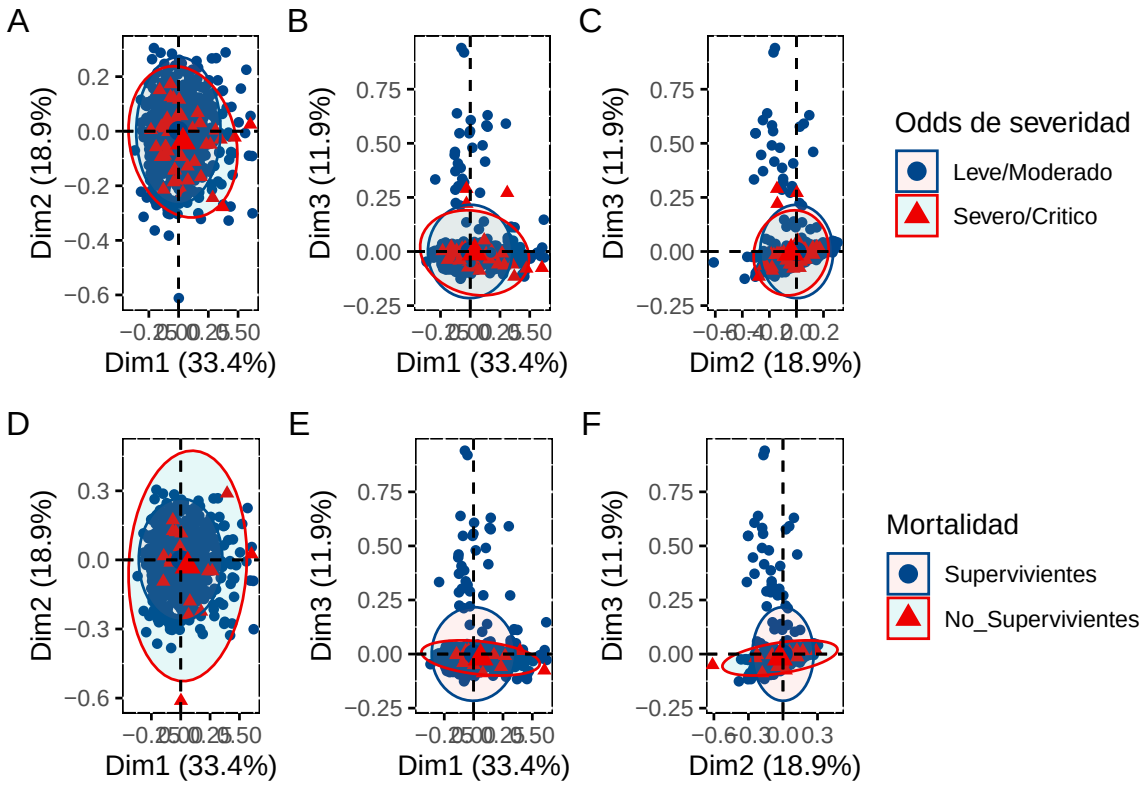


Figura A.8: Gráfico entre las 3 primeras componentes según severidad o mortalidad.

```
## null device
##           1
```

A.5. Asociación de la edad con enfermar grave o muerte por COVID-19

En este apartado se estudia la asociación que existe entre la edad y la severidad o muerte de la enfermedad COVID-19. En este caso, no se agrupara la variable edad en diferentes tramos, sino que se hará el estudio teniendo en cuenta todas las edades (sin considerar valores extremos).

```
# Se define la nueva variable agrupando en
# pacientes severos (grupos 1 y 2) y no severos (grupos 3 y 4)
covidcohort %<>% mutate(
  Disease_severity_group_binary = case_when(
```

```
    Disease_severity %in% c(1, 2) ~ "No_Severo",
    Disease_severity %in% c(3, 4) ~ "Severo"
  )
)

covidcohort$Disease_severity_group_binary <- factor(
  covidcohort$Disease_severity_group_binary,
  levels = c("No_Severo", "Severo")
)

# Se cogen todas las edades menos las extremas
# Se considera extremo tener 0 años o 100
covidcohort %<>% dplyr::filter(Age != 0 & Age != 100)
```

En este caso, se hace una regresión no lineal splines, este tipo de regresión es una extensión de la regresión polinómica y de las funciones steps (paso a paso). Para hacerlo usaremos la función `bs` de la librería `splines` (Bates & Venables, 2019).

Aplicaremos esta función al factor edad (`Age`) y lo que haremos será dividir el rango de esta variable en 3 subintervalos. Para esto pasamos el argumento `knots = c(15, 50, 65)`. Para cada subintervalo se ajustará una función polinómica de grado 1 (`degree = 1`). Los límites de la edad están en 1 y 99 años (`Boundary.knots = c(1, 99)`. Todo ello irá introducido dentro de la fórmula del modelo lineal general.

En el modelo ajustado para la severidad de la enfermedad de COVID-19 la variable objetivo es la que hemos creado anteriormente `Disease_severity_group_binary` y las variables explicativas son `Age`, `Any_comorbidities`, `Kazahk_ethnicity` y `White_blood_cells` que son las variables que aparecían asociadas a enfermar gravemente en la Tabla 2.2.

```
fit_severity <- glm(
  Disease_severity_group_binary ~ splines::bs(Age,
                                  knots = c(15, 50, 65),
                                  Boundary.knots = c(1, 99),
                                  degree = 1) +
    Kazakh_ethnicity +
    Any_comorbidities +
    White_blood_cells,
```

```
  data = covidcohort,
  family = binomial())
```

El modelo ajustado para la mortalidad, la variable `Deaths` viene explicada por la variable `Age` y `Sex` (asociadas en la Tabla 2.4).

```
fit_mortality <- glm(
  Deaths ~ splines::bs(Age,
                       knots = c(15, 50, 65),
                       Boundary.knots = c(1, 99),
                       degree = 1) +
    Sex,
  data = covidcohort,
  family = binomial())
```

```
# Normalizamos la ODDS al grupo de 15 a 49 años
OR_severity <- exp(coef(fit_severity))[2:5] %>%
  as.vector() / exp(coef(fit_severity))[3] %>%
  as.vector()

OR_mortality <- exp(coef(fit_mortality))[2:5] %>%
  as.vector() / exp(coef(fit_mortality))[3] %>%
  as.vector()
```

Calculamos a continuación el estadístico C con la función `Cstat` de la librería `DescTools` (Signorell, 2021). Este estadístico es una medida de la bondad del ajuste de una regresión logística o para otro modelo de clasificación. Es equivalente al área que hay debajo de la curva ROC. A mayor valor del estadístico C, mejor es el modelo ajustado. El valor oscila entre 0.5 y 1.

```
# C-statistics of fitted models
library(DescTools)
cat("El valor del estadístico C para el modelo ajustado
    para la severidad del COVID-19 es", DescTools::Cstat(fit_severity), "\n",
    "El valor del estadístico C para el modelo ajsutado
    para la mortalidad por COVID-19 es", DescTools::Cstat(fit_mortality), ".")
```

```
## El valor del estadístico C para el modelo ajustado
##     para la severidad del COVID-19 es 0.8145032
```

```
##  El valor del estadístico C para el modelo ajsutado
##     para la mortalidad por COVID-19 es 0.9284892 .
```

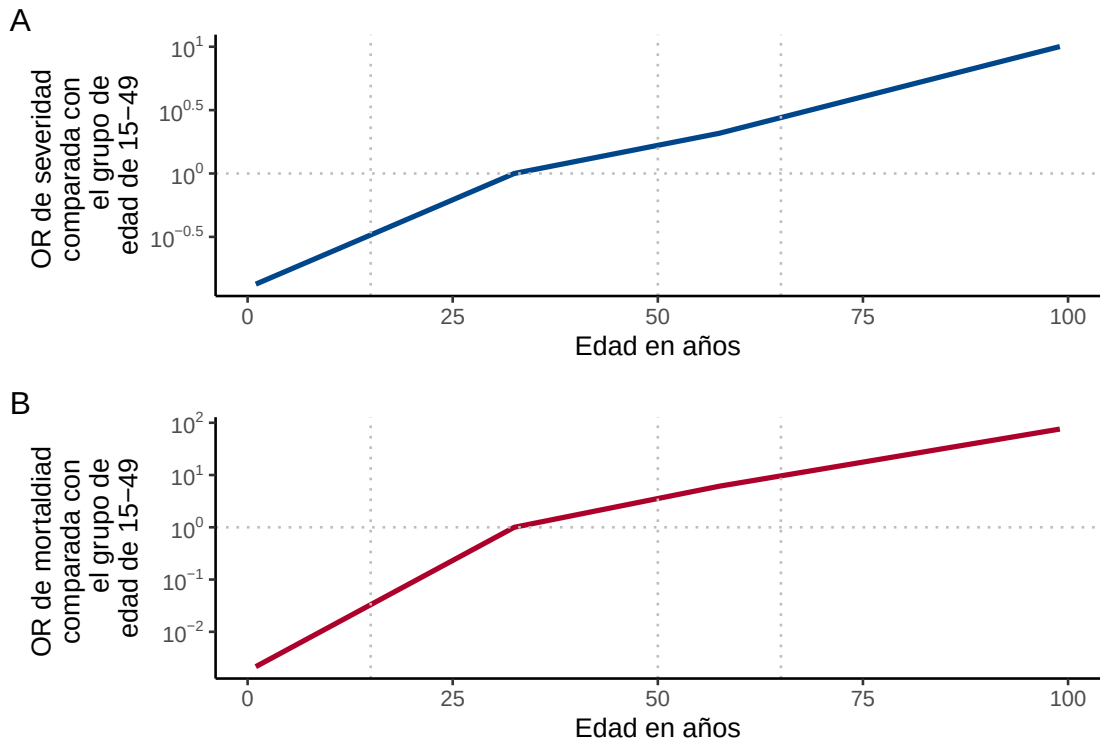
Como vemos, salen valores muy altos que nos indican que estamos ante buenos modelos.

Ahora nuestra intención es ver como afecta la edad a la enfermedad bajo estudio, esto podemos verlo a través de los valores que va tomando la Odds Ratio conforme avanza la edad. Haremos un gráfico para estudiarlo.

```r
# Gráfico
library(scales)
library(ggsci)
# Creamos un data frame en el que guardemos
# los extremos y los puntos medios
# de los intervalos 15-49 y 50-65
# con el OR asociado a cada etapa
ORdf <- tibble(Age = c(1, 32.5, 57.5, 99),
               ORadj = OR_severity)

# A continuación se crea el gráfico
p <- ORdf %>% ggplot() +
  geom_line(aes(x = Age, y = ORadj),
            size = 1,
            color = "#00468BFF")

p <- p + geom_vline(xintercept = c(15, 50, 65),
                    linetype = "dotted",
                    color = "gray",
                    size = 0.5) +
  geom_hline(yintercept = c(1),
             linetype = "dotted",
             color = "gray",
             size = 0.5)

p_severity_OR <- p + scale_y_continuous(trans = log10_trans(),
                                        breaks = trans_breaks(
```

```
                                          "log10",
                                          function(x) 10^x),
                                        labels = trans_format(
                                          "log10",
                                          math_format(10^.x))) +
  theme_classic() +
  xlab("Edad en años") +
  ylab("OR de severidad \ncomparada con
       el grupo de \nedad de 15-49")

# El mismo procedimiento para la mortalidad
ORdf <- tibble(Age = c(1, 32.5, 57.5, 99),
               ORadj = OR_mortality)

p <- ORdf %>%
  ggplot() +
  geom_line(aes(x = Age,
                y = ORadj),
            size = 1,
            color = "#AD002AFF")

p <- p +
  geom_vline(xintercept = c(15, 50, 65),
             linetype = "dotted",
             color = "gray",
             size = 0.5) +
  geom_hline(yintercept = c(1),
             linetype = "dotted",
             color = "gray",
             size = 0.5)

p_mortality_OR <- p + scale_y_continuous(trans = log10_trans(),
                                         breaks = trans_breaks(
                                           "log10",
```

```
                                          function(x) 10^x),
                                        labels = trans_format(
                                          "log10",
                                          math_format(10^.x))) +
  theme_classic() +
  xlab("Edad en años") +
  ylab("OR de mortaldiad \ncomparada con
       el grupo de \nedad de 15-49")

library(patchwork)
patchwork <- p_severity_OR / p_mortality_OR
patchwork + plot_annotation(tag_levels = "A")
```




```
dev.off()
```

```
## null device
##           1
```

Efectivamente, el riesgo de morir o enfermar gravemente por COVID-19 aumenta en personas expuestas a una edad mayor. Al estar comparado con la OR del grupo de 15 a 49 años en el punto 32.5 (punto medio del intervalo) la OR vale exactamente 1. Por tanto, existe una fuerte

asociación entre la muerte y enfermedad grave por COVID-19 en edades altas.

A.6. Factores de riesgo asociados a la muerte o a enfermar gravemente por COVID-19 según el sitio clínico

```
covidcohort %<>%
  mutate(Disease_severity_group_binary = case_when(
  Disease_severity %in% c(1, 2) ~ "No_Severo",
  Disease_severity %in% c(3, 4) ~ "Severo"
))

covidcohort$Disease_severity_group_binary <- factor(
  covidcohort$Disease_severity_group_binary,
  levels = c("No_Severo", "Severo"))
# model fitting
fit_severity <- glm(Disease_severity_group_binary ~ Age +
                      Kazakh_ethnicity + Any_comorbidities +
                      White_blood_cells,
                    data = covidcohort,
                    family = binomial())
fit_mortality <- glm(Deaths ~ Age + Sex,
                     data = covidcohort,
                     family = binomial())

# model characteristics
broom::tidy(fit_severity)
```

```
## # A tibble: 5 x 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         -6.49     0.791      -8.21  2.28e-16
## 2 Age                  0.0532   0.00949     5.61  2.08e- 8
## 3 Kazakh_ethnicity1   -0.320    0.361      -0.887 3.75e- 1
## 4 Any_comorbidities1   0.759    0.361       2.10  3.55e- 2
## 5 White_blood_cells    0.0972   0.0629      1.55  1.22e- 1
```

```
broom::tidy(fit_mortality)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -10.7      1.33       -8.06 7.46e-16
## 2 Age            0.106    0.0171      6.20 5.54e-10
## 3 Sex1           1.75     0.561       3.13 1.75e- 3
```

No se han encontrado diferencias significativas con respecto a los resultados obtenidos de los estudios principales que se encuentran en las Secciones 2.5.2 y 2.5.3.

A.7. Estudio de la asociación del índice de Masa Corporal con enfermar gravemente o morir por COVID-19

El 91 % de los datos de la variable `IBM`, que recoge el Índice de Masa Corporal (IMC) de cada individuo, eran valores perdidos. Es por esta razón por la que no se incluyó esta variable en los estudios de asociación que aparecen dentro de la Sección 2.5. Pero, debido a que es una medida importante, vamos a imputar los datos que faltan para poder hacer el estudio de la asociación del IMC (`BMI`) como el riesgo de enfermar gravemente o morir por COVID-19.

La imputación de estos valores se realiza con la librería `missRanger` (Mayer, 2021). Este paquete utiliza un enfoque no paramétrico para la imputación de los valores perdidos basado en `Random Forest`.

```
# IMC antes de imputar los datos
covidcohort$BMI %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   11.00   21.00   24.00   25.26   29.00   68.00     977
```

```
# IMC después de imputar los datos
imc_imp$BMI %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   23.00   25.00   25.07   27.00   68.00
```

Las medidas se han mantenido pero el número de valores perdidos se ha hecho 0.

Se va preparando el conjunto de datos para obtener las características de la variable `BMI` en

pacientes diagnosticados por COVID-19 según la gravedad o supervivencia.

```
imc_imp%<>%
  mutate(Disease_severity_group = case_when(
    Disease_severity == 1 ~ "Asintomatico/Leve",
    Disease_severity == 2 ~ "Moderado",
    Disease_severity %in% c(3,4) ~ "Severo/Critico")
)
imc_imp$Disease_severity_group <- factor(
  imc_imp$Disease_severity_group,
  levels = c("Asintomatico/Leve",
             "Moderado",
             "Severo/Critico"))
```

```
tbimc1 <- rbind(
  imc_imp %>%
    finalfit::summary_factorlist(
      dependent = "Disease_severity_group",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  cbind(
  imc_imp %>%
    filter(BMI < 18.5) %>%
    finalfit::summary_factorlist(
      dependent = "Disease_severity_group",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  "Severo/Critico" = 0),
```

```
  imc_imp %>%
    filter(BMI >= 18.5 & BMI < 25) %>%
    finalfit::summary_factorlist(
      dependent = "Disease_severity_group",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  imc_imp %>%
    filter(BMI >= 25 & BMI < 30) %>%
    finalfit::summary_factorlist(
      dependent = "Disease_severity_group",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  imc_imp %>%
    filter(BMI >= 30) %>%
    finalfit::summary_factorlist(
      dependent = "Disease_severity_group",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F)
)


tbimc1[2:5,1]  = c("Menos de 18.5",
                   "18.5-24.9",
```

```
                    "25-29.9 (sobrepeso)",
                    "Mas de 30 (obesidad)")
knitr::kable(tbimc1,
             caption = "\\label{table033}Tabla de características
             del IMC según la gravedad de la enfermedad por COVID-19.",
             align = "c")
```

Tabla A.3: Tabla de características del IMC según la gravedad de la enfermedad por COVID-19.

label	levels	Asintomatico/Leve	Moderado	Severo/Critico	Total	p
BMI	Median (IQR)	24.0 (23.0 to 26.0)	26.0 (24.0 to 28.0)	27.5 (26.0 to 32.0)	25.0 (23.0 to 27.0)	<0.001
Menos de 18.5	Median (IQR)	18.0 (15.0 to 18.0)	18.0 (14.5 to 18.0)	0	18.0 (15.0 to 18.0)	0.715
18.5-24.9	Median (IQR)	23.0 (22.0 to 24.0)	23.0 (22.0 to 24.0)	24.0 (24.0 to 24.0)	23.0 (22.0 to 24.0)	0.026
25-29.9 (sobrepeso)	Median (IQR)	26.0 (25.0 to 27.0)	26.0 (25.0 to 28.0)	27.0 (26.0 to 27.8)	26.0 (25.0 to 27.0)	0.039
Mas de 30 (obesidad)	Median (IQR)	32.0 (31.0 to 35.0)	32.0 (31.0 to 33.0)	33.0 (32.0 to 34.2)	32.0 (31.0 to 34.0)	0.324

```
tbimc2 <- rbind(
  imc_imp %>%
    finalfit::summary_factorlist(
      dependent = "Deaths",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  cbind("label" = "Menos de 18.5",
        "levels"= "Median (IQR)",
        "0" = median(imc_imp$BMI[imc_imp$BMI[which(
```

```
          imc_imp$Deaths=="0")]<18]),
        "1" = ifelse(median(imc_imp$BMI[imc_imp$BMI[which(
          imc_imp$Deaths=="1")]<18])=="NA", 0),
        "Total" = median(imc_imp$BMI[which(
          imc_imp$BMI<18)]),
        "p" = NA),
  imc_imp %>%
    filter(BMI >= 18.5 & BMI < 25) %>%
    finalfit::summary_factorlist(
      dependent = "Deaths",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  imc_imp %>%
    filter(BMI >= 25 & BMI < 30) %>%
    finalfit::summary_factorlist(
      dependent = "Deaths",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
      column = TRUE,
      na_include = F),
  imc_imp %>%
    filter(BMI >= 30) %>%
    finalfit::summary_factorlist(
      dependent = "Deaths",
      explanatory = "BMI",
      p = T,
      cont = "median",
      total_col = TRUE,
```

```
      column = TRUE,
      na_include = F)
)


tbimc2[2:5,1]  = c("Menos de 18.5",
                   "18.5-24.9",
                   "25-29.9 (sobrepeso)",
                   "Mas de 30 (obesidad)")
knitr::kable(tbimc2,
             caption = "\\label{table034}Tabla de características
             del IMC según la mortalidad por COVID-19.")
```

Tabla A.4: Tabla de características del IMC según la mortalidad por COVID-19.

label	levels	0	1	Total	p
BMI	Median (IQR)	25.0 (23.0 to 27.0)	28.0 (26.0 to 33.0)	25.0 (23.0 to 27.0)	<0.001
Menos de 18.5	Median (IQR)	15	NA	14.5	NA
18.5-24.9	Median (IQR)	23.0 (22.0 to 24.0)	24.0 (24.0 to 24.0)	23.0 (22.0 to 24.0)	0.242
25-29.9 (sobrepeso)	Median (IQR)	26.0 (25.0 to 27.0)	26.5 (26.0 to 28.0)	26.0 (25.0 to 27.0)	0.214
Mas de 30 (obesidad)	Median (IQR)	32.0 (31.0 to 34.0)	34.0 (33.0 to 35.5)	32.0 (31.0 to 34.0)	0.011

Ahora hacemos la regresión logística simple y múltiple.

```
imc_imp %<>%
  mutate(Disease_severity_group_binary = case_when(
  Disease_severity %in% c(1,2) ~ "No_Severo",
  Disease_severity %in% c(3,4) ~ "Severo"))


imc_imp$Disease_severity_group_binary <- factor(
  imc_imp$Disease_severity_group_binary,
```

```
  levels = c("No_Severo", "Severo"))



imc_imp %<>% mutate(BMI_group = case_when(
  BMI < 25 ~ "Normal",
  BMI >= 25 & BMI < 30 ~ "25-30 (sobrepeso)",
  BMI >= 30  ~ "30 o mas (obesidad)")
)


imc_imp$BMI_group <- factor(imc_imp$BMI_group,
                            levels = c("Normal",
                                       "25-30 (sobrepeso)",
                                       "30 o mas (obesidad)"))


explanatory <- c("Age","Sex",  "Any_comorbidities",
                 "White_blood_cells", "BMI_group")
explanatory_multi <- c("Age", "Any_comorbidities",
                       "White_blood_cells", "BMI_group")
dependent <- "Disease_severity_group_binary"


tbimc3 <- imc_imp %>%
  finalfit::finalfit(dependent,
                     explanatory,
                     explanatory_multi)
knitr::kable(tbimc3,
             caption = "\\label{table035}Regresión logística bivariada de la
             asociación del IMC con la probabilidad de
             enfermedad grave por COVID-19 en Kazajistán.",
             align = c("c", "c", "l", "r", "c", "c"))
```

Tabla A.5: Regresión logística bivariada de la asociación del IMC con la probabilidad de enfermedad grave por COVID-19 en Kazajistán.

	Dependent: Disease_severity_group_binary		No_Severe	Severo	OR (univariable)	OR (multivariable)
1	Age	Mean (SD)	36.8 (17.7)	58.7 (18.6)	1.07 (1.05-1.08, p<0.001)	1.04 (1.01-1.06, p=0.001)
7	Sex	0	567 (96.6)	20 (3.4)	-	-
8		1	457 (94.6)	26 (5.4)	1.61 (0.89-2.96, p=0.116)	-
2	Any_comorbidities	0	627 (97.8)	14 (2.2)	-	-
3		1	397 (92.5)	32 (7.5)	3.61 (1.94-7.06, p<0.001)	1.91 (0.95-4.00, p=0.074)
9	White_blood_cells	Mean (SD)	6.6 (2.3)	7.2 (3.0)	1.10 (0.98-1.24, p=0.098)	1.11 (0.98-1.25, p=0.100)
6	BMI_group	Normal	498 (99.2)	4 (0.8)	-	-
4		25-30 (sobrepeso)	455 (94.6)	26 (5.4)	7.11 (2.75-24.25, p<0.001)	3.60 (1.10-16.32, p=0.055)
5		30 o mas (obesidad)	71 (81.6)	16 (18.4)	28.06 (9.97-100.07, p<0.001)	6.47 (1.43-35.69, p=0.020)

```
explanatory <- c("Age","BMI_group")
explanatory_multi <- c("Age", "BMI_group")
dependent <- "Deaths"


tbimc4 <- imc_imp %>% finalfit::finalfit(dependent, explanatory, explanatory_multi)
```

```
knitr::kable(tbimc4,
             caption = "\\label{table036}Regresión logística bivariada de la
             asociación del IMC con la muerte
             por COVID-19 en Kazajistán.")
```

Tabla A.6: Regresión logística bivariada de la asociación del IMC con la muerte por COVID-19 en Kazajistán.

	Dependent: Deaths		0	1	OR (univariable)	OR (multivariable)
1	Age	Mean (SD)	37.2 (17.8)	68.6 (13.3)	1.10 (1.07-1.13, p<0.001)	1.10 (1.06-1.15, p<0.001)
4	BMI_group	Normal	501 (99.8)	1 (0.2)	-	-
2		25-30 (sobrepeso)	471 (97.9)	10 (2.1)	10.64 (2.03-195.60, p=0.024)	2.02 (0.32-39.18, p=0.525)
3		30 o mas (obesidad)	79 (90.8)	8 (9.2)	50.73 (9.13-947.63, p<0.001)	1.68 (0.17-38.84, p=0.682)

En la regresión logística simple vemos que tener sobrepeso u obesidad es un factor de riesgo para enfermar gravemente o morir por COVID-19. Sin embargo, en la regresión logística múltiple solo la obesidad está asociada a una mayor gravedad de la enfermedad COVID-19.

Bibliografía

Ahrens, W., & Pigeot, I. (2014). *HandBook of epidemiology*. Springer.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). *Rmarkdown: Dynamic documents for r*. https://github.com/rstudio/rmarkdown

Babenko, D. (2021). *COVID-19-in-KZ*. GitHub. https://github.com/dimbage/COVID-19-in-KZ

Bates, Douglas M., & Venables, W. N. (2019). *Splines: Regression splines functions and classes*. https://cran.r-project.org/src/contrib/Archive/splines/

Decorps, J. P. (2021). *EpiStats: Tools for epidemiologists*. https://CRAN.R-project.org/package=EpiStats

Gallego Iborra, A., Moreno Muñoz, G., & Castillo Aguas, G. (2012). Estudios analíticos observacionales (cohortes y casos-controles): Investigando asociaciones causales. *Formación Activa En Pediatría de Atención Primaria*, *5*(4), 227–233. https://fapap.es/files/639-869-RUTA/FAPAP4_2012_07.pdf

Gómez-Gómez, M., Danglot-Banck, C., Huerta, A., & García, T. (2003). El estudio de casos y controles: Su diseño, análisis e interpretación, en investigación clínica. *Rev Mex Pediatr*, *70*(5), 257–263. https://www.medigraphic.com/pdfs/pediat/sp-2003/sp035h.pdf

Harrison, E., Drake, T., & Ots, R. (2021). *Finalfit: Quickly create elegant regression results tables and plots when modelling*. https://github.com/ewenharrison/finalfit

Henquin, R. (2013). *Epidemiología y estadística para principiantes*. Corpus Libros Médicos y Científicos; Corpus Editorial.

Hernández-Ávila, M., Garrido-Latorre, F., & López-Moreno, S. (2000). Diseño de estudios epidemiológicos. *Salud Pública de México*, *42*, 144–154. https://www.scielosp.org/pdf/spm/2000.v42n2/144-154/es

Kassambara, A. (2021). *Rstatix: Pipe-friendly framework for basic statistical tests*. https://rpkgs.datanovia.com/rstatix/

Kestenbaum, B. (2009). *Epidemiology and biostatistics: An introduction to clinical research*. Springer.

Kirk, M. (2013). *Analysis of Public Health Data* (A. Housen Tambri y Richardson, Ed.). Australian National University; RPubs. https://rpubs.com/Mindy_20/case_control

Lazcano-Ponce, E., Salazar-Martínez, E., & Hernández-Ávila, M. (2001). Estudios epidemiológicos de casos y controles. Fundamento teórico, variantes y aplicaciones. *Salud Pública de México*, *43*(2), 135–150. http://www.scielo.org.mx/pdf/spm/v43n2/a09v43n2.pdf

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Luque-Calvo, P. L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. http://destio.us.es/calvo

Luque-Calvo, P. L. (2019). *Cómo crear Tablas de información en R Markdown*. http://destio.us.es/calvo

Mayer, M. (2021). *MissRanger: Fast imputation of missing values*. https://github.com/mayer79/missRanger

Mirón Canelo, J. A., & Alonso Sardón, M. (2008). Medidas de frecuencia, asociación e impacto en investigación aplicada. *Medicina Y Seguridad Del Trabajo*, *54*(211), 93–102. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0465-546X2008000200011

Molina Arias, M. (2016). Diseños híbridos. *Evidencias En Pediatría. Asociación Española de Pediatría*. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322016000100021

Molina Arias, M., & Ochoa Sangrador, C. (2013). Estudios observacionales (I). Estudios transversales.Medidas de frecuencia. Técnicas de muestreo. *Evidencias En Pediatría. Asociación Española de Pediatría*, 9–72. https://evidenciasenpediatria.es/files/41-12105-RUTA/72Fundamentos.pdf

Molina Arias, M., & Ochoa Sangrador, C. (2014). Estudios observacionales (II). Estudios de cohortes. *Evidencias En Pediatría. Asociación Española de Pediatría*, 10–14. https://evidenciasenpediatria.es/files/41-12164-RUTA/014Fundamentos.pdf

Moreno-Altamirano, A., López-Moreno, S., & Corcho-Berdugo, A. (2000). Principales medidas en epidemiología. *Salud Pública de México*, *42*, 337–348. https://www.scielosp.org/article/spm/2000.v42n4/337-348/es/

Olsen, J., Christensen, K., Murray, J., & Ekbom, A. (2010). *An introduction to epidemiology for health professionals*. Springer.

Pallarés Mestre, J. (2016). *La metodología cuantitativa aplicada al estudio de la reincidencia en menores infractores* [PhD thesis, Universitat Jaume I]. https://www.tdx.cat/handle/10803/432779#page=1

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Royo-Bordonada, M. Á., Damián, J., Pérez-Gómez, B., Rodríguez-Artalejo, F., Villar Álvarez, F., López-Abente, G., Imaz-Iglesia, I., Castilla Catalán, J., González-Enriquez, J., & Martín Moreno, J. M. (2009). *Método epidemiológico*. https://repisalud.isciii.es/bitstream/handle/20.500.12105/5271/M%c3%a9todoEpidemiol%c3%b3gico_2009.pdf?sequence=1&isAllowed=y

RStudio Team. (2015). *RStudio: Integrated development environment for r*. RStudio, Inc. http://www.rstudio.com/

Signorell, A. (2021). *DescTools: Tools for descriptive statistics*. https://CRAN.R-project.org/package=DescTools

Solís Sánchez, G., & Orejas Rodríguez-Arango, G. (1999). Epidemiología y metodología científica aplicada a la pediatría (VI): Confusión e interacción. *An Esp Pediatr*, *51*, 91–96. https://www.aeped.es/sites/default/files/anales/51-1-23.pdf

Stevenson, M., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., Sebastiani, P., & Solymos, P. (2021). *EpiR: Tools for the analysis of epidemiological data*. https://CRAN.R-project.org/package=epiR

Therneau, T. M. (2020). *Survival: Survival analysis*. https://github.com/therneau/survival

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2021). *Ggplot2: Create elegant data visualisations using the grammar of graphics*. https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Xie, Y. (2020). *Knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/

Yegorov, S., Goremykina, M., Ivanova, R., Good, S. V., Babenko, D., Shevtsov, A., MacDonald, K. S., & Zhunussov, Y. (2021a). Appendix: Epidemiology, clinical characteristics, and virologic

features of COVID-19 patients in Kazakhstan: A nation-wide retrospective cohort study. *The Lancet Regional Health - Europe*. https://www.thelancet.com/cms/10.1016/j.lanepe.2021.100096/attachment/e1e63a9b-358b-4fc3-890b-a9c0629e173b/mmc1.pdf

Yegorov, S., Goremykina, M., Ivanova, R., Good, S. V., Babenko, D., Shevtsov, A., MacDonald, K. S., & Zhunussov, Y. (2021b). Epidemiology, clinical characteristics, and virologic features of COVID-19 patients in Kazakhstan: A nation-wide retrospective cohort study. *The Lancet Regional Health - Europe*. https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(21)00073-9/fulltext#seccesectitle0025