

# Estudio de la mortalidad entre grupos de pacientes con artritis reumatoide: análisis de Electronic Health Records

Marta Uceda Martin  
Master en Bioestadística y Bioinformática  
Área 2

**Profesora Consultora: Nuria Perez Álvarez**  
**Profesor responsable de la asignatura: Marc Maceira Duch.**

08 Junio 2021



Esta obra está sujeta a una licencia de Reconocimiento [3.0 España de Creative Commons](https://creativecommons.org/licenses/by/3.0/es/)

## FICHA DEL TRABAJO FINAL

<b>Título del trabajo:</b>	<i>Estudio de la mortalidad entre grupos de pacientes con artritis reumatoide: análisis de Electronic Health Records</i>
<b>Nombre del autor:</b>	<i>Marta Uceda Martin</i>
<b>Nombre del consultor/a:</b>	<i>Nuria Perez Alvarez</i>
<b>Nombre del PRA:</b>	<i>Marc Maceira Duch</i>
<b>Fecha de entrega (mm/aaaa):</b>	06/2021
<b>Titulación:</b>	<i>Master en bioestadística y bioinformática</i>
<b>Área del Trabajo Final:</b>	Análisis de datos
<b>Idioma del trabajo:</b>	Español
<b>Número de créditos:</b>	3
<b>Palabras clave</b>	<i>Electronic Health Records, artritis reumatoide, farmacovigilancia, SQL, R.</i>
<b>Resumen del Trabajo:</b>	
<p>Los registros electrónicos de salud (o Electronic Health Records) constituyen un nuevo tipo de fuente de datos para estudios no intervencionales. La terapia biológica inmunomoduladora usada para tratar la artritis reumatoide presenta ciertos efectos adversos asociados a su uso. Aunque estos efectos adversos son conocidos, existen pocos estudios comparativos entre los distintos medicamentos biológicos en términos de mortalidad. En este estudio se comparó la mortalidad (en el hospital y según los registros de la seguridad social) entre grupos de pacientes de artritis tratados con los distintos medicamentos biológicos. Se estudiaron también los mismos indicadores de mortalidad entre grupos de pacientes de artritis tratados con y sin medicamento biológico. Sólo se encontraron diferencias significativas en cuanto a la mortalidad según los registros de la seguridad social entre pacientes con y sin medicamento biológico. En este caso el Odds Ratio fue de 2.52, siendo 2.52 veces más probable sobrevivir si estás en el grupo sin medicamento biológico. Estos resultados deben ser interpretados con cautela puesto que el estudio presenta ciertas limitaciones, como que los grupos de pacientes obtenidos eran muy pequeños para algunos casos y que los resultados no permiten establecer causalidad. No se han encontrado estudios similares con este u otro tipo de datos. Aún así, los registros electrónicos de salud, se presentan como importantes fuente de datos que permitirán complementar los otros tipos de estudios ya existentes (intervencionales y no intervencionales). Es por ello que su estudio está en continua expansión desde hace una década.</p>	

**Abstract:**

Electronic Health Records are a new data source for non-interventional studies. Biological disease-modifying anti-rheumatic therapy used to treat the rheumatoid arthritis has some associated adverse events. Even though some of these adverse events are known, there are few studies comparing mortality among the different biological drugs. In this study, we compared mortality (in hospital, and according to the social security registries) among groups of patients treated with the different biological drugs. The same indicators for mortality were studied between group of patients treated with and without biological drugs. No significant differences in mortality were found, except for mortality according to the social security registries for patients treated with or without biological drugs, with an Odds Ratio of 2.52, meaning that survival is 2.52 times more likely when a patient is not treated with a biological drug. Those results should be interpreted with caution, as this study has some limitations, such as the obtained cohort was small in number for some cases, and those results do not allow to establish a causality. No similar studies were found with this kind or another type of data. Nevertheless, the Electronic Health Records are an important source of data that will complement other type of existing studies (interventional and non-interventional), and its study has been constantly growing for over a decade.

# Índice

<b>LISTA DE FIGURAS.....</b>	<b>A</b>
<b>LISTA DE TABLAS.....</b>	<b>B</b>
<b>1 RESUMEN .....</b>	<b>1</b>
<b>2 INTRODUCCIÓN.....</b>	<b>1</b>
2.1 CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO .....	1
2.1.1 Descripción general.....	1
2.1.2 Justificación del Trabajo Fin de Master (TFM).....	3
2.2 OBJETIVOS DEL TRABAJO .....	3
2.2.1 Objetivos generales.....	3
2.2.2 Objetivos específicos.....	3
2.2.3 Adaptaciones de los objetivos específicos.....	3
2.3 ENFOQUE Y MÉTODO SEGUIDO .....	4
2.4 PLANIFICACIÓN DEL TRABAJO .....	4
2.4.1 Tareas.....	4
2.4.2 Calendario .....	5
2.5 BREVE SUMARIO DE CONTRIBUCIONES Y PRODUCTOS OBTENIDOS.....	5
2.6 BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA .....	6
<b>3 ESTADO DEL ARTE .....</b>	<b>6</b>
3.1 ELECTRONIC HEALTH RECORDS.....	6
3.1.1 Historia .....	6
3.1.2 Usos de los EHR .....	7
3.1.3 Limitaciones: sesgos en los datos .....	9
3.1.4 Limitaciones: políticas de protección de datos y privacidad.....	9
3.2 LA BASE DE DATOS MIMIC.....	10
3.3 LA ARTRITIS REUMATOIDE.....	12
<b>4 METODOLOGÍA.....</b>	<b>13</b>
4.1 ACCESO A LOS DATOS.....	13
4.2 DESCARGA E INSTALACIÓN DE MIMIC.....	13
4.3 SELECCIÓN DE LA COHORTE .....	14
4.3.1 Selección de la cohorte con medicamento biológico.....	14
4.3.2 Selección de la cohorte sin medicamento biológico .....	15
4.4 ANÁLISIS DE DATOS .....	16
<b>5 RESULTADOS .....</b>	<b>17</b>
5.1 COMPARACIÓN DE LA MORTALIDAD ENTRE PACIENTES CON AR TRATADOS CON LOS DISTINTOS MEDICAMENTOS BIOLÓGICOS .....	17
5.1.1 Definición de la cohorte con medicamento biológico.....	17
5.1.2 Mortalidad en el hospital (“hospital_expire_flag”).....	19
5.1.3 Mortalidad según los registros de la seguridad social (“expire_flag”).....	19
5.2 COMPARACIÓN DE LA MORTALIDAD ENTRE PACIENTES CON AR TRATADOS CON MEDICAMENTOS BIOLÓGICOS O NO BIOLÓGICOS .....	20
5.2.1 Definición de la cohorte sin medicamento biológico.....	20
5.2.2 Mortalidad en el hospital (“hospital_expire_flag”).....	21
5.2.3 Mortalidad según los registros de la seguridad social. (“expire_flag”).....	22
<b>6 DISCUSIÓN .....</b>	<b>23</b>
6.1 SOBRE LOS RESULTADOS .....	23
6.2 SOBRE LOS USOS, VENTAJAS Y LIMITACIONES DE LOS EHR .....	24

<b>7</b>	<b>VALORACIÓN ECONÓMICA</b> .....	<b>24</b>
<b>8</b>	<b>CONCLUSIONES</b> .....	<b>25</b>
8.1	CONCLUSIONES.....	25
8.2	LÍNEAS DE FUTURO.....	25
8.3	SEGUIMIENTO DE LA PLANIFICACIÓN.....	27
<b>9</b>	<b>GLOSARIO</b> .....	<b>I</b>
<b>10</b>	<b>BIBLIOGRAFÍA</b> .....	<b>II</b>
<b>11</b>	<b>ANEXOS</b> .....	<b>IV</b>
11.1	ANEXO I: CERTIFICADO DEL CURSO “DATA OR SPECIMENS ONLY RESEARCH”, COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM).....	IV
11.2	ANEXO II: CODIGO SQL PARA LA SELECCIÓN DE LAS COHORTE.....	V
11.3	ANEXO III: CÓDIGO R PARA EL ANÁLISIS DE LAS COHORTE Y ESTUDIO DE LAS DIFERENCIAS EN LA MORTALIDAD.....	XI
11.3.1	<i>Comparación de la mortalidad entre pacientes con AR tratados con los distintos medicamentos biológicos.....</i>	<i>XI</i>
11.3.2	<i>Comparación de la mortalidad entre pacientes con AR tratados con medicamentos biológicos o no biológicos.....</i>	<i>XXIII</i>

## **Lista de figuras**

**Figura 1:** Gráfico del número de citas que incluyen los términos Electronic Health Record, por año.

**Figura 2:** Esquema del calendario propuesto para la realización del TFM.

**Figura 3:** Esquema representativo de una base de datos con EHR y los diferentes usos que se pueden dar a los datos.

**Figura 4:** Listado de medicamentos biológicos para la AR (en las cajas rojas) junto con sus respectivas dianas terapéuticas.

**Figura 5:** Esquema del proceso de selección de la cohorte de pacientes con medicamento biológico.

## Lista de tablas

**Tabla 1:** Visión general de las tablas contenidas en la base de datos MIMIC-III.

**Tabla 2:** Descripción de la cohorte de pacientes con medicamento biológico.

**Tabla 3:** Tabla de contingencia sobre el estatus del indicador de mortalidad “hospital\_expire\_flag”, para cada uno de los medicamentos biológicos para tratar la AR

**Tabla 4:** Tabla de contingencia sobre el estatus del indicador de mortalidad “expire\_flag”, para cada uno de los medicamentos biológicos para tratar la AR.

**Tabla 5:** Descripción de la cohorte de pacientes sin medicamento biológico.

**Tabla 6:** Tabla de contingencia sobre el estatus del indicador de mortalidad “hospital\_expire\_flag”, para los dos tipos de medicamentos (biológicos y no biológicos) para tratar la AR.

**Tabla 7:** Tabla de contingencia sobre el estatus del indicador de mortalidad “expire\_flag”, para los dos tipos de medicamentos (biológicos y no biológicos) para tratar la AR.



# 1 Resumen

Los registros electrónicos de salud (o Electronic Health Records) constituyen un nuevo tipo de fuente de datos para estudios no intervencionales. La terapia biológica inmunomoduladora usada para tratar la artritis reumatoide presenta ciertos efectos adversos asociados a su uso. Aunque estos efectos adversos son conocidos, existen pocos estudios comparativos entre los distintos medicamentos biológicos en términos de mortalidad. En este estudio se comparó la mortalidad (en el hospital y según los registros de la seguridad social) entre grupos de pacientes de artritis tratados con los distintos medicamentos biológicos. Se estudiaron también los mismos indicadores de mortalidad entre grupos de pacientes de artritis tratados con y sin medicamento biológico. Sólo se encontraron diferencias significativas en cuanto a la mortalidad según los registros de la seguridad social entre pacientes con y sin medicamento biológico. En este caso el Odds Ratio fue de 2.52, siendo 2.52 veces más probable sobrevivir si estás en el grupo sin medicamento biológico. Estos resultados deben ser interpretados con cautela puesto que el estudio presenta ciertas limitaciones, como que los grupos de pacientes obtenidos eran muy pequeños para algunos casos y que los resultados no permiten establecer causalidad. No se han encontrado estudios similares con este u otro tipo de datos. Aún así, los registros electrónicos de salud, se presentan como importantes fuente de datos que permitirán complementar los otros tipos de estudios ya existentes (intervencionales y no intervencionales). Es por ello que su estudio está en continua expansión desde hace una década.

## 2 Introducción

### 2.1 Contexto y justificación del trabajo

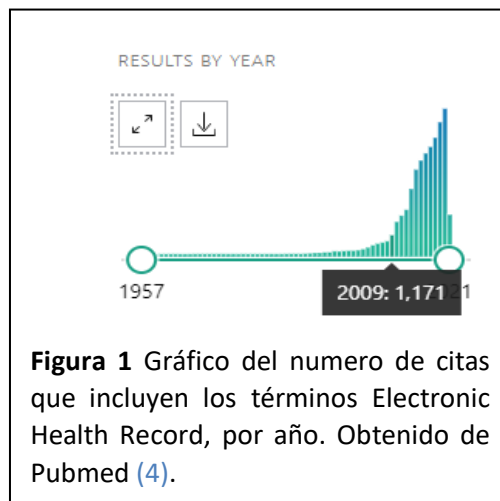
#### 2.1.1 Descripción general

La tecnología digital está transformando nuestra realidad en todos los aspectos de nuestro día a día, y el terreno de la salud no es una excepción. Dentro de este campo, el proceso de digitalización ha incluido la creación de los “Registros Electrónicos de Salud” (1, 2).

Los “Registros Electrónicos de Salud”, del inglés “Electronic Health Records” (EHR) son la versión digital de los clásicos informes médicos en papel. Estos poseen la ventaja de ser un registro a tiempo real de la información relativa a la salud de un paciente, pudiendo contener información sobre su historial médico, diagnósticos, prescripciones médicas (pasadas y presentes), resultados de análisis de laboratorio, carné de vacunaciones, alergias, imágenes de diagnóstico, notas del personal sanitario, etc (2, 3).

El uso de los EHR ha aumentado en los últimos años, y paralelo al uso ha aumentado también su estudio. Una muestra de ello son los resultados obtenidos al hacer una búsqueda simple en Pubmed con los términos “Electronic Health Records” (4), y es que el número de publicaciones que incluyen esos términos ha crecido exponencialmente desde 2009, como se puede ver en **Figura 1**.

Los EHR fueron creados originalmente con el propósito de facilitar los procesos de reembolso y copago que existen en Estados Unidos (EE. UU) (3), pero han ido evolucionando progresivamente. En la actualidad, el objetivo primario de los EHR es tener un registro del proceso y los resultados de cada una de las intervenciones que se hacen desde los distintos establecimientos sanitarios sobre un paciente. Esto hace que se generen base de datos dinámicas y en continuo crecimiento, que pueden proporcionar información interesante para su estudio. Entender en qué consisten estas bases de datos y saber bien cómo explotar e interpretar la información contenida en ellos puede proporcionar grandes ventajas a los procesos de toma de decisiones y cuidados del paciente (1).



Concretamente, en el área de la farmacovigilancia, se ha visto que el uso de EHR ayuda a disminuir los eventos adversos (AEs, del inglés “adverse events”) y errores de medicación (MEs, del inglés “medication errors”). Dentro de esta misma área, los EHR pueden funcionar como herramienta de detección de AEs no previamente observados durante los ensayos clínicos (por su baja incidencia o por necesitar un periodo de exposición prolongado) (1, 5). Tradicionalmente, el rastreo de efectos adversos una vez el medicamento ha salido al mercado se ha fundamentado en las notificaciones espontáneas por parte de sanitarios, y ciertos estudios señalan que este tipo de eventos pueden estar infrarreportados (6, 7). Los EHR se presentan como importantes potenciales aliados en este sector, ya que los datos contenidos en ellos pueden aportar información sobre riesgos asociados al uso de ciertos medicamentos o facilitar la detección de otros AEs no previamente identificados y espontáneamente notificados (5).

El riesgo aumentado de infecciones es un efecto adverso conocido de la terapia inmunomoduladora usada para tratar la artritis reumatoide (AR), y constituye un porcentaje importante de morbimortalidad en pacientes con esta afección. Sin embargo, existen pocos estudios que hagan comparaciones entre medicamentos biológicos en términos de morbimortalidad, y los que existen presentan limitaciones (8).

Con este trabajo pretendemos aportar evidencia sobre la utilidad de los EHR como fuente de información clínica. Para ello, estudiaremos la diferencia en mortalidad por todas las causas y mortalidad por septicemia entre pacientes de AR tratados con los diferentes medicamentos biológicos. De la misma manera, dado que dichos medicamentos presentan diferentes mecanismos de acción, realizaremos el mismo estudio, pero agrupando medicamentos en función del mecanismo de acción, para observar si existen diferencias en la mortalidad.

### 2.1.2 Justificación del Trabajo Fin de Master (TFM)

Como se ha mencionado anteriormente, el uso de los EHR está en continua expansión y sus ventajas e inconvenientes están siendo aún discutidos. Si bien el uso de estas bases de datos presenta limitaciones, éstas no deben ser motivo de exclusión para su uso, ya que un correcto manejo de ellas y de la información obtenida de las mismas podría tener un conlleva numerosos beneficios que aconsejan su fomento. Entre ellos cabe citar su potencial impacto en la calidad de la atención sanitaria brindada al paciente, la reducción de riesgos y costes sanitarios, etc.

Desde que me licencié en Farmacia llevo desarrollando mi actividad profesional en el ámbito de la farmacovigilancia, trabajando con distintos tipos de datos provenientes de diferentes fuentes (notificaciones espontaneas, bases de datos originadas a partir de programas de apoyo a pacientes, foros de internet, redes sociales, etc). En la gran mayoría de casos se trataba de datos no estructurados que había que procesar para introducirlos en la base de datos de farmacovigilancia de forma estructurada. Muy frecuentemente trabajamos con informes médicos en papel (escaneados o en formato .pdf) de los que se extrae de forma manual o semimanual la información necesaria para las notificaciones de farmacovigilancia. Desde el punto de vista de la explotación de datos, el estudio de los EHR se presenta como una oportunidad tanto a nivel académico como profesional por ser este un terreno novedoso para la comunidad científica y por su potencial de aplicación en mi terreno laboral.

## **2.2 Objetivos del Trabajo**

### 2.2.1 Objetivos generales

- A.1: Aportar nueva evidencia sobre el interés del uso de los EHR como fuente de información.
- A.2: Discutir las ventajas y limitaciones de los EHR.

### 2.2.2 Objetivos específicos

- B.1: Estudiar diferencia en la mortalidad por todas las causas y mortalidad por sepsis entre grupos de pacientes adultos con AR tratados con los distintos medicamentos biológicos.
- B.2: Estudiar diferencia en la mortalidad por todas las causas y mortalidad por sepsis entre pacientes adultos con AR tratados con medicamentos biológicos, agrupados por mecanismo de acción.
- B.3: Elaborar un breve estado del arte sobre los EHR.
- B.4: Estudiar la metodología de análisis de datos procedentes de EHR; preparar un programa de análisis con Rmarkdown.

### 2.2.3 Adaptaciones de los objetivos específicos

A lo largo del desarrollo del presente trabajo fue necesario realizar adaptaciones sobre los objetivos específicos.

La primera de ellas es que, debido a que la causa de muerte no viene dada en la base de datos MIMIC, el estudio de la mortalidad por sepsis entre pacientes tratados con medicamentos biológicos y entre pacientes con medicamento biológico y no biológico no podrá ser analizado.

Después de la selección de la cohorte (con medicamento biológico, apartado 4.3.1) se observó que la mayoría de los pacientes están siendo tratados con medicamentos que comparten mecanismo de acción. El punto B.2 no pudo ser realizado, ya que solamente se obtuvo un grupo con una muestra suficientemente grande.

Alternativamente a estas comparaciones que no han podido realizarse, se extrajo una cohorte de pacientes adultos con AR que no estaban recibiendo tratamiento biológico y se comparó la mortalidad de este grupo y la de pacientes de AR tratados con algún medicamento biológico.

Además, se vio que en la base de datos existen dos indicadores de mortalidad, uno que indica la mortalidad en hospital y otra según los registros de la seguridad social (SS, que tienen en cuenta también la mortalidad en el hospital). Puesto que existen estos indicadores, los estudiamos por separado.

Teniendo en cuenta todas las adaptaciones previamente detalladas, los objetivos específicos quedarían adaptados de la siguiente manera:

- B.1': Estudiar diferencia en la mortalidad (en el hospital y según los registros de la SS) entre grupos de pacientes adultos con AR tratados con los distintos medicamentos biológicos.
- B.2': Estudiar diferencia en la mortalidad (en el hospital y según los registros de la SS) entre pacientes adultos con AR, tratados con los diferentes tipos de medicamentos para la AR: biológicos y no biológicos.
- B.3': Elaborar un breve estado del arte sobre los EHR.
- B.4': Estudiar la metodología de análisis de datos procedentes de EHR; preparar un programa de análisis con Rmarkdown.

## **2.3 Enfoque y método seguido**

La base de datos utilizada es la "Medical Information Mart for Intensive Care III" (MIMIC-III), por no disponer de datos propios y ofrecer éste acceso gratuito a la comunidad científica e investigadora. Para la selección de cohorte, los datos serán explotados mediante SQL, ya que esta base de datos está estructurada en diferentes tablas. La información de interés será aunada en una sola tabla para poder realizar los tests estadísticos con Rstudio.

## **2.4 Planificación del Trabajo**

### **2.4.1 Tareas**

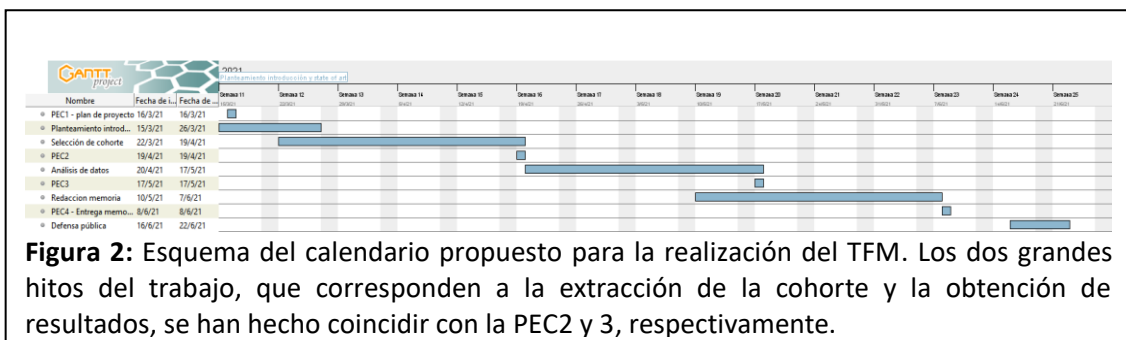
En cuanto al tratamiento de los datos se refiere, después de haber reunido los requisitos y conseguir acceso a la base de datos MIMIC se deberán seleccionar las cohortes

necesarias, asegurándose que la información obtenida sea adecuada para el proceso de análisis de datos que vendrá después.

La memoria del TFM debe recoger todo el proceso anteriormente descrito, junto con el contexto teórico y las conclusiones que se destilan del estudio realizado. Esta memoria, además de entregarse como documento escrito, deberá defenderse ante un tribunal, lo que implica también la preparación de los soportes digitales.

## 2.4.2 Calendario

La primera fase del proyecto consistió en realizar los cursos y reunir los requisitos para poder solicitar el acceso a los datos. Una vez hecho esto, el permiso fue concedido a los pocos días hábiles. A continuación se pasó a la fase de selección de la cohorte, a la que se le quiso dedicar suficiente tiempo para asegurar que estuviese correctamente realizada para no entorpecer fases posteriores. La fecha límite de esta fase se hizo coincidir con una de las entregas intermedias (PEC) de la memoria. Después se pasó a la fase de análisis de datos y resultados, cuya realización se hizo coincidir también con otra de las entregas intermedias. La redacción de la introducción de la memoria se planificó para las primeras fases del calendario, entendiendo que podría sufrir modificaciones a medida que se ahondase en los conocimientos sobre el tema. El resto de la memoria se fue redactando a medida que se iba avanzando en las diferentes fases. La entrega final de la memoria fue el 8 de junio, y su defensa está planificada para las semanas posteriores. (**Figura 2**)



## 2.5 Breve resumen de contribuciones y productos obtenidos

- Plan de trabajo: documento de presentación del proyecto.
- Memoria: documento actual.
- Producto: En función del resultado final y la calidad de este, podría discutirse el interés de publicar los resultados en forma de artículo científico.
- Presentación virtual: Los resultados contenidos en la memoria serán presentados frente a un tribunal científico-académico para su evaluación.
- Código SQL del proceso de selección de las cohortes.
- Código R de análisis de los datos.
- Certificado del curso “Data or Specimens Only Research”, ofrecido por la Collaborative Institutional Training Initiative (CITI Program).

## 2.6 Breve descripción de los otros capítulos de la memoria

- Estado del arte: introducción a la historia de los EHR y discusión sobre sus ventajas y límites. Presentación de la base de datos MIMIC y pequeño resumen de la AR y sus tratamientos.
- Metodología: Descripción del proceso de acceso a datos, de la creación de la estructura de las tablas e inclusión de los datos en las mismas. Descripción del proceso de selección de las cohortes. Presentación de los métodos estadísticos usados para el análisis de los datos.
- Resultados: Resultados del estudio de las diferencias de mortalidad (para dos indicadores distintos) entre pacientes de AR tratados con los distintos medicamentos biológicos. También se muestran las comparaciones de mortalidad entre pacientes de AR con y sin medicamento biológico.
- Discusión: valoración de los resultados obtenidos.
- Valoración económica: estimación aproximada de las implicaciones económicas del presente trabajo.
- Conclusiones: discusión de los resultados en el contexto de los EHR y valoración de las posibles líneas de futuro.

# 3 Estado del arte

## 3.1 Electronic Health records

### 3.1.1 Historia

Existen evidencias de la existencia de registros médicos desde la edad antigua, aunque no empezaron a usarse de forma regular hasta principio del siglo XX. Tradicionalmente, los registros médicos tenían formato papel, guardándose en archivadores dentro de armarios y por lo general solo existía una copia. Más tarde, en la década de los sesenta y los setenta, los avances en la tecnología computacional permitieron sentar las bases para la aparición, unos años después, de los primeros EHR. Estos a menudo fueron creados para facilitar los sistemas de reembolso y copago existentes en EE. UU. o para favorecer la organización del hospital, y muy pocos contenían información médica de los pacientes.

Las limitaciones de los informes en papel fueron haciéndose más evidentes de manera progresiva, por lo que ciertas instituciones empezaron a impulsar el cambio hacia los registros electrónicos. Sin embargo, esta transición se alargó y retrasó debido a los grandes costes que esto suponía para la época, la baja aceptación por parte de los médicos, los errores al registrar los datos (que se hacía de forma manual) y la falta de alicientes. Todas estas razones hicieron que sólo la información clave fuese registrada electrónicamente, y hubo una fase en la que los archivos electrónicos y en papel coexistieron y se complementaron.

El uso de los EHR fue popularizándose a medida que se normalizaban los nuevos avances tecnológicos: de manera destacada, el hardware y el software necesarios se volvieron

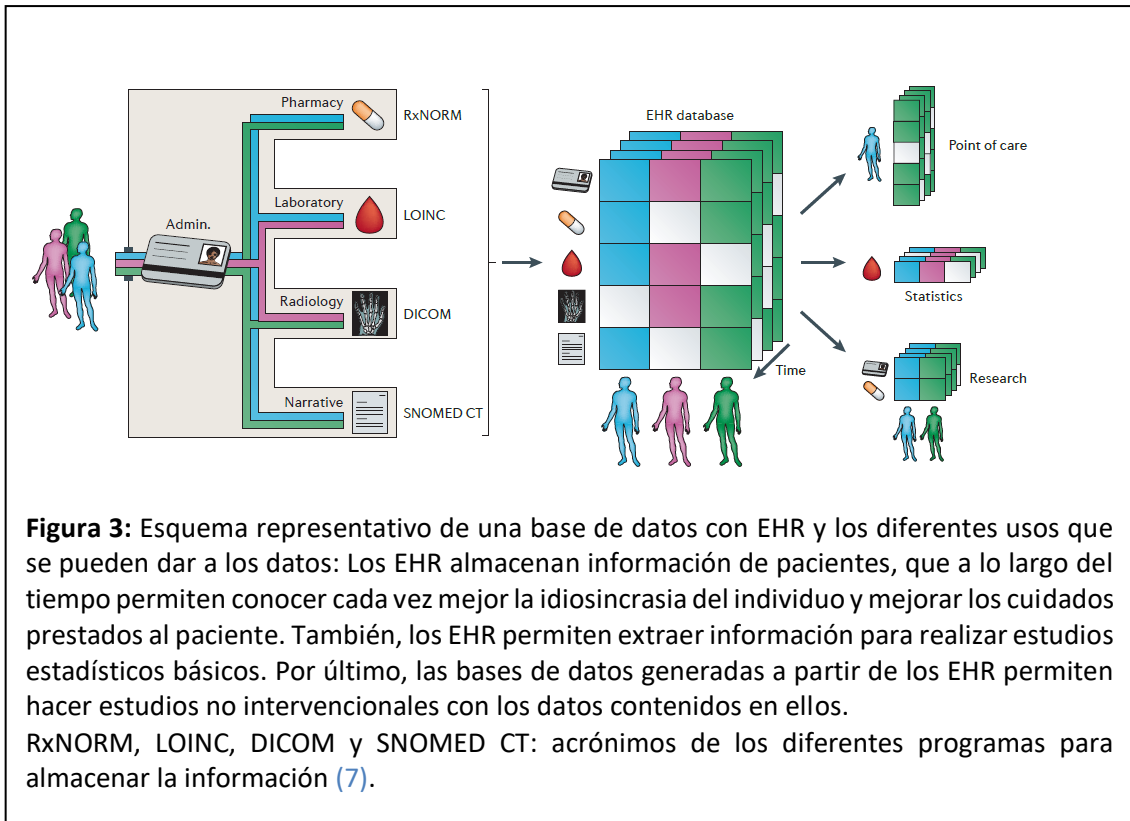
más potentes y accesibles económicamente, se establecieron interacciones entre los sistemas de EHR y otros dispositivos médicos que permitieron que la información quedara registrada automáticamente (eliminando la entrada manual de datos), se desarrollaron sistemas para la consulta de información paralela (como manuales médicos de referencia, listado de medicamentos, etc.), se permitió la firma electrónica del facultativo dentro del sistema de EHR, etc.

Actualmente, muchos programas que integran EHR cuentan también con un sistema de apoyo en la toma de decisiones que interactúa con el facultativo y que puede lanzar recordatorios sobre las interacciones del medicamento o señalar alergias del paciente. El uso de este tipo de registros, que comenzó en los hospitales (especialmente en las unidades de críticos), se encuentra actualmente expandido también entre la atención primaria, las residencias (de ancianos, de rehabilitación, psiquiátricas, etc.) y otros establecimientos sanitarios, lo que facilita enormemente la comunicación entre las distintas entidades. Todo esto repercute de manera directa en la calidad del servicio ofrecido al paciente. Además, los programas informáticos basados en los EHR pueden contar con una interfaz a través de la cual el paciente puede acceder a sus propios datos médicos, que resulta igualmente beneficioso en un modelo de medicina basado en el paciente (3).

### 3.1.2 Usos de los EHR

Los usos y aplicaciones de los EHR han ido evolucionando y definiéndose a medida que los avances en tecnología e investigación han permitido abrir nuevos campos de actuación para estos. En la actualidad existen tres grandes áreas de aplicación donde el uso de los EHR es relevante, y aún se sigue estudiando su potencial (7):

- Mejoras en la atención sanitaria al paciente: en el proceso de toma de decisiones, el facultativo debe elegir la mejor opción para un paciente concreto y sus circunstancias en función de la evidencia científica disponible en ese momento. Como se ha avanzado previamente, los sistemas informáticos que incorporan EHR ayudan en este proceso, ya que pueden analizar la totalidad de registros que existen en la base de datos para ese paciente y, mediante un sistema de apoyo a la toma de decisiones, lanzar alertas o recordatorios al facultativo, informando sobre alergias (medicamentosas o alimentarias) y duplicidades en prescripciones, así como recordatorios sobre la necesidad de algún tipo de intervención periódica. Además, algunos cuentan también con un sistema de alerta de detección precoz cuando los datos muestran indicios de alguna patología emergente.



- **Investigación:** el estudio de cohortes extraídas de los EHR permite el establecimiento de correlaciones, como las que pueden darse entre dos condiciones clínicas distintas, evento-enfermedad, medicamento-evento, etc. Esto puede ser especialmente interesante en la detección y estudio de eventos con baja incidencia y/o desconocidos, puesto que la información registrada se va acumulando (1). El establecimiento de ciertas correlaciones permite hacer predicciones a futuro (siempre que las variables predictoras estén bien escogidas), como por ejemplo el riesgo de enfermedad coronaria o ciertos tipos de cáncer atendiendo a variables demográficas como edad, sexo, uso de tabaco, etc. Las investigaciones con EHR suponen un nuevo tipo de investigación dentro de los estudios no intervencionales.
- **Estadística:** los EHR contienen datos que pueden ser de interés simplemente por motivos administrativos, sin que haya un estudio de investigación detrás como los previamente mencionados. Algunos ejemplos de datos estadísticos relevantes que permiten conocer son el número diario de admisiones hospitalarias, cuántas veces se ha prescrito un medicamento en concreto o el número de veces que se ha hecho un test de laboratorio.



### 3.1.3 Limitaciones: sesgos en los datos

Debido a que los EHR pueden contener varios sesgos, a la hora de trabajar con ellos hay ciertos elementos que deben ser tenidos en cuenta para evaluar si nuestros datos son adecuados para el estudio que queremos realizar.

El primero es que un evento *tiene que darse*, y ese evento debe ser medible y medido. Esto, que puede resultar obvio, depende de muchos factores: de la necesidad de hacer esa medición (basada en protocolos o en el juicio médico), de que exista interacción paciente-profesional de la salud y de los distintos alicientes existentes para que los profesionales de salud realicen dicha medición.

El evento tiene que ser registrado en el sistema, pero no todas las consultas están dotadas de estos sistemas informáticos; de hecho, el porcentaje de unidades de salud que incluyen su uso varía mucho entre países y en función de si se trata de hospitales o de consultas de atención primaria. Este acto se ve también influenciado por alicientes al personal sanitario para realizar el registro (ya sean económicos o de otra clase, si los hubiese), por la existencia de protocolos o por la posibilidad de que los pacientes u otros profesionales de la salud tengan acceso a las anotaciones (lo cual puede influenciar la manera en la que se lleva a cabo el registro e incluso condicionar si se hace o no).

Por otro lado, para poder trabajar con EHR es necesario extraer la información de los sistemas informáticos que la albergan. Este acto tiene también sus entresijos técnicos y puede afectar a los datos obtenidos, ya que los sistemas informáticos funcionan de manera diferente según la herramienta de extracción utilizada y los detalles de este proceso a menudo no son públicos por razones de propiedad intelectual.

La información extraída tiene que realmacenarse en una base de datos para su posterior análisis, y esta, a su vez, debe tener capacidad suficiente para albergar la información extraída. A la hora de analizar los datos hay que tener en cuenta si la información proviene de un único sistema o de varios, ya que pueden existir (o no) protocolos para la anotación de los datos y, en caso de haberlos, estos pueden variar de una base a otra. Del mismo modo, es necesario saber si estos protocolos están unificados y si ha habido cambios durante el periodo de tiempo que concierne a nuestros datos. A menudo los investigadores sólo cuentan con un subconjunto de la base de datos completa (“need to know principle”), por lo que existe el riesgo de perder información al crear el subconjunto si la información sobre la que se trabaja está relacionada con otra que no se encuentre disponible.

Por último, aunque esto no es un problema global en el análisis de datos y tampoco es exclusivo de los EHR, los métodos elegidos para el análisis van a tener un impacto también en los resultados (9).

### 3.1.4 Limitaciones: políticas de protección de datos y privacidad

Existen normas y regulaciones para garantizar la seguridad y privacidad de la información de los pacientes. Estas son:

- HIPAA (EE. UU.): ley creada con el objetivo de proteger la información relativa a la salud de los pacientes. También incluye especificaciones sobre la responsabilidad sobre la protección de los datos (10).
- ISO 13606 (Europa): normativa para asegurar la interoperabilidad de los EHR, incluye también directivas sobre la privacidad y seguridad (11).
- ISO 27799 (Europa): normativa relativa a los estándares de seguridad de la información y la gestión de su seguridad (11).

Cuando se usan EHR para usos secundarios, como el que concierne al presente trabajo, es importante que la información esté desidentificada o, como mínimo, pseudoanonimizada. Se entiende por “desidentificar” el proceso de eliminar o modificar los identificadores de la persona, de tal manera que el paciente no pueda ser re-identificado. Por su parte, anonimizar consiste en reemplazar la información personal por un pseudónimo, (esto permite que el paciente pueda ser identificado bajo ciertas circunstancias. Esto es importante para evitar que se haga un mal uso de la información, aunque la confidencialidad total no puede ser garantizada siempre que exista el registro.

### 3.2 La base de datos MIMIC

La “Medical Information Mart for Intensive Care” (MIMIC) es una base de datos creada con la intención de hacer accesible a la comunidad científica e investigadora datos de pacientes reales, lo que favorece la investigación en diversos campos. Ejemplos de investigaciones desarrolladas gracias a esta base de datos son los estudios retrospectivos no intervencionales o el desarrollo de algoritmos automáticos de detección y predicción de ciertos eventos (12).

MIMIC contiene información anónima de más de 40.000 pacientes admitidos en la unidad de cuidados intensivos del Beth Israel Deaconess Medical Center (Boston, EE. UU). Esta base de datos, que a fecha 1 de abril de 2021 se encuentra en su versión MIMIC-III v1.4, recoge datos de pacientes admitidos entre junio de 2001 y octubre de 2012 tales como información demográfica del paciente, resultados de análisis de laboratorio, medicación a la admisión y administrada durante la hospitalización, medidas de constantes vitales, notas del personal sanitario y un largo etcétera (13).

MIMIC-III engloba información obtenida de dos sistemas informáticos diferentes usados en la unidad de críticos de hospital: CareVue (CV, utilizado de 2001 a 2008) y MetaVision (MV, utilizado a partir de 2008). Este es un punto a tener en cuenta en función de la información que vayamos a analizar, ya que existen datos que necesitan ser procesados e interpretados de manera diferente dependiendo del sistema de origen. Si bien la columna ITEMID requiere una consideración especial por este motivo, esto no tiene impacto en los datos necesarios para realizar el presente estudio.

En la **Tabla 1** podemos encontrar una visión general del número y el nombre de las tablas de la base de datos, así como una breve descripción de cada una de ellas:

**Tabla 1:** Visión general de las tablas contenidas en la base de datos MIMIC-III. Obtenido de Johnson et al., 2016 (13).

Table name	Description
ADMISSIONS	Every unique hospitalization for each patient in the database (defines HADM_ID).
CALLOUT	Information regarding when a patient was cleared for ICU discharge and when the patient was actually discharged.
CAREGIVERS	Every caregiver who has recorded data in the database (defines CGID).
CHARTEVENTS	All charted observations for patients.
CPTEVENTS	Procedures recorded as Current Procedural Terminology (CPT) codes.
D_CPT	High level dictionary of Current Procedural Terminology (CPT) codes.
D_ICD_DIAGNOSES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes relating to diagnoses.
D_ICD_PROCEDURES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes relating to procedures.
D_ITEMS	Dictionary of local codes ('ITEMIDs') appearing in the MIMIC database, except those that relate to laboratory tests.
D_LABITEMS	Dictionary of local codes ('ITEMIDs') appearing in the MIMIC database that relate to laboratory tests.
DATETIMEEVENTS	All recorded observations which are dates, for example time of dialysis or insertion of lines.
DIAGNOSES_ICD	Hospital assigned diagnoses, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.
DRGCODES	Diagnosis Related Groups (DRG), which are used by the hospital for billing purposes.
ICUSTAYS	Every unique ICU stay in the database (defines ICUSTAY_ID).
INPUTEVENTS_CV	Intake for patients monitored using the Philips CareVue system while in the ICU, e.g., intravenous medications, enteral feeding, etc.
INPUTEVENTS_MV	Intake for patients monitored using the iMDSoft MetaVision system while in the ICU, e.g., intravenous medications, enteral feeding, etc.
OUTPUTEVENTS	Output information for patients while in the ICU.
LABEVENTS	Laboratory measurements for patients both within the hospital and in outpatient clinics.

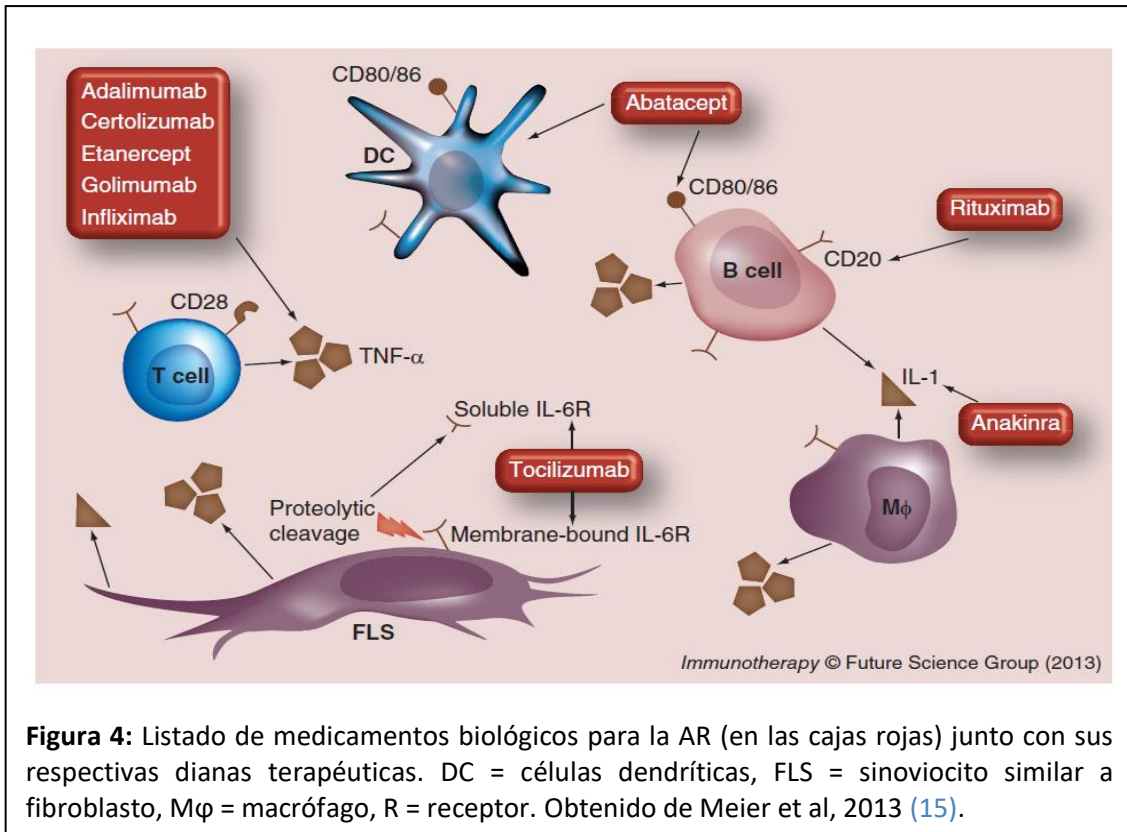
Table name	Description
MICROBIOLOGYEVENTS	Microbiology culture results and antibiotic sensitivities from the hospital database.
NOTEEVENTS	Deidentified notes, including nursing and physician notes, ECG reports, radiology reports, and discharge summaries.
PATIENTS	Every unique patient in the database (defines SUBJECT_ID).
PRESCRIPTIONS	Medications ordered for a given patient.
PROCEDUREEVENTS_MV	Patient procedures for the subset of patients who were monitored in the ICU using the iMDSoft MetaVision system.
PROCEDURES_ICD	Patient procedures, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.
SERVICES	The clinical service under which a patient is registered.
TRANSFERS	Patient movement from bed to bed within the hospital, including ICU admission and discharge.

### 3.3 La Artritis Reumatoide

La artritis reumatoide es una enfermedad de carácter autoinmune que afecta al 0.24-1% de la población en países desarrollados, mostrando una mayor prevalencia en mujeres y población anciana (14). El objetivo de la terapia farmacológica para tratar esa enfermedad es conseguir la remisión clínica del paciente, aliviando el dolor y frenando la degeneración articular que produce junto a otras complicaciones.

Dentro del arsenal terapéutico se encuentran en primera línea los analgésicos y antiinflamatorios para el tratamiento de los síntomas. Para tratar el origen de la enfermedad es necesario recurrir a terapia inmunomoduladora, dentro de la cual distinguimos entre los inmunomoduladores clásicos y la terapia biológica inmunomoduladora.

Los medicamentos biológicos representan normalmente el último escalón terapéutico, empleados cuando otras terapias no han sido efectivas o cuando la enfermedad se encuentra más avanzada. El presente trabajo se centra en este último grupo de medicamentos, que se han desarrollado en las últimas dos décadas y cuyo uso está algo menos extendido que el de las otras opciones farmacológicas (Figura 4) (15).



## 4 Metodología

### 4.1 Acceso a los datos

La base de datos MIMIC, a pesar de estar desidentificada, contiene información privada relativa a la salud y los cuidados de pacientes, por lo que está sujeta a las normas de protección de la HIPAA y debe ser tratada con adecuado cuidado y respeto. En consecuencia, para acceder a los datos se deben cumplir ciertos requisitos.

Para conseguir acceso a los datos hay que completar con una calificación de al menos el 90% el curso “Data or Specimens Only Research” [ANEXO I], ofrecido por la Collaborative Institutional Training Initiative (CITI Program). El certificado obtenido debe aportarse en la solicitud de acceso a los datos, donde se debe justificar además la necesidad de acceso a los mismos (fines académicos y/o de investigación) mediante una breve descripción del proyecto a realizar.

La solicitud fue presentada el 12 de marzo, y el acceso a los datos fue concedido el 16 de marzo de 2021.

### 4.2 Descarga e instalación de MIMIC

La página web de Physionet (16) cuenta con un tutorial para la descarga de la base de datos MIMIC, donde se incluyen instrucciones para la descarga e instalación de PostgreSQL (PSQL) para Windows. Por tanto, se ha elegido este software en su versión

PostgreSQL 10 para crear y almacenar la base de datos. El manejo de los datos se hizo a través de Dbeaver (v21.0.2), una plataforma que ofrece un entorno más atractivo e intuitivo para la gestión de bases de datos.

La página web de physionet cuenta con un link a un repositorio donde podemos encontrar el mastercode para la creación de las tablas que van a albergar los datos de MIMIC (17). Para la creación de las tablas e inclusión de los datos en las mismas seguimos el tutorial de la pagina web de usando la SQL Shell de PSQL. Una vez creadas las tablas y pobladas con los datos, continuamos la selección de la cohorte a través de Dbeaver, que conectamos con nuestra base de datos creada en PSQL

## 4.3 Selección de la cohorte

### 4.3.1 Selección de la cohorte con medicamento biológico

Tras un estudio preliminar de las diferentes tablas, detectamos que la información que nos interesa para este estudio se encuentra repartida en las tablas ADMISSIONS, PATIENTS, DIAGNOSES\_ICD y NOTEEVENTS.

La base de datos MIMIC cuenta con casi 60.000 registros y un peso de unos 43 Gb, y se encuentra almacenada localmente en el PC. Por problemas de memoria de disco duro eliminamos las tablas que no vamos a necesitar y, a partir de las tablas previamente mencionadas (tablas madre), creamos otras que contengan sólo las columnas de interés (tablas hijas). Esto se hizo con la intención de eliminar las tablas madre, pero finalmente no hizo falta. Las tablas hijas fueron fusionadas en una sola tabla, de la cual seleccionaremos esta cohorte.

Los criterios de inclusión en esta cohorte son los siguientes:

- Pacientes adultos: La edad del paciente a la admisión no viene directamente reportada, pero puede ser calculada con la fecha de nacimiento (columna DOB) y la fecha de admisión (columna ADMITTIME), ambas recogidas de la tabla ADMISSIONS. Para los pacientes donde existe más de una admisión, nos quedamos sólo con la última de ellas. Para ello, calculamos la edad del paciente en cada una de las admisiones y eliminamos todas las admisiones salvo aquella en la que el paciente tenía mayor edad (correspondiente a la última admisión).
- Pacientes con AR → De los pacientes adultos, se seleccionaron solo aquellos con el código de diagnóstico 7140 (código correspondiente a la AR según la clasificación ICD-9-CM (18)) recogido en la columna ICD9\_CODE de la tabla DIAGNOSES.
- Pacientes tratados con algún medicamento biológico para el tratamiento de la AR: La medicación recibida a la admisión viene recogida en la columna "TEXT" de la tabla NOTEEVENTS. Esta columna contiene las notas del personal médico, y la información contenida en ella está no estructurada. Por ello, buscamos patrones de texto que contengan los nombres de nuestros medicamentos, por principio activo y marca comercial, con el fin de abarcar todas las formas posibles de

referirse a las sustancias de interés. Para ello se investigó también la existencia de biosimilares, pero estos no existían para ninguno de nuestros medicamentos en el periodo de tiempo que comprende nuestra base de datos. Los medicamentos de interés en el estudio son: infliximab, etanercept, golimumab, adalimumab, certolizumab, anakinra, tocilizumab, abatacept y rituximab. Después de esta primera selección se revisó manualmente la columna “TEXT”, para poder discriminar entre situaciones que puedan nombrar el medicamento que no correspondan a nuestro criterio de inclusión (medicamento a la admisión), como pueden ser alergias medicamentosas o prescripciones pasadas. Se añadió una columna nueva para registrar el medicamento a la admisión y, finalmente, se eliminaron las entradas que no tenían ningún registro.

Un esquema representativo del proceso de selección de la cohorte de pacientes con medicamento biológico puede ser observado en **Figura 5**.

El código SQL utilizado para la selección de la cohorte con medicamento biológico puede ser consultado en el **ANEXO II**.

#### 4.3.2 Selección de la cohorte sin medicamento biológico

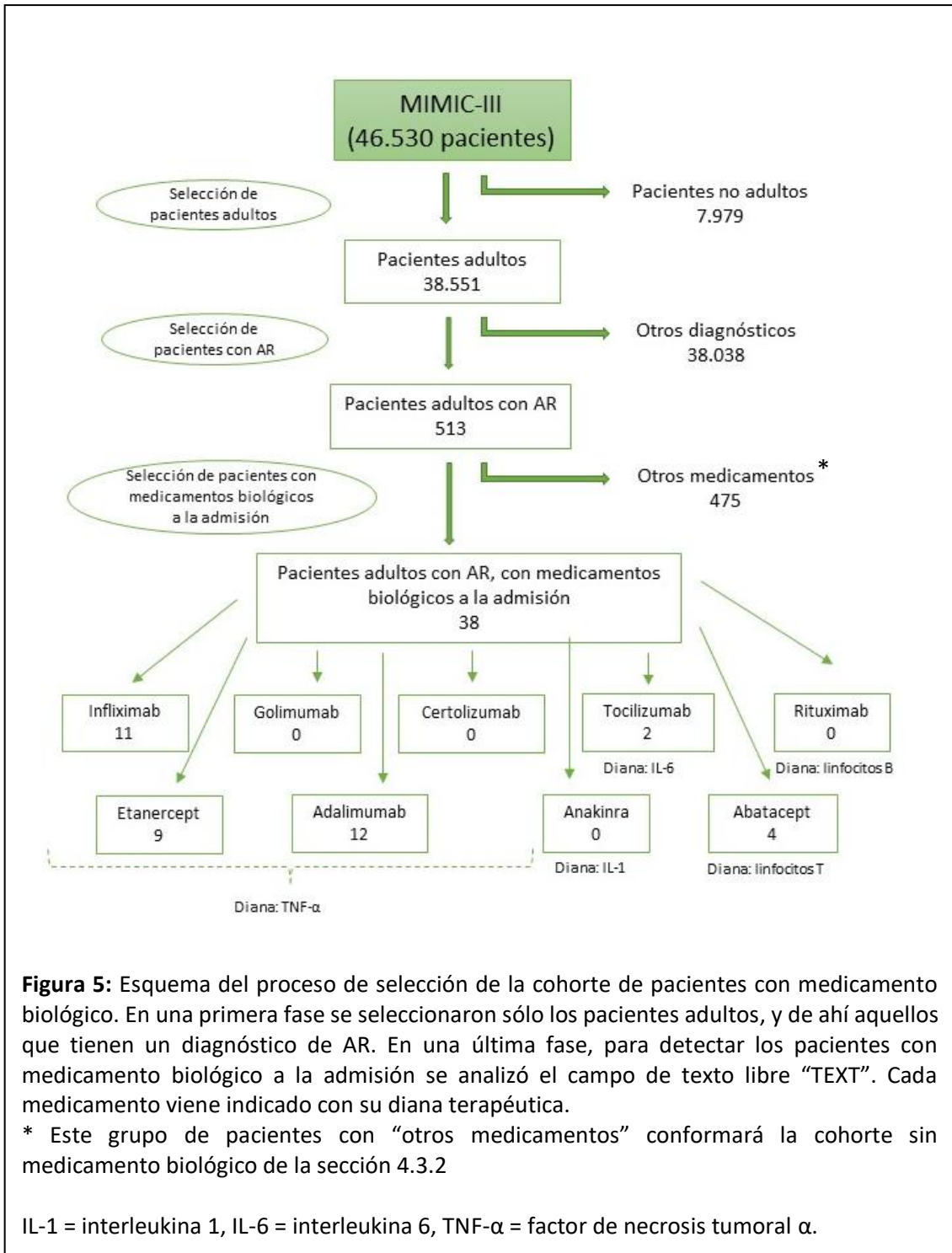
A partir de la tabla que combina las columnas de interés mencionadas en el apartado anterior, volvimos a seleccionar los pacientes adultos con diagnóstico de AR. Para seleccionar los pacientes sin medicamento biológico, eliminamos aquellos que habían sido seleccionados en el apartado anterior (**Figura 5**).

Durante el proceso de selección de esta cohorte se identificó un grupo de pacientes con una edad de 300 años o más, algo que es completamente ilógico. Esto es debido al proceso de desidentificación que han sufrido los datos, ya que ha sido modificada para los pacientes mayores de 89 años edad. Según la información proporcionada por MIMIC, la media de este grupo de pacientes es de 91.4 años. Para que los datos reflejen una situación más parecida a la realidad, la edad de todos los pacientes con 300 años o más se modificó para reflejar la media edad de ese grupo de pacientes.

Igual que en el caso de la cohorte anterior, para los pacientes que han tenido más de una admisión nos quedamos sólo con la última admisión siguiendo el método antes descrito.

El código SQL utilizado para la selección de la cohorte sin medicamento biológico puede ser consultado en el **ANEXO II**.





#### 4.4 Análisis de datos

Las tablas de la base de datos MIMIC fueron inicialmente procesadas usando PSQL a través de la interfaz Dbeaver en las versiones previamente mencionadas. De este modo se obtuvo una sola tabla con toda la información de interés que pudiésemos trabajar con R. Los test estadísticos se llevaron a cabo con Rstudio (R version 4.0.3 [2020-10-10])



Para describir las cohortes seleccionadas estudiamos la media de edad y la desviación estandar, expresadas en años, y la proporción entre hombres y mujeres, expresada en porcentajes y representada con un gráfico de barras.

La base de datos nos ofrece dos indicadores de mortalidad, “hospital\_expire\_flag”, que muestra la mortalidad en el hospital, y “expire\_flag”, que muestra la mortalidad según los registros de la SS e incluye también la mortalidad en el hospital. Dado que disponemos de estos dos indicadores de mortalidad, los estudiamos por separado.

Para estudiar la diferencia de mortalidad entre grupos, el test estadístico planteado es el test de Chi cuadrado, ya que este permite ver si existe independencia o no entre dos variables categóricas representadas en una tabla de contingencia. El test de Chi cuadrado suele ser más adecuado para tamaños de muestra relativamente grandes. Si nos enfrentamos a un tamaño de muestra pequeño (es decir, cuando para alguna de las celdas de la tabla de contingencia el número de sujetos es  $<5$ ), se empleará como alternativa el test de Fisher, que permite igualmente observar independencia o no entre dos variables categóricas pero está más adaptado a estos casos. Si la tabla de contingencia sobre la que vamos a realizar el test estadístico (test de Chi cuadrado o test de Fisher) es de un tamaño mayor de  $2 \times 2$ , a la función estadística utilizada en Rstudio se le añadirá “simulate.p.value = TRUE” para analizar este tipo de datos (que realiza la simulación de Monte Carlo).

El nivel de significancia para los test estadísticos será de  $\alpha = 0.05$ . En caso de encontrar diferencias significativas entre grupos, realizaremos el Odds Ratio para calcular la probabilidad de ocurrencia del evento, que en nuestro estudio es “muerte”.

El código R utilizado para el análisis de los datos puede ser consultado en el **ANEXO III**.

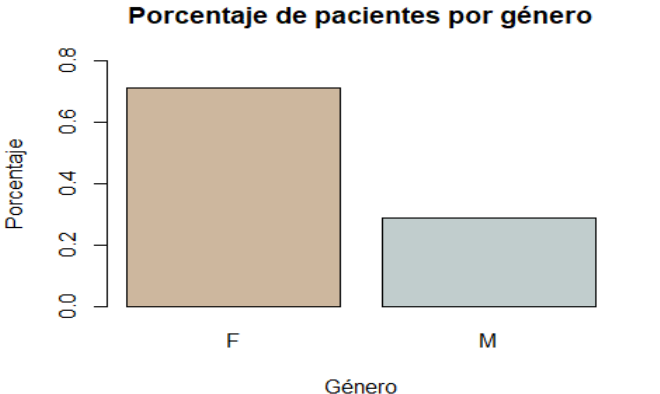
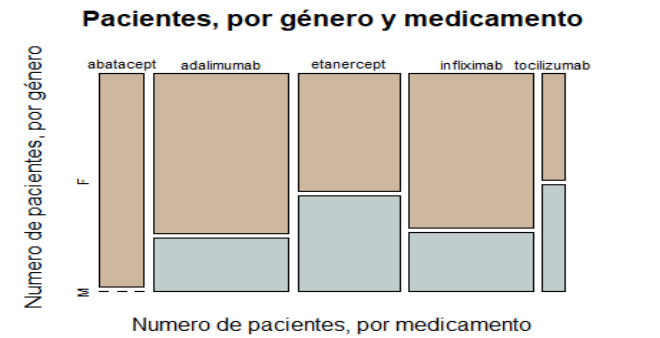
## 5 Resultados

### 5.1 Comparación de la mortalidad entre pacientes con AR tratados con los distintos medicamentos biológicos

#### 5.1.1 Definición de la cohorte con medicamento biológico

La media de edad de los pacientes es de 67.58 años, con una desviación estandar de 10.49 años, y el 71% son mujeres, lo cual concuerda bastante con la descripción demográfica de los pacientes con AR.

No todos los medicamentos biológicos antirreumáticos están presente en nuestra cohorte, y entre los que sí, hay algunos (infliximab, adalimumab, etanercept) que se dan con más frecuencia que otros (**Figura 5**). La **Tabla 2** muestra un resumen visual de la descripción de nuestra cohorte.

<p>Edad (años):</p> <p>Media: 67.58 Desv. Est.: 10.49</p> <p>.....(a)</p> <p>Género:</p> <p>F M 0.71 0.29</p> <p>.....(b)</p>	<p style="text-align: center;"><b>Porcentaje de pacientes por género</b></p>  <p style="text-align: center;">(c)</p>																		
<p>Medicamento, por género:</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>F</th> <th>M</th> </tr> </thead> <tbody> <tr> <td>Abatacept</td> <td>4</td> <td>0</td> </tr> <tr> <td>Adalimumab</td> <td>9</td> <td>3</td> </tr> <tr> <td>Etanercept</td> <td>5</td> <td>4</td> </tr> <tr> <td>Infliximab</td> <td>8</td> <td>3</td> </tr> <tr> <td>Tocilizumab</td> <td>1</td> <td>1</td> </tr> </tbody> </table> <p>.....(d)</p>		F	M	Abatacept	4	0	Adalimumab	9	3	Etanercept	5	4	Infliximab	8	3	Tocilizumab	1	1	<p style="text-align: center;"><b>Pacientes, por género y medicamento</b></p>  <p style="text-align: center;">(e)</p>
	F	M																	
Abatacept	4	0																	
Adalimumab	9	3																	
Etanercept	5	4																	
Infliximab	8	3																	
Tocilizumab	1	1																	
<p><b>Tabla 2:</b> Descripción de la cohorte de pacientes con medicamento biológico</p> <p>(a) Media y desviación estándar de la edad de la cohorte, expresada en años.</p> <p>(b) Proporción entre hombres y mujeres de la cohorte, expresada en porcentaje. F = mujer, M = hombre</p> <p>(c) Gráfico de barras expresando la proporción entre hombres y mujeres, expresada en porcentaje. F = mujer, M = hombre.</p> <p>(d) Número de pacientes para cada uno de los medicamentos biológicos, separados por género. F = mujer, M = hombre.</p> <p>(e) Representación gráfica de la tabla en (e). F = mujer, M = hombre.</p>																			

Se comprobó la homogeneidad de la edad de los grupos en función del género y/o el medicamento a la admisión y no se encontraron diferencias significativas (ANEXO III).

### 5.1.2 Mortalidad en el hospital (“hospital\_expire\_flag”)

En la **Tabla 3** podemos observar la tabla de contingencia que recoge los datos sobre el indicador “hospital\_expire\_flag” para cada uno de los medicamentos biológicos de nuestro estudio.

**Tabla 3:** Tabla de contingencia sobre el estatus del indicador de mortalidad “hospital\_expire\_flag” para cada uno de los medicamentos biológicos para tratar la AR.

	Abatacept	Adalimumab	Etanercept	Infliximab	Tocilizumab
No fallecido	4	10	9	10	2
Fallecido	0	2	0	1	0

Test de Fisher (con simulación de Monte Carlo), p-value = 0.8561

Podemos observar que, para algunas celdas de la tabla, el contaje es menor a 5 o incluso 0. Es por eso que el test estadístico más apropiado es el test de Fisher (con la simulación de Monte Carlo).

La p-value obtenida es mayor a 0.05, por lo que la hipótesis nula de homogeneidad entre grupos no puede ser descartada; es decir, no se observan diferencias estadísticamente significativas en cuanto a mortalidad entre grupos para este indicador.

### 5.1.3 Mortalidad según los registros de la seguridad social (“expire\_flag”).

En la **Tabla 4** podemos observar la tabla de contingencia que recoge los datos sobre el indicador “expire\_flag” para cada uno de los medicamentos biológicos de nuestro estudio.

**Tabla 4:** Tabla de contingencia sobre el estatus del indicador de mortalidad “expire\_flag”, para cada uno de los medicamentos biológicos para tratar la AR.

	Abatacept	Adalimumab	Etanercept	Infliximab	Tocilizumab
No fallecido	4	8	6	8	2
Fallecido	0	4	3	3	0

Test de Fisher (con simulación de Monte Carlo), p-value = 0.8056

Podemos observar que, para algunas celdas de la tabla, el conteo es menor a 5 o incluso 0, por lo que realizamos el test de Fisher (con la simulación de Monte Carlo).

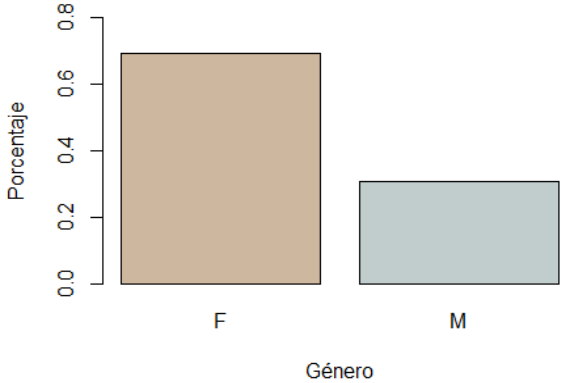
La p-value obtenida es mayor a 0.05, por lo que la hipótesis nula de homogeneidad entre grupos no puede ser descartada; es decir, no se observan diferencias estadísticamente significativas en cuanto a mortalidad entre grupos para este indicador.

## 5.2 Comparación de la mortalidad entre pacientes con AR tratados con medicamentos biológicos o no biológicos.

### 5.2.1 Definición de la cohorte sin medicamento biológico

La media de edad de los paciente es de 71.38 años, con una desviación estandar de 12.71 años, y el 69% son mujeres, lo cual concuerda bastante con la descripción demográfica de los pacientes con AR y está también en consonancia con los resultados obtenidos para la cohorte con medicamento biológico.

Se observa una gran diferencia en el número de pacientes de los diferentes grupos con y sin medicamento biológico, lo cual se considera normal y esperable por el lugar que ocupa este tipo de terapias en los protocolos de tratamiento de esta enfermedad. La **Tabla 5** muestra un resumen visual de la descripción de la cohorte sin medicamento biológico.

<p>Edad (años):</p> <p>Media: 71.38 Desv. Est.: 12.71</p> <p>.....(a)</p> <p>Género:</p> <table border="0"> <tr> <td>F</td> <td>M</td> </tr> <tr> <td>0.69</td> <td>0.31</td> </tr> </table> <p>.....(b)</p>	F	M	0.69	0.31	<p style="text-align: center;"><b>Porcentaje de pacientes por género</b></p>  <p style="text-align: center;">(c)</p>					
F	M									
0.69	0.31									
<p>Tipo de medicamento, por género:</p> <table border="0"> <thead> <tr> <th></th> <th>F</th> <th>M</th> </tr> </thead> <tbody> <tr> <td>Biológico</td> <td>27</td> <td>11</td> </tr> <tr> <td>No biológico</td> <td>330</td> <td>145</td> </tr> </tbody> </table> <p>.....(d)</p>		F	M	Biológico	27	11	No biológico	330	145	<p><b>Tabla 5:</b> Descripción de la cohorte de pacientes sin medicamento biológico.</p> <p><b>(a)</b> Media y desviación estándar de la edad de la cohorte, en años. F = mujer, M = hombre.</p> <p><b>(b)</b> Proporción entre hombres y mujeres, en porcentaje.</p> <p><b>(c)</b> Gráfico de barras expresando la proporción entre hombres y mujeres, en porcentaje. F = mujer, M = hombre.</p> <p><b>(d)</b> Número de pacientes para cada tipo de medicamento (biológico y no biológico), separados por género. F = mujer, M = hombre.</p>
	F	M								
Biológico	27	11								
No biológico	330	145								

Se comprobó la homogeneidad de la edad de los grupos en función del género y/o el medicamento a la admisión y no se encontraron diferencias significativas (**ANEXO III**).

### 5.2.2 Mortalidad en el hospital (“hospital\_expire\_flag”)

En la **Tabla 6** podemos observar la tabla de contingencia que recoge los datos sobre el indicador “hospital\_expire\_flag” para cada tipo de medicamento.

**Tabla 6:** Tabla de contingencia sobre el estatus del indicador de mortalidad “hospital\_expire\_flag”, para los dos tipos de medicamentos (biológicos y no biológicos) para tratar la AR.

	Biológico	No biológico
No fallecido	35	407
Fallecido	3	68

Test de Fisher, p-value = 0.3369

Podemos observar que, para algunas celdas de la tabla, el conteo es menor a 5 o incluso 0, por lo que realizamos el test de Fisher.

La p-value obtenida es mayor a 0.05, por lo que la hipótesis nula de homogeneidad entre grupos no puede ser descartada; es decir, no se observan diferencias estadísticamente significativas en cuanto a mortalidad entre grupos para este indicador.

### 5.2.3 Mortalidad según los registros de la seguridad social. (“expire\_flag”)

En la **Tabla 7** podemos observar la tabla de contingencia que recoge los datos sobre el indicador “expire\_flag” para cada tipo de medicamento.

**Tabla 7:** Tabla de contingencia sobre el estatus del indicador de mortalidad “expire\_flag” para los dos tipos de medicamentos (biológicos y no biológicos) para tratar la AR.

	Biológico	No biológico
No fallecido	28	250
Fallecido	10	225

Test de Chi-cuadrado. P-value = 0.019

Odds Ratio: 2.52

En este caso todas las celdas tiene un contaje mayor a 5, por lo que el test de Chi cuadrado sería viable. Atendiendo al p-value obtenido en ambos tests (menor a 0.05), podemos descartar la hipótesis nula de homogeneidad entre grupos, concluyendo que sí existen diferencias estadísticas significativas. El Odds Ratio obtenido es de 2.52.

Puesto que en todos los casos anteriores hemos realizado el tet de Fisher, se ha vuelto a realizar aquí para poder comparar entre resultados obtenidos con el mismo test estadístico, siempre bajo la consideración de que, aunque el test de Chi cuadrado es mas adecuado para este tipo de datos, el de Fisher no es incorrecto. En este caso, observamos también diferencias estadísticamente significativas entre grupos. [ANEXO III]

Nota: el código R utilizado para toda la seccion de resultados puede ser consultado en el ANEXO III.

## 6 Discusión

### 6.1 Sobre los resultados

De todas las comparaciones de mortalidad realizadas, sólo se han encontrado diferencias significativas entre medicamentos biológicos y no biológicos atendiendo al indicador de mortalidad “expire\_flag”. El Odds ratio obtenido de esa comparativa es de 2.52, lo que indica que es 2.52 veces mas probable sobrevivir (o no fallecer) si se está en el grupo sin medicamento biológico.

La cohorte con medicamento biológico obtenida mostraba un numero de pacientes inferior a 5 para muchas de las celdas de las tablas de contingencia (**Tabla 3, Tabla 4**). Como hemos mencionado al principio de este trabajo, los medicamentos biológicos representan la ultima alternativa terapéutica para tratar la AR, por lo que puede ser esperable un grupo de población menor. Además, casi todos estos medicamentos fueron aprobados por la Food and Drug Administration (FDA) para su comercialización entre los años 2000-2010; puesto que la base de datos MIMIC abarca un periodo de tiempo entre 2001 y 2012, el bajo número de pacientes con este tipo de terapia podría deberse a que no ha transcurrido el tiempo suficiente para que el uso de esos medicamentos se extendiese y este tipo de pacientes apareciese en nuestra base de datos.

Tras una búsqueda bibliográfica, no hemos estudios comparativos sobre la mortalidad entre grupos de pacientes de AR con y sin medicamento biológico. Sólo se ha encontrado el abstract de una publicación en la que no se hallaron diferencias en la mortalidad entre esos dos grupos pacientes de AR. Puesto que el texto completo no pudo ser obtenido, los detalles de la publicación no pudieron ser analizados (19).

Los resultados obtenidos deben ser analizados con cautela, ya que solo se ha establecido una correlación entre medicamento biológico y aumento de la mortalidad para uno solo

de los indicadores (“expire\_flag”). Dado que la causalidad de esta asociación no puede ser establecida mediante el presente trabajo, será necesario realizar más estudios, probablemente intervencionales, que además incluyan el estudio de indicadores de salud del paciente y/o valores de laboratorio. Se plantean ciertas hipótesis que pueden explicar este fenómeno de diferencias en la mortalidad: ya que la terapia biológica representa el último escalón terapéutico para tratar la AR, es probable que estos pacientes presenten una forma más avanzada de la enfermedad. También es sabido que la terapia biológica aumenta la incidencia de algunos efectos adversos de tipo cardiovascular o infeccioso (8) que podrían justificar este aumento de la mortalidad.

## 6.2 Sobre los usos, ventajas y limitaciones de los EHR

El presente trabajo supone un ejemplo más de como los EHR pueden colaborar en la investigación médica, orientada en este caso hacia la farmacovigilancia. Existen también muchos otros proyectos orientados hacia la medicina preventiva o el tratamiento de enfermedades, por ejemplo.

La existencia de una base de datos de acceso público como MIMIC ha permitido que este y otros estudios hayan podido realizarse. No obstante, además de los sesgos mencionados en el apartado del estado del arte, hay que considerar que, en este caso, los datos provienen de una unidad de un único centro. En consecuencia, existe un sesgo asociado a las características demográficas de la población cubierta por el hospital y otro asociado a las características de pacientes que requieren de una estancia en cuidados intensivos.

En cualquier caso, las iniciativas tipo MIMIC son de gran interés para la investigación biomédica, ya que su estudio puede traer grandes avances en el conocimiento. Existe aún el debate sobre la protección de datos y los derechos de los pacientes sobre su propia información (a pesar de estar desidentificada y la persona física no pueda ser identificable). Con todo, siempre que esa cuestión pueda resolverse adecuadamente, la realización y expansión de este tipo de proyectos debe ser apoyada.

## 7 Valoración económica

Del presente trabajo no se espera ningún producto que genere ingresos económicos directos ni tampoco ha requerido realizar ningún desembolso monetario. Sin embargo, hay ciertos puntos que pueden valorarse dentro de este apartado:

- Horas de trabajo: El TFM supone 3 créditos dentro del plan de estudios del Master en Bioestadística y Bioinformática de la Universitat Oberta de Catalunya (UOC). Cada crédito está estimado en unas 125 horas de dedicación, lo que suponen unas 375 horas de trabajo no remunerado. Teniendo en cuenta el Salario Mínimo interprofesional (SME) actual, el tiempo de dedicación correspondería a aproximadamente 1500 € de trabajo remunerado. Si tuviésemos en cuenta el salario medio de un bioestadístico, esta cantidad sería aún mayor.



- Materiales: No ha habido que hacer ningún desembolso directo para la realización del trabajo, pero es necesario tener en cuenta que ya se disponía previamente de algunos materiales utilizados: ordenador portátil, pantalla accesoria y ratón.
- Licencias y accesos: los programas utilizados para el análisis de datos (Rstudio, PSQL y DBeaver) son de acceso gratuito al público. La licencia del paquete Microsoft (Word, Power Point) es ofrecida por la universidad. Las suscripciones a las revistas científicas que han permitido la documentación del trabajo también son ofrecidas por la universidad.
- Datos: al no tener datos propios sobre los que trabajar, hemos tenido que recurrir a una fuente externa. El proyecto MIMIC ha permitido que esto fuese posible gracias a la cesión de los datos de forma gratuita.
- Conexión a internet.

## 8 Conclusiones

### 8.1 Conclusiones

Este estudio arroja nueva evidencia sobre la utilidad de los EHR en la detección de eventos adversos de medicamentos, ya se trate de screenings rutinarios para la detección y notificación de los mismos o de investigaciones de otros eventos desconocidos o poco estudiados. Los estudios realizados a partir de EHR vienen a complementar los otros estudios clásicos, tanto intervencionales (como los ensayos clínicos) como no intervencionales (como los estudios observacionales): en ningún caso se pretende sustituirlos, ya que todos ellos presentan limitaciones que pueden ser cubiertas en gran medida por otro tipo de estudio diferente. Por ejemplo, los ensayos clínicos presentan condiciones de experimentación muy protocolizadas, pero a menudo estas condiciones están demasiado idealizadas y no son representativas de la realidad fuera de los ensayos clínicos, y el número de pacientes incluidos tiende a ser menor del que podemos encontrar en una base de datos tipo EHR.

En nuestro caso, no se han encontrado estudios previos sobre diferencias en mortalidad entre medicamentos biológicos para el tratamiento de la AR, por lo que es difícil extraer conclusiones más allá de los resultados obtenidos. Se requieren más estudios de este y otro tipo, realizados con otro tipo de datos para poder concluir y establecer esta diferencia en términos de mortalidad aquí observada.

### 8.2 Líneas de futuro

Los EHR son bases de datos estructuradas en casi la totalidad de sus campos, lo que permite poder obtener información contenida en ellos de una forma rápida. Sin embargo, los EHR suelen tener un campo de texto libre (comúnmente llamado "comentarios", "notas" o, en este caso, "text") donde el profesional de la salud puede dejar comentarios en forma de lenguaje natural. Pese a que por lo general se anima a usar este campo lo menos posible, a menudo resulta ser el más rico en información, ya

que permite capturar los matices de la historia clínica del paciente y el razonamiento para cada una de las estrategias utilizadas en el proceso de cuidados. Este campo suele ser el más difícil de analizar computacionalmente, ya que su contenido es altamente heterogéneo, a veces no sigue las normas gramaticales y contiene mucho vocabulario, abreviaturas y acrónimos específicos del área de conocimiento y del autor (7). Los datos aquí contenidos suponen un nicho de datos no estructurados dentro de una base de datos estructurada.

Tradicionalmente, la farmacovigilancia se ha sustentado en el análisis de diferentes fuentes (notificaciones espontáneas, literatura, análisis clínicos, etc) de forma manual por parte de personal experto en la materia. En los últimos años, sobre todo a partir de 2012, han aflorado los estudios que incluyen “Natural Language Processing” (NLP) en el estudio de AEs aplicado a EHR. Estos estudios utilizan técnicas de análisis estadístico y “machine learning” para detectar AEs y otras situaciones reportables (desde el punto de vista de la farmacovigilancia) dentro de datos no estructurados, a menudo ayudándose de otras técnicas que facilitan el establecimiento de clasificaciones de afirmaciones y/o relaciones temporales en el lenguaje incluido en estos campos (6).

Nosotros hemos estudiado la mortalidad en pacientes con AR tratados con diferentes medicamentos inmunomoduladores, donde uno de los criterios de inclusión (medicamento a la admisión) estaba recogido en la base de datos de forma no estructurada. Para poder discriminar entre casos en los que el medicamento estuviese prescrito a la admisión de otras situaciones, hubo que realizar una revisión manual del campo. Aplicar NLP a este paso habría permitido automatizar y agilizar este paso.

Igualmente, el estudio de mortalidad sobre EHR está facilitado por la existencia de una clave dicotómica que indica si el paciente está vivo o no (en nuestro caso, reflejada en las columnas HOSPITAL\_EXPIRE\_FLAG y EXPIRE\_FLAG de la tabla ADMISSIONS). Si quisiésemos estudiar algún evento no recogido de forma estructurada en alguna columna, podríamos recurrir nuevamente a las técnicas de NLP para estudiar el campo de texto libre. En ese caso sería también interesante aplicar un filtro más para estrechar la búsqueda en el proceso de selección de la cohorte, como podría ser algún valor de laboratorio que estuviese directamente relacionado con el evento (por ejemplo, infección – proteína C reactiva).

Actualmente existen pocos proyectos como MIMIC que pretendan favorecer el acceso a los EHR para promover su estudio e investigación. En estos casos, la información proviene de una sola unidad de un centro en concreto. Sería muy interesante que estos proyectos existieran a nivel multicentro y multiunidad, ya que estos datos presentan también un sesgo derivado de la zona demográfica que cubre el hospital. Para que este tipo de iniciativas pueda llevarse a cabo, las unidades participantes deberían usar el mismo programa informático, u otros distintos pero que sean compatibles entre sí (es decir, que el tipo y estructura de los datos, así como la forma de registrarlos, sea la misma). No se han encontrado referencias bibliográficas sobre este tipo de proyectos, aunque sí existen otros que están desarrollando herramientas para facilitar la integración de diferentes tipos de bases de datos en una sola (20).

Existen líneas de investigación que han desarrollado algoritmos capaces de predecir la ocurrencia de un evento en un paciente a partir de la información recogida en los EHR

(20). La creación de estos algoritmos debe realizarse de forma individual para la base de datos y el evento concreto. Este proceso, que al principio puede hacerse largo y costoso, permite automatizar después la detección de la incidencia o el riesgo de incidencia del evento, lo cual puede implicar un beneficio directo para el paciente en cuanto a cuidados se refiere. Si quisiéramos aplicar ese algoritmo a una base de datos diferente, tendría que llevarse a cabo una fase de adaptación y validación para los nuevos datos, pues es muy probable que el formato de los datos cambie y el algoritmo no sea óptimo en ese nuevo contexto.

### 8.3 Seguimiento de la planificación

La planificación del trabajo presentada en el plan de proyecto al principio del semestre y recogida en el apartado 2.4 (**Figura 2**) ha podido cumplirse de forma muy satisfactoria. Durante la realización del presente trabajo se han presentado algunas dificultades que han generado retrasos respecto a la planificación entre hitos, pero las entregas intermedias del TFM, que se han hecho coincidir con los grandes hitos (selección de la cohorte, análisis de datos) han podido ser realizadas y entregadas convenientemente.

Se valora positivamente la realización del calendario en el plan de proyecto, ya que el hecho de fragmentar la realización del TFM y el haber puesto una fecha para cada uno de los fragmentos ha ayudado a la organización del presente proyecto, lo que a su vez ha reducido el estrés asociado a este tipo de tareas.

## 9 Glosario

AE: evento adverso (del inglés adverse event)

AR: Artritis Reumatoide

CV: CareVue

EE. UU: Estados Unidos

EHR: Electronic Health Records

FDA: Food and Drug Administration

HIPAA: Health Insurance Portability and Accountability Act

ME: error de medicación (del inglés medication error)

MIMIC: Medical Information Mart for Intensive Care

MV: MetaVision

NPL: Natural Language Processing

PQSL: Postgres SQL

SME: Salario Mínimo interprofesional

SQL: Structured Query Language

SS: Seguridad Social

TFM: Trabajo Fin de Master

UOC: Universitat Oberta de Catalunya

## 10 Bibliografía

- (1) Campanella, P., Lovato, E., Marone, C., Fallacara, L., Mancuso, A., Ricciardi, W., & Specchia, M. (2015). The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. *The European Journal Of Public Health*, 26(1), 60-64. doi: 10.1093/eurpub/ckv122.
- (2) Health IT (<https://healthit.gov>). Consultado el 11-mar-2021.
- (3) Evans, R. (2016). Electronic Health Records: Then, Now, and in the Future. *Yearbook Of Medical Informatics*, 25(S 01), S48-S61. doi: 10.15265/iys-2016-s006.
- (4) Pubmed (<https://pubmed.ncbi.nlm.nih.gov/>). Consultado el 20-mar-2021.
- (5) Celi, L., Moseley, E., Moses, C., Ryan, P., Somai, M., Stone, D., & Tang, K. (2014). From Pharmacovigilance to Clinical Care Optimization. *Big Data*, 2(3), 134-141. doi: 10.1089/big.2014.0008.
- (6) Luo, Y., Thompson, W., Herr, T., Zeng, Z., Berendsen, M., & Jonnalagadda, S. et al. (2017). Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Safety*, 40(11), 1075-1089. doi: 10.1007/s40264-017-0558-6.
- (7) Jensen, P., Jensen, L., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395-405. doi: 10.1038/nrg3208.
- (8) Rutherford, A., Subesinghe, S., Hyrich, K., & Galloway, J. (2018). Serious infection across biologic-treated patients with rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register for Rheumatoid Arthritis. *Annals Of The Rheumatic Diseases*, annrheumdis-2017-212825. doi: 10.1136/annrheumdis-2017-212825.
- (9) Verheij, R., Curcin, V., Delaney, B., & McGilchrist, M. (2018). Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *Journal Of Medical Internet Research*, 20(5), e185. doi: 10.2196/jmir.9134.
- (10) U.S. Department of Health and Human Services (<https://www.hhs.gov/>). Consultado el 11 abril 2021.
- (11) International Organization for Standardization (<https://www.iso.org/>). Consultado el 11 abril 2021.
- (12) Lee, J., Scott, D., Villarroel, M., Clifford, G., Saeed, M., & Mark, R. (2011). Open-access MIMIC-II database for intensive care research. 2011 Annual International Conference Of The IEEE Engineering In Medicine And Biology Society. doi: 10.1109/iembs.2011.6092050
- (13) Johnson, A., Pollard, T., Shen, L., Lehman, L., Feng, M., & Ghassemi, M. et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1). doi: 10.1038/sdata.2016.35
- (14) England, B., & Mikuls, T. (2021). Epidemiology of, risk factors for, and possible causes of rheumatoid arthritis. Consultado el 17 Abril 2021 - <https://www.uptodate.com/contents/epidemiology-of-risk-factors-for-and-possible-causes-of-rheumatoid-arthritis#:~:text=Estimates%20of%20RA%20prevalence%20in,persons%20%5B2%2C4%5D>.

- (15) Meier, F., Frerix, M., Hermann, W., & Müller-Ladner, U. (2013). Current immunotherapy in rheumatoid arthritis. *Immunotherapy*, 5(9), 955-974. doi: 10.2217/imt.13.94
- (16) Physionet (<https://mimic.physionet.org/>). Consultado el 23-mar-2021
- (17) MIMIC Code Repository (<https://github.com/MIT-LCP/mimic-code>). Consultado el 23-mar-2021
- (18) The web's free ICD-9-CM Medical Codig Referenrence (<http://www.icd9data.com>). Consultado el 1 abril 2021
- (19) Rodriguez-Rodriguez, L., Leon, L., Ivorra-Cortes, J., Gómez, A., Ramon Lamas, J., & Pato, E. et al. (2016). Treatment in rheumatoid arthritis and mortality risk in clinical practice: the role of biologic agents. *Clin Exp Rheumatol*. Nov-Dec 2016;34(6):1026-1032. Epub 2016 Oct 7.
- (20) Armengol, MA., (2020). Promotion, Integration, Management and Processing of Critical Inpatients' Open Big Data Repositories (Tesis doctoral). Universidad Politecnica de Madrid, Madrid, España.

# 11Anexos

## 11.1 Anexo I: Certificado del curso “Data or Specimens Only Research”, Collaborative Institutional Training Initiative (CITI Program)

### COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM) COMPLETION REPORT - PART 1 OF 2 COURSEWORK REQUIREMENTS\*

\* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- Name: Maria Uceda Martin (ID: 9971528)
- Institution Affiliation: Massachusetts Institute of Technology Affiliates (ID: 1912)
- Institution Email: mucedad@uoc.edu
- Institution Unit: biostatistics
  
- Curriculum Group: Human Research
- Course Learner Group: Data or Specimens Only Research
- Stage: Stage 1 - Basic Course
  
- Record ID: 41402266
- Completion Date: 09-Mar-2021
- Expiration Date: 08-Mar-2024
- Minimum Passing: 90
- Reported Score\*: 92

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and its Principles (ID: 1127)	08-Mar-2021	3/3 (100%)
History and Ethics of Human Subjects Research (ID: 498)	08-Mar-2021	4/5 (80%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	08-Mar-2021	5/5 (100%)
Records-Based Research (ID: 5)	09-Mar-2021	3/3 (100%)
Genetic Research In Human Populations (ID: 6)	09-Mar-2021	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	09-Mar-2021	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	09-Mar-2021	3/5 (60%)
Conflicts of Interest In Human Subjects Research (ID: 17464)	09-Mar-2021	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	09-Mar-2021	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing Institution identified above or have been a paid Independent Learner.

Verify at: [www.citiprogram.org/verify/?k1d438348-2bod-4a50-be2e-589a4e436a47-41402266](http://www.citiprogram.org/verify/?k1d438348-2bod-4a50-be2e-589a4e436a47-41402266)

Collaborative Institutional Training Initiative (CITI Program)  
Email: [support@citiprogram.org](mailto:support@citiprogram.org)  
Phone: 888-529-5929  
Web: <https://www.citiprogram.org>

## 11.2 Anexo II: Código SQL para la selección de la cohorte

### 11.2.1 Anexo II.1: Selección de la cohorte con medicamento biológico:

```
-- creación tabla temp_admissions únicamente con las columnas de interés de la tabla madre
ADMISSIONS --
create table mimiciii.temp_admissions as (
select ADMISSIONS.HADM_ID, ADMISSIONS.SUBJECT_ID, ADMISSIONS.ADMITTIME,
ADMISSIONS.HOSPITAL_EXPIRE_FLAG
FROM mimiciii.admissions);

-- creación tabla temp_patients unicamente con las columnas de interés de la tabla madre PATIENTS --
create table mimiciii.temp_patients as(
select patients.subject_id, patients.gender, patients.dob, patients.expire_flag
from mimiciii.patients);

-- creación tabla temp_diagnoses unicamente con las columnas de interés de la tabla madre DIAGNOSES
--
create table mimiciii.temp_diagnoses_icd as(
select DIAGNOSES_ICD.subject_id, diagnoses_icd.hadm_id, DIAGNOSES_ICD.ICD9_CODE
from mimiciii.diagnoses_icd);

-- creación tabla temp_noteevents unicamente con las columnas de interés de la tabla madre
NOTEEVENTS --
create table mimiciii.temp_noteevents as(
select noteevents.subject_id, noteevents.hadm_id, noteevents.text
from mimiciii.noteevents);

-- Numero de pacientes adultos con diagnóstico de AR, que potencialmente podrían haber recibido
alguno de los medicamentos en el momento de admision (ésto deberá ser revisado manualmente
leyendo la columna noteevents.text) --
select count(distinct temp_admissions.subject_id)
from temp_admissions
INNER JOIN temp_patients ON temp_admissions.SUBJECT_ID = temp_patients.SUBJECT_ID
INNER JOIN temp_diagnoses_icd ON temp_admissions.HADM_ID = temp_diagnoses_icd.HADM_ID
INNER JOIN temp_noteevents ON temp_admissions.HADM_ID = temp_noteevents.HADM_ID
where age(temp_admissions.ADMITTIME, temp_patients.DOB) >= interval '18 years'
and temp_diagnoses_icd.ICD9_CODE = '7140'
and (temp_noteevents.TEXT LIKE '%nflixima%' or temp_noteevents.TEXT LIKE '%emicad%'
or temp_noteevents.TEXT like '%tanercep%' or temp_noteevents.TEXT like '%mbre%'
or temp_noteevents.TEXT like '%dalimuma%' or temp_noteevents.TEXT like '%umir%'
or temp_noteevents.TEXT like '%olimuma%' or temp_noteevents.TEXT like '%impon%'
or temp_noteevents.TEXT like '%ertolizum%' or temp_noteevents.TEXT like '%imzi%'
or temp_noteevents.TEXT like '%nakinr%' or temp_noteevents.TEXT like '%inere%'
or temp_noteevents.TEXT like '%ocilizum%' or temp_noteevents.TEXT like '%ctemr%'
or temp_noteevents.TEXT like '%batacep%' or temp_noteevents.TEXT like '%renci%'
or temp_noteevents.TEXT like 'ituxima%' or temp_noteevents.TEXT like 'ituxa%')
;

-- fusionamos todas las columnas en una sola tabla, a partir de la cual trabajaremos --
drop table if exists mimiciii.cohort2;
create table mimiciii.cohort2 as (
select
temp_admissions.hadm_id, temp_admissions.subject_id, temp_admissions.admittime,
temp_admissions.hospital_expire_flag,
temp_patients.gender, temp_patients.dob, temp_patients.expire_flag,
```



```

temp_diagnoses_icd.icd9_code,
temp_noteevents.text
from temp_admissions
INNER JOIN temp_patients ON temp_ADMISSIONS.SUBJECT_ID = temp_PATIENTS.SUBJECT_ID
INNER JOIN temp_DIAGNOSES_ICD ON temp_ADMISSIONS.HADM_ID =
temp_DIAGNOSES_ICD.HADM_ID
INNER JOIN temp_NOTEEVENTS ON temp_ADMISSIONS.HADM_ID = temp_NOTEEVENTS.HADM_ID
where age(ADMITTIME, DOB) >= interval '18 years'
and ICD9_CODE = '7140'
and (TEXT LIKE '%nflixima%' or TEXT LIKE '%emica%'
or TEXT like '%tanercep%' or TEXT like '%mbre%'
or TEXT like '%dalimuma%' or TEXT like '%umir%'
or TEXT like '%olimuma%' or TEXT like '%impon%'
or TEXT like '%ertolizuma%' or TEXT like '%imzi%'
or TEXT like '%nakinr%' or TEXT like '%inere%'
or TEXT like '%ocilizuma%' or TEXT like '%ctemr%'
or TEXT like '%batacep%' or TEXT like '%renci%'
or TEXT like '%ituxima%' or TEXT like '%ituxa%')
);

```

```

-- numero de admisiones por paciente, del grupo de cohorte inicialmente seleccionado
select count(distinct hadm_id), subject_id
from cohort2
group by subject_id
order by count(distinct hadm_id) desc
;

```

```

-- Extraemos sólo aquellos que tienen más de una admisión
select count(distinct hadm_id), subject_id
from cohort2
group by subject_id
having count (distinct hadm_id) >1
;

```

```

-- para estos pacientes, calculamos la edad a la admisión, para las distintas admisiones
-- con el objetivo de quedarnos sólo con la última admisión y eliminar el resto
select distinct hadm_id, subject_id, age(admittime, dob)
from cohort2
where subject_id = 6317 or subject_id = 6448 or subject_id = 21021 or subject_id = 73200 or
subject_id = 97181 or subject_id = 98494
order by age(admittime, dob) desc ;

```

```

-- las líneas correspondientes a las siguientes admisiones deben ser eliminadas:
-- 194396 122673 116629 142627 131171 172815 167753 114109
delete from cohort2
where hadm_id = 194396
or hadm_id = 122673
or hadm_id = 116629
or hadm_id = 142627
or hadm_id = 131171
or hadm_id = 172815
or hadm_id = 167753
or hadm_id = 114109;

```

```

-- para detectar la medicacion a la admisión hay que revisar la columna TEXT
-- creamos una tabla que extraemos en formato CVS para su análisis manual
select hadm_id, subject_id, text

```

```
from cohort2
order by hadm_id
;
```

```
-- Creamos nueva columna "medatadmin" para introducir el medicamento a admisión
```

```
alter table cohort2
```

```
add column medatadmin varchar;
```

```
-- añadimos el medicamento para cada una de las admisiones
```

```
update cohort2 set medatadmin = 'infliximab' where hadm_id = 102557;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 102781;
update cohort2 set medatadmin = 'none' where hadm_id = 103889;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 104948;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 105900;
update cohort2 set medatadmin = 'none' where hadm_id = 105916;
update cohort2 set medatadmin = 'none' where hadm_id = 108644;
update cohort2 set medatadmin = 'none' where hadm_id = 108849;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 111821;
update cohort2 set medatadmin = 'tocilizumab' where hadm_id = 114292;
update cohort2 set medatadmin = 'none' where hadm_id = 117191;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 118208;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 118694;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 124983;
update cohort2 set medatadmin = 'none' where hadm_id = 125603;
update cohort2 set medatadmin = 'tocilizumab' where hadm_id = 125687;
update cohort2 set medatadmin = 'none' where hadm_id = 126393;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 127116;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 128729;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 130710;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 134410;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 135596;
update cohort2 set medatadmin = 'none' where hadm_id = 137091;
update cohort2 set medatadmin = 'abatacept' where hadm_id = 137866;
update cohort2 set medatadmin = 'none' where hadm_id = 139354;
update cohort2 set medatadmin = 'none' where hadm_id = 145875;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 147084;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 147181;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 148470;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 148982;
update cohort2 set medatadmin = 'none' where hadm_id = 149465;
update cohort2 set medatadmin = 'none' where hadm_id = 151223;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 151364;
update cohort2 set medatadmin = 'none' where hadm_id = 156834;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 157798;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 158759;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 163289;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 163473;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 165557;
update cohort2 set medatadmin = 'none' where hadm_id = 167198;
update cohort2 set medatadmin = 'none' where hadm_id = 168324;
update cohort2 set medatadmin = 'abatacept' where hadm_id = 170490;
update cohort2 set medatadmin = 'adalimumab' where hadm_id = 173430;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 174968;
update cohort2 set medatadmin = 'abatacept' where hadm_id = 175411;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 176556;
update cohort2 set medatadmin = 'none' where hadm_id = 177212;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 177754;
```

```

update cohort2 set medatadmin = 'etanercept' where hadm_id = 178896;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 179638;
update cohort2 set medatadmin = 'none' where hadm_id = 179805;
update cohort2 set medatadmin = 'none' where hadm_id = 182238;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 186443;
update cohort2 set medatadmin = 'none' where hadm_id = 187411;
update cohort2 set medatadmin = 'none' where hadm_id = 187944;
update cohort2 set medatadmin = 'infliximab' where hadm_id = 189028;
update cohort2 set medatadmin = 'none' where hadm_id = 189862;
update cohort2 set medatadmin = 'none' where hadm_id = 193878;
update cohort2 set medatadmin = 'none' where hadm_id = 194004;
update cohort2 set medatadmin = 'etanercept' where hadm_id = 194116;
update cohort2 set medatadmin = 'abatacept' where hadm_id = 195809;

```

-- comprobamos que el medicamento ha sido capturado para todos los pacientes y no hay errores de transcripción:

```

select count(distinct subject_id) from cohort2;
select subject_id, hadm_id
from cohort2
where medatadmin is null;
select count(distinct subject_id), medatadmin
from cohort2
group by medatadmin;

```

-- de la tabla previamente creada (cohort2) creamos otra que contenga exclusivamente la información de los pacientes seleccionados que están siendo tratados con alguno de los medicamentos biológicos

```

create table mimiciii.cohort2_selected as(
select hadm_id , subject_id, admittime, expire_flag, hospital_expire_flag, age(admittime,
dob), gender, dob, icd9_code, medatadmin
from cohort2
where medatadmin = 'abatacept'
or medatadmin = 'adalimumab'
or medatadmin = 'etanercept'
or medatadmin = 'infliximab'
or medatadmin = 'tocilizumab'
);

```

-- para el estudio de los datos en Rstudio, seleccionamos las columnas de la tabla, extrayendo de la columna "age" sólo los años de edad del paciente. La tabla será exportada en formato .csv para su análisis

```

select * from cohort2_selected;
select distinct(subject_id), hadm_id, hospital_expire_flag, expire_flag, (extract (year from
age)), gender, icd9_code, medatadmin
from cohort2_selected;

```

### 11.2.2 Anexo II.2: Selección de la cohorte con medicamento sin biológico:

-- De las tablas temporales previamente creadas, creamos tabla con todos los pacientes adultos de AR

```
drop table if exists cohortRA;
create table mimiciii.cohortRA as (
select
temp_admissions.hadm_id,
temp_admissions.subject_id,
temp_admissions.admittime,
temp_patients.dob,
temp_patients.gender,
age (temp_admissions.admittime, temp_patients.dob),
temp_admissions.hospital_expire_flag,
temp_patients.expire_flag,
temp_diagnoses_icd.icd9_code,
temp_noteevents.text
from temp_admissions
INNER JOIN temp_patients ON temp_ADMISSIONS.SUBJECT_ID = temp_PATIENTS.SUBJECT_ID
INNER JOIN temp_DIAGNOSES_ICD ON temp_ADMISSIONS.HADM_ID =
temp_DIAGNOSES_ICD.HADM_ID
INNER JOIN temp_NOTEEVENTS ON temp_ADMISSIONS.HADM_ID = temp_NOTEEVENTS.HADM_ID
where age(ADMITTIME, DOB) >= interval '18 years'
and ICD9_CODE = '7140');
```

-- comprobación del numero de pacientes

```
select * from cohortra limit 5;
select count(distinct subject_id) from cohortra; /*513 pacientes */
```

-- Creamos una columna para indicar si el paciente está recibiendo medicamento biológico o no

```
alter table cohortra
add column medatadmin varchar;
```

-- completamos la columna antes creada con 'biologic' para los pacientes incluidos en la cohorte con medicamento biológico y 'nonbiologic' para aquellos no incluidos.

```
update cohortRA set medatadmin = 'biologic' where subject_id in (
select subject_id from cohort2_selected
);
update cohortRA set medatadmin = 'nonbiologic' where subject_id not in (
select subject_id from cohort2_selected
);
```

-- comprobaciones

```
select count(subject_id)
from cohortra
where medatadmin is null; /* 0 pacientes */
select count(distinct subject_id)
from cohortra
where medatadmin like 'biologic'; /* 38 pacientes */
select count(distinct subject_id)
from cohortra
where medatadmin like 'nonbiologic'; /* 475 pacientes */
```

-- Calculamos cuantos pacientes tienen más de una admisión, para después eliminar entradas y quedarnos solo con la última (paciente tiene más edad)

```
select count(distinct hadm_id), subject_id
from cohortRA
group by subject_id
```

```

having count (distinct hadm_id) >1; /* 85 pacientes con más de 1 admisión */
select distinct hadm_id, subject_id, age
from cohortRA
where subject_id in (
select subject_id
from cohortRA
group by subject_id
having count (distinct hadm_id) >1
)
order by age ; /* mostramos la edad del paciente para cada una de las admisiones */

-- se observan que hay pacientes con 300 años o mas, se extraen para analizar estas entradas
select hadm_id, subject_id, dob, admittime, age, text
from cohortRA
where subject_id in (
select subject_id
from cohortRA
group by subject_id
having count (distinct hadm_id) >1
)
and age(admittime, dob) >= interval '300 years'
order by subject_id ;
/* tras análisis vemos que este valor es generado por el proceso de desidentificación de los datos, se
harán correcciones posteriores */

--pequeña query para calcular edad máxima de cada uno de los pacientes
select max (age)
from cohortra
group by subject_id;

-- eliminamos todas las entradas que no corresponden con la edad máxima, así nos quedamos sólo
con la última hadm_id.
delete from cohortra
where age not in (
select max (age)
from cohortra
group by subject_id);

-- Hacemos las comprobaciones:
select count(distinct hadm_id)
from cohortRA
group by subject_id
having count (distinct hadm_id) >1; /* No hay pacientes con más de una admisión */
select count(distinct subject_id)
from cohortra;

-- Nota: corrección para la edad paciente: hay pacientes que tienen 300 años o mas, esto es debido al
proceso de deidentificación de los datos
-- Cambiamos esa edad por 91.4 años, que es la media de edad de esos pacientes, según MIMIC.
update cohortra
set age = interval '91 years 4 months 26 days'
where age >= interval '300 years';

-- exportamos la cohorte AR (pacientes adultos con AR)
select distinct(subject_id), hadm_id, hospital_expire_flag, expire_flag, (extract (year from
age)), gender, medatadmin, icd9_code
from cohortra;

```

## 11.3 Anexo III: Código R para el análisis de las cohorte y estudio de las diferencias en la mortalidad.

### 11.3.1 Comparación de la mortalidad entre pacientes con AR tratados con los distintos medicamentos biológicos

- **Importación de los datos y adaptación para el formato necesario.**

Importamos los datos:

```
library(readr)
cohort2_selected <- read_csv("C:/Users/Marta/Desktop/TFM/PEC3/cohort2_selected_202104211802.csv")

##
## -- Column specification -----
## cols(
##   subject_id = col_double(),
##   hadm_id = col_double(),
##   hospital_expire_flag = col_double(),
##   expire_flag = col_double(),
##   date_part = col_double(),
##   gender = col_character(),
##   icd9_code = col_double(),
##   medatadmin = col_character()
## )

# View(cohort2_selected_202104211802)
```

Pasamos a “factor” algunas columnas que venian como numérico, para análisis posteriores. También cambiamos el nombre a la columna “date\_part” que corresponde a la edad en años, para facilitar el análisis:

```
cohort2_selected$hospital_expire_flag <- factor(cohort2_selected$hospital_expire_flag)
cohort2_selected$expire_flag <- factor(cohort2_selected$expire_flag)
cohort2_selected$gender <- factor(cohort2_selected$gender)
cohort2_selected$medatadmin <- factor(cohort2_selected$medatadmin)
cohort2_selected$age <- cohort2_selected$date_part

str(cohort2_selected)

## tibble [38 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ subject_id      : num [1:38] 97181 74624 90801 97637 53181 .
## ..
## $ hadm_id         : num [1:38] 147084 105900 118694 165557 170490 ...
## $ hospital_expire_flag: Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
## $ expire_flag      : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 2 1 1 1 ...
## $ date_part        : num [1:38] 72 78 61 57 62 67 67 74 74 66 .
## ..
```

```
## $ gender          : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1
1 1 1 ...
## $ icd9_code       : num [1:38] 7140 7140 7140 7140 7140 7140 7
140 7140 7140 7140 ...
## $ medatadmin      : Factor w/ 5 levels "abatacept","adalimumab
",,..: 2 4 2 2 1 1 3 2 3 4 ...
## $ age             : num [1:38] 72 78 61 57 62 67 67 74 74 66 .
..
## - attr(*, "spec")=
## .. cols(
## ..   subject_id = col_double(),
## ..   hadm_id = col_double(),
## ..   hospital_expire_flag = col_double(),
## ..   expire_flag = col_double(),
## ..   date_part = col_double(),
## ..   gender = col_character(),
## ..   icd9_code = col_double(),
## ..   medatadmin = col_character()
## .. )
```

```
summary(cohort2_selected)
```

```
##   subject_id      hadm_id      hospital_expire_flag expire_flag
## Min.   : 717      Min.   :102557    0:35                0:28
## 1st Qu.:31315    1st Qu.:126044    1: 3                1:10
## Median :57042    Median :148726
## Mean   :57828    Mean   :148652
## 3rd Qu.:84824    3rd Qu.:174584
## Max.   :99660    Max.   :195809
##   date_part      gender   icd9_code      medatadmin      age
## Min.   :44.00    F:27   Min.   :7140    abatacept : 4    Min.   :44.
00
## 1st Qu.:62.00    M:11   1st Qu.:7140    adalimumab :12   1st Qu.:62.
00
## Median :67.00                Median :7140    etanercept : 9    Median :67.
00
## Mean   :67.58                Mean   :7140    infliximab :11   Mean   :67.
58
## 3rd Qu.:75.50                3rd Qu.:7140    tocilizumab: 2    3rd Qu.:75.
50
## Max.   :86.00                Max.   :7140                Max.   :86.
00
```

- **Descripción de la cohorte seleccionada (pacientes con los distintos medicamentos biológicos)**

- o Género

Tablas descriptivas del número de pacientes divididos por género, en números absolutos y en porcentaje:

```
table(cohort2_selected$gender)
```

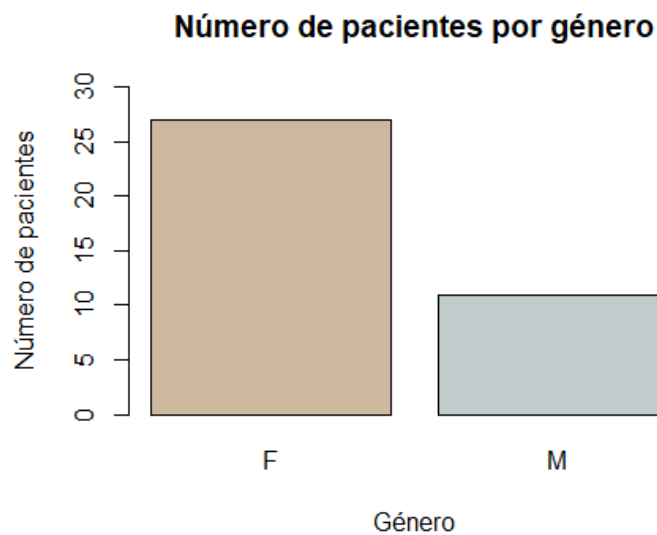
```
##
## F M
## 27 11

round(prop.table(table(cohort2_selected$gender)), 2)

##
## F M
## 0.71 0.29
```

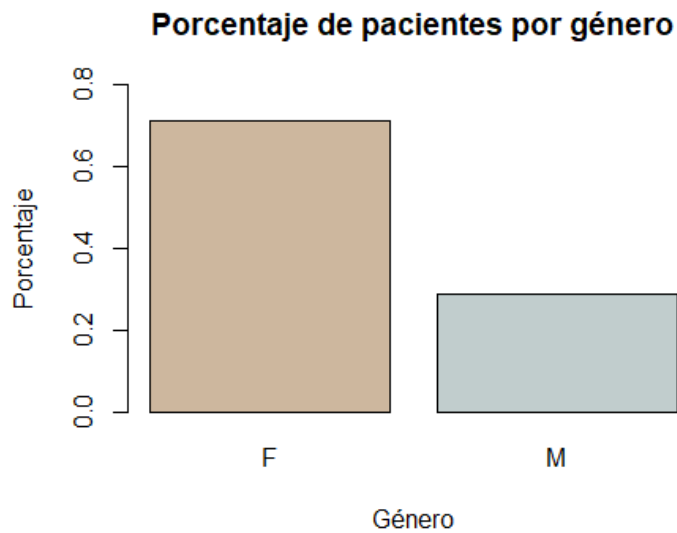
Gráficos de barras de las tablas anteriores (descripción visual del número de pacientes divididos por género, en números absolutos y en porcentaje)

```
barplot(table(cohort2_selected$gender), col= c("bisque3", "azure3"), main = "Número de pacientes por género", ylab = "Número de pacientes", xlab = c("Género"), ylim = c(0, 30))
```



```
barplot(prop.table(table(cohort2_selected$gender)), col= c("bisque3", "azure3"), main = "Porcentaje de pacientes por género", ylab = c("Porcentaje"), xlab = c("Género"), ylim = c(0, 0.8))
```





- Edad

Media y mediana, relativo a la edad y desviación estandar de la población seleccionada (pacientes con medicamentos biológicos):

```
mean(cohort2_selected$age)
```

```
## [1] 67.57895
```

```
median(cohort2_selected$age)
```

```
## [1] 67
```

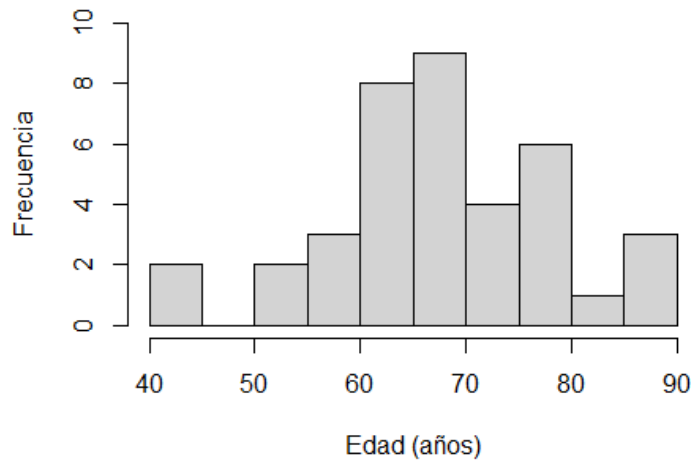
```
sd(cohort2_selected$age)
```

```
## [1] 10.48972
```

Histograma, representación visual de la distribución de las distintas edades de la población seleccionada:

```
hist(cohort2_selected$age, main="Distribución de la edad de los pacientes", xlab = "Edad (años)", ylab = "Frecuencia", ylim = c(0, 10))
```

### Distribución de la edad de los pacientes



- Estatus "hospital\_expire\_flag" (fallecido en el hospital)

Tabla descriptiva del número de pacientes fallecidos en el hospital, y su correspondiente gráfico de barras, en números absolutos.

```
table(cohort2_selected$hospital_expire_flag)
```

```
##  
##  0  1  
## 35  3
```

```
barplot(table(cohort2_selected$hospital_expire_flag), main = "Estatus  
'hospital_expire_flag', nº pacientes", ylab = "Numero de pacientes", x  
lab = "Estatus 'hospital_expire_flag'", col = c("grey", "black"), ylim  
= c(0, 35))
```

### Estatus 'hospital\_expire\_flag', nº pacientes

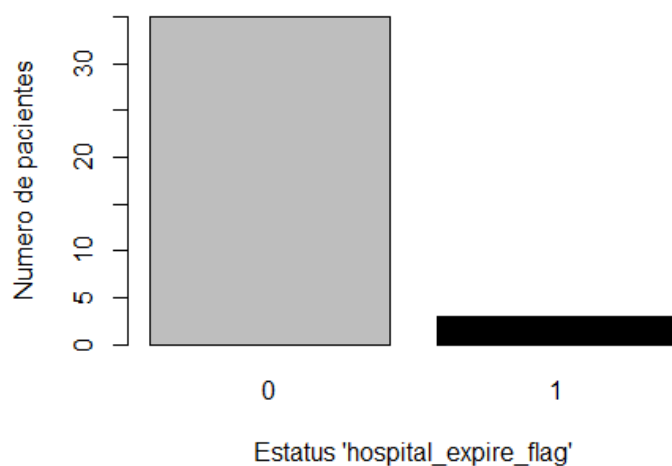
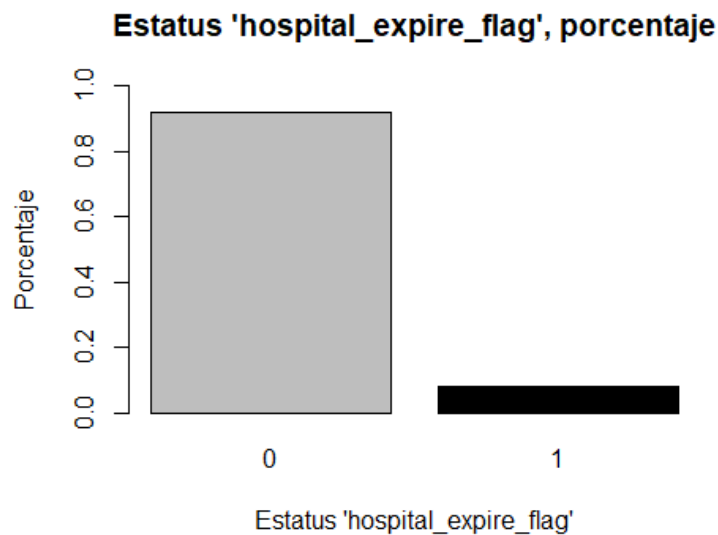


Tabla descriptiva del número de pacientes fallecidos en el hospital, y su correspondiente gráfico de barras, en porcentaje.

```
round(prop.table(table(cohort2_selected$hospital_expire_flag)), 2)

##
##  0  1
## 0.92 0.08

barplot(prop.table(table(cohort2_selected$hospital_expire_flag)), main = "Estatus 'hospital_expire_flag', porcentaje", ylab = "Porcentaje", xlab = "Estatus 'hospital_expire_flag'", col = c("grey", "black"), ylim = c(0, 1))
```



- Estatus "expire flag" (fallecido, según los registros de la seguridad social)

Tabla descriptiva del número de pacientes fallecidos (según los registros de la seguridad social), y correspondiente gráfico de barras, en números absolutos.

```
table(cohort2_selected$expire_flag)

##
##  0  1
## 28 10

barplot(table(cohort2_selected$expire_flag), main = "Estatus 'expire_flag', nº pacientes", ylab = "Numero de pacientes", xlab = "Estatus 'expire_flag'", col = c("grey", "black"), ylim = c(0, 35))
```

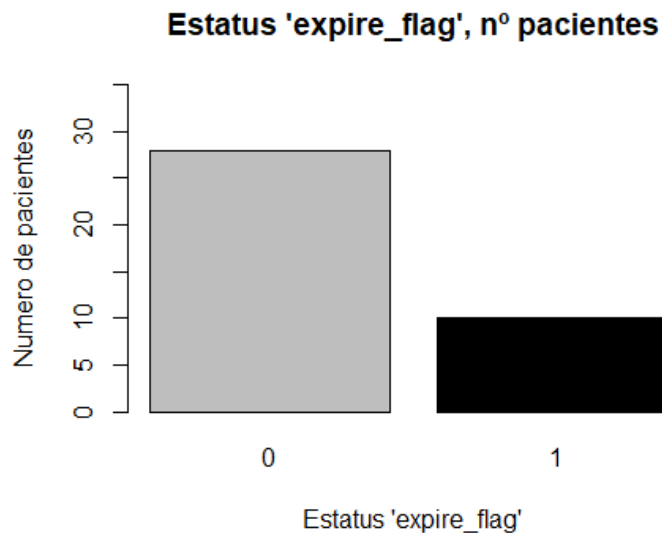
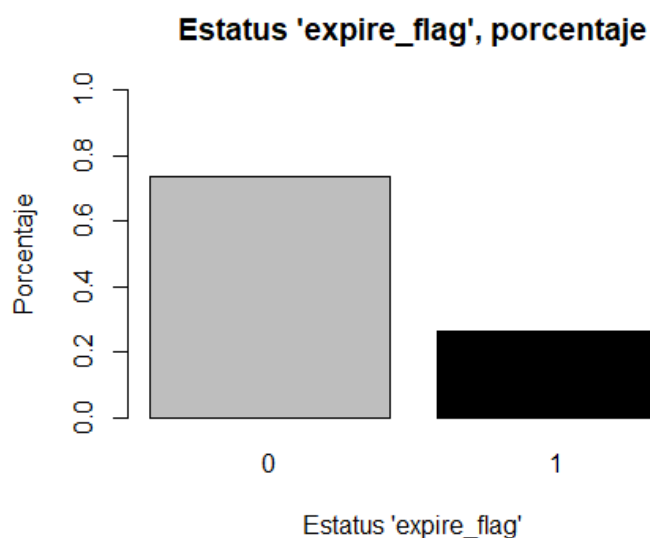


Tabla descriptiva del número de pacientes fallecidos (según los registros de la seguridad social), y correspondiente gráfico de barras, en porcentaje.

```
round(prop.table(table(cohort2_selected$expire_flag)), 2)
```

```
##
##      0      1
## 0.74 0.26
```

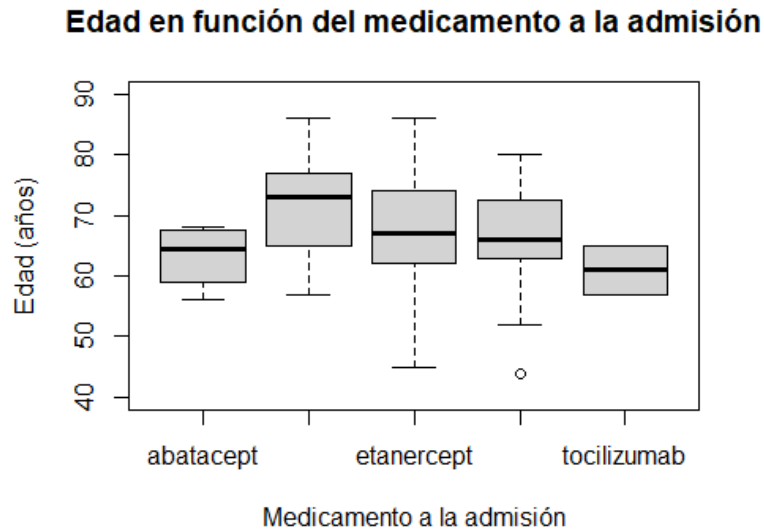
```
barplot(prop.table(table(cohort2_selected$expire_flag)), main = "Estat
us 'expire_flag', porcentaje", ylab = "Porcentaje", xlab = "Estatus 'e
xpire_flag'", col = c("grey", "black"), ylim = c(0, 1))
```



- Estudio de la homogeneidad entre grupos, atendiendo a edad, género y medicamento a la admisión

Estudio de la homogeneidad de la edad de los pacientes, entre diferentes grupos de medicamentos:

```
boxplot(age ~ medatadmin, cohort2_selected, ylim = c(40, 90), xlab = c("Medicamento a la admisión"), ylab = c("Edad (años)"), main = c("Edad en función del medicamento a la admisión"))
```



```
summary(aov(age ~ medatadmin, cohort2_selected))
```

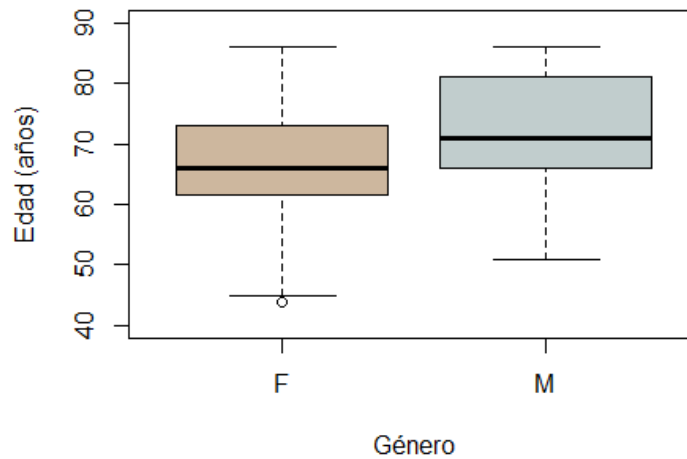
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## medatadmin  4    451   112.8    1.028  0.407
## Residuals 33   3620   109.7
```

El gráfico boxplot no refleja grandes diferencias de edad entre grupos de medicamentos. El rest ANOVA realizado confirma la hipótesis de homogeneidad entre grupos.

Estudio de la homogeneidad de la edad de los pacientes, entre grupos de género:

```
boxplot(age ~ gender, cohort2_selected, ylim = c(40, 90), xlab = c("Género"), ylab = c("Edad (años)"), main = c("Edad en función del género"), col = c("bisque3", "azure3"))
```

### Edad en función del género



```
summary(aov(age ~ gender, cohort2_selected))
```

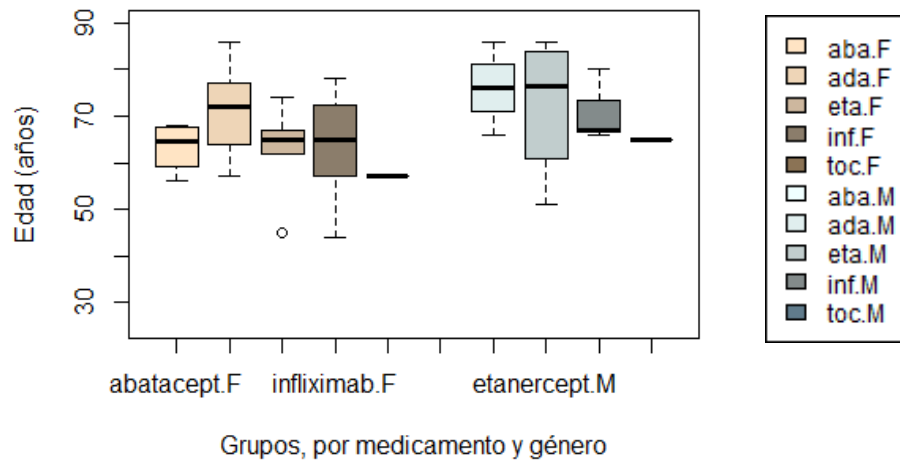
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## gender      1    354    354.4   3.433 0.0721 .
## Residuals  36   3717    103.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El gráfico boxplot no refleja grandes diferencias de edad entre grupos definidos por el género del paciente. El rest ANOVA realizado confirma la hipótesis de homogeneidad entre grupos.

Estudio de la homogeneidad de la edad de los pacientes, entre grupos diferenciados por el género y el medicamento a la admisión:

```
boxplot(age ~ medatadmin+gender, cohort2_selected, main="Edad por medicamento y género", ylab = "Edad (años)", xlab = "Grupos, por medicamento y género", col=c("bisque1", "bisque2", "bisque3", "bisque4", "burlywood4", "azure1", "azure2", "azure3", "azure4", "lightskyblue4"), ylim = c(25, 90))
legend("bottomright", "grupos", c("aba.F", "ada.F", "eta.F", "inf.F", "toc.F", "aba.M", "ada.M", "eta.M", "inf.M", "toc.M"), fill = c("bisque1", "bisque2", "bisque3", "bisque4", "burlywood4", "azure1", "azure2", "azure3", "azure4", "lightskyblue4"), xpd=TRUE, inset=c(0, -0.0), cex=.75)
```

### Edad por medicamento y género



```
summary(aov(age ~ medatadmin+gender, cohort2_selected))
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## medatadmin  4    451   112.8    1.119 0.3650
## gender      1    393   393.3    3.901 0.0569 .
## Residuals  32   3227   100.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El gráfico boxplot no refleja grandes diferencias de edad entre grupos definidos por el género del paciente y su medicamento a la admisión. El rest ANOVA realizado confirma la hipótesis de homogeneidad entre grupos.

Tablas descriptivas del número de pacientes por medicamento a la admisión y género, dadas en números absolutos y en porcentaje, junto a gráfico descriptivo representativo de la tabla dada en números absolutos

```
table(cohort2_selected$medatadmin, cohort2_selected$gender)
```

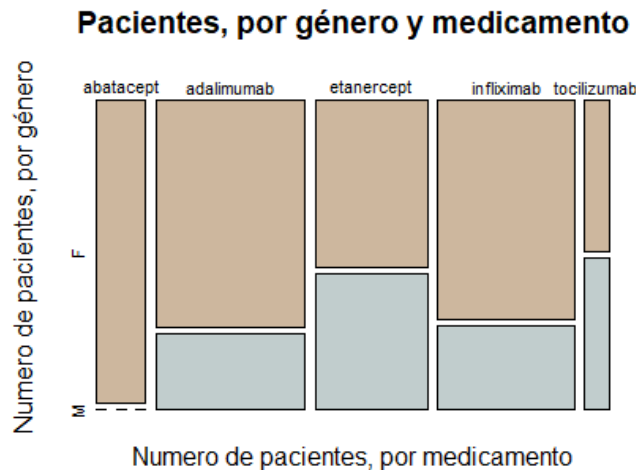
```
##
##           F M
## abatacept  4 0
## adalimumab 9 3
## etanercept  5 4
## infliximab  8 3
## tocilizumab 1 1
```

```
round(prop.table(table(cohort2_selected$medatadmin, cohort2_selected$gender)), 2)
```

```
##
##           F    M
## abatacept 0.11 0.00
## adalimumab 0.24 0.08
## etanercept 0.13 0.11
```

```
## infliximab 0.21 0.08
## tocilizumab 0.03 0.03
```

```
mosaicplot(table(cohort2_selected$medatadmin, cohort2_selected$gender)
, main = "Pacientes, por género y medicamento", xlab = "Numero de pacie
ntes, por medicamento", ylab = "Numero de pacientes, por género", col
= c("bisque3", "azure3"))
```



- Estudio estadístico de la mortalidad, en el hospital y según los registros de la seguridad social, entre distintos medicamentos.
  - o Mortalidad en el hospital ("hospital\_expire\_flag")

Tabla de contingencia mostrando el estatus "hospital\_expire\_flag" para los distintos medicamentos. Estatus "hospital\_expire\_flag": 0=not expired, 1=expired

```
table(cohort2_selected$hospital_expire_flag, cohort2_selected$medatadmin)
```

```
##
##      abatacept adalimumab etanercept infliximab tocilizumab
## 0         4         10         9         10         2
## 1         0          2          0          1          0
```

Dado que hay varias celdas en la tabla que muestran un contejo inferior a 5, se le da preferencia al test de Fisher frente al test de Chi cuadrado, por ser éste más adecuado para nuestros datos.

```
# chisq.test(cohort2_selected$hospital_expire_flag, cohort2_selected$medatadmin, simulate.p.value = TRUE)
```

```
fisher.test(cohort2_selected$hospital_expire_flag, cohort2_selected$medatadmin)
```

```
##
## Fisher's Exact Test for Count Data
##
```



```
## data: cohort2_selected$hospital_expire_flag and cohort2_selected$m
edatadmin
## p-value = 0.8592
## alternative hypothesis: two.sided

fisher.test(cohort2_selected$hospital_expire_flag, cohort2_selected$m
edatadmin, simulate.p.value = TRUE)

##
## Fisher's Exact Test for Count Data with simulated p-value (based o
n
## 2000 replicates)
##
## data: cohort2_selected$hospital_expire_flag and cohort2_selected$m
edatadmin
## p-value = 0.8561
## alternative hypothesis: two.sided
```

Este test parte sobre la hipótesis nula de homogeneidad entre grupos. Dado que el valor de p obtenido es mayor al valor de “alpha” establecido (0.05), concluimos que no existen diferencias significativas entre medicamentos biológicos en cuanto a la mortalidad en hospital.

- Mortalidad según los registros de la seguridad social (“expire flag”)

Tabla de contingencia mostrando el estatus “expire\_flag” para los distintos medicamentos. Estatus “expire\_flag”: 0=not expired, 1=expired

```
table( cohort2_selected$expire_flag, cohort2_selected$medatadmin)

##
##      abatacept adalimumab etanercept infliximab tocilizumab
## 0          4          8          6          8          2
## 1          0          4          3          3          0
```

Dado que hay varias celdas en la tabla que muestran un contaje inferior a 5, se le da preferencia al test de Fisher frente al test de Chi cuadrado, por ser éste más adecuado para nuestros datos.

```
#chisq.test(cohort2_selected$expire_flag, cohort2_selected$medatadmin,
simulate.p.value = TRUE)
```

```
fisher.test(cohort2_selected$expire_flag, cohort2_selected$medatadmin)

##
## Fisher's Exact Test for Count Data
##
## data: cohort2_selected$expire_flag and cohort2_selected$medatadmin
## p-value = 0.8044
## alternative hypothesis: two.sided

fisher.test(cohort2_selected$expire_flag, cohort2_selected$medatadmin,
simulate.p.value = TRUE)
```

```
##
## Fisher's Exact Test for Count Data with simulated p-value (based on
## 2000 replicates)
##
## data: cohort2_selected$expire_flag and cohort2_selected$medatadmin
## p-value = 0.8056
## alternative hypothesis: two.sided
```

Este test parte sobre la hipótesis nula de homogeneidad entre grupos. Dado que el valor de p obtenido es mayor al valor de “alpha” establecido (0.05), concluimos que no existen diferencias significativas entre medicamentos biológicos en cuanto a la mortalidad, según los registros de la seguridad social.

### 11.3.2 Comparación de la mortalidad entre pacientes con AR tratados con medicamentos biológicos o no biológicos.

#### - Importación de datos y adaptación para el formato necesario

Importamos los datos

```
library(readr)
cohortRA <- read_csv("cohortra_202105040000.csv")

##
## -- Column specification -----
## -----
## cols(
##   subject_id = col_double(),
##   hadm_id = col_double(),
##   hospital_expire_flag = col_double(),
##   expire_flag = col_double(),
##   date_part = col_double(),
##   gender = col_character(),
##   medatadmin = col_character(),
##   icd9_code = col_double()
## )

# View(cohortra_202105040000)
```

Cambio de formato de los datos que lo requieren

```
cohortRA$hospital_expire_flag <- factor(cohortRA$hospital_expire_flag)
cohortRA$expire_flag <- factor(cohortRA$expire_flag)
cohortRA$gender <- factor(cohortRA$gender)
cohortRA$medatadmin <- factor(cohortRA$medatadmin)
cohortRA$age <- cohortRA$date_part

str(cohortRA)

## tibble [513 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ subject_id      : num [1:513] 51459 30064 4471 90066 54585 .
## ..
```

```

## $ hadm_id          : num [1:513] 186203 172342 108658 129683 16
7198 ...
## $ hospital_expire_flag: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
2 1 1 ...
## $ expire_flag      : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1
2 1 1 ...
## $ date_part        : num [1:513] 86 52 37 79 88 55 65 62 59 83
...
## $ gender           : Factor w/ 2 levels "F","M": 1 1 1 2 1 2 1
2 1 1 ...
## $ medatadmin       : Factor w/ 2 levels "biologic","nonbiologic
": 2 2 2 2 2 2 2 2 2 ...
## $ icd9_code        : num [1:513] 7140 7140 7140 7140 7140 7140
7140 7140 7140 7140 ...
## $ age              : num [1:513] 86 52 37 79 88 55 65 62 59 83
...
## - attr(*, "spec")=
## .. cols(
## ..   subject_id = col_double(),
## ..   hadm_id = col_double(),
## ..   hospital_expire_flag = col_double(),
## ..   expire_flag = col_double(),
## ..   date_part = col_double(),
## ..   gender = col_character(),
## ..   medatadmin = col_character(),
## ..   icd9_code = col_double()
## .. )

```

**summary**(cohortRA)

```

##   subject_id      hadm_id      hospital_expire_flag  expire_flag
## Min.   : 62      Min.   :100227    0:442              0:278
## 1st Qu.:16843    1st Qu.:124300    1: 71              1:235
## Median :32504    Median :149165
## Mean   :43000    Mean   :148813
## 3rd Qu.:68217    3rd Qu.:173430
## Max.   :99660    Max.   :199418
##   date_part      gender      medatadmin      icd9_code      age
## Min.   :28.0     F:357    biologic      : 38      Min.   :7140      Min.   :28
.0
## 1st Qu.:62.0     M:156    nonbiologic:475  1st Qu.:7140      1st Qu.:62
.0
## Median :73.0
## Mean   :71.1
## 3rd Qu.:81.0
## Max.   :91.0

```

- **Descripción de la cohorte sin medicamento biológico**

o Selección del subset

```
cohortRA_nonbiologic <- subset(cohortRA, medatadmin == 'nonbiologic',  
select = c(subject_id, hadm_id, hospital_expire_flag, expire_flag, age  
, gender, medatadmin))
```

o Género

Tablas descriptivas del número de pacientes divididos por género, en números absolutos y en porcentaje:

```
table(cohortRA_nonbiologic$gender)
```

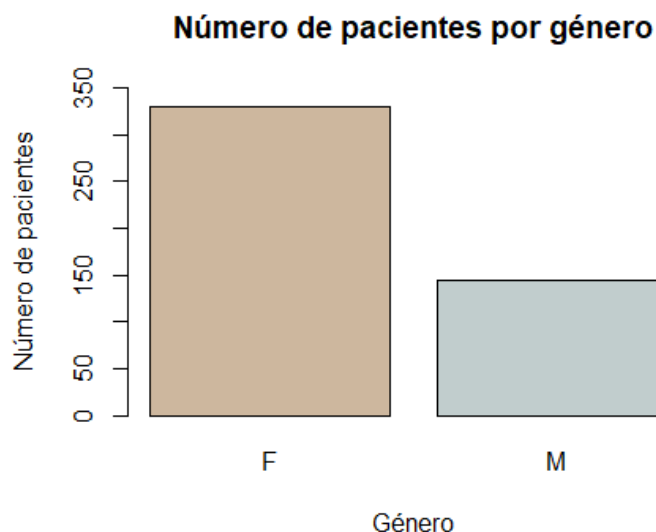
```
##  
##   F   M  
## 330 145
```

```
round(prop.table(table(cohortRA_nonbiologic$gender)), 2)
```

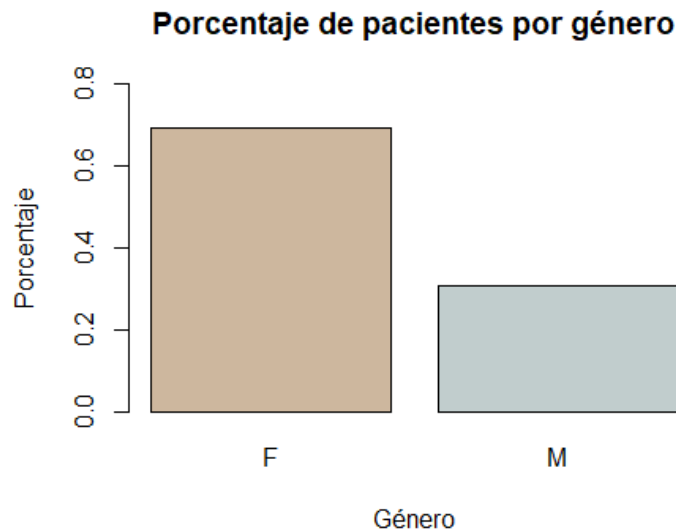
```
##  
##   F   M  
## 0.69 0.31
```

Gráficos de barras de las tablas anteriores (descripción visual del número de pacientes divididos por género, en números absolutos y en porcentaje)

```
barplot(table(cohortRA_nonbiologic$gender), col= c("bisque3", "azure3"),  
main = "Número de pacientes por género", ylab = "Número de paciente  
s", xlab = "Género", ylim = c(0, 350) )
```



```
barplot(prop.table(table(cohortRA_nonbiologic$gender)), col= c("bisque3", "azure3"), main = "Porcentaje de pacientes por género", ylab = "Porcentaje", xlab = "Género", ylim = c(0, 0.80))
```



- Edad

Media y mediana, relativo a la edad y desviación estandar de la población seleccionada:

```
mean(cohortRA_nonbiologic$age)
```

```
## [1] 71.38316
```

```
median(cohortRA_nonbiologic$age)
```

```
## [1] 74
```

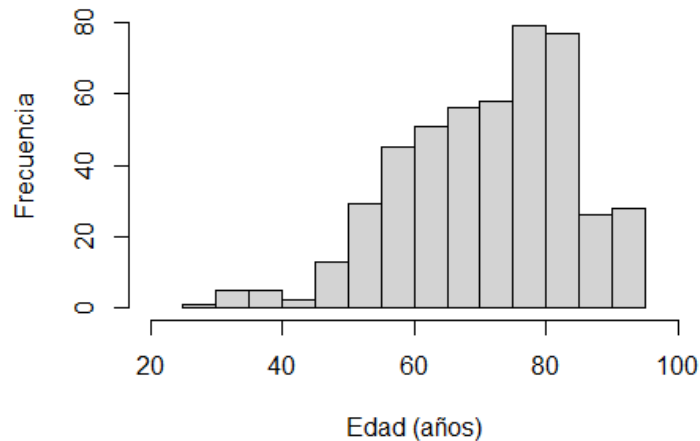
```
sd(cohortRA_nonbiologic$age)
```

```
## [1] 12.71418
```

Histograma, representación visual de la distribución de las distintas edades de la población seleccionada (pacientes sin medicamento biológico):

```
hist(cohortRA_nonbiologic$age, main="Distribución de la edad de los pacientes", xlab = "Edad (años)", ylab = "Frecuencia", xlim = c(20, 100), ylim = c(0, 85))
```

## Distribución de la edad de los pacientes



Nota: recordamos que hay pacientes que inicialmente mostraban una edad de 300 años. esto es debido al proceso de desidentificación, en el que las fechas de nacimiento de pacientes de más de 89 años fueron modificadas. La edad media de este grupo de población era de 91.4 años y la edad de estos pacientes fue modificada para reflejar ese número.

- Estatus "hospital\_expire\_flag" (fallecido en el hospital)

Tabla descriptiva del número de pacientes fallecidos en el hospital, y su correspondiente gráfico de barras, en números absolutos.

```
table(cohortRA_nonbiologic$hospital_expire_flag)
```

```
##  
##  0  1  
## 407 68
```

```
barplot(table(cohortRA_nonbiologic$hospital_expire_flag), main = "Estatus 'hospital_expire_flag', nº pacientes", ylab = "Numero de pacientes", xlab = "Estatus 'hospital_expire_flag'", col = c("grey", "black"), ylim = c(0, 500))
```

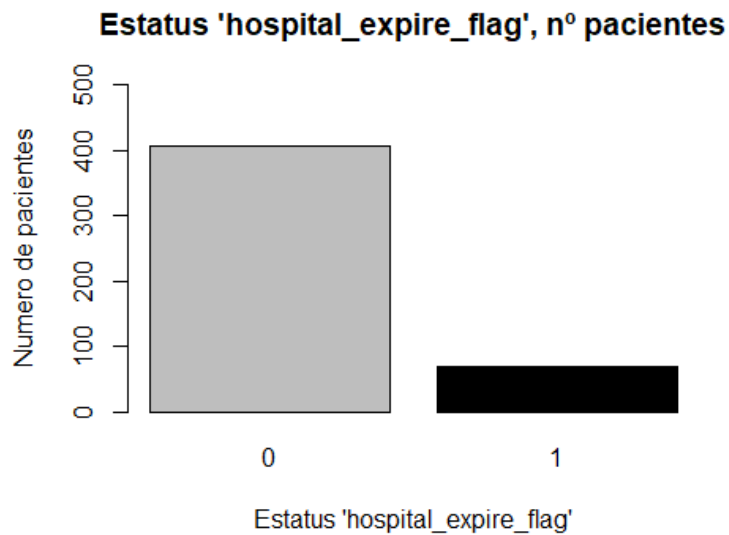


Tabla descriptiva del número de pacientes fallecidos (según los registros de la seguridad social), y correspondiente gráfico de barras, en porcentaje.

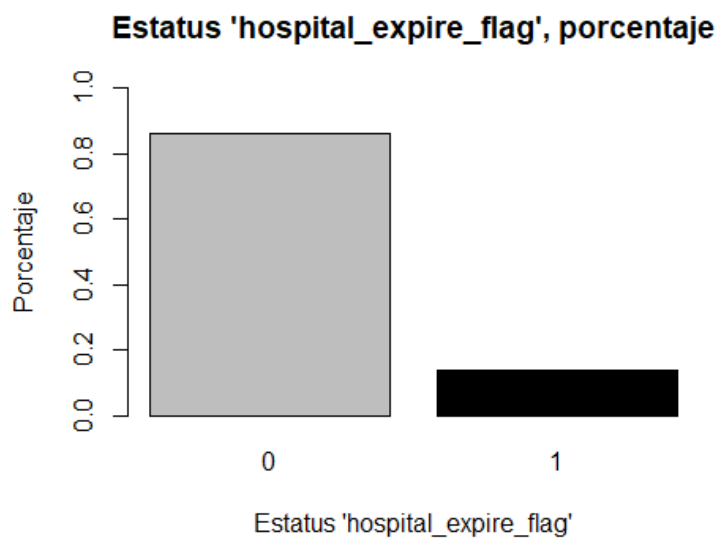
```

round (prop.table(table(cohortRA$hospital_expire_flag)), 2)

##
##      0      1
## 0.86 0.14

barplot(prop.table(table(cohortRA$hospital_expire_flag)), main = "Estatus 'hospital_expire_flag', porcentaje", ylab = "Porcentaje", xlab = "Estatus 'hospital_expire_flag'", col = c("grey", "black"), ylim = c(0, 1))

```



- Estatus "expire\_flag" (fallecido según los registros de la seguridad social)

Tabla descriptiva del número de pacientes fallecidos (según los registros de la seguridad social), y correspondiente gráfico de barras, en números absolutos.

```
table(cohortRA_nonbiologic$expire_flag)

##
##  0  1
## 250 225

barplot(table(cohortRA_nonbiologic$expire_flag), main = "Estatus 'expire_flag', nº pacientes", ylab = "Numero de pacientes", xlab = "Estatus 'expire_flag'", col = c("grey", "black"), ylim = c(0, 500))
```

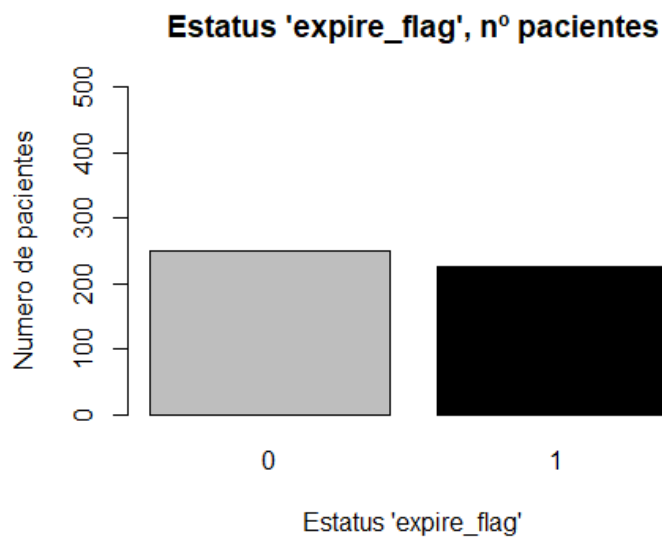


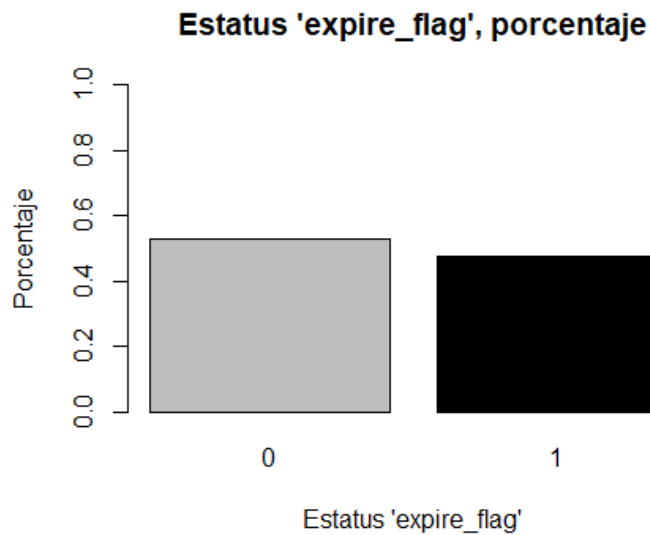
Tabla descriptiva del número de pacientes fallecidos (según los registros de la seguridad social), y correspondiente gráfico de barras, en porcentaje.

```
round(prop.table(table(cohortRA_nonbiologic$expire_flag)), 2)

##
##  0  1
## 0.53 0.47

barplot(prop.table(table(cohortRA_nonbiologic$expire_flag)), main = "Estatus 'expire_flag', porcentaje", ylab = "Porcentaje", xlab = "Estatus 'expire_flag'", col = c("grey", "black"), ylim = c(0, 1))
```

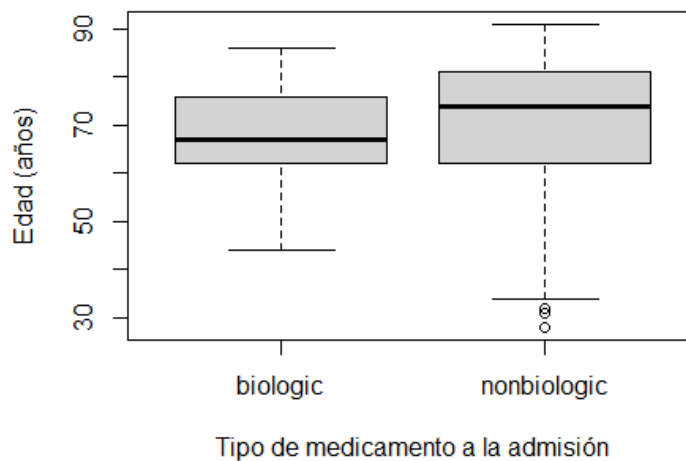




- Estudio de la homogeneidad entre grupos, atendiendo a edad, género y tipo de medicamento a la admisión

Estudio de la homogeneidad de la edad de los pacientes, entre diferentes tipos de medicamento:

```
boxplot(age ~ medatadmin, cohortRA, xlab = 'Tipo de medicamento a la admisión', ylab = 'Edad (años)')
```



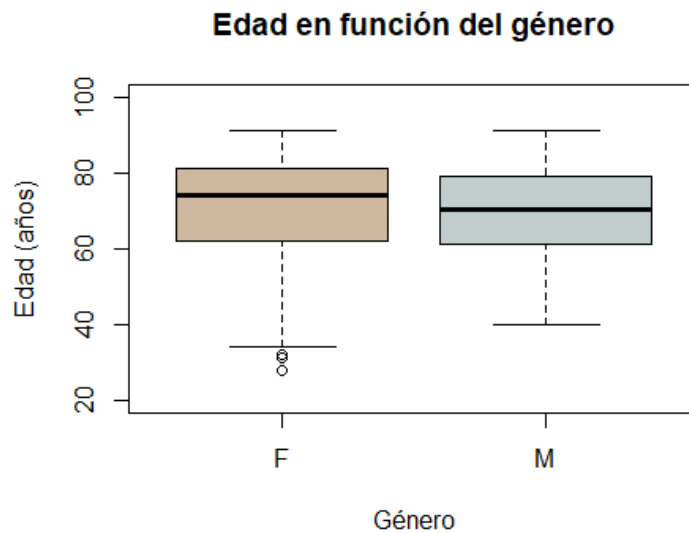
```
summary(aov(age ~ medatadmin, cohortRA))
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## medatadmin  1    509    509.2   3.225 0.0731 .
## Residuals 511  80694    157.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El gráfico boxplot no refleja grandes diferencias de edad entre tipos de medicamento. El test ANOVA realizado confirma la hipótesis de homogeneidad entre grupos.

Estudio de la homogeneidad de la edad de los pacientes, entre grupos de género:

```
boxplot(age ~ gender, cohortRA, yin = c(20, 100), la = c("Género"), ylab = c("Edad (años)"), main = c("Edad en función del género"), col = c("bisque3", "azure3"))
```



```
summary(aov(age ~ gender, cohortRA))
```

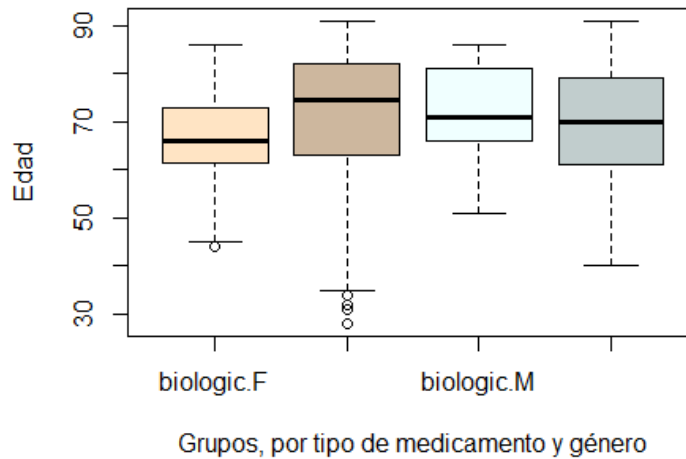
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## gender      1    278   278.3    1.757  0.186
## Residuals 511  80924   158.4
```

El gráfico boxplot no refleja grandes diferencias de edad entre grupos definidos por el género del paciente. El test ANOVA realizado confirma la hipótesis de homogeneidad entre grupos.

Estudio de la homogeneidad de la edad de los pacientes, entre grupos diferenciados por el género y el medicamento a la admisión:

```
boxplot(age ~ medatadmin+gender, cohortRA, main="Edad por género y medicamento", ylab = "Edad", xlab = "Grupos, por tipo de medicamento y género", col=c("bisque1", "bisque3", "azure1", "azure3"))
```

### Edad por género y medicamento



```
summary(aov(age ~ medatadmin + gender, cohortRA))
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## medatadmin  1    509   509.2    3.230 0.0729 .
## gender      1    285   285.1    1.808 0.1793
## Residuals 510  80408   157.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El gráfico boxplot no refleja grandes diferencias de edad entre grupos definidos por el género del paciente y su medicamento a la admisión. El test ANOVA realizado confirma la hipótesis de homogeneidad entre grupos.

Tabla de contingencia, reflejando reparto de medicamentos biológicos y no biológicos entre géneros

```
table(cohortRA$medatadmin, cohortRA$gender)
```

```
##
##           F    M
## biologic   27  11
## nonbiologic 330 145
```

- **Estudio estadístico de la mortalidad, en el hospital y según los registros de la seguridad social, entre distintos tipos de medicamentos.**
  - o Mortalidad en el hospital (“hospital\_expire\_flag”)

Tabla de contingencia mostrando el estatus “hospital\_expire\_flag” para los dos tipos de medicamentos. Estatus “hospital\_expire\_flag”: 0=not expired, 1=expired

```
table(cohortRA$hospital_expire_flag, cohortRA$medatadmin)
```

```
##
##   biologic nonbiologic
## 0       35       407
## 1        3        68
```

Dado que hay varias celdas en la tabla que muestran un contejo inferior a 5, se le da preferencia al test de Fisher frente al test de Chi cuadrado, por ser éste más adecuado para nuestros datos.

```
# chisq.test(cohortRA$hospital_expire_flag, cohortRA$medatadmin)

fisher.test(cohortRA$hospital_expire_flag, cohortRA$medatadmin)

##
## Fisher's Exact Test for Count Data
##
## data: cohortRA$hospital_expire_flag and cohortRA$medatadmin
## p-value = 0.3369
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5878131 10.1693430
## sample estimates:
## odds ratio
##  1.947146
```

Este test parte sobre la hipótesis nula de homogeneidad entre grupos. Dado que el valor de p obtenido es mayor al valor de “alpha” establecido (0.05), concluimos que no existen diferencias significativas entre medicamentos biológicos y no biológicos en cuanto a la mortalidad en hospital.

- Mortalidad según los registros de la seguridad social (“expire flag”)

Tabla de contingencia mostrando el estatus “expire\_flag” para los dos tipos de medicamentos. Estatus “expire\_flag”: 0=not expired, 1=expired

```
table(cohortRA$expire_flag, cohortRA$medatadmin)

##
##      biologic nonbiologic
##  0         28         250
##  1         10         225
```

Al contrario que en casos anteriores, no hay celdas que muestren un número inferior a 5, por lo que podríamos realizar el test de Chi cuadrado. Dado que en todos los casos anteriores hemos realizado el test de Fisher, lo haremos aquí también, con el objetivo de poder comparar resultados realizados con el mismo test.

```
chisq.test(cohortRA$expire_flag, cohortRA$medatadmin)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: cohortRA$expire_flag and cohortRA$medatadmin
## X-squared = 5.4625, df = 1, p-value = 0.01943

fisher.test(cohortRA$expire_flag, cohortRA$medatadmin)

##
## Fisher's Exact Test for Count Data
##
```

```
## data: cohortRA$expire_flag and cohortRA$medatadmin
## p-value = 0.01686
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
## 1.154631 5.940476
## sample estimates:
## odds ratio
## 2.515948
```