UNIVERSIDAD DE SALAMANCA

Facultad de Ciencias, Departamento de Estadística

Grado en Estadística

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Trabajo Fin de Grado

Análisis e implementación del modelo común de datos OMOP y posterior estudio estadístico aplicado a datos de muestras de pacientes de SMD (Síndrome Mielodisplásico)

Autora:

Alicia Lojo Iglesias

Tutores:

Javier Martínez Elicegui

Angela Villaverde Ramiro

María Jesús Rivas López

2021

UNIVERSIDAD DE SALAMANCA

Facultad de Ciencias, Departamento de Estadística

Grado en Estadística

Trabajo Fin de Grado

Análisis e implementación del modelo común de datos OMOP y posterior estudio estadístico aplicado a datos de muestras de pacientes de SMD (Síndrome Mielodisplásico)

MARTINEZ ELICEGUI FRANCISCO JAVIER - 17437279J

Digitally signed by MARTINEZ ELICEGUI FRANCISCO JAVIER - 17437279J
Date: 2021.07.05 13:43:54 +02'00'

Autora:

Alicia Lojo Iglesias

Tutores:

Javier Martínez Elicegui

Angela Villaverde Ramiro

Firmado por VILLAVERDE RAMIRO ANGELA - 06587720Z el día 05/07/2021 con un certificado emitido por AC FNMT Usuarios

María Jesús Rivas López

RIVAS LOPEZ MARIA JESUS - 07956354X

Firmado digitalmente por RIVAS LOPEZ MARIA JESUS - 07956354X
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-07956354X, givenName=MARIA JESUS, sn=RIVAS LOPEZ, cn=RIVAS LOPEZ MARIA JESUS - 07956354X
Fecha: 2021.07.06 11:32:25 +02'00'

2021

Resumen

Uno de los principales objetivos de la medicina actual consiste en alcanzar mediante el aprendizaje, la evidencia médica. Para ello es necesario el almacenado en bases de datos de los datos para un posterior análisis estadístico. El modelo común de datos OMOP proporciona un modelo estandarizado para poder almacenar estos datos y facilitar su análisis. Mediante pacientes de síndromes mielodisplásicos (SMDs), un tipo de anomalía producida en las células de la sangre, se estudiará el proceso de extracción, transformación y carga de los datos en este tipo de base relacional, a la par que se realizarán análisis de supervivencia con ellos.

Palabras clave: Síndromes Mielodisplásicos, Modelo Común de Datos, OMOP, ETL, Análisis de supervivencia

Abstract

One of the main objectives of modern-day medicine is to achieve medical evidence through learning. In order to accomplish this, it is necessary to store the data in databases for further statistical analysis. The OMOP common data model provides a standardised model for storing these data and facilitating their analysis. By studying patients with myelodysplastic syndromes (MDS), a type of blood cell abnormality, the process of extracting, transforming and loading data into this type of relational database will be investigated, whilst survival analyses will be carried out with them.

Keywords: Myelodysplastic syndromes, Common Data Model, OMOP, ETL, Survival analysis.

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

Desde hace más de una década, uno de los principales objetivos de la medicina es la mejora de la **evidencia clínica** mediante el desarrollo de un sistema de atención medica de aprendizaje (Olsen, Aisner, & McGinnis, 2007).

El valor fundamental de este sistema es la recopilación de datos clínicos sobre pacientes durante la asistencia médica, teniendo como objetivo su análisis para poder obtener conclusiones extrapolables. Consecuentemente, se alcanzaría una evidencia global que sería empleada en posteriores informes de prácticas clínicas.

En 2007, The Institute of Medicine's Roundtable on Evidence-Based Medicine se marcó como meta que, en el pasado 2020, el 90% de las decisiones clínicas estuvieran respaldadas por información clínica precisa y actualizada que presentase la mayor evidencia posible (Olsen et al., 2007). Pese a que este objetivo no ha sido alcanzado en su totalidad, nos encontramos en un momento único en la historia de la salud, debido a la unión del conocimiento médico y la ciencia de datos

En este punto de unión intervienen, sobre todo, dos perfiles académicos; los especialistas clínicos y los analistas de datos. Esto se debe a que, en el arduo camino desde los datos de un paciente hasta la evidencia, es necesario el conocimiento en diversas ramas. Se requieren competencias estadísticas para conseguir extraer de los datos la información necesaria para dar respuestas coherentes a las preguntas clínicas. Por su parte, la informática es necesaria para implementar los algoritmos que organizan y analizan la gran cantidad de datos que se obtienen a lo largo del tiempo. Y de manera fundamental, son necesarios los conocimientos clínicos para evaluar los resultados y aplicarlos en la práctica clínica. Por lo tanto, es necesario un trabajo cohesionado para que se pueda llegar desde un punto a otro del proceso sin que la información se corrompa.

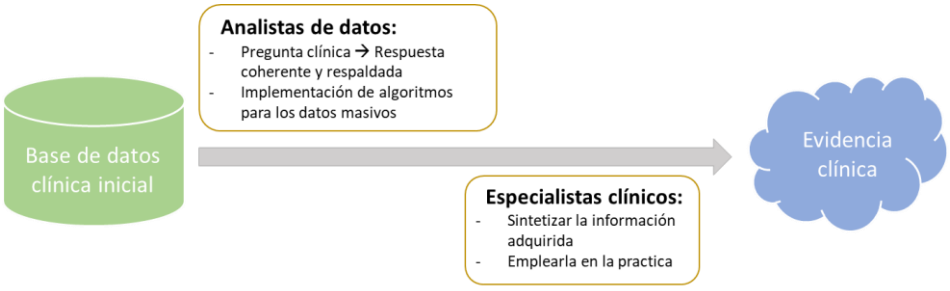


Ilustración 1: *Esquema sobre la participación de los especialistas en el paso de los datos a la evidencia clínica.*

El análisis de estos datos clínicos en numerosas ocasiones se complica debido a la diversa cantidad de bases de datos que existen (prácticamente tantas como centros médicos). Este problema, derivado de la falta de unificación, provoca dificultades de compresión por parte de los analistas a la hora de trabajar con las bases, así como la inconsistencia que presentan algunas a la hora de respaldar los enfoques analíticos. Por consiguiente, gran parte de los análisis se centra en el estudio de una única base de datos, ya que cada análisis se debe amoldar a cada base (Overhage, Ryan, Reich, Hartzema, & Stang, 2012).

A raíz de esta problemática surge la necesidad de crear un **modelo de datos común** que unifique y codifique los distintos conceptos médicos de forma que se elimine una de las principales adversidades que surgen en este proceso. De esta forma, se evitaría que cada estudio sea individualizado para cada base de datos.

1.1.Modelo de Datos Común OMOP

La Observational Medical Outcomes Partnership (OMOP) es una asociación público-privada que fue creada con objeto del desarrollo de la tecnología y los métodos necesarios para perfeccionar el uso de los datos observacionales, para maximizar el beneficio y minimizar el riesgo de los productos farmacéuticos (Stang et al., 2010).

Para ello, se planteó un Modelo de Datos Común (CDM) en el que estaría estandarizada tanto la estructura como el contenido y la semántica. Este CDM fue probado en 10 bases observacionales distintas con resultados favorables. Concluyendo en que la implementación del modelo común podría mejorar en los siguientes aspectos (Overhage, Ryan, Reich, Hartzema, & Stang, 2012):

- Comprensión de los datos, minimizando la confusión acerca de la organización de estos.
- Uso de un único sistema de codificación aplicable a todas las bases con ínfimos cambios para la implementación de los análisis.
- Capacidad de reproducción de los análisis por distintos investigadores, lo que facilitaría la compresión de estos métodos y el metaanálisis para la extracción de información adicional.

Debido a estos convenientes resultados, quedaba claro que este modelo no podía limitarse a ser un proyecto temporal, pero surgieron distintas barreras en las cuales era necesario un apoyo colaborativo. De esta forma, se puso en marcha la comunidad OHDSI (Observational Heralth Data Science and Informatics), que pretende el progreso de la técnica puntera de CDM OMOP.

Para la mejora del proyecto anterior se determinó que debería haber una colaboración en (Observational Health Data Sciences and Informatics, 2019):

- La estandarización de los datos y vocabularios, así como en las convenciones de la creación de ETL (Extract - Transform - Load), lo que da lugar a una mayor confianza en la calidad de los datos ya que existirá una coherencia tanto en la estructura como en la semántica y el contenido.
- La investigación de los métodos que facilitan la investigación.
- Las aplicaciones clínicas para llegar a la evidencia médica.

Otro de los puntos importantes de este proyecto internacional creado en 2014, es la característica de ciencia abierta que posee.

1.1.1. Ciencia abierta

El término de “ciencia abierta” se define según Anglada & Abadal (2018) como un cambio en el paradigma de la manera de hacer ciencia, cambiando sus métodos pero no sus motivaciones y objetivos.

Esta nueva forma de ciencia 2.0 pretende que la investigación científica sea accesible para todos los niveles de la sociedad, considerando como su fundamento que la ciencia debe ser abierta, colaborativa y hecha por y para la sociedad. Entendiendo el término de abierta, como libre y gratuita.

Debido a que es un término reciente y su interés generalizado ha surgido hace relativamente poco, su definición e implementación está cambiando y evolucionando de forma constante. Lo que provoca que existan numerosas representaciones sobre los elementos que forman esta ciencia abierta.

Ilustración 2: *Elementos de la ciencia abierta.*

Los elementos que aparecen en todas las representaciones son el acceso abierto, los datos abiertos, revisión por partes (*open peer review* [1]) y el software libre (Masuzzo & Martens, 2017).

Recapitulando y centrando el tema en el CDM OMOP, ha adoptado la naturaleza de ciencia abierta debido a que uno de sus principales objetivos es generar evidencia médica a gran escala.

Mediante la ciencia abierta; que facilita la utilización y verificación con el código abierto, los estándares abiertos, tanto de la estructura de la base como del vocabulario estandarizado, los datos abiertos (debido a la sensibilidad de los datos, esta parte no es totalmente abierta, se cuenta con conjuntos de datos mapeados por OMOP o datos simulados que sirven de ejemplo) y un discurso abierto a través de la comunidad OHDSI se logra alcanzar este modelo de ciencia.

1.1.2. Estructura de modelo común

A lo largo del proceso en el que un paciente recibe atención médica, sus datos clínicos son recogidos y almacenados en bases de datos dando lugar a un número cada vez mayor de datos sobre pacientes, lo que se denomina como *Big Heath Data*.

Estos datos, son recopilados en distintas bases de datos dependiendo de sus necesidades primarias, lo que implica que los datos para una investigación deben recogerse de diversas fuentes, compararse y contrastarse para comprobar el posible sesgo de captura, lo que implica el análisis de varias bases de datos a la par.

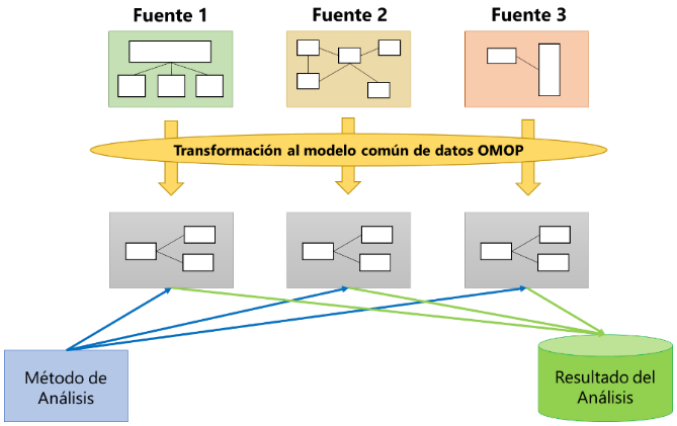


Ilustración 3: *Transformación de bases de distintas fuentes en el modelo OMOP /Basado en: Observational Heath Data Sciences and Informatics (OHDSI).*

[1] Sistema en el que se basan las revistas científicas para decidir si aceptan o rechazan un artículo en el cual el evaluador y el autor conocen mutuamente sus identidades.

Este problema quedaría paliado con un modelo común de datos, ya que al armonizarse los datos en un único arquetipo se garantizaría poder aplicar de forma sistemática los métodos de investigación produciendo así resultados comparables y reproducibles.

El CDM OMOP está diseñado para la recogida de datos observacionales de forma que puedan identificarse poblaciones de pacientes con intervenciones médicas y sus resultados; ser caracterizadas poblaciones con particularidades sociodemográficas similares con el fin de que se puedan predecir estos resultados y estimadas el efecto de las intervenciones.

Para poder alcanzar de forma óptima estos objetivos (Observational Health Data Sciences and Informatics, 2019), este tipo de base de datos cuenta con un diseño en el cual, los datos están organizados con intención de ser analizados, y no para cubrir las necesidades de proveedores. Los datos se encuentran protegidos, ya que cuentan con información sobre las identidades de los pacientes.

La base de datos está estructurada de forma en la que existen una serie de dominios, relacionados a través de un identificador único de paciente. El modelo está centrado en la persona, es decir, que todas las tablas de eventos clínicos están conectadas con la tabla PERSON, exceptuando las tablas de datos del sistema de salud estandarizadas.

Se puede realizar en cualquier base de datos relacional, es decir, no requiere una tecnología específica, y está diseñada de forma en la que el tamaño de los datos fuente puede ser muy diverso (escalabilidad).

EL CDM cuenta con 37 tablas que quedan representadas en la *Figura 4.* En ellas quedan recogidos todos los datos sobre eventos clínicos, vocabulario, metadatos, datos del sistema y economía de la salud y resultados.

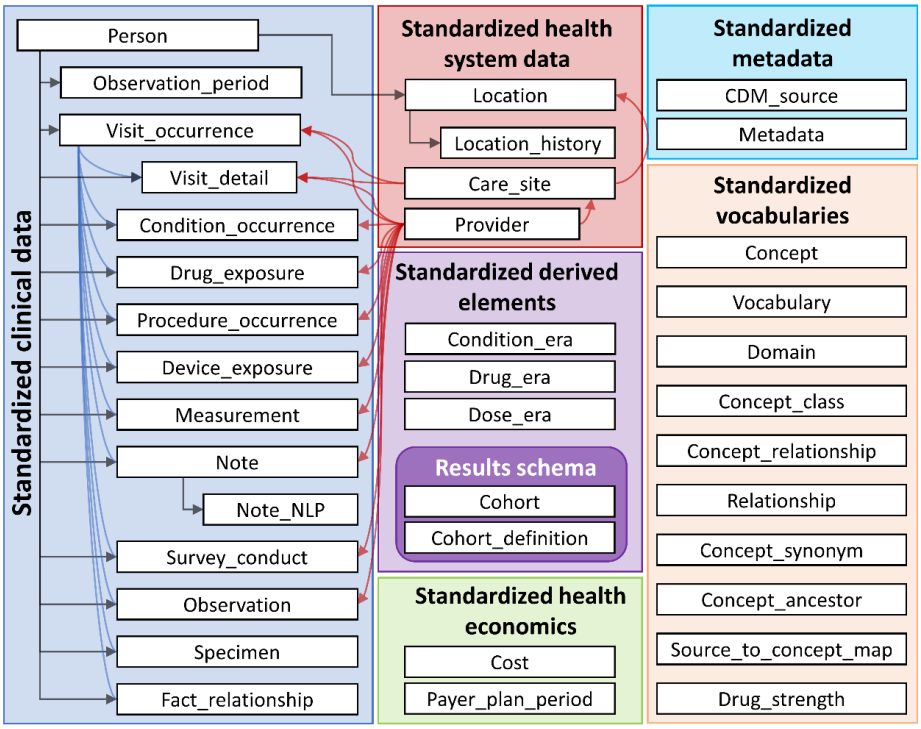


Ilustración 4: *Visión general de las tablas de CDM / The Book of OHDSI*

Dentro de cada una de las tablas existen diferentes campos, los cuales son cubiertos con los datos de cada uno de los pacientes. De forma general, estos campos siguen la siguiente nomenclatura (Observational Health Data Sciences and Informatics, 2019):

- [Evento]_ID: Identificador único para cada registro que establece relaciones entre tablas de eventos
- [Evento]_CONCEPT_ID: En la que se recoge el código del concepto ya estandarizado.
- [Evento]_SOURCE_CONCEPT_ID: En la que se recoge el código del concepto antiguo (se entrará exhaustivamente en este detalle en el punto siguiente *Vocabulario estandarizado*).
- [Evento]_TYPE_CONCEPT_ID: Recoge el código estandarizado del tipo de concepto.
- [Evento]_SOURCE_VALUE: Código o cadena de texto libre que recoge la representación del evento.

De esta forma, los datos de los pacientes quedarían recogidos por fila, siendo cada columna cada campo de la tabla a la que se refieran los datos, y estando conectadas las tablas a través de la columna PERSON_ID. Como se ve reflejado en el ejemplo, los datos son recogidos en forma de vocabulario estandarizado.

1.1.3. Vocabulario estandarizado

Uno de los principales inconvenientes de la unificación universal de bases radica en el propio vocabulario que se adopta en esas bases de datos, ya que cada concepto clínico puede ser recogido con diversas nomenclaturas o en diversos idiomas (por ejemplo; el Paracetamol, se denomina también como Acetaminofén). El CDM OMOP pretende no solo la armonización en un formato estandarizado, sino también un contenido estándar riguroso (Observational Health Data Sciences and Informatics, 2019)

La conceptualización de vocabularios médicos es una práctica que se remonta a la época medieval, y como es lógico, desde ese momento, se ha ampliado cuantiosamente el número de conceptos existentes. Algunos vocabularios están mantenidos por organizaciones mundiales, como puede ser la OMS y cada gobierno local genera versiones adecuadas a su país. Esto provoca que existan distintas minerías de datos sobre vocabularios.

El CDM proporciona no solo una normalización de los vocabularios, sino un depósito común de todos los vocabularios utilizados.

En la imagen de la derecha, aparecen recogidas las organizaciones de la cuales parte el vocabulario estandarizado de OMOP. Como se puede apreciar, la que cuenta mayor número de conceptos es la SNOMED (Standard Nomenclature of Medicine) organización multinacional.

Para poder acceder a estos vocabularios estandarizados, ha sido creada la plataforma ATHENA[2], donde se pueden encontrar tanto el código asociado al concepto como la jerarquía de ese concepto.

A cada concepto se almacena en la tabla CONCEPT. Algunos de los campos existentes en la tabla son los siguientes:

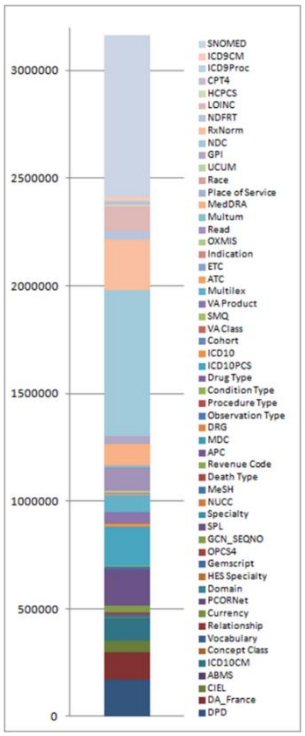


Ilustración 5: *Tabla de referencia del concepto único/ OMOP Common Data Model CDM ExtractTransformLoad ETL Tutorial*

[2] ATHENA: Plataforma online: https://athena.ohdsi.org/

- CONCEPT_ID: Donde se almacena la clave principal, es decir, el código que se va a asociar al concepto en el resto de las tablas.
- CONCEPT_NAME: Se almacena la descripción del concepto
- DOMAIN_ID: Se le asocia un dominio que dirige a que tabla y campo del CDM se registra en evento clínico.
- VOCABULARY_ID: Organización de la que parte el concepto en un origen.
- CONCEPT_CODE: Código dado por la organización recogida en el VOCABULARIO_ID, del que parte el nuevo código.

De forma en la que en el CDM de OMOP quedarían recogidos el nuevo vocabulario como su relación con el antiguo, cuando es el caso, quedando todo almacenado en una misma base a la vez que unificado y normalizado.

En conclusión, para poder aplicar mejores decisiones en el futuro teniendo en cuenta las decisiones tomadas en un pasado es necesario tener mecanismo que facilite la recogida de esta información.

El Modelo Común de Datos de OMOP proporciona una plataforma para poder normalizar y estandarizar estos datos facilitando la obtención de conclusiones posteriores, objetivo que se pretenderá demostrar en el presente trabajo. Para poder lograrlo, se tomarán datos de pacientes de Síndrome Mielodisplásico.

1.2. Síndromes Mielodisplásicos

Los síndromes mielodisplásicos (SMDs) son alteraciones que ocurren cuando las células productoras de sangre de la médula ósea se convierten en células anormales (displásicas), lo que produce que se reduzca el número de uno o varios tipos de células en sangre (glóbulos rojos, glóbulos blancos y plaquetas). Existen varios tipos de SMDs, dependiendo de los tipos de células afectadas y otros factores (American Cancer Society, 2018b).

Estas células displásicas es habitual que mueran más temprano que las células normales y que el organismo destruya células sanguíneas no defectuosas, produciendo que la persona tenga escasez de células normales. La anemia (escasez de glóbulos rojos) es el hallazgo más común producido por esta enfermedad.

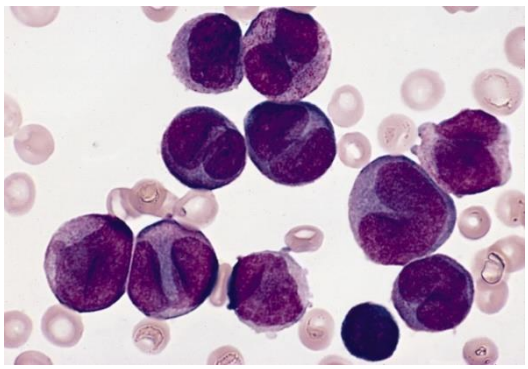


Ilustración 6: *Frotis sanguíneo de un varón de 47 años con leucemia mieloide aguda / The Armed Forces Institute of Pathology-AFIP (WIKIMEDIA)*

Una parte de los pacientes con SMDs llega a progresar a leucemia mieloide aguda (AML), un tipo de cáncer de crecimiento rápido. Debido a esto, en el pasado, el SMD era conocido como preleucemia, y como la mayoría de los pacientes no derivan en AML, era considerado una enfermedad de bajo potencial maligno, cosa que finalmente se reconsideró debido al conocimiento más profundo de esta patología. En la actualidad el SMD es considerado un tipo de cáncer (American Cancer Society, 2018b).

Pese a que son considerados de esta forma, no son un tipo de cáncer al uso. Los canceres sólidos, como pueden ser los de pulmón o de mama, se caracterizan por tres propiedades: se expanden de forma incontrolada, no respetan los límites del propio órgano y pueden causar metástasis en órganos

lejanos de la infección primaria. En cambio, los síndromes mieoldisplasicos no hacen esto; es decir, lo que se teme no es la expansión del cáncer, sino el fallo de los órganos; la sangre no se infiltra en otros órganos y no causan metástasis (Asociacion Española de Afectados por Linfoma, n.d.).

Entre las principales causas conocidas que pueden dañar el ADN y producir esta patología se encuentra la radiación, tanto radiación por rayos X o radioactividad, como por la propia radiación que provoca la Tierra. Otro causante de estas lesiones en el ADN es el benceno (Asociación Española de Afectados por Linfoma, n.d.).

Los SMDs son una enfermedad de difícil respuesta clínica debido en parte a su complejidad y a las escasas posibilidades de terapia específica, ya que, suele presentarse en individuos de edad avanzada (entorno a los 70 años) y con enfermedades adicionales (Jiménez, 2016).

1.2.1. Clasificación FAB y OMS

Inicialmente, la clasificación de los SMDs fue establecida por un grupo Franco-Américo-Británico (FAB) en 1982, manteniéndose en vigor hasta la revisión realizada en 2002 por la OMS (Sanz G., 2012). La clasificación FAB está basada en criterios morfológicos de cinco subgrupos: anemia refractaria (AR), anemia refractaria con sideroblastos en anillo (ARSA), anemia refractaria con exceso de blastos (AREB), anemia refractaria con exceso de blastos en transformación (AREB-t) y leucemia mielomonocítica crónica (LMMC).

Esta clasificación fue base en la revisión de 2002 para la nueva clasificación OMS (WHO) que combina morfología, citoquímica y citogenética (Jiménez, 2016)., clasificación que se ha ido actualizando hasta la más reciente en 2016. Esta clasificación cuenta con los siguientes subgrupos: SMD con displasia multilinaje (MDS-MLD), SMD con displasia unilinaje (MDS-SLD), SMD con sideroblastos en anillo (MDS-RS), SMD con exceso de blastos (MDS-EB), SMD con del(5q) aislada y SMD, no clasificable (MDS-U).

1.2.2. Pronósticos IPSS y IPSS-R

Debido a la diversidad de enfermedades, los SMD presentan gran variedad de pronósticos. Esto sumado a la edad avanzada de los pacientes y a la presencia de comorbilidades dificulta significativamente la elección del tratamiento para un paciente en concreto, siendo clave la adaptabilidad de los tratamientos para evitar riesgos. A consecuencia de eso se ha realizado un estudio cooperativo internacional cuyo resultado fue definido en el Sistema Pronóstico Internacional (IPSS).

Este sistema determina que las variables que tienen un mayor impacto pronóstico son: el porcentaje de blastos en Médula Ósea, el análisis citogenético y el número de citopenias periféricas. De esta forma, quedan discriminados 4 grupos de riesgo para la supervivencia y la supervivencia libre de evolución leucémica: Bajo, Intermedio 1 (Int-1), Intermedio 2 (Int-2) y Alto (Belli & Larripa, 2007).

El IPSS ha sido recientemente revisado por el Grupo de Trabajo Internacional para la determinación del pronóstico en los SMD (IWG-PM). En este nuevo IPSS-R (Revisado) se proponen 5 grupos (Muy Bajo, Bajo, Intermedio, Alto y Muy Alto) y se recuentan los neutrófilos y el número de plaquetas (Belli et al., 2014).

Para poder diagnosticar y pronosticar en función de las clasificaciones anteriores se deben realizar diversos estudios, como el de Médula Ósea, Sangre Periférica, Cariotipo y FISH[3], estudios realizados en los pacientes del presente análisis.

[3] FISH: Hibridación *in situ* con fluorescencia

2. Objetivos

El objetivo principal del presente trabajo es observar la potencia de implementar el modelo común de datos OMOP en la base de datos proporcionada por el IBSAL sobre pacientes de SMD (Síndrome mielodisplásico). Se tiene este objetivo con el fin de concluir en un resultado positivo o negativo acerca del uso de este tipo de bases de datos.

Como segundo objetivo se tiene el análisis de los datos anteriores mediante técnicas estadísticas de supervivencia. Pretendiendo, principalmente, corroborar con los datos las conclusiones que aparecen en la literatura sobre los SMDs.

De esta forma, se presentarán todas las fases llevadas a cabo por un estadístico al realizar un estudio; preparación de los datos, codificación, análisis y extracción de conclusiones.

3. Materiales

3.1. Dataset

El Instituto de Investigación de Salamanca (IBSAL) ha proporcionado para este trabajo una base de datos de Síndrome Mielodisplásico de 2882 pacientes de hospitales de toda España. Algunos de estos individuos cuentan con varias muestras recogidas a lo largo del tiempo, constituyendo de esta forma una base formada por 3241 resultados.

Las variables que forman este estudio pueden ser explicadas en dos secciones; una de variables las cuales corresponden a datos sobre los pacientes, y otra mayoría, que corresponden a datos sobre las muestras de estos pacientes.

En los datos sobre los pacientes contamos con información tanto del individuo como del equipo médico por el cual ha sido tratado; datos que son confidenciales a la par que irrelevantes para este estudio y por tanto se prescindirá de ellos (*Nombre y Apellidos, Iniciales, Responsable médico, Afiliación* y *e-mail*).

Los hospitales asignan diversos identificadores a sus pacientes, estos también quedan recogidos en la base. Ocurre lo mismo con las muestras de estos individuos, a las que se les designa un numero de cultivo, número de congelación y un id de laboratorio. Todos estos identificadores, tanto los de la muestra como los de los pacientes, son necesarios para poder general y unir los datos que corresponden a cada individuo, así como para poder trazar las uniones entre las diferentes tablas que van a formar la base de datos OMOP.

Entre los datos que forman la base se encuentran los necesarios para realizar las clasificaciones descritas en el apartado 1.2.1 (las clasificaciones FAB y OMS), y pronostico IPSS y IPSS-R, descrito en el apartado 1.2.2. Como se mencionó con anterioridad, entre ellos se encuentran los estudios de **Médula Ósea, Sangre Periférica, Cariotipo, FISH** y **Secuenciación.**

En el estudio de **Médula Ósea** se recogen los datos de *Fecha del estudio*, *Porcentaje de blastos* (formas muy inmaduras de células sanguíneas), número de *Bastones de Auer*[4], porcentaje de glóbulos rojos primitivos en *Sideroblastos en anillo* (células que contienen anillos de depósitos de hierro alrededor del núcleo), numero de células precursoras que muestran *Displasia* y el *Tipo de displasia* que presentan si puede distinguirse, valoración de rasgos de mielodisplasia mediante el *Porcentaje de Diseritropoyesis*, *Disgranulopoyesis* y *Distrombopoyesis* y la presencia o no de *Fibrosis*.

En el de **Sangre Periférica** se incluyen los datos sobre la *Fecha del estudio, Porcentaje de blastos*, la cantidad de *Hemoglobina*, *Plaquetas*, *Neurofitos*, *Leucocitos*, *Monocitos*, *Linfocitos*, *Eosinófitos*, *Basofitos*, *Incremento de la secreción de eritropoyetina* (*EPO*) y de *Ferritina* presente en la muestra, el *Riesgo citogenético IPSS y IPSS-R*, la presencia de citopenias: *Anemia, Trombocitopenia y Neutropenia* y el recuento de estas.

En el estudio de **Cariotipo** tenemos la *Fecha del estudio*, *Resultado del estudio* y *la Formula ISCN (Internacional System for Human Cytogenetics Nomenclature) 2013 & 2016*; y en el de **FISH** la

[4] Los bastones de Auer (o cuerpos de Auer) son cuerpos de inclusión citoplasmáticos cristalinos grandes que a veces se observan en células blásticas mieloides durante la leucemia mieloide aguda, la leucemia promielocítica aguda y los síndromes mielodisplásicos de alto grado y los trastornos mieloproliferativos (Wikipedia contributors, 2021a).

Fecha del estudio, el *Resultado del estudio* y el *Tipo de sonda*[5], *Resultado de la sonda* y el *Porcentaje de alteración* de estas. Para el estudio de Secuenciación se recogen datos acerca del *Material de secuenciación* empleado, el *Identificador de la muestra en el laboratorio* y el *tipo de panel Illumina*.

También son realizados otros estudios en los cuales se recogen datos que se emplean en el diagnóstico y evolución de la enfermedad. Entre ellos se encuentra el **estudio aCGH** (Array Comparative Genomic Hybridization) que permite detectar cambios en el número de copias, tanto deleciones como duplicaciones, en el genoma de forma rápida y con alta resolución (Repáraz & Torreira, 2018). De este estudio tenemos tanto el resultado de este como su fórmula (hg19 y hg38). Se realiza también el **estudio de expresión Affymetrix con Microarrays** de los que se obtienen dos plataformas de expresión, el **estudio de ARN secuencial**, **estudio Sanger**[6] con los resultados en los genes SF3B1 y SF3F2, **estudio Roche, estudio metilación** y **estudio WES**[7] (Whole Exome Sequencing).

Acompañado de los estudios, se recogen los proyectos en los que se encuentra cada una de las muestras de los pacientes que forman parte del estudio.

Contamos con datos sobre el tipo de diagnóstico del paciente, si es secundario de que tipo es, la fecha del diagnóstico inicial del paciente, sobre la progresión de la enfermedad tanto de SMD como de LAM y la fecha del último control del paciente y el estado, si está vivo o muerto.

Los datos sobre los pacientes y las muestras se encuentran recogidos en distintos Excel; todos estos datos deben ser tratados y recogidos en la base de datos OMOP y para ello es necesario que se realice un proceso de ETL (Extract, Transform and Load).

3.2. Librerías de R

Para el procedimiento de ETL se emplea el programa de R, *RStudio*, siendo necesaria la previa instalación de las siguientes librerías para que esto pueda producirse:

- ***Lubridate:*** Contiene funciones necesarias para el tratamiento de datos del tipo fecha de forma sencilla (Spinu, Grolemund, & Wickham, 2021).
- ***Tidyverse:*** Conjunto de paquetes formado por *dyplr, ggplot, tibble, readr, tidyr* y *purr* de gran utilidad no solo para la carga y el tratamiento de los datos, sino también para la elaboración de graficas (Wickham et al., 2021).
- ***Readxl:*** Librería necesaria para cargar en R datos procedentes de Excel (Wickham & Bryan, 2019).
- ***Date.table:*** Permite la unión de datos de gran tamaño (Dowle & Srinivasan, 2021).

[5] FISH usa segmentos de una única hebra de ADN que son tintados, con una sustancia fluorescente que puede ligarse a un cromosoma específico; estos segmentos de ADN son llamados sondas (Wikipedia contributors, 2021b).

[6] Consiste en la sinterización de forma secuencial de una hebra de ADN complementaria a una hebra de cadena simple (que se utiliza como molde), en presencia de ADN polimerasa, los cuatro 2'-deoxinucleótidos que componen la secuencia del ADN (dATP, dGTP, dCTP y dTTP) y cuatro dideoxinucleótidos (ddATP, ddGTP, ddCTP y ddTTP) (Garrigues, 2017)

[7] *Secuenciación del exoma:* técnica que consiste en la determinación de secuencia de la parte codificante del DNA genómico (Grupo español de síndromes mielodisplásicos, 2017)

- ***RMySQL***: Utilizada para establecer las conexiones entre RStudio y MySQL (Ooms, James, DebRoy, Horner, & Wickham, 2020). [MySQL es una plataforma que será descrita en el apartado 4.1.3. en la que se expone el proceso de carga y conexión con la base de datos].

- ***Rjson:*** Convierte objetos R en objetos de tipo JSON[8] y viceversa (Couture-Beil, 2018).

Y para el posterior análisis de supervivencia también se va a utilizar como herramienta RStudio, para ello se instalan las siguientes librerías:

- ***Survival:*** Contiene las funciones necesarias para realizar los análisis de supervivencia: curvas Kaplan-Meier, modelos paramédicos y modelos de Cox.

- ***Survminer:*** Permite dibujar los gráficos de supervivencia**.**

[8] JSON: JavaScript Object Notation, se trata de un lenguaje sencillo utilizado para el intercambio de datos entre sistemas. En este trabajo es usado para poder conectar la base de datos OMOP de MySQL con R, ya que en este lenguaje son guardadas las claves para establecer esas conexiones.

4. Metodología

Como se ha introducido en el apartado anterior, para que los datos comiencen a formar parte de la base de datos OMOP, así como, se pueda realizar el posterior análisis estadístico de supervivencia, es necesario que pasen por un proceso de ETL (Extract, Transform and Load).

En este bloque temático se describirá este procedimiento, seguido de un marco estadístico sobre las técnicas de análisis necesarias para obtener los resultados del estudio.

4.1. ETL (Extract – Transform – Load)

ETL es un tipo de integración de datos empleada para combinar datos de diversas fuentes. Consta de tres fases; la extracción de los datos de origen, la transformación de estos en un formato que pueda ser analizado, y la carga o almacenamiento en una base de datos.

En la primera fase se extraen los datos desde los sistemas de origen, se revisan para verificar que estos cumplen la estructura esperada y una vez hecho esto, se pasa a la siguiente fase de transformación. Es importante que, en este primer paso, el impacto sobre los datos sea el mínimo posible. En la segunda fase, se realza una depuración de los datos a la par que una codificación de los mismos, en la cual, la información inicial es tratada para convertirse en el modelo estandarizado que podrá ser finamente cargado en el paso de Load, paso final de la ETL (Shang, Adams, & Hassan, 2012).

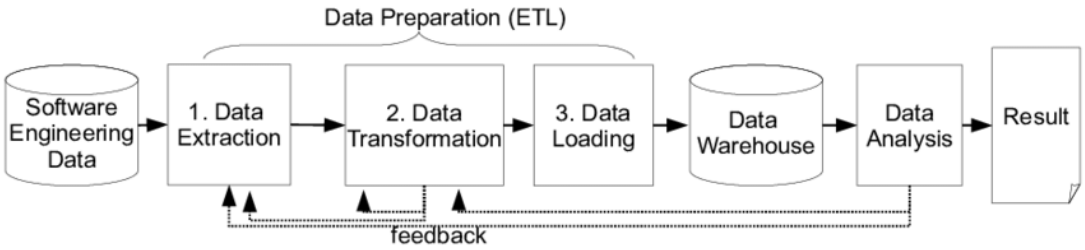


Ilustración 7: *Pipeline del proceso desde la obtención de los datos hasta los resultados, recalcando la fase de procesado de los datos ETL/ Journal of Systems and Software - JSS*

En el presente estudio, todo el proceso de extracción, transformación y carga será realizado a través del programa de R, RStudio. En este apartado, se intenta describir lo más detalladamente el proceso ejecutado para tratar los datos.

4.1.1. Extract

La información, tanto de los pacientes como de las muestras, es recogida en diferentes Excel dependiendo de los conceptos que se traten (por ejemplo, uno de ellos cuenta con las clasificaciones FAB y OMS u otro con los datos sobre los proyectos).

En este caso, se parte de once documentos; los cuales, cuentan con columnas o variables comunes, como pueden ser los identificadores del paciente que permiten la unión de los Excel, y con las variables con la información específica que recogería cada uno de ellos. Estas bases son importadas en forma de dataframes.

Como se menciona en la parte descriptiva del proceso de ETL, en la fase de Extract es necesario comprobar y revisar que los datos cumplen con la estructura esperada; por ejemplo, que los data-frames generados tengan el mismo número de filas, que corresponde al número de muestras del estudio, o que los nombres de las variables comunes sean iguales para que la futura unión de los datos en un único dataframe sea posible. Para poder lograr esto, se ha realizado un proceso de depurado de los datos, eliminando filas vacías, y de renombrado.

4.1.2. Transform

Una vez terminado el proceso de extracción de los datos, se pasa a la fase más importante de la ETL, el tratamiento los datos. En esta parte lo que se pretende es, unir estos once Excel proporcionados por el IBSAL, formando así un único dataframe del cual se puedan manipular directamente las variables, y así realizar las trasformaciones descritas en este apartado.

Tanto como para poder aunar los dataframes como para tener un identificador único para cada paciente y para cada muestra, es necesario crear dos nuevas columnas; una que será el **PERSON_ID** y otra el **SPECIMEN_ID**. Para crear estos identificadores empleamos los ids que aparecen en los datos. Este paso es necesario solo porque todos los identificadores proporcionados tienen pacientes o muestras en las que existen datos faltantes.

Al realizar este proceso, puede que existan muestras y pacientes que tengan el mismo id, esto no produciría errores, ya que lo importante para que la base de datos sea estable es que los identificadores no se repitan en la columna de la cual se toma referencia. Una vez creados, se procedería a la unión de los dataframes mediante estas dos columnas, consiguiendo así un gran dataframe del cual se irán extrayendo y clasificando las variables en las distintas tablas de OMOP, proceso que se describirá a continuación.

Para poder seguir el modelo común de datos OMOP es necesario realizar un esquema previo en el cual se van a ir clasificando las diferentes variables dentro de las distintas tablas que forman la estructura de la base de datos.

Comenzamos por la tabla **PERSON**, cuya clave primaria es PERSON_ID. Las claves primarias son identificadores únicos de cada fila de la tabla, necesarios en las bases de datos relacionales.

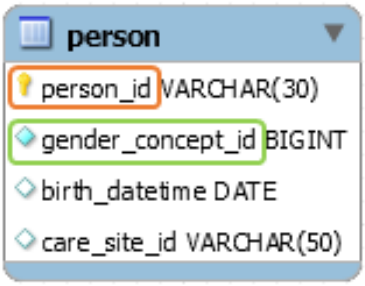
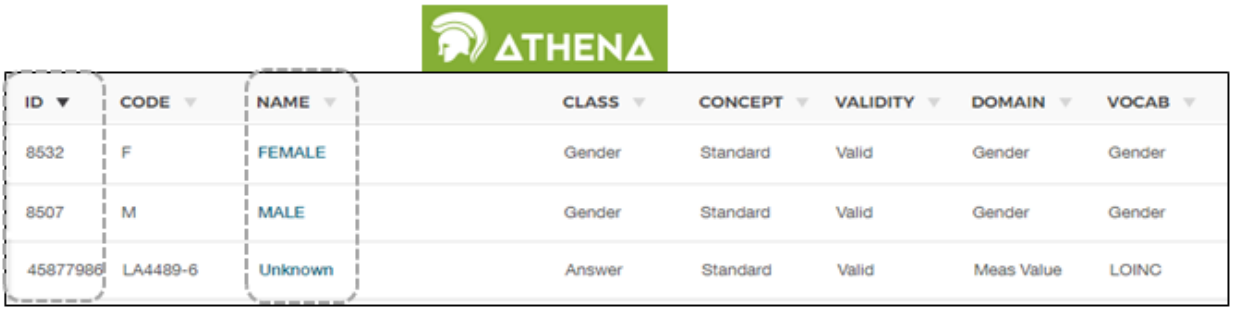


Ilustración 8: *Representación gráfica del campo PERSON del CDM del estudio*

Una de las columnas que forma esta tabla es GENDER_CONCEP_ID (donde se almacena el género del paciente), en la cual se comienza a usar el vocabulario estandarizado de OMOP, por ende, se ilustrará el procedimiento con este caso, siendo este desarrollo, extrapolable para el resto de las variables que necesiten ser estandarizadas.

Tal como se ha descrito en el punto 1.1.3, buscamos la información asociada al concepto en ATHENA:



ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
8532	F	FEMALE	Gender	Standard	Valid	Gender	Gender
8507	M	MALE	Gender	Standard	Valid	Gender	Gender
45877986	LA4489-6	Unknown	Answer	Standard	Valid	Meas Value	LOINC

Ilustración 9: *Screenshot de la obtención de códigos a través de la plataforma ATHENA*

Teniendo estas codificaciones, podemos sustituir los valores e introducirlos de esta forma en la columna GENDER_CONCEP_ID de la tabla.

Continuando con la descripción de la tabla PERSON, también se incluye la fecha de nacimiento. Las fechas en el CDM son recogidas tal cual aparecen, es decir, no se les asocia ningún código.

Por último, se recogen los hospitales por los que pasa el paciente en la columna de CARE_SITE. Como un individuo puede empezar el estudio en un centro y posteriormente ser trasladado a otro, los hospitales quedan recogidos por orden de más antiguo a más actual separados por comas. Debido a que los nombres de los hospitales no se encuentran codificados en la plataforma ATHENA, es preciso generarlos. Para ello, se decide que todos ellos empiecen por "3200000***" continuado con los números que se le asignan a cada hospital.

Esto no solo ocurre con los nombres de los hospitales, sino que existen algunos conceptos muy específicos que no cuentan con códigos tampoco, por lo que se crea un diccionario local que reúne todos estos casos.

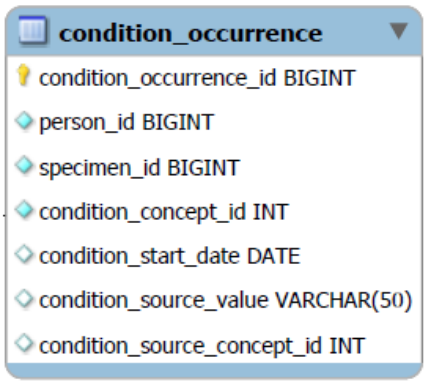
Se adjunta un ejemplo de dos pacientes incluidos en esta tabla:

Tabla 1: *Tabla ejemplo de dos casos almacenados en el campo PERSON*

PERSON

PERSON_ID	*GENDER_CONCEP_ID*	*BIRTH_DATETIME*	*CARE_SITE*
1234	8532	01/03/2000	3200000003
5678	8507	01/06/1990	3200000001,3200000005

En la tabla de **CONDITION_OCURRENCE** se recogen los diagnósticos, signos o síntomas de una condición observada por un experto o reportada por el paciente (Observational Health Data Sciences and Informatics, 2019c). A consecuencia de esto, una gran parte de las variables de este estudio serán recogidas en ella.



La tabla CONDITION_OCURRENCE; esta cuenta con diversos campos, pero solo serán utilizados CONDITION_SOURCE_CONCEPT_ID, donde se recogen los códigos de los valores de las variables, CONDITION_CONCEPT_ID, donde se almacenan los códigos de esas variables, CONDITION_START_DATE, la fecha en la que ocurre el evento y CONDITION_SOURCE_VALUE, para los valores de las variables que no es necesario que sean codificadas.

Ilustración 10: *Representación gráfica del campo CONDITION_OCURRENCE del CDM del estudio*

Para facilitar la explicación se utilizará el ejemplo de las variables *Subtipo Mieloide*, en la que se recoge el tipo de leucemia mieloide que presentan las muestras, y la *Clasificación FAB* que se le asigna a cada una.

Para ello, buscamos en ATHENA los códigos correspondientes:

Tabla 2: *Representación las codificaciones extraídas de ATHENA para dos variables ejemplo de la tabla CONDITION OCURRENCE*

NOMBRE	DESCRIPCIÓN	CÓDIGO	VALORES	ESTÁNDAR
SUBTIPO MIELOIDE	-	4212320	-	SNOMED
	Síndrome mielodisplásico	138994	MDS	SNOMED
	No Síndrome mielodisplásico	3100000008	NOT MDS	-
	Neoplasias mieloproliferativas	4019110	MPN	SNOMED
	Mielodisplásico/Neoplasias mieloproliferativas	36311295	MDS/MPN	LOINC
	Leucemia linfoblástica aguda	36311304	AML	LOINC
	Leucemia linfoblástica aguda secundaria	3100000009	sec AML	-
	Mielofibrosis	133169	MF	SNOMED
	Sano	6308878	HEALTHY CONTROL	LOINC
	Desconocido	45877986	UNKNOWN	LOINC
CLASIFICACIÓN FAB	-	3100000025	-	-
	Anemia refractaria	3476381	RA	NEBRASKA LEXICON
	Anemia refractaria con sideroblastos en anillo	4003186	RARS	SNOMED
	Anemia refractaria con exceso de blastos	3428970	RAEB	NEBRASKA LEXICON
	Anemia refractaria con exceso de blastos en trasformación	3472647	RAEB-t	NEBRASKA LEXICON
	Leucemia mielomonocítica crónica	3289053	CMML	NEBRASKA LEXICON

En la columna de CONDITION_CONCEPT_ID van a ser introducidos los códigos de todos los campos que existen en esta tabla. Siendo particularizado al ejemplo, los códigos que aparecen son los dos que corresponden a *Subtipo mieloide* y a la *clasificación FAB*. Asociado a cada uno de ellos, en la columna CONDITION_SOURCE_CONCEPT_ID aparece el código que le responde a la variable, es decir, el resultado de cada una.

Al recoger de esta forma los datos, tanto el PERSON_ID como el SPECIMEN_ID se repetirían. En concreto, el SPECIMEN_ID, clave foránea, se repetirá tantas veces como variables sean recogidas en la tabla. Como se ha mencionado con anterioridad, en las bases de datos relacionales es necesaria la existencia de una clave primaria que identifique a cada fila de forma única, por lo tanto, se precisa de la utilización, tanto en esta tabla como en las que siguen este esquema, de un ID que se genere automáticamente y distinga las filas, en esta tabla se le denomina CONDITION_OCURRENCE_ID.

Han sido seleccionados dos pacientes con sus respectivas muestras para ejemplificar como quedaría formada la tabla:

Tabla 3: *Ejemplificación de dos casos de la tabla CONDITION_OCURRENCE*

CONDITION_ OCURRENCE_ID	PERSON_ID	SPECIMEN_ID	CONDITION_ CONCEPT_ID	CONDITION_SOURCE_ CONCEPT_ID
22709	10068	27692	4212320	138994
22710	10068	27692	3100000025	3428970
22711	10068	29275	4212320	45877986
22712	10068	29275	3100000025	45877986
22713	39779	100514	4212320	138994
22714	39779	100514	3100000025	45877986

A diferencia de lo que ocurre en la tabla PERSON, en la que se clasifican las variables de forma más intuitiva; correspondiéndole a cada variable, una columna donde almacenar los códigos, la clave del modelo común de datos OMOP reside en este tipo de tabla. Una tabla que contiene el id de las muestras de los individuos repetido tantas veces como variables existan, quedando una tabla resumida en número de columnas, pero extensa en número de filas.

El modelo es establecido de esta forma debido a la gran ventaja asociada a esta estructura, en la que, si necesita ser ampliada a nuevos campos, no sería preciso modificar su forma a nivel columnas, sino solo habría que añadir nuevas filas con los códigos de esos campos a incluir.

Se ha mencionado que la clave foránea de la tabla anterior es SPECIMEN_ID, conectada con la tabla **SPECIMEN**.

En esta tabla, fue necesario añadir nuevas columnas a las ya existentes en el modelo OMOP para poder recoger toda nuestra información.

La tabla recoge todos los identificadores asociados a las muestras (número de cultivo, congelación, …), cada uno en una columna de la tabla. Datos que no necesitan ser codificados, ya que se tratan de códigos en sí.

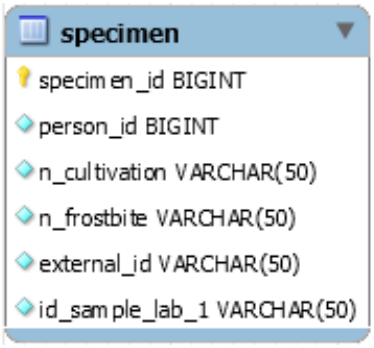


Ilustración 11: *Representación gráfica del campo SPECIMEN del CDM del estudio*

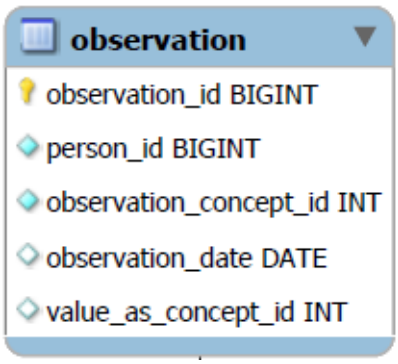


Ilustración 12: *Representación gráfica del campo OBSERVATION del CDM del estudio*

Siguiendo el mismo proceso que la tabla CONDITION_OCURRENCE se forma la tabla **OBSERVATION**, en la cual únicamente se incluye, en este caso, el Status del paciente, es decir, si se encuentra en el momento del estudio vivo o ha fallecido. La tabla está formada por las columnas VALUE_AS_CONCEPT_ID, que recoge los valores codificados de Status, OBSERVATION_CONCEPT_ID, que almacena únicamente el código de esta variable, ya que se trata de la única y la fecha del ultimo control en OBSERVATION_DATE. Como se ve reflejado en la imagen de la tabla, su clave primaria es OBSERVATION_ID

Esta tabla es de suma importancia para los análisis estadísticos de supervivencia posteriores, ya que recoge el estado del paciente en el momento del estudio, crucial para ellos.

La tabla **MEASUREMENT** recoge las medidas de los tratamientos que son llevados a cabo.

Se trata, junto a la tabla de OBSERVATION, de la que contiene mayor información, ya que la mayoría de los datos de este estudio son medidas.

Los campos de esta tabla que contendrán datos serán los siguientes: MEASUREMENT_CONCEPT_ID, en el que se guardan los códigos de las variables de esta tabla, OPERATOR_CONCEPT_ID, donde se recogen signos de mayor menos e igual, necesarios para alguna de las variables de esta tabla, VALUE_AS_NUMBER, que almacena el valor en sí, MEASUREMENT_DATE, la fecha en la que se recoge la medida y VALUE_AS_CONCEPT_ID, donde se guardan variables de medición que no se recogen en forma de número, sino de concepto codificado.

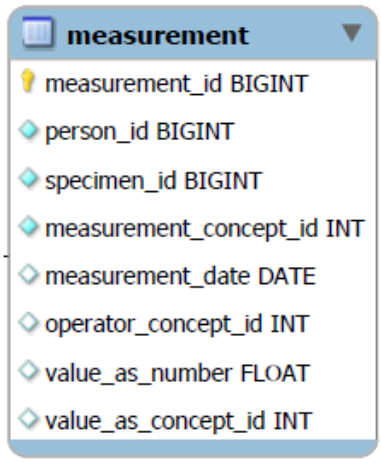


Ilustración 13: *Representación gráfica del campo MEASUREMENT del CDM del estudio*

Las medidas, en este estudio, son utilizadas para poder darle una clasificación y un tipo a cada paciente; es decir, para darle valores a variables recogidas en OBSERVATION.

Una de las ventajas que presenta el modelo OMOP es la posibilidad de añadir tablas conectadas con formatos similares. Actualmente, este modelo está preparado para la recogida de datos clínicos sobre pacientes una vez estos llegan al centro médico en forma de consulta y se le receta un determinado fármaco.

El presente estudio analiza a detalle muestras de pacientes de SMD, recogiendo no solo datos de consulta sino análisis, por lo que es necesario crear determinadas tablas que recojan los estudios a los que son sometidos los pacientes y el proyecto al cual pertenecen estos.

La tabla que recoge los estudios, **STUDIES** no solo cuenta con el tipo de estudio realizado y su resultado, sino también con las fechas de esos estudios, los materiales empleados e información codificada útil para ellos. Esta tabla tiene como clave el SPECIMEN_ID

Ocurre lo mismo con la tabla de **PROJECTS**, tabla pequeña en la que solamente se recoge el proyecto al que pertenece cada muestra, por lo tanto, esta será su clave primaria, SPECIMEN_ID.

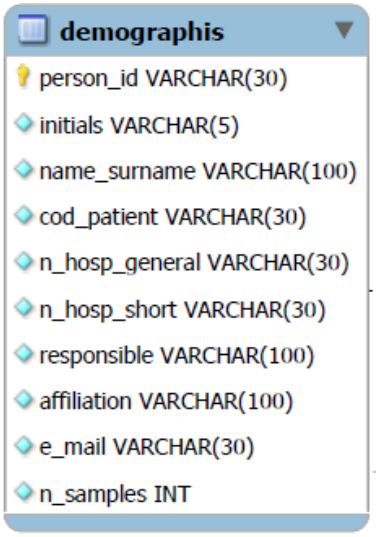


Ilustración 14: *Representación gráfica del campo DEMOGRAPHIS del CDM del estudio*

Seguimos por la tabla **DEMOGRAPHIS**, tabla que, como las anteriores, es creada desde cero siguiendo el modelo. Es necesario crearla debido a que en la estructura base de OMOP no existe ninguna tabla que recoja la información demográfica de los pacientes debido a la protección de datos.

En esta parte del estudio, como lo que se pretende es observar la capacidad de almacenamiento de datos de este modelo, si se utilizara a modo de ejemplo, a pesar de que, como se indica en la primera parte de este apartado (3.1.), estos datos no serán empleados en el análisis estadístico posterior.

La clave primaria para esta tabla será el PERSON_ID.

La tabla **NOTE** recoge anotaciones realizadas sobre algunos aspectos o eventos que le sucede al paciente.

Cuenta con los campos: NOTE_TEXT, donde se recoge el texto libre; NOTE_EVENT_ID, donde aparece reflejado el nombre de la tabla a la que corresponde la nota, por ejemplo, la variable *Causa de Muerte* va asociada a la tabla anterior, OBSERVATION, por lo tanto, se recogería este nombre sin codificar; y NOTE_EVENT_FIELD_CONCEPT_ID donde se almacena el código OMOP que les corresponde a las notas que se recogen.

Como las notas informan sobre datos del paciente la clave de esta tabla es PERSON_ID.

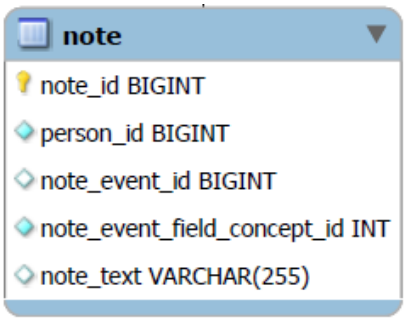


Ilustración 15: *Representación gráfica del campo NOTE del CDM del estudio*

Una vez almacenados los datos del estudio en los distintos campos del Modelo se puede pasar al último paso del proceso, en el cual se comienzan a cargar los datos en la base.

4.1.3. Load

Para el proceso de carga de los datos es necesaria la utilización de MySQL, un gestor de bases de datos del tipo relacional. Las bases de datos relacionales son bases de datos en las que los datos aparecen almacenados en registros organizados en campos siendo relacionadas a través de alguna variable, es decir, el tipo de base de datos del CMD que se ha sido descrito en el aparatado anterior.

El primer paso que se debe realizar para la carga de los datos es la creación la estructura OMOP en el propio MySQL. OMOP proporciona un script en el mediante la función CREATE_TABLE del leguaje de SQL se van creando todas las tablas posibles del Modelo. En el caso del estudio, solo serán cargadas las tablas necesarias para la recogida de datos, es decir, las tablas que se han descrito en el aparatado *Transform.* Como se ha mencionado en ese apartado también, ha sido necesario crear de forma manual algunas tablas, así como campos dentro de ellas.

Una vez está creado el "esqueleto" de la base de datos, es necesario un proceso de inicialización y conexión de RStudio con MySQL. En este proceso se emplea el lenguaje *json* para el almacenamiento de las claves necesarias para poder acceder la estructura de MySQL. Y mediante la función *dbConnect* realizaría este proceso.

Pasamos a la parte real de carga de los datos, en la que es crucial el comando *dbGetQuery* que permite escribir consultas y recuperar datos en lenguaje SQL.

A través de este comando, se crea una función que, introduciendo el dataframe donde se encuentran los datos codificados que corresponde a la tabla que queremos rellenar, el nombre de esa tabla y la o las claves que tiene se cargan los datos en el campo de MySQL.

Esta función creada, **defaud_load** se encarga no solo de adaptar e insertar los datos, sino que también los actualiza y genera un *report* con las diferencias de lo que se cambia en la base de datos.

```
default_load <- function(df, table_name, key, modify=TRUE, con){
    db_ <- dbGetQuery(con,  sprintf("select * from %s", table_name))
# QA checking
    if (modify)
      QA_checking(table_name, db_, key)}
```

Este proceso de carga es necesario realizarlo con cada una de las tablas que queramos rellenar. Para ejemplificar el código usaremos el utilizado para la tabla PERSON:

```
default_load(outPerson, "person", "person_id", modify= F, con)
```

De esta forma, quedaría finalizado el proceso de ETL, quedando los datos guardados siguiendo el Modelo Común de Datos OMOP en la plataforma MySQL.

4.2. Marco Estadístico Teórico

Una vez creada la base de datos estandarizada, en la cual se recogen las características médicas de los pacientes y de sus muestras, se puede proceder al análisis estadístico de estos datos. Para poder realizarlo, ha sido necesario un previo estudio de los métodos de análisis de supervivencia que se van a emplear. En esta sección del documento se describirán las técnicas para la estimación de la supervivencia, así como se definirán algunos conceptos básicos. Para ello se tiene n pacientes u observaciones y s eventos, sabiendo que existen $n - s$ censuras. El número de variables será p y cada una de ellas tendrá k categorías.

4.2.1. Conceptos básicos en análisis de supervivencia

Para para poder profundizar en el análisis de supervivencia y sus métodos es necesario describir algunos conceptos básicos como el *tiempo de supervivencia o tiempo de fallo* y la *censura.*

Cuando se realiza un estudio se marca un evento puntual, el *tiempo de supervivencia* se puede definir como el tiempo que transcurre desde la entrada de un individuo al estudio hasta que ocurre este evento, abandona el estudio o lo finaliza.

En los análisis de supervivencia la situación más favorable es poder observar de forma completa el tiempo de aparición del suceso de interés en todos los individuos, pero esta situación en la que los datos no están censurados no suele ser común. Los datos están *censurados* cuando, a lo largo del tiempo del estudio, se pierde la pista del paciente, es decir, no se cuenta con más datos del paciente. Existen distintos tipos de censura:

- **Censura tipo I:** Se trata de individuos que no han presentado el evento que se desea observar en el tiempo fijado del estudio.
- **Censura tipo II:** Ocurre cuando el investigador decide prolongar el estudio hasta que se produzcan un determinado número de fallos de las *n* posibles, siendo las que siguen sin presentar el evento las observaciones censuradas. Tanto en este tipo de censura como en el caso anterior, están controladas por el investigador, ya que dependen del tiempo de estudio que se ha marcado.
- **Censura aleatoria:** Este tipo de censura se da debido a que los individuos abandonan el estudio o debido a otra causa no relacionada con el evento de interés, por lo tanto, se da sin control, de forma aleatoria.

4.2.2. Introducción a los modelos de supervivencia

Una vez establecidos los aspectos básicos sobre supervivencia se puede proceder a la explicación teórica de los conceptos relativos al análisis.

4.2.2.1. Función de supervivencia

La función de supervivencia $S(t)$ puede ser definida como la probabilidad de que un individuo sobreviva por lo menos hasta un tiempo t:

$$S(t) = P(T > t) \tag{4.1}$$

Sea el tiempo de supervivencia T una variable continua no negativa, T >=0, con F(t) como función de distribución y f(t) función de densidad. La función de supervivencia S(t) es:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f(u)du \qquad (4.2)$$

Si T es discreta con función de probabilidad $f(t_j) = P(T = t_j),\ j = 1,2, \dots y\ t_1 < t_2$ .... Entonces, su función de supervivencia sería la siguiente:

$$S(t) = P(T \geq t) = \sum_{t_j \geq t} f(t_j) \qquad (4.3)$$

La función de supervivencia continua es empleada en los análisis paramétricos y la discreta en los no paramétricos; tanto un tipo como el otro de análisis será descrito en los siguientes apartados.

Es necesario mencionar también que, S(t) es una función monótona decreciente, ya que la probabilidad de sobrevivir a tiempo cero es uno y la de sobrevivir un tiempo infinito es nula.

$$S(0) = 1 \text{ y } S(t) = 0\ cuando\ t \to\ \infty$$

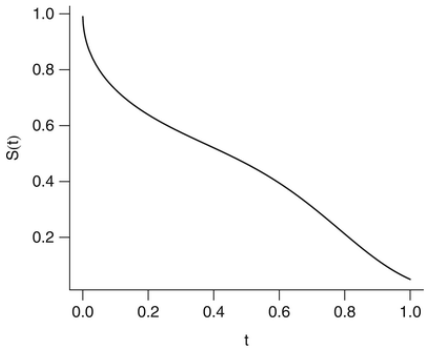


Ilustración 16: *Función de supervivencia / Regression Modeling Strategies*

4.2.2.2. Función de riesgo (Hazard Function)

La función de riesgo instantáneo Hazard Function se trata de un concepto importante en supervivencia ya que indica la probabilidad instantánea de que, si un individuo está en el estudio a tiempo t, le ocurra el evento en un instante de tiempo muy próximo a t.

Se define como:

$$h(t) = lim_{\Delta t \to 0} \frac{P(t < T \leq t + t\Delta | T > t)}{\Delta t} \qquad (4.4)$$

En el caso discreto:

$$h(t_j) = P(T = t_j | T \geq t_j),\ \ j = 1,2, \dots \qquad (4.5)$$

En el caso continuo, si se aplica la definición de probabilidad condicional a la ecuación de la función se obtiene:

$$h(t) = lim_{\Delta t \to 0} \frac{\frac{P\big((t < T \leq t + \Delta t) \cap (T > t)\big)}{P(T > t)}}{\Delta t} = lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{P(T > t)\Delta t} \tag{4.6}$$

Dado que:

$$P(t < T \leq t + \Delta t) = \int_{t}^{t+\Delta t} f(u)du = F(t + \Delta t) - F(t) \tag{4.7}$$

Sustituyendo esto en la ecuación anterior y aplicando la definición de derivada, se tiene que:

$$h(t) = lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{P(T > t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)} \tag{4.8}$$

Donde $f(t)$ es la función de densidad en t, dado que esta es igual a menos esa derivada $f(t) = \frac{\partial S(t)}{\partial t}$:

$$\frac{\partial \log\, S(t)}{\partial t} = \frac{\partial S(t)/\partial t}{S(t)} = -\frac{f(t)}{S(t)} \tag{4.9}$$

La función de riesgo también podría expresarse como:

$$h(t) = -\frac{\partial \log\, S(t)}{\partial t} \tag{4.10}$$

Una característica destacable de la función de riesgo es que puede intentar asociarse con la forma de alguna función de riesgo paramétrica, lo que puede ser útil para plantear un modelo paramétrico.

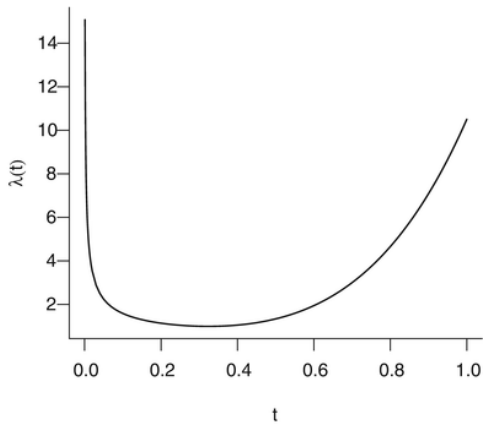


Ilustración 17: *Función de riesgo / Regression Modeling Strategies*

El gráfico anterior representa la función de riesgo asociada a la función de supervivencia dibujada en la *Ilustración 16***.** Se puede observar la relación entre las fases de más riesgo, al principio y en torno a 0.6, donde la función de supervivencia descendía.

Asociado a este término, aparece la función de riesgo acumulado. Se denota por $H(t)$ y puede ser descrita como la suma de todos los riesgos instantáneos hasta el momento t.

En el caso de que T sea continua, se define como:

$$H(t) = \int_0^t h(u)du \qquad (4.11)$$

Utilizando la función de supervivencia, obtenemos:

$$S(t) = \exp\{-H(t)\} \quad \rightarrow \quad H(t) = -\log S(t) \qquad (4.12)$$

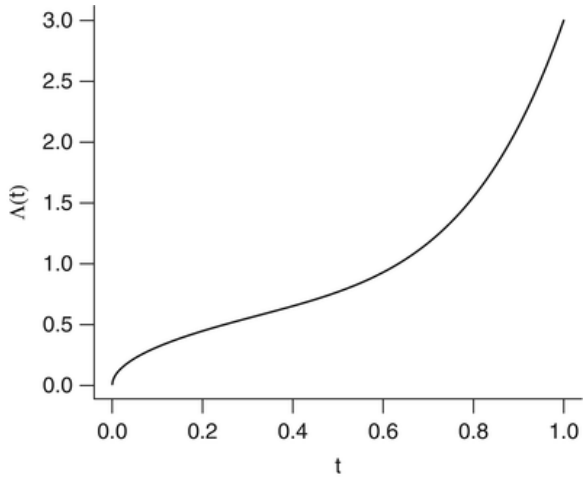


Ilustración 18: *Función de riesgo acumulada / Regression Modeling Strategies*

Esta función es creciente, es decir, el riesgo acumulado solo puede mantenerse o incrementarse.

En el caso de que T sea discreta con valores $t_1 < t_2 < \cdots$:

$$H(t) = \sum_{t_j \leq t} h(t_j) \qquad (4.13)$$

En este caso discreto, no existiría una relación entre la función de riesgo acumulado y la función de supervivencia.

Existen diversas formas de estimar la función de supervivencia, en siguiente apartado se describirán métodos paramétricos y no paramétricos. La regresión de Cox se trata de un modelo semiparamétrico, ya que cuenta con una parte paramétrica y una no paramétrica, por lo que se tratará de forma individualizada.

Además, se incluirá en cada uno de los apartados el código necesario para realizar el modelo en RStudio, ya que como ha sido descrito con anterioridad, es la herramienta utilizada para realizar los análisis.

4.2.3. Modelos No Paramétricos de supervivencia

Los modelos no paramétricos permiten interpretar los datos obtenidos sin tener que asumir ningún modelo probabilístico concreto para los tiempos de supervivencia y las funciones descritas en los apartados anteriores, siendo estimadas directamente de los datos, sin tener que realizar grandes supuestos previos al modelo.

Además, estos métodos no paramétricos nos permiten trabajar con datos censurados, que, aunque su información este incompleta, son útiles y no deben ser desestimados.

4.2.3.1. Estimador de Kaplan-Meier

El estimador Kaplan-Meier se basa en la descomposición de la curva de supervivencia en un producto de probabilidades condicionadas.

$$S(t) = \prod_{t_i \leq t)} (1 - {}^{d_{t_i}}/_{r_{t_i}}) \qquad (4.14)$$

donde d_{t_i} es el número de muertes a tiempo t_i y r_{t_i} son las personas que se encuentran en riesgo inmediatamente antes de ese tiempo.

De esta forma, la probabilidad de sobrevivir más de un determinado tiempo t es el producto de las probabilidades condicionadas en todos los tiempos t_i menores a t.

Siendo esto representado en forma de tabla:

Tabla 4: *Proceso para estimar la supervivencia mediante Kaplan-Meier*

$\boldsymbol{Tiempo\ evento}$	$\boldsymbol{Nº\ indiv\ en\ riesgo}$	$\boldsymbol{Nº\ Eventos}$	$\boldsymbol{Riesgo}$	$\boldsymbol{Fraccion\ de\ supervivencia}$	$\boldsymbol{\hat{S}(t)}$
$(\boldsymbol{t_1, t_2}]$	r_1	d_1	${}^{d_1}/_{r_1}$	$S_1 = 1 - {}^{d_1}/_{r_1}$	S_1
$(\boldsymbol{t_2, t_3}]$	r_2	d_2	${}^{d_2}/_{r_2}$	$S_2 = 1 - {}^{d_2}/_{r_2}$	$S_1 \cdot S_2$
$(\boldsymbol{t_3, t_4}]$	r_3	d_3	${}^{d_3}/_{r_3}$	$S_3 = 1 - {}^{d_3}/_{r_3}$	$S_1 \cdot S_2 \cdot S_3$
$\boldsymbol{\dots}$	…	…	…	…	…
$(\boldsymbol{t_{n-1}, t_n}]$	r_n	d_n	${}^{d_n}/_{r_n}$	$S_n = 1 - {}^{d_n}/_{r_n}$	$S_1 \cdot S_2 \cdot \dots \cdot S_{n-1} \cdot S_n$

Sea $r(t_i)$ el número de individuos en riesgo y $d(t_i)$ el número de eventos ocurridos hasta el instante t_i, el estimador Kaplan-Meier se define como:

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{r((t_i) - d(t_i)}{r(t_i)} \qquad (4.15)$$

Normalmente, la representación gráfica de esta estimación aparece acompaña por su intervalo de confianza. En el caso de muestras grandes a tiempo fijo t se distribuye aproximadamente normal. En este estudio se emplearán intervalos de confianza del 95%, por lo que con un $\alpha = 0{,}05$ tendríamos:

$$\hat{S}_{KM}(t) \pm z_{0,975} SE\left(\hat{S}_{KM}(t)\right) = \hat{S}_{KM}(t) \pm 1{,}96\sqrt{V\left(\hat{S}_{KM}(t)\right)} \tag{4.16}$$

Siendo la varianza del estimador igual a:

$$V\left(\hat{S}_{KM}(t)\right) = \hat{S}^2{}_{KM}(t) \sum_{t_i \leq t} \frac{d(t_i)}{n(t_i)[n(t_i) - d(t_i)]} \tag{4.17}$$

Una vez calculadas las curvas de supervivencia es frecuente querer comparar la supervivencia entre grupos de individuos. Para contrastar si existen diferencias significativas entre estos grupos es necesario plantear un contraste de hipótesis; para resolverlo, se utiliza el test estadístico Log-Rank.

Donde la hipótesis nula es la igualdad entre las funciones de supervivencia de los grupos que se desea comparar y la hipótesis alternativa la diferencia:

$$\begin{cases} H_0: \ S_1(t) = S_2(t) \\ H_1: \ S_1(t) \neq S_2(t) \end{cases} \tag{4.18}$$

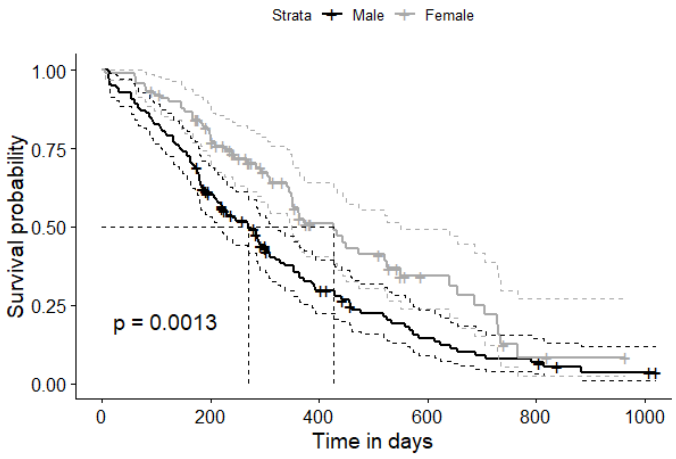


Ilustración 19: *Curva ejemplo Kaplan-Meier*

4.2.3.2. Test Log-Rank

El test logarítmico Log-Rank compara el número de eventos observados de cada grupo (O_i) con el número de eventos esperados en el grupo (E_i) empleando para ello el estadístico Chi-cuadrado con k-1 grados de libertad, donde k es el número de grupos.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{4.19}$$

Tanto este test como el método de Kaplan-Meier son métodos univariantes, es decir, describen la supervivencia sin tener en cuenta el efecto conjunto de las variables predictoras sobre la variable respuesta.

4.2.3.3. Modelos No paramétricos en R

Para calcular la estimación de las curvas de supervivencia de Kaplan-Meier en R se puede utilizar la función **`surfit()`** incluida en el paquete **survival** descrito en el apartado *de Librerías de R.*

Para esta función, sus principales argumentos son el objeto supervivencia, creado con la función Surv. A la derecha de esta fórmula, se debe añadir **`~1`** si simplemente se quiere observar supervivencia de toda la población, y añadir **`~`** con el nombre de las covariables categóricas, si se quiere ver la comparación de la curva entre varias cohortes. Se añadiría también a la función el data frame donde están los datos y el tipo de estimador "kaplan-meier".

```
survfit(Surv(time, status) ~ sex, data = data, Type = "Kaplan-meier")
```

Si imprimimos los resultados de esta función, obtendríamos un pequeño resumen sobre las curvas, con el número de observaciones, eventos y la mediana y sus límites de confianza de la supervivencia, pudiendo obtener un resumen más detallado con summary.

Para graficar estos resultados se utiliza la función **`ggsurvplot()`** paquete **Survminer.** En la que se puede indicar que muestre el p valor de la prueba Log-Rank, con el que se compararían los grupos (*pval = TRUE*). Para calcular este test de forma individual se utiliza la función **`survdiff()`**

```
ggsurvplot(fit, # Función de supervivencia
           pval = TRUE, # p-valor Test Log-Rank
           conf.int = TRUE, # INTERVALOS DE CONFIANZA
           risk.table = TRUE, # Añade la tabla de riesgo
           surv.median.line = "hv", # Especifica mediana de supervivencia)
```

Añadiendo el argumento *fun* a esta función se puede representar con "*cumhaz*" la función de riesgo acumulado.

4.2.4. Modelos Paramétricos de supervivencia

Los métodos paramétricos pretenden que los datos de los análisis se aproximen a las funciones proporcionadas por las distintas distribuciones, para así ser empleadas como funciones de supervivencia.

4.2.3.1. Distribución Exponencial

La distribución exponencial expresa un riesgo constante a lo largo del tiempo.

Sea X una variable aleatoria continua, real y positiva se dice que sigue una distribución exponencial de parámetro θ con $\theta \geq 0$, siendo su función de densidad:

$$f(x) = \theta e^{-\theta x} \tag{4.20}$$

por lo que, X~$Exp(\theta)$.

Esto llevado a la función de supervivencia sería igual a:

$$S(t) = \int_t^{\infty} f(u)du = \int_t^{\infty} \theta e^{-\theta u} du = e^{-\theta u}|_t^{\infty} = e^{-\theta t} \qquad (4.21)$$

Y la función de riesgo:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta e^{-\theta t}}{e^{-\theta t}} = \theta \qquad (4.22)$$

Siendo esta constante, lo que implica que el riesgo es el mismo a lo largo del tiempo.

4.2.3.2. Distribución Weibull

La distribución de Weibull se trata de una generalización del modelo exponencial y puede ser considerada la distribución paramétrica más utilizada en análisis de supervivencia.

Siendo t una variable aleatoria continua, β el parámetro de forma y λ el parámetro de escala tenemos que la función de densidad de Weibull es igual a:

$$f(t;\beta,\lambda) = \beta\lambda t^{\beta-1} \exp\{-\lambda t^{\beta}\} \qquad \text{con } \beta > 0,\ \lambda > 0 \text{ y } t \geq 0 \qquad (4.23)$$

por lo que, T~$Weibull(\beta,\lambda)$.

En la función de superveniencia para esta distribución sería:

$$S(t) = \int_t^{\infty} f(u)du = \int_t^{\infty} \beta\lambda u^{\beta-1} \exp\{-\lambda u^{\beta}\} du = -\exp\{-\lambda u^{\beta}\}|_t^{\infty} = \exp\{-\lambda t^{\beta}\} \qquad (4.24)$$

Su función de riesgo:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\beta\lambda t^{\beta-1} \exp\{-\lambda t^{\beta}\}}{\exp\{-\lambda t^{\beta}\}} = \beta\lambda t^{\beta-1} \qquad (4.25)$$

La interpretación de esta función de riesgo sería la siguiente:

$$h'(t) = (\beta-1)\beta\lambda t^{\beta-2} = \begin{cases} h'(t) > 0, si\ \beta > 1\ ;\ Riesgos\ monótonos\ crecientes \\ h'(t) < 0, si\ \beta < 1\ ;\ Riesgos\ monótonos\ decrecientes \\ h'(t) = 0, si\ \beta = 1;\ h(t) = \lambda\ y\ se\ trasformaría\ en\ una\ exponencial \end{cases} \qquad (4.26)$$

4.2.3.2. Criterio de información Akaike

Para evaluar como de bien se ajusta un modelo se utiliza el método matemático de criterio de información Akaike (AIC). Para determinar el valor de esta información utiliza la estimación de máxima verosimilitud y el número de variables independientes.

$$AIC = 2p - 2\ln(L) \tag{4.27}$$

Donde:

- p es el número de variables independientes que se usan
- L es la estimación de la probabilidad de que el modelo haya reproducido los valores observados (verosimilitud)

Para saber si comparando dos modelos uno es significativamente mejor, es necesario realizar el cálculo de sus dos AIC. Cuando las diferencias entre ellos son superiores a 2 unidades, el que tenga el valor más bajo será el que mejor se ajuste en comparación con el oro de forma significativa.

4.2.3.3. Modelos paramétricos en R

Para estimar la función de supervivencia en R a través de estas distribuciones es necesario utilizar la función **`frexsurvreg()`.** Los parámetros son estimados por máxima verosimilitud.

Las variables necesarias para implementar esta función son, por orden, el objeto fórmula, dónde se incluye la función **`Surv`**[9] seguida de `~1` para el caso de la distribución total de la población, como para las no paramétricas, y las covariables para obtenerla por grupos. Se añade también, el data-frame con los datos del análisis, el ancho de los intervalos simétricos, que por defecto es 0,95, y la distribución, que en este caso seria "exp" para la distribución exponencial y "weibull" para la distribución Weibull.

En la salida en consola obtendríamos con esta función tanto la estimación de máxima verosimilitud con sus intervalos de confianza y su desviación estándar, como el número de datos observados t censurados, el tiempo total de riesgo y los grados de libertad.

```
flexsurvreg(Surv(x,y)~1, data = data, cl = 0.95, dist =”exp”)
```

4.2.5. Regresión de Cox

El modelo de regresión de Cox (Cox, 1972) permite estimar la relación que existe entre un conjunto de variables explicativas, llamadas *covariables,* con la tasa instantánea del suceso de interés, es decir, la función de riesgo.

Este modelo evalúa simultáneamente el efecto de varios factores sobre la supervivencia y se expresa mediante la función de riesgo:

$$h(t;\ x_1, x_2, \dots, x_p) = h_0(t) \cdot \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\} \tag{4.28}$$

donde;

- $h(t)$ es la función de riesgo determinada por las p covariables.
- Los coeficientes β_i miden el impacto de esas covariables.

[9] El objeto Surv es la combinación de información entre la censura y los tiempos **`Surv(Time, Event)`**, donde Time es el tiempo hasta la observación y Event donde 1 es la ocurrencia del evento y 0 la censura. La mayoría de las funciones de supervivencia utilizan este objeto.

- Y el término $h_0(t)$ es el riesgo basal, que involucra al tiempo, no a las variables predictoras.

El modelo de Cox se describe como un modelo semiparamétrico debido a que cuenta con dos componentes: la función $h_0(t)$ que depende del tiempo se estima de forma no paramétrica, a la que se le denomina función de riesgo base o riesgo basal, y el exponencial del predictor lineal, $\exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}$, donde está el vector de parámetros β se estima mediante la maximización de la función de verosimilitud parcial.

La función de riesgo basal es la función de riesgo para un individuo con variables predictoras iguales a 0 (grupo control), ya que $h(t) = e^0 h_0(t) = h_0(t)$, y no es necesario conocerla para estimar los parámetros del modelo, pero una vez advertidas sus estimaciones, se puede conocer con ella, la supervivencia de cualquier perfil de un individuo.

Debido a que existen datos incompletos, los parámetros del modelo de Cox no podrían estimarse siguiendo el método común de máxima verosimilitud, ya que se desconoce la función de riesgo. Cox propone para resolver esto el **método de estimación de verosimilitud parcial** (partial likelihood).

$$L_{t_{(i)}}(\beta_1, \dots, \beta_p) = \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{i(j)}\right)}{\sum_{l \epsilon R(t_{(i)})} \exp\left(\sum_{j=1}^{p} \beta_j x_{i(j)}\right)} \quad (4.29)$$

Cox también puede ser denominado Modelo de riesgos proporcionales, ya que compara el riesgo de dos individuos.

Teniendo los riesgos de un individuo $h_1(t) = h_0(t)\exp(\beta \cdot 1) = h_0(t)\exp(\beta)$ que recibe el tratamiento, y uno $h_2(t) = h_0(t)\exp(\beta \cdot 0) = h_0(t)$ placebo, se obtiene la **razón de riesgos instantáneos** (Hazard ratio) respecto a estos dos individuos.

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t)\exp(\beta)}{h_0(t)} = \exp(\beta) \quad (4.30)$$

Por lo tanto, se puede afirmar que:

$$log\left(\frac{h_1(t)}{h_2(t)}\right) = \beta \quad (4.31)$$

Es decir, el parámetro β puede ser interpretado como el logaritmo entre los riesgos instantáneos de un apersona tratada y no otra no tratada.

4.2.5.1. Regresión de Cox en R

Para realizar el modelo semiparamétrico de Cox se utiliza la función **`coxph()`** que se encuentra en el paquete **survival.** Los parámetros que se deben incluir son la formula, creada también con la función **`Surv()`,** los datos y el método, que por defecto utiliza "efron", el más eficiente computacionalmente.

En la salida por consola obtenemos, la importancia estadística, los coeficientes de regresión, las razones de riesgo y sus intervalos de confianza y la significación estadística del modelo.

```
coxph(Surv(time, status) ~ sex, data = data, method = “efron”)
```

En el caso de que la variable estudiada tenga más de 2 variables R crea las k-1 variables Dummy correspondientes. Por norma, R toma como variable referencia la más pequeña si se trata de números y la primera por orden alfabético si las categorías tienen etiquetas.

Si se quiere, se puede realizar un análisis de regresión de Cox multivariante añadiendo más variables a la formula del modelo.

```
coxph(Surv(time, status) ~ sex + age, data = data, method = “efron”)
```

Una vez el modelo ha sido ajustado a los datos, se puede graficar la proporción de supervivencia prevista para un grupo de riesgo en particular. Para ello, es necesario calcula la proporción de supervivencia mediante la función **`survfit()`.**

```
ggsurvplo(survfit(coxph(Surv(time, status) ~ sex + age, data = data,
          method = “efron”)))
```

5. Resultados y Discusión

En el siguiente apartado, se reflejarán los resultados obtenidos tanto en la parte de bases de datos descrita en la parte *4.1. ETL*, como de los análisis de supervivencia que se estudian en el apartado *4.2.*

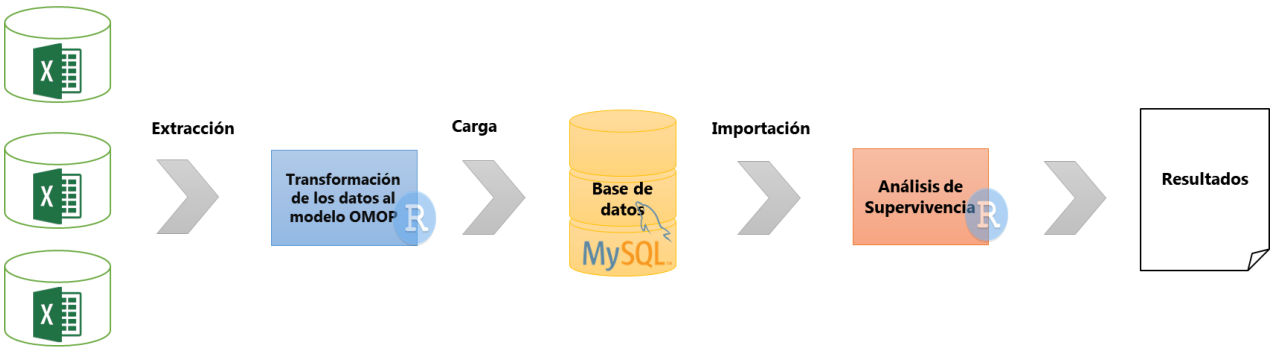


Ilustración 20: *Diagrama sobre el proceso ejecutado en el estudio.*

Como se muestra en el esquema, el proceso ha sido el siguiente: importación de los Excel iniciales en R, transformación de los datos en códigos y estructuración de ellos en el Modelo Común y la carga en la base de datos.

Una vez que los datos se encuentran en la base, los necesarios para el análisis de supervivencia son importados a R, cambiando los códigos por sus valores correspondientes para poder obtener los resultados de forma clara y sacar las conclusiones de los análisis.

5.1. Resultados CDM OMOP

En el apartado de Extracción, Transformación y Carga de los datos, ha sido explicado el desarrollo que forma la base de datos, proceso que se ha ido siguiendo para poder obtener los datos almacenados siguiendo el Modelo común de Datos de OMOP. La plataforma MySQL permite obtener el esquema de la base de datos que se muestra en la *Ilustración 13* de la página siguiente.

Se cuenta con una base de datos de 9 tablas relacionadas a través de las claves primarias de PERSON y SPECIMEN con las claves foráneas del resto de las tablas.

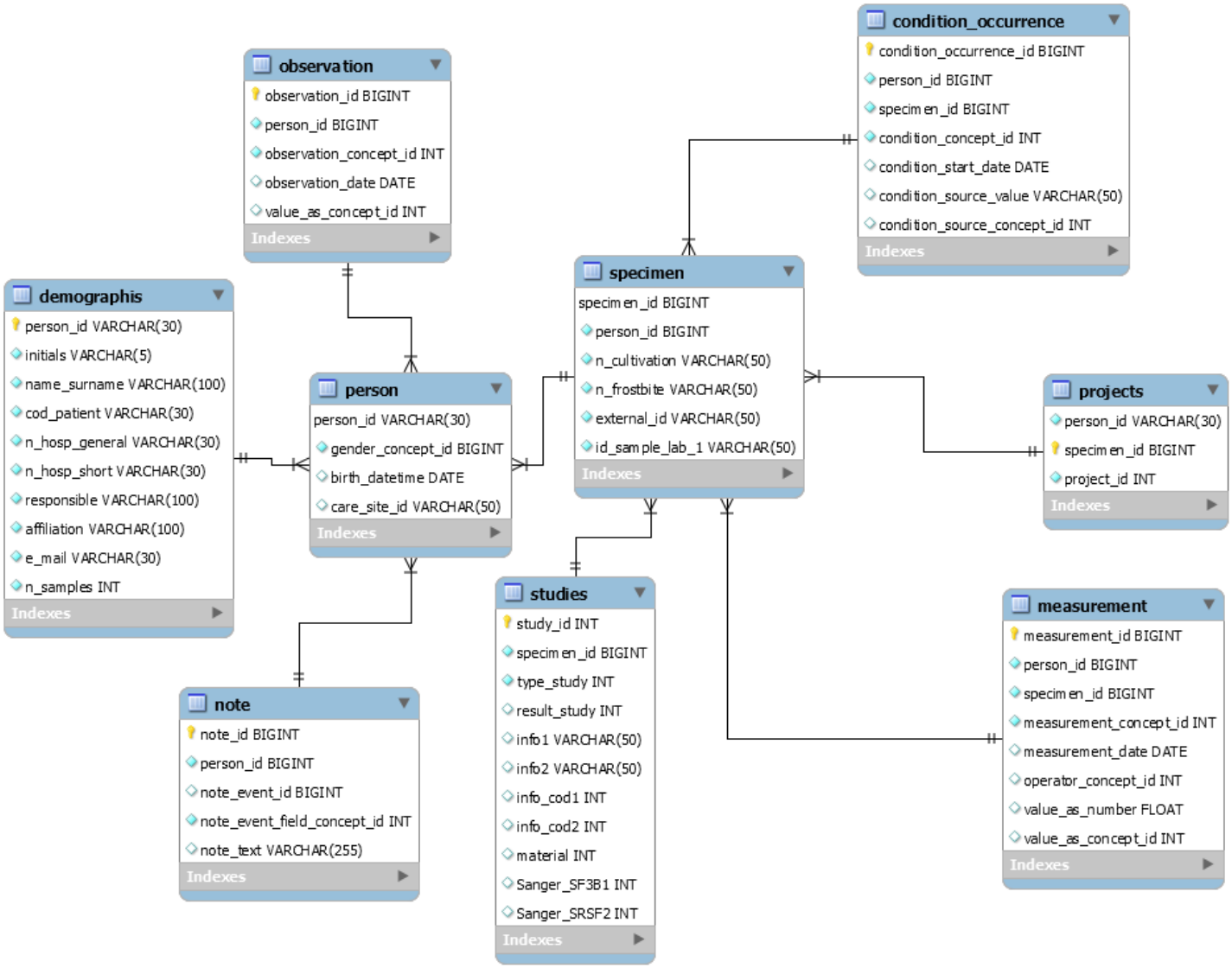


Ilustración 21: *Diagrama de la base de datos OMOP obtenida en el estudio.*

5.2. Análisis descriptivos

En esta parte, se realizará una caracterización de los datos descritos en el apartado *Data set* mediante análisis descriptivos. Apartado en el que se recoge que la población del estudio está formada por pacientes de toda España de síndromes mielodisplásico recogidas desde 1993 hasta principios del 2020.

Para los análisis, tanto descriptivos como de supervivencia, se emplearán únicamente las historias clínicas de **2231** pacientes, ya que las restantes a las 2882 ha sido eliminadas debido a que no se cuenta, a fecha del estudio, con los datos completos sobre estos individuos.

Como ha sido dicho en la introducción, una de las principales dificultades del estudio de síndromes mieoldisplasicos es la avanzada edad a la que estos aparecen. Por lo tanto, la gran mayoría de los pacientes que forman la base de datos son pacientes senior de edades entre los 76 y 85 años, como refleja la siguiente gráfica:

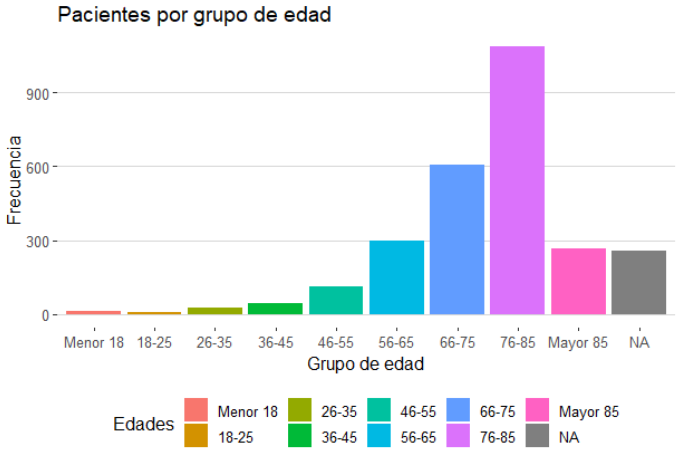


Figura 1: *Diagrama de barras por grupos de edad.*

El 61% de los pacientes de estudio son hombres, lo que puede dar un indicio de que este tipo de patología puede ser superior en el género masculino. Posteriormente, se estudiará si resulta significativa esta característica con relación a la supervivencia de los pacientes.

Tabla 5: *Porcentajes por género.*

Género	**Hombres**	**Mujeres**
Frecuencias relativas	60,5%	39,4%

En algunos pacientes es necesario realizar un traslado de un hospital a otro y esto puede ser debido a diversas causas, en este estudio un 6% de los individuos fue desplazado a otro hospital, siendo en su mayoría desplazados al Hospital Universitario de Salamanca.

Debido a que el estudio comenzó en los años 90, cuando se utilizaba la clasificación FAB, los pacientes cuentan con esa clasificación, habiendo sido actualizada a la de las décadas siguientes, la clasificación OMS. Estas clasificaciones catalogan los tipos de síndromes mielodisplásicos, basándose en las observaciones realizadas en las células de la médula ósea, como se ha descrito en el punto *1.2*. Es por ello por lo que numerosas variables del estudio no van a ser empleadas en el análisis, ya que son ya utilizadas para esta clasificación.

Puesto que se cuenta con la clasificación más actual, OMS 16, se van a describir a los individuos tomando como referencia esta clasificación.

Tabla 6: *Porcentajes por la clasificación OMS 2016.*

Clasificación OMS 2016	Frecuencias relativas
SMD-MDL / SMD-RS	25.09%
SMD-EB (SMD con exceso de blastos)	20.15%
SMD-RS (SMD con sideroblastos en anillo)	17.95%
CMML (Leucemia mielomonocítica crónica)	8.12%
SMD-MLD (SMD con displasia multilinaje)	7.87%
MF (Mielofibrosis)	7.43%
SMD-SLD / SMD-RS	4.69 %
SMD with isolated del(5q)	3.86%
AML (Leucemia mieloide aguda)	2.98%
SMD-SLD (SMD con displasia unilinaje)	0.93%
SMD-U (SMD, no clasificable)	0.29 %

En la tabla se observa que la mayoría de los pacientes que forman el estudio se clasifican en una categoría compartida entre *SMD con displasia multilinaje* y *SMD con sideroblastos en anillo,* esta categoría existe debido a que no se sabe exactamente a cuál de los dos tipos de MSD pertenece.

Según la literatura encontrada, el tipo de SMD más común es el SMD-MLD, seguido por los de tipo SMD-EB (*SMD con exceso de blastos*) lo que podría coincidir con los datos del estudio (American Cancer Society, 2018a).

En estos pacientes también se recoge si han progresado a un SMD de riesgo mayor o si finalmente el SMD deriva en AML. Un 52% de la población del estudio progresa a un SMD de riesgo mayor, pero solo un 13,9% progresa a leucemia mieloide aguda, como se recoge en la introducción, son escasos los pacientes que pasan de SMD a AML. En cambio, si observamos únicamente a ese 52% que progresa a un SMD de mayor riesgo, el 62,8% de ellos desarrolla AML.

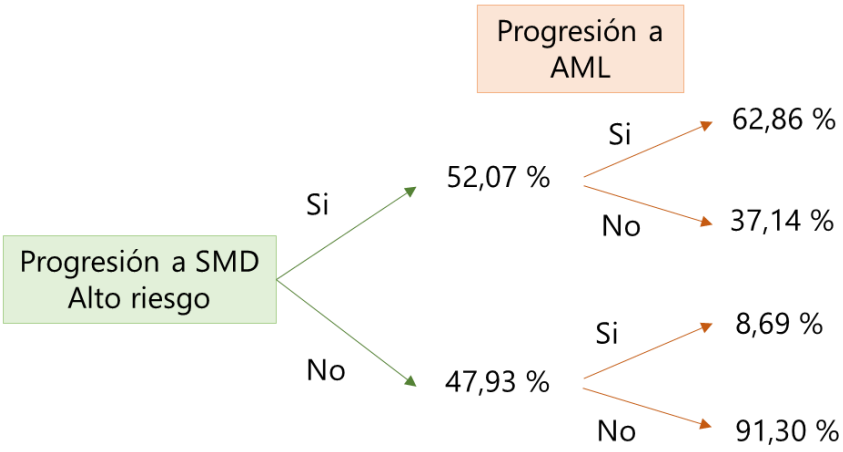


Ilustración 22: *Representación de los porcentajes de progresión a SMD de alto riesgo y AML.*

Para poder clasificar el riesgo del MDS se utiliza la clasificación IPSS. En la base se recogen tanto la clasificación normal como la revisada, utilizaremos esta última para los estudios, ya que es más actual.

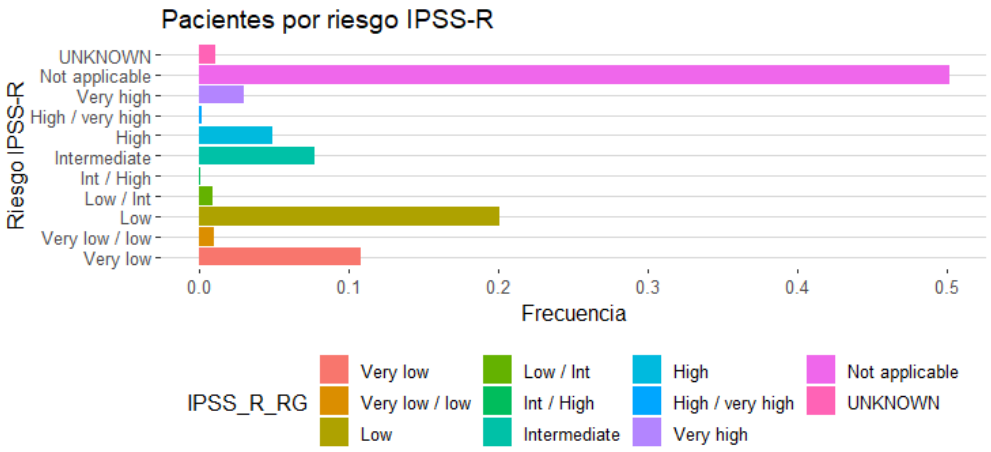


Figura 2: *Diagrama de barras por riesgo IPSS-R.*

Prácticamente la mitad de los pacientes del estudio no pudieron ser categorizados con el riesgo IPSS-R. Debido posiblemente a que la ausencia de resultados en algunas variables dificulta su clasificación. Obviando esto, la mayoría de los pacientes del estudio tienen SMD de riesgos bajos.

Una vez estudiadas las características demográficas y de riesgo de la población podemos pasar a analizar la supervivencia de estos pacientes mediante los métodos descritos en el apartado teórico de la metodología.

5.3. Análisis no paramétricos

Observamos en la curva de supervivencia obtenida mediante el método no paramétrico de Kaplan-Meier que existen un gran número de pacientes en el estudio que están censurados. Pese a esto se puede observar que en los 2.5 primeros años en los cuales el paciente presenta SMD es cuando la probabilidad de supervivencia se ve más afectada, estabilizándose a partir de los 5 años.

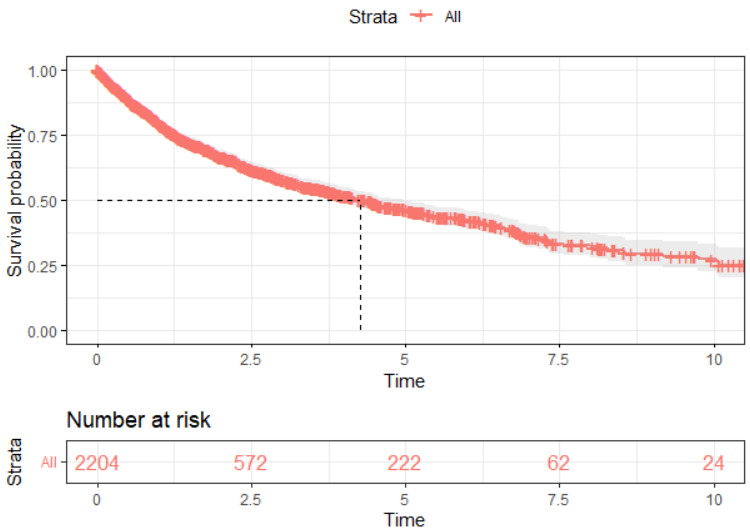


Figura 3: *Curva Kaplan-Meier.*

Habiendo observado en los análisis descriptivos una mayor presencia de pacientes de SMD del género masculino, parece preciso comparar si existen diferencias significativas entre las supervivencias de los pacientes por género.

Gracias al test Log-Rank sabemos que existen diferencias significativas entre las dos curvas, ya que se obtiene un $p < 0.0001$. Se aprecia que la diferencia es favorable para las mujeres, ya que sus probabilidades son siempre superiores a las de los hombres.

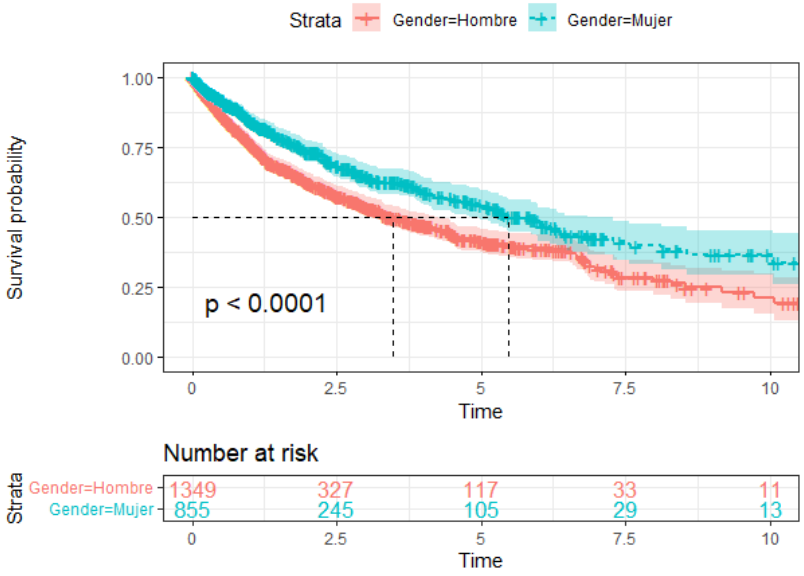


Figura 4: *Curvas K-M por género.*

Algunos pacientes son derivados a otros hospitales durante su enfermedad, observando en el gráfico que existe un riesgo acumulado mayor a lo largo del tiempo cuando los pacientes no son trasladados a otro hospital. Esto se debe a que, por norma, cuando los pacientes son trasladados lo hacen a hospitales más especializados en su perfil, conociendo que la mayoría de los traslados se producen al Hospital Universitario de Salamanca.

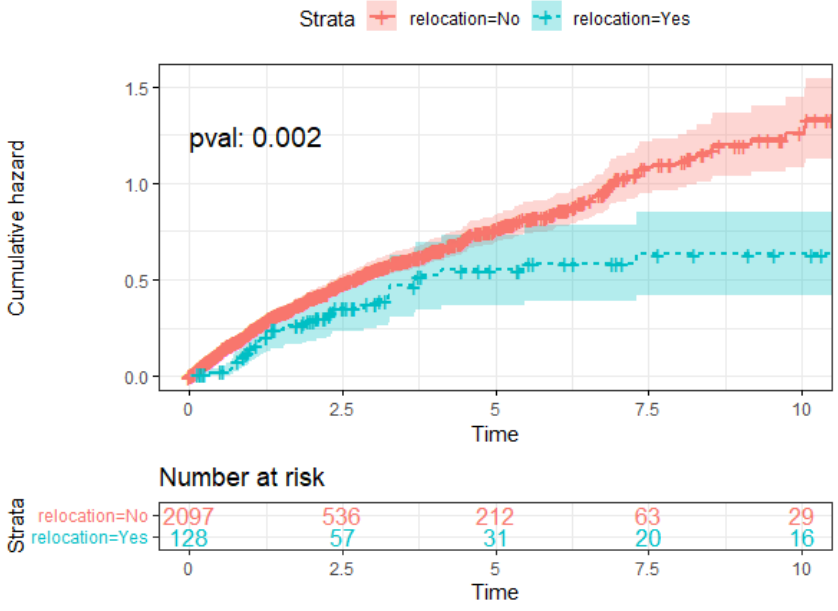


Figura 5: *Curvas K-M por traslado.*

Como es lógico, los pacientes que entran en el estudio como SMD y su enfermedad progresa a AML tienen un riesgo mucho mayor de defunción. Siendo la curva de los pacientes de AML muy pronunciada en los 2 primeros años.

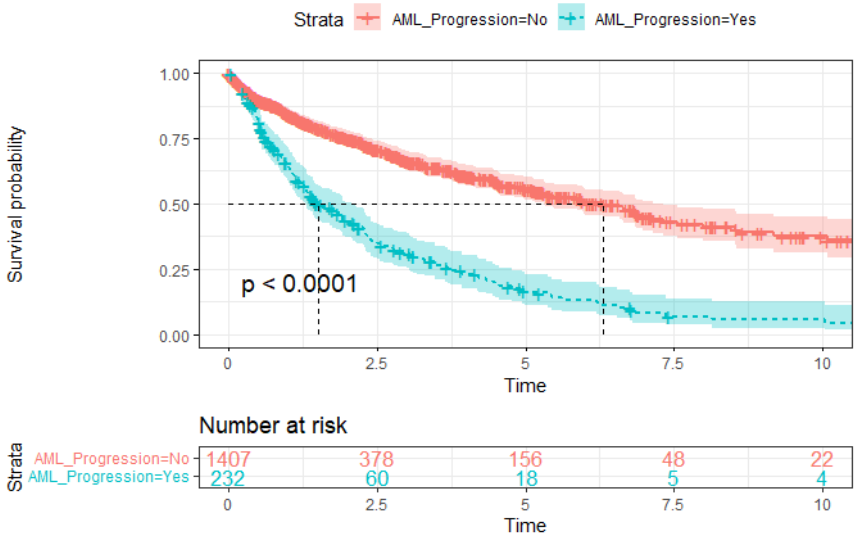


Figura 6: *Curva K-M por progresión a AML.*

Debido a que en los análisis descriptivos se observa que los grupos de edad más jóvenes tienen muy pocos pacientes, para analizar las gráficas de supervivencia de Kaplan-Meier se decide crear un nuevo grupo de menores de 55 años.

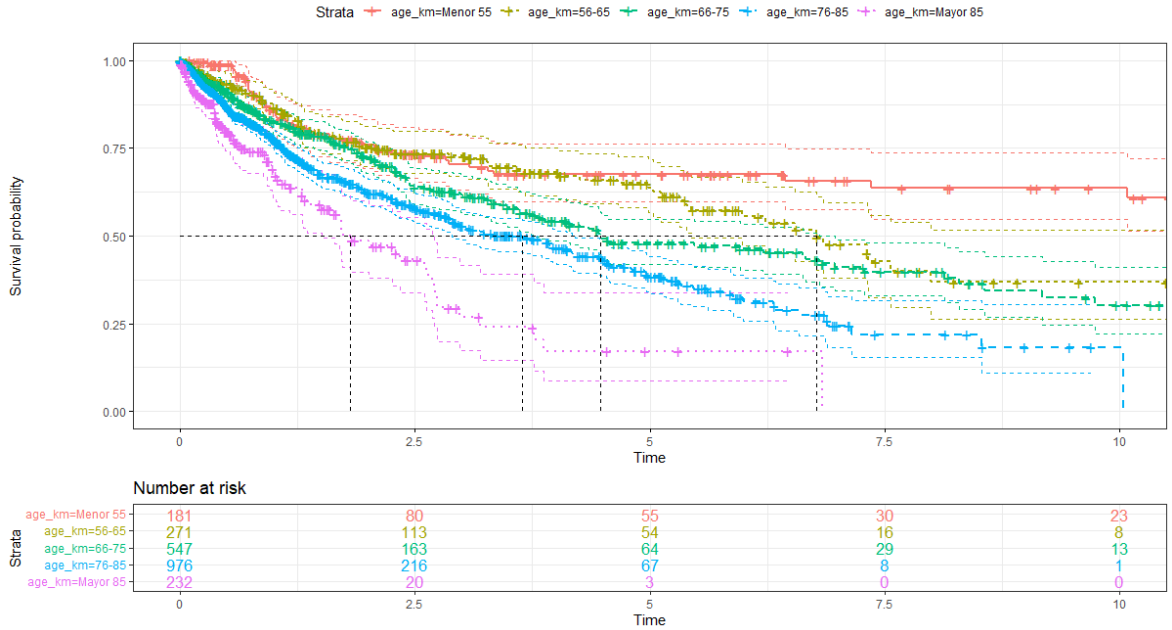


Figura 7: *Curvas K-M por grupos de edad.*

Observamos en el grafico anterior como cuanto mayor sea la edad, más pronunciada es la curva de supervivencia, siendo bastante altas las probabilidades de supervivencia en los 2.5 primeros años para los grupos menores de 65. Cabe destacar también, que las curvas de menores de 55 y del grupo de 56 a 65 se solapan durante los 5 primeros años con la patología.

La supervivencia para los mayores de 65 cae en picado desde el principio, no sobreviviendo la mitad de los pacientes a los 5 años para ninguno de los fragmentos de edad.

Un gráfico que puede ser muy interesante es el que relaciona la curva de supervivencia con el tipo de SMD diagnosticado. Seleccionando a los pacientes con los tipos con mayor presencia obtenemos que existe una mayor posibilidad a morir en los primeros años de pacientes con MDS-EB.

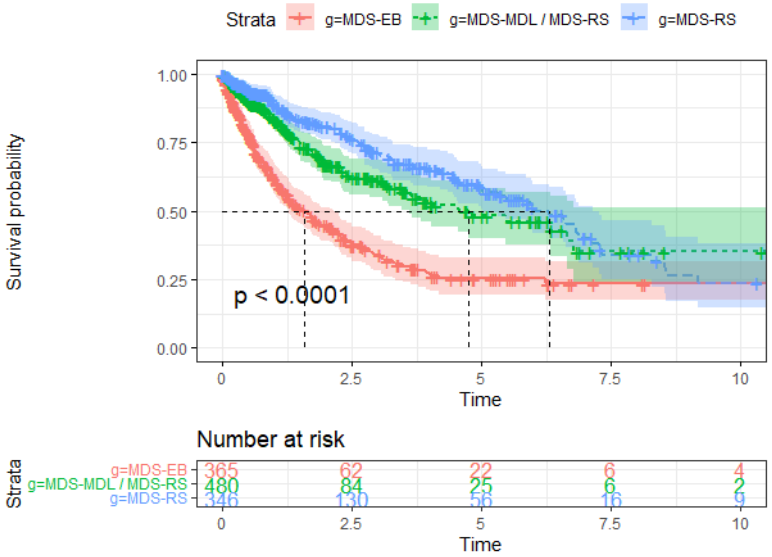


Figura 8: *Curvas K-M por tipo de SMD diagnosticado.*

Junto con estos análisis realizados, también se ha estudiado la relación con el resto de las variables, como por ejemplo tipo de diagnóstico, variable que se codifica como primario cuando no se conoce el motivo por el que determinada enfermedad aparece y secundario si se conoce la causa. De esta comparación entre las curvas, se obtiene que no existen diferencias significativas entre ellas, y que, por lo tanto, seria indiferente para la probabilidad de supervivencia el conocimiento de la causa de la enfermedad.

5.4. Análisis paramétricos

Se ha querido también comprobar si la curva de supervivencia de los pacientes que forman este estudio se ajusta algún de las funciones paramétricas.
Vista la curva de supervivencia por grupos de edad, los pacientes son juntados por edad menor que 65, *Adultos*, y mayores de 65, *Ancianos*. Una vez agrupados de forma dicotómica, se estudia si la función de supervivencia de la variable edad se ajusta a la distribución exponencial.

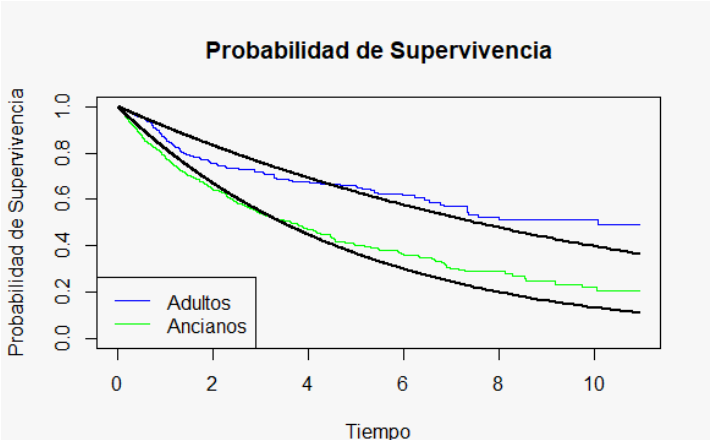


Figura 9: *Curvas de supervivencia ajustadas a la distribución exponencial por grupos de edad.*

Se observa que el ajuste no es demasiado exacto en ninguna de las dos curvas.

En cambio, cuando ajustamos la función de supervivencia de la variable edad con la distribución de Weibull obtenemos que, gráficamente, el ajuste es mejor.

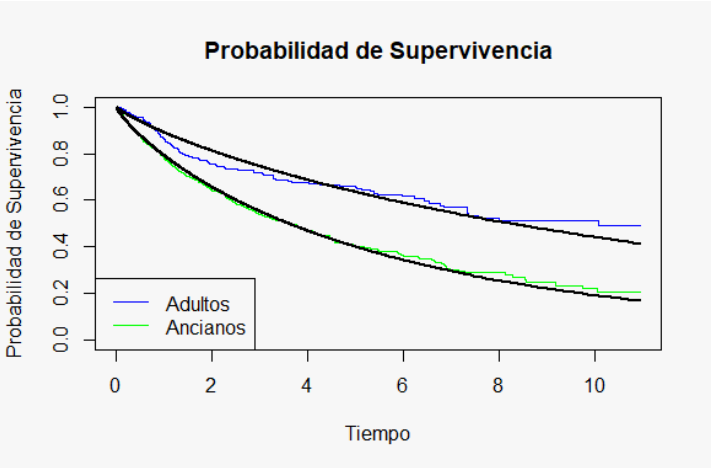


Figura 10: *Curvas de supervivencia ajustadas a la distribución Weibull por grupos de edad.*

Mediante el método matemático descrito en el apartado de *Análisis Paramétrico* AIC, se demostrará no solo gráficamente el mejor ajuste de la curva Weibull a la función de supervivencia de nuestro modelo. Esto será realizado también en RStudio.

Tabla 7: *Valores de análisis paramétrico AIC.*

	weibull	**exp**
AIC	3913.792	3946.550
Log-Lik	-1953.896	-1971.275

Se aprecia que las diferencias entre las dos distribuciones son significativas, ya que difieren en más de 2 unidades. La distribución de Weibull es la menor, por lo tanto, se confirma lo observado en las gráficas, es el modelo que mejor se ajustaría.

5.5. Regresión de Cox

Para estudiar el modelo de Cox consideramos en un principio la estimación de los parámetros asociados a todas las variables; edad, tipo de diagnóstico, clasificación OMS 16, profesión SMD de alto riesgo, progresión AML y relocalización. La edad se utiliza de forma dicotómica como se apunta en la parte de análisis paramétrico.

Una vez obtenido el modelo con todas las variables se eliminan las variables no significativas y se obtiene que el modelo va a estar formado únicamente por la edad, el género y la progresión a AML.

```
coxph(formula = Surv(os_time, os_status) ~ Age + Gender + AML_Progression,
   data = km_data2)
```

Observando en el grafico que el único efector "protector" es ser mujer, ya que su Hazard ratio es inferior a 1, siendo muy destacable como aumenta en gran medida, 3.2 puntos, el progreso de la enfermedad a AML para el riesgo de no supervivencia, y una edad superior a los 65 en algo más que dos puntos.

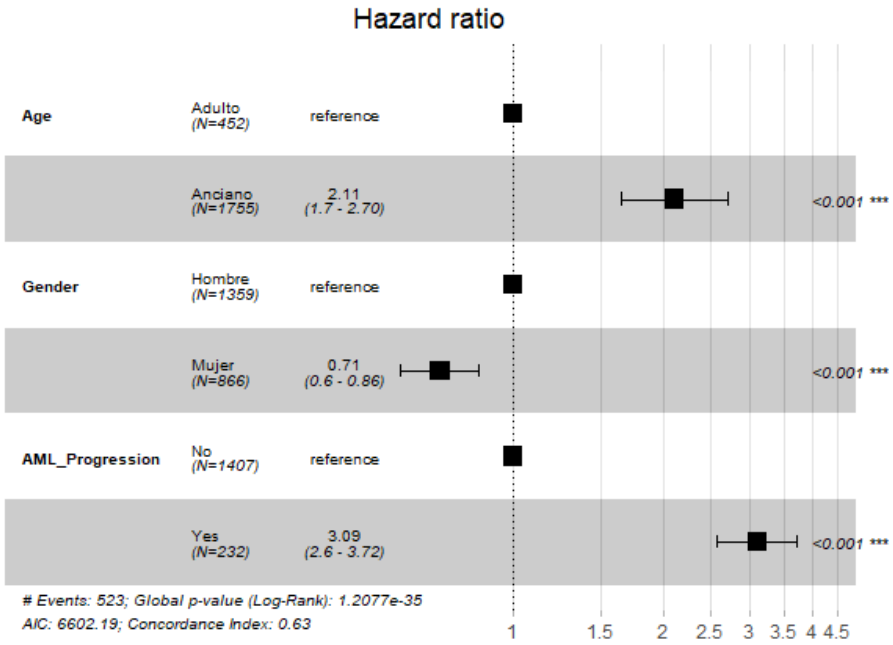


Figura 11: *Representación gráfica de los Hazard Ratio.*

Se realiza una validación del supuesto de riesgos proporcionales para el ajuste del modelo de Cox. Para ello se emplea el test de Schoenfeld que proporciona los residuos para cada covariable escalados contra el tiempo.

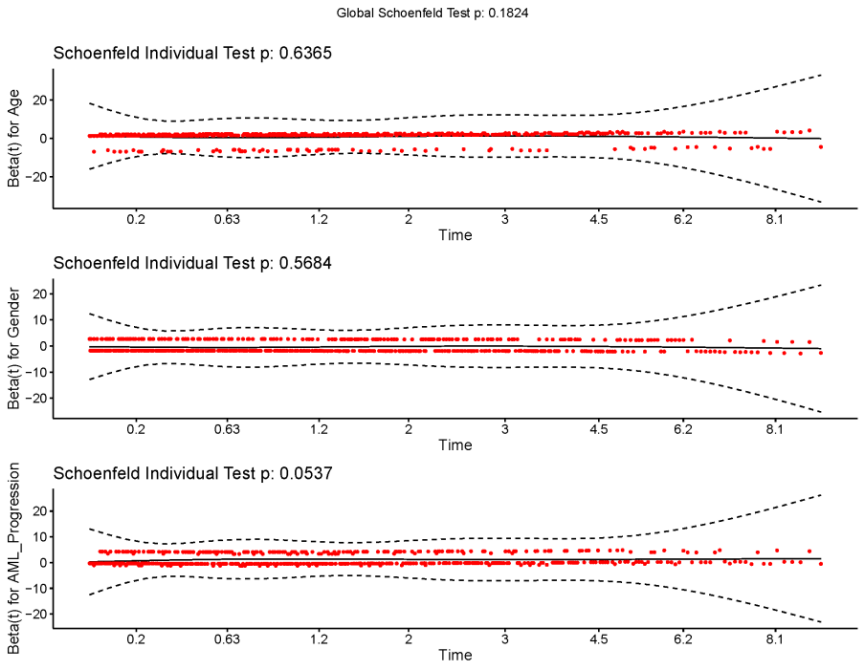


Figura 12: *Test de Schoenfeld.*

Se advierte que el supuesto de riesgos proporcionales está respaldado por las variables Edad, Genero y la progresión a AML, es decir, que el cociente entre sus riesgos es independiente del tiempo.

Teniendo en cuenta esto, el modelo de Cox se podría emplear para predecir casos particulares, lo que se acerca al termino de medicina personalizada. Para ello se usará como ejemplo la comparación entre personas con edades distintas, género distintos y personas con progreso a AML diferente.

```
datosnuevo <- data.frame(Age = c("Anciano", "Anciano"),
                         Gender = c("Mujer", "Hombre"),
                         AML_Progression = c("Yes", "Yes"))
Pred <- survfit(res.cox, datosnuevo)
```

Este es el código utilizado para la gráfica de personas con género distinto, para obtener el resto solo que habría que ir cambiando los fatores a criterio.

Se observa en los gráficos su coherencia con los Hazard ratio respecto a la curva basal (que es la calculada mediante K-M), ya que, teniendo las mismas condiciones, el género femenino determina positivamente la curva de supervivencia, ocurriendo lo contrario con la edad superior a 65 años y la progresión a AML, que causa un efecto muy negativo en las curvas de supervivencia.

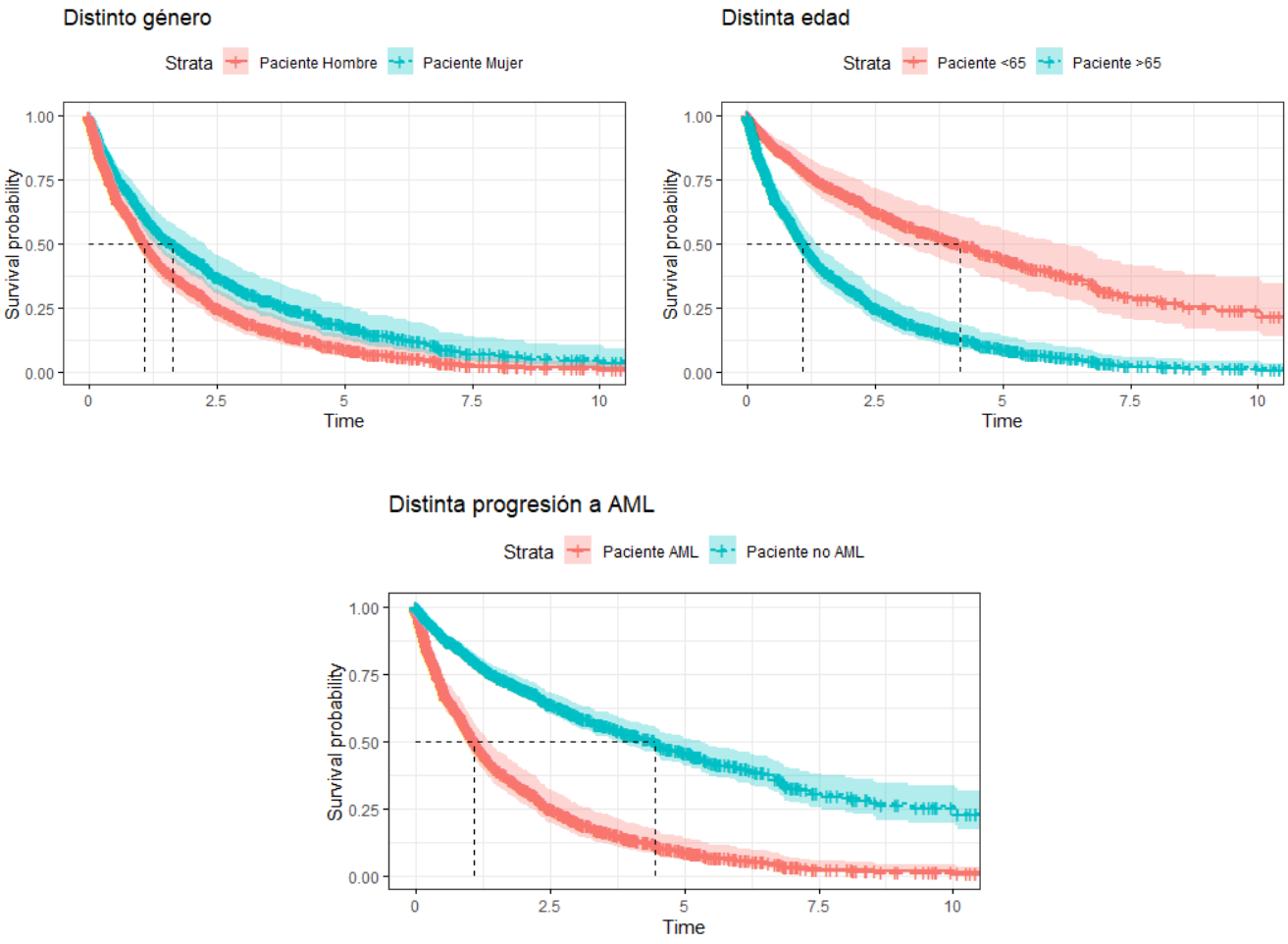


Figura 13: *Comparación de curvas de supervivencia empleando el modelo de Cox.*

6. Conclusiones

Como análisis final y concluyendo de esta forma el estudio cabe valorar tanto el uso del modelo común de datos OMOP, así como los resultados del análisis de supervivencia realizado en pacientes con Síndrome Mielodisplásico.

El Modelo Común de Datos OMOP permite la unificación de bases de datos en un solo modelo facilitando la recogida, tratamiento y análisis posterior de los datos. Esto conlleva una absoluta mejora en el proceso de consecución de evidencia clínica; siendo patentes las numerosas las ventajas médicas que supondría si este modelo común se comenzase a globalizar de forma rápida, cosa que se está produciendo gracias a la comunidad OHDSI, comunidad en constante actuación.

Otra ventaja muy importante de este modelo es la forma "vertical" en la que se almacenan los datos, ya que esto facilita la incorporación de nuevas variables sin modificar la estructura de las tablas, ya que el número de campos que formaría cada una no se ve afectado.

Como complicaciones derivadas de todo el proceso de extracción, trasformación y carga de los datos, existe la dificultad, como analista de datos o estadístico, de la búsqueda de los códigos asociados a conceptos. En este proceso es necesaria la intervención de un experto en la materia del estudio.

Proponiendo una continuación futura para el CDM se puede realizar una ampliación de las tablas que forman el modelo estándar. Pese a la posibilidad de adicionar nuevos campos a la base sí que se ve necesario la ampliación del modelo a algunas ramas de la medicina. Esto ya ha ocurrido recientemente con tablas que recogen datos sobre la secuenciación, lo que lleva a la deducción de que en un periodo corto-medio de tiempo el modelo común de datos será mayor.

En resumen, son numerosos los grupos médicos que deciden unirse a este tipo de ciencia puntera para la recogida de datos, haciendo necesaria la presencia y unión de grupos de trabajo con conocimientos estadísticos e informáticos.

Concluyendo de esta forma con la parte de base de datos del proceso, particularizando al análisis estadístico de los datos podemos resumir los resultados en las siguientes líneas:

- Los pacientes con SMD tienen una supervivencia bastante alta en los 4 primeros años desde que la enfermedad es diagnosticada.
- Como es lógico, a medida que la edad en la que aparece la patología aumenta, también aumenta la probabilidad de no supervivencia.
- Una de las grandes problemáticas del SMD es la capacidad que tienen de derivar en una enfermedad con mayor tasa de mortalidad como es el AML. Claro está que los pacientes que progresan a esta enfermedad disminuyen notablemente sus probabilidades se sobrevivir durante los primeros años de enfermedad.
- Coincidiendo con la literatura encontrada, los pacientes que presentan el tipo de SMD con exceso de blastos, son los que más riesgo tienen de que su enfermedad derive en AML, por lo tanto, su curva de supervivencia es muy pronunciada.
- La supervivencia en pacientes trasladados de un hospital a otro es mayor. Un 93,7% de los pacientes desplazados lo hacen al Hospital Universitario de Salamanca, lo que indica la clara mejora al tratar a los pacientes en este hospital.

- La variable género femenino influye positivamente en la supervivencia a la enfermedad, en cambio la progresión a AML, influye de forma muy negativa.
- Sabemos que los factores que influyen en el cálculo de la supervivencia son la edad, el sexo y la progresión a AML, sabiendo que este modelo global esta respaldad por el supuesto de riesgos proporcionales.

7. Bibliografía

American Cancer Society. (2018a). Tipos de síndromes mielodisplásicos.

American Cancer Society. (2018b). What Are Myelodysplastic Syndromes? Retrieved from https://www.cancer.org/es/cancer/sindrome-mielodisplasico/acerca/que-es-sindrome-mielodisplasico.html

Anglada, L., & Abadal, E. (2018). ¿ Qué es la ciencia abierta? *Anuario ThinkEPI*, *12*, 292–298.

Asociacion Española de Afectados por Linfoma, M. y L. (AEAL). (n.d.). Qué son los Síndromes Mielodisplásicos. Retrieved from http://www.aeal.es/sindromes-mielodisplasicos-espana/4-que-son-los-sindromes-mielodisplasicos/

Belli, C. B., Bestach, Y. S., Prates, M. V., Sakamoto, F., Alfonso, G., Rosenhain, M., … Larripa, I. B. (2014). Aplicación del Sistema Pronóstico Internacional revisado (IPSS-R) en 511 pacientes con Síndromes Mielodisplásicos de población Argentina. *Hematología*, *18*(1), 17–25.

Belli, C. B., & Larripa, I. B. (2007). CLASIFICACIONES Y SISTEMAS PRONOSTICOS EN SINDROMES MIELODISPLASICOS. *Sociedad Iberoamericana de Informacion Científica*. Retrieved from https://www.siicsalud.com/des/expertoimpreso.php/71509

Couture-Beil, A. (2018). *JSON for R*. Retrieved from https://cran.r-project.org/package=rjson

Cox, D. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Dowle, M., & Srinivasan, A. (2021). *Extension of "data.frame."* Retrieved from https://r-datatable.com

Fernández, P. (1995). Análisis de supervivencia. Unidad de Epidemiología Clínica y Bioestadística. Cad Aten Primaria, 130-135.

Garrigues, F. (2017). Sanger: Estrategia de secuenciación de Primera Generación.

Grupo español de síndromes mielodisplásicos. (2017). *Guía de aplicacion clínica de las secuenciación masiva en síndromes mielodisplásicos y leucemia mielomonocítica crónica*.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Jiménez, S. I. (2016). Síndrome mielodisplásico. Un reto en medicina clínica-hematología. *Acta Médica Colombiana*, *41*(1), 16–18.

Masuzzo, P., & Martens, L. (2017). Do you speak open science? Resources and tips to learn the language. *PeerJ Preprints*.

Observational Health Data Sciences and Informatics. (2019a). Observational Medical Outcomes Partnership. In *The Book of OHDSI* (pp. 5–6).

Observational Health Data Sciences and Informatics. (2019b). Standardized Vocabularies. In *The Book of OHDSI* (pp. 55–73).

Observational Health Data Sciences and Informatics. (2019c). The Common Data Model. In *The Book of OHDSI* (pp. 31–53).

Ooms, J., James, D., DebRoy, S., Horner, H., & Wickham, J. (2020). *RMySQL: Database Interface and*

"MySQL" Driver for R. Retrieved from https://cran.r-project.org/package=RMySQL

Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., & Stang, P. E. (2012). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, *19*(1), 54–60.

Pol, A. L. P. (1993). Modelo de regresión de Cox: ejemplo numérico del proceso de estimación de parámetros. Psicothema, 387-402.

Repáraz, A., & Torreira, C. (2018). Arrays de CGH para el estudio de la discapacidad intelectual y trastornos del espectro autista. *Ed. Cont. Lab. Clin*, *37*, 9–16.

Sanz G. (2012). En Guías españolas de diagnóstico y tratamiento de los síndromes mielodisplásicos. . Abril , Vol 97, sup 5: 5-6. *Haematologica*, *97*(5), 5–6.

Sánchez-Cantalejo Ramírez, E., & Sánchez-Cantalejo Castañeda, J. (2014). Análisis de supervivencia. In Serie Cuadernos metodológicos.

Shang, W., Adams, B., & Hassan, A. E. (2012). Using Pig as a data preparation language for large-scale mining software repositories studies: An experience report. *Journal of Systems and Software - JSS*, *85*.

Spinu, V., Grolemund, G., & Wickham, H. (2021). *Make Dealing with Dates a Little Easier*. Retrieved from https://lubridate.tidyverse.org

Stang, P. E., Ryan, P. B., Racoosin, J. A., Overhage, J. M., Hartzema, A. G., Reich, C., … Woodcock, J. (2010). Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine*, *153*(9), 600–606.

Wickham, H., Averick, M., Bryan, J., Winston D'Agostino McGowan, L. C., François, R., Grolemund, G., … Miller, E. (2021). *Easily Install and Load the "Tidyverse."*

Wickham, H., & Bryan, J. (2019). *Read Excel Files*. Retrieved from https://readxl.tidyverse.org

Wikipedia contributors. (2021a). Auer rod. Retrieved from Wikipedia, The Free Encyclopedia website: https://en.wikipedia.org/w/index.php?title=Auer_rod&oldid=1015841619

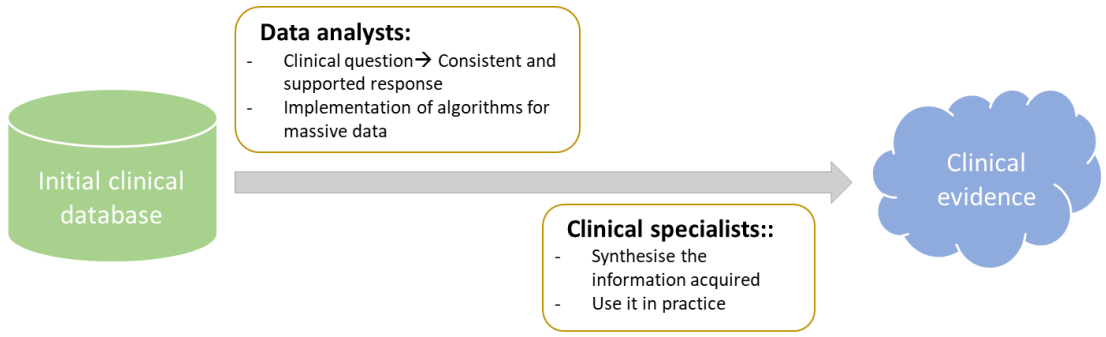
Wikipedia contributors. (2021b). Hibridación fluorescente in situ. Retrieved from Wikipedia, The Free Encyclopedia website website: https://es.wikipedia.org/w/index.php?title=Hibridación_fluorescente_in_situ&oldid=133336439.

8. SUMMARY

Introduction

One of the main goals of medicine in recent decades has been that of obtaining clinical evidence through the study of previous data. This requires a compilation of data in order to perform analyses that will provide extrapolable conclusions.

This is made possible through the involvement of two distinct academic profiles: data analysts, who try to organise and analyse large amounts of data by extracting answers to the questions posed by clinical specialists, who are responsible for evaluating and applying the results in practice.



This process is not straightforward due to the diversity of database types that exist; This causes difficulties, both in terms of analysis and understanding on the part of the researchers and making it necessary to adapt the analysis to each database. As a result of this problematic / problem, the need arises to create a common data model.

The Observational Medical Outcomes Partnership (OMOP) was responsible for conducting a pilot test to analyse the benefits of using a common model (CDM). The objective of the study was to improve the difficulties raised in the previous paragraph and the results were favourable. This resulted in the need to create a community to continue the progress of the project, OHDSI (Observational Health Data Science and Informatics).

An important characteristic of this type of relational database is that it is built as "open science", i.e. it is free and open for everyone to use, as its main objective is to generate large-scale medical evidence.

Its structure is designed for the collection of observational data on both medical interventions and their outcomes. It contains a series of dominoes that are interrelated through the unique patient identifier. The standard model is made up of numerous tables that collect the different data. These data are coded by means of a standardised vocabulary that is collected in the ATHENA platform.

One of the advantages associated with the use of the OMOP model is the capacity to store all of the vocabularies employed and not just a standardisation of these. And the possibility of adding new variables thanks to its structure, without having to modify the number of fields in the tables that make up the standard.

In order to demonstrate the potential of using the CDM OMOP, data on patients with Myelodysplastic Syndromes (MDS) will be used.

MDS are disorders that occur when blood-producing cells become dysplastic, that is, abnormal cells, resulting in a reduction in the number of one or more types of cells in the blood. Sometimes this disease progresses into acute myeloid leukaemia (AML), a fast-growing type of cancer, which

is why in the past this disease was known as pre-leukaemia. Today we know that the majority of patients do not progress into AML and therefore it is considered a disease of low malignant potential, differing from solid cancers in that the concern is not its spread but rather organ failure.

There are several types of MDS; the first FAB classification was established in 1982 and is based on morphological criteria. This classification served as a reference for the current WHO classification, which combines morphology, cytochemistry and cytogenetics.

To evaluate the prognosis of the disease, the variables that have the greatest impact are the percentage of blasts in bone marrow, cytogenetic analysis and the number of peripheral cytopenias.

The IPSS (International Prognostic System) is used which, depending on whether it is the standard or the revised version, uses 4 or 5 risk groups respectively for the classification.

In order to classify the diagnosis of the type of MDS as well as its risk prognosis, several studies must be performed, which include Bone Marrow, Peripheral Blood, Karyotype and FISH (Fluorescence In Situ Hybridisation), studies which have been conducted on the individuals in this analysis and which will be stored in the CDM OMOP.

Objectives

The main objective of this work is to observe the potential of implementing the OMOP common data model in the database provided by the IBSAL on MDS (Myelodysplastic Syndrome) patients. This objective is intended to conclude a positive or negative result regarding the use of this type of database.

The second objective is to analyse the above data using statistical survival techniques. The main aim is to use the data to corroborate the conclusions that appear in the literature on MDS.

Materials

The *Instituto de Investigación de Salamanca* (IBSAL) has provided for this study a database containing a total of 2882 myelodysplastic syndrome patients from hospitals all over Spain. Some of these individuals have several samples collected over time, thus constituting a database comprising 3241 results.

The patient data contains information on both the individual and the medical team by which he/she has been treated; this data is confidential and irrelevant for this study and will therefore be disregarded.

Among the data that form the database are those necessary to carry out the FAB and WHO classifications, and IPSS and IPSS-R prognoses. As mentioned above, these include Bone Marrow, Peripheral Blood, Karyotype, FISH and Sequencing studies.

Other studies are also carried out in which data are collected and used in the diagnosis and evolution of the disease.

Together with the studies, it contains the projects in which each of the samples of the patients who are part of the study are found.

We have data on the patient's type of diagnosis, whether it is secondary, and what type it is, the date of the patient's initial diagnosis, the disease progression of both MDS and LAM, the date of the patient's last follow-up and their status, whether the patient is alive or dead.

The data on the patients and their samples are collected in different Excel files; all these data must be processed and collected in the OMOP database and for this purpose it is necessary to carry out an ETL (Extract, Transform and Load) process.

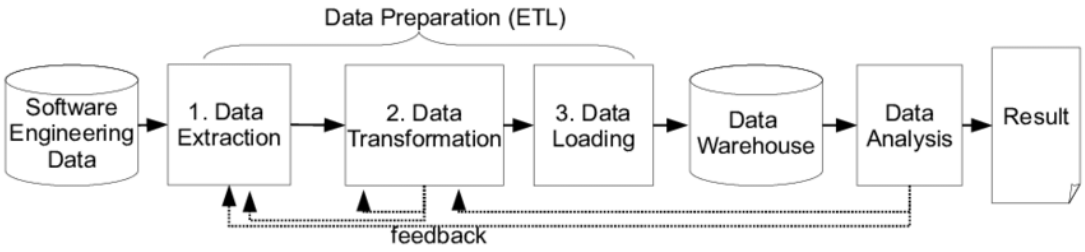
To carry out this ETL, the R program, RStudio, is used, being necessary the previous installation of the libraries: Lubridate, Tidyverse, Readxl, Date.table, RMySQL and Rjson.

The Survival and Survminer libraries are also necessary, both to conduct the survival analyses and to graph them.

Methodology

In order to explain the process carried out for data loading and analysis, it is necessary to divide this section into two parts. The first part describes the ETL (extraction, transformation and loading) process, and the second part outlines the concepts and theoretically analyses the survival analysis used in the study.

ETL is a type of data integration used to combine data from different sources. It consists of three phases; extracting the source data, transforming it into a format that can be analysed, and uploading or storing it in a database.



In the first phase, the data are extracted from the source systems, checked to verify that they comply with the expected structure, and once this is done, the next phase of transformation takes place. It is important that, in this first step, the impact on the data is kept to a minimum. In the second phase, data cleansing and coding is performed, in which the initial information is processed to become the standardised model that can be finely loaded in the Load step, the final step of the ETL.

Taken to the specific case of the study, which aims to demonstrate the storage capacity of OMOP databases, we start from a set of documents that need to be unified in a single dataframe. unified in a single dataframe. This ETL process will be carried out using RStudio. Once all the all the variable data together, the concepts are encoded thanks to the ATHENA platform, while these concepts are also encoded by the ATHENA platform and at the same time they are stored following the common structure of the OMOP model. In order to be able to load the data into MySQL, it is necessary to create a function that stores and reports the modifications in the database.

Once the data is loaded, the survival analysis can begin. Among the most important basic concepts of this type of analysis are the survival time, which is the time that elapses from the entry of an individual into the study until this event occurs, the individual leaves the study or the study ends, and censoring, which occurs when, over the time of the study, the patient is lost track of, i.e., no more data is available on the patient.

The survival function can be defined as the probability that an individual survives to at least time t, having the following form for continuous variables:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f(u)du \tag{8.1}$$

And for discrete variables:

$$S(t) = P(T \geq t) = \sum_{t_j \geq t} f(t_j) \tag{8.2}$$

The instantaneous risk function Hazard Function indicates the instantaneous probability that, if an individual is in the study at time *t*, the event will occur at an instant of time very close to *t*. It is defined as:

$$h(t) = lim_{\Delta t \to 0} \frac{P(t < T \leq t + t\Delta | T > t)}{\Delta t} \tag{8.3}$$

In the discrete case:

$$h(t_j) = P(T \geq t_j), \ \ j = 1,2, \ldots \tag{8.4}$$

Non-parametric models allow the interpretation of the data obtained without having to assume any specific probabilistic model for the survival times and the functions described in the previous sections, being estimated directly from the data, without having to make major assumptions prior to the model. assumptions prior to the model.

Moreover, these non-parametric methods allow us to work with censored data, which, although their information is incomplete, are useful and should not be disregarded.

The Kaplan-Meier estimator is based on the decomposition of the survival curve into a conditional probability product.

$$S(t) = \prod_{t_i \leq t)} (1 - {}^{d_{t_i}}\!/_{r_{t_i}}) \tag{8.5}$$

where d_{t_i} is the number of deaths in time and are the people at risk immediately before that time.

The logarithmic Log-Rank test compares the number of observed events in each group (O_i) with the number of expected events in the group (E_i) using the Chi-square statistic with k-1 degrees of freedom, where k is the number of groups.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{8.6}$$

Both this test and the Kaplan-Meier method are univariate methods, i.e. they describe survival without taking into account the joint effect of the predictor variables on the response variable. response variable.

To calculate the estimation of Kaplan-Meier survival curves in R, **the surfit()** function included in the survival package described in the R Libraries section can be used.

```
survfit(Surv(time, status) ~ sex, data = data, Type = "Kaplan-meier")
```

The parametric methods aim to approximate the data from the analyses to the functions provided by the different distributions, so that they can be used as survival functions.

The exponential distribution expresses a constant risk over time; its survival function is equal to function is equal to:

$$S(t) = \int_t^{\infty} f(u)du = \int_t^{\infty} \theta e^{-\theta u} du = e^{-\theta u}|_t^{\infty} = e^{-\theta t} \tag{8.7}$$

The Weibull distribution is a generalisation of the exponential model and can be considered the most widely used parametric distribution in survival analysis.

The survival function for this distribution would be:

$$S(t) = \int_t^{\infty} f(u)du = \int_t^{\infty} \beta\lambda u^{\beta-1} \exp\{-\lambda u^{\beta}\}\, du = -\exp\{-\lambda u^{\beta}\}|_t^{\infty} = \exp\{-\lambda t^{\beta}\} \tag{8.8}$$

To estimate the survival function in R across these distributions, it is necessary to use the **frexsurvreg()** function. The parameters are estimated by maximum likelihood.

```
flexsurvreg(Surv(x,y)~1, data = data, cl = 0.95, dist ="exp")
```

The Cox model is described as a semi-parametric model because it has two components: the time-dependent function is estimated non-parametrically, which is called the baseline risk function, and the exponential of the linear predictor, where the parameter vector is estimated by maximising the partial likelihood function.

It allows estimating the relationship between a set of explanatory variables, called covariates, with the instantaneous rate of the event of interest, i.e. the hazard function.

This model simultaneously assesses the effect of several factors on survival and is expressed by the hazard function:

$$h(t;\, x_1, x_2, \ldots, x_p) = h_0(t) \cdot \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\} \tag{8.9}$$

where:

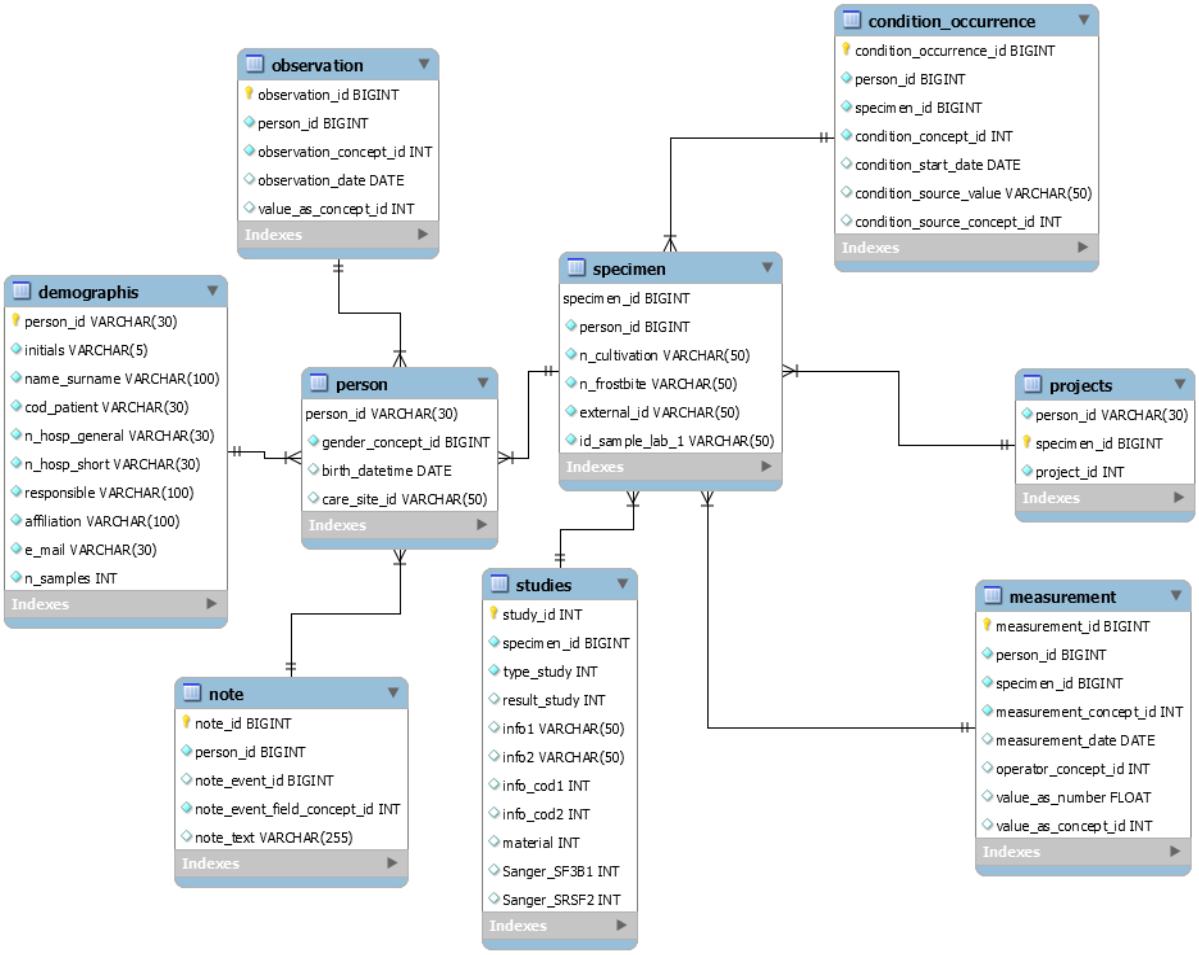
- $h(t)$ is the hazard function determined by the p covariates.
- The coefficients β_i measure the impact of these covariates.
- And the term $h_0(t)$ is the baseline risk, which involves time, not the predictor variables.

The **coxph()** function found in the **survival** package is used to run the Cox semi-parametric model. The parameters to be included are the formula, also created with the **Surv()** function, the data and the method, which defaults to the more computationally efficient "efron".

```
coxph(Surv(time, status) ~ sex, data = data, method = “efron”)
```

Results and discussion

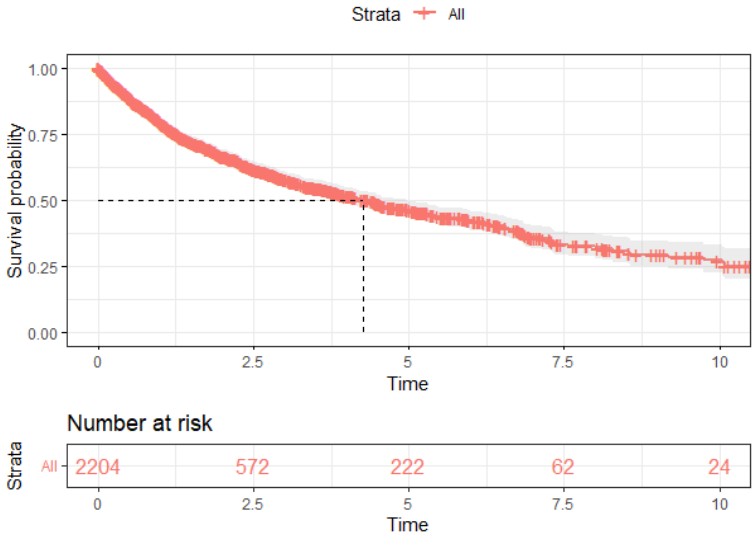
In the section on Extraction, Transformation and Uploading of the data, the development that forms the database has been explained, a process that has been followed in order to obtain the data stored following the OMOP Common Data Model as shown in the image.



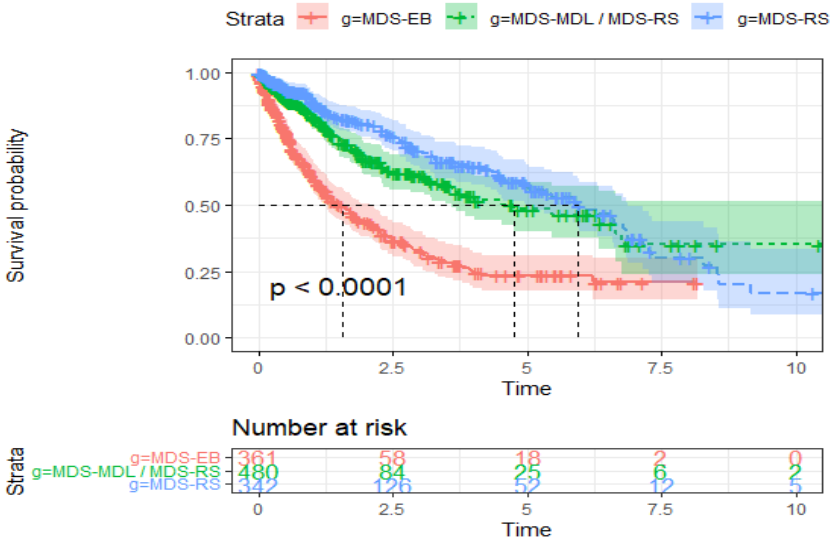
The descriptive analyses that characterise our study population show that the majority of the individuals are over 65 years of age, male, and that the most recurrent types of MDS are MDS with multilineage dysplasia and MDS with ring syderoblasts and MDS with excess blasts. It is noteworthy that most individuals do not progress to AML and that the IPSS-R risk is generally low.

In the non-parametric survival analyses we obtained with Kaplan-Meier that there are significant differences in the survival of individuals of different genders, being better in women, in hospital transfer, with transfer being favourable, and in the progression to AML, as is logical, the survival of those who do not progress is greater.

The overall Kaplan-Meier shows that there are a large number of patients in the study who are censored. Despite this, it can be seen that the probability of survival is most affected in the first 2.5 years in which the patient has MDS, stabilising after 5 years.

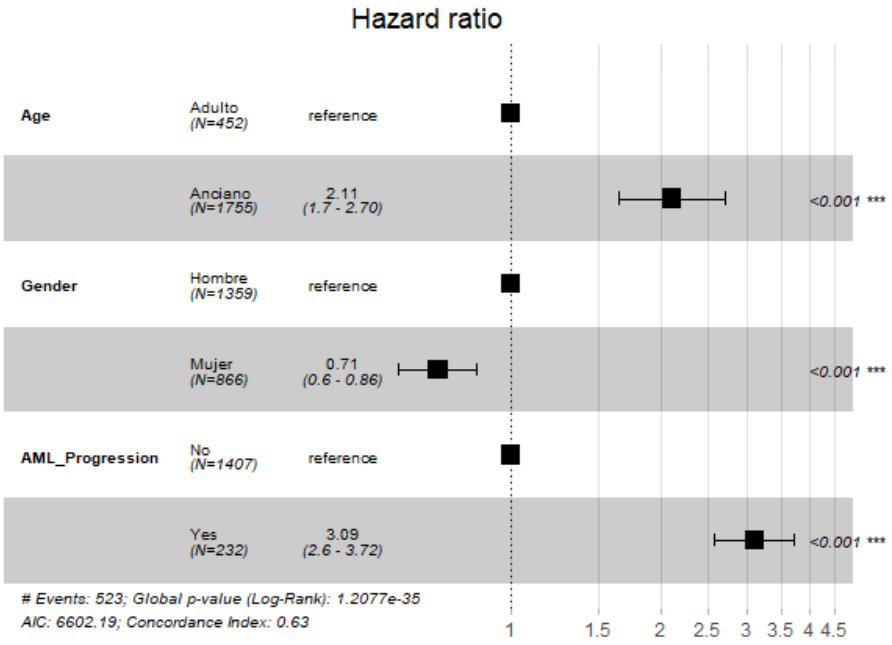


A graph that can be very interesting is the one that relates the survival curve to the type of MDS diagnosed. By selecting patients with the most prevalent types, we obtain that there is a greater chance of death in the first years of patients with MDS-EB.



We also wanted to check whether the survival curve of the patients in this study fits any of the parametric functions. Using the AIC mathematical method, we know that the Weibull distribution fits better than the exponential distribution.

The Cox model was run and the significant variables were age, gender and AML progression. We observe that the proportional hazards assumption, calculated thanks to the Schoenfeld test, is supported by the variables Age and Gender and progression to AML, i.e. that the risk ratio of these variables is independent of time.



We observe in the graph that the only "protective" effector is being a woman, as her Hazard ratio is less than 1, being very remarkable as it greatly increases, 3.2 points, the progression of the disease to AML for the risk of non-survival, and an age over 65 by slightly more than two points.

Conclusions

As a conclusion to the first objective, the OMOP Common Data Model allows the unification of databases in a single model, facilitating the collection, processing and subsequent analysis of the data. This leads to an absolute improvement in the process of obtaining clinical evidence; being evident the numerous medical advantages that this common model would entail if it started to be globalised quickly, something that is happening thanks to the OHDSI community, a community in constant action.

As a second objective in the survival analysis, patients with MDS have a fairly high survival in the first 4 years after the disease is diagnosed, patients who progress to AML have a marked decrease in their chances of survival during the first years of disease, patients with the MDS type with excess blasts have the highest risk of their disease progressing to AML, therefore, their survival curve is very steep y the factors influencing the survival calculation are age, sex and progression to AML, knowing that this model is supported by the proportional hazards assumption, and that the survival curve is very steep. is supported by the proportional Hazards assumption.