
# Tratamiento de missings en estudios de cohorte

Trabajo de Final de Máster

**Máster en Bioinformática y Bioestadística**

**Autor:** Mar Harmut Prats

**Tutor/a:** Núria Pérez

**Fecha de entrega:** 24 de mayo de 2017

# Tabla de contenido

# Parte I

## 1. Introducción

### 1.1. Motivación del trabajo

La presencia de datos faltantes, en adelante *missings*, en ensayos clínicos, estudios longitudinales y más específicamente en los estudios de cohorte son un problema habitual y recurrente con el cual los investigadores deben lidiar. Estos *missings* son susceptibles de ser causa de un sesgo en el análisis y una pérdida de poder estadístico.

Las razones por las cuales es posible que se produzca la ausencia de datos puedes de diversa índole. En estudios de cohorte, en los cuales suele diseñarse un protocolo que contempla diferentes visitas durante el periodo de tiempo que dura el estudio, es habitual perder la pista de los pacientes o bien por no comparecencia, abandono o, en los casos más graves, muerte del paciente. Otras causas por las cuales pueden producirse estos *missings* son: fallos en los instrumentos de medida, no contestan a una serie de preguntas o responden con la opción no sabe/no contesta incorporada en el cuestionario, etc.

Los estudios de cohorte se caracterizan principalmente por que los sujetos de estudio se eligen de acuerdo con la exposición de interés. La recogida de datos de manera reiterada durante el período de seguimiento de los pacientes proporciona una información muy valiosa para los investigadores, relacionada con los cambios experimentados, a raíz de la exposición a un factor determinado, en diferentes estadios de tiempo en grupos de individuos.

Así pues, se propone un proyecto consistente en la gestión y análisis de una base de datos correspondiente al Registro de Seguridad SORASE, llevado a cabo por el Grupo Internacional Ferrer, S.A., iniciado en 2013 en México, por tratarse de un registro de tipo longitudinal de una cohorte de pacientes tratada con un fármaco combinado (*polypill*), para la prevención secundaria de eventos cardiovasculares, siendo necesaria una correcta gestión de los datos faltantes existentes.

La adecuada planificación del análisis de *missings* se traduce en los siguientes objetivos planteados para el desarrollo de esta tesis:

- Revisión bibliográfica acerca de los diferentes tipos de datos faltantes.
- Investigar acerca de las metodologías y estrategias principalmente utilizadas para el manejo de *missings*, de manera general y más concretamente en estudios de cohorte.
- Aplicación de los métodos introducidos en la gestión y análisis del Registro SORASE en el entorno del paquete estadístico SPSS 23 para Windows.

## 1.2. Definición de enfermedad cardiovascular, tipología y tratamiento habitual

Las enfermedades cardiovasculares (ECV) constituyen un conjunto de trastornos que afectan al corazón y a los vasos sanguíneos, representando la principal causa de muerte a nivel mundial. Dichas enfermedades presentan una mayor incidencia en países de ingresos medio-bajos (más del 80% de las defunciones por esta causa se producen en esos países). Generalmente son debidas a la acumulación de colesterol y grasa en las paredes de los vasos sanguíneos (arterosclerosis), provocando su estrechamiento y, a largo plazo, generando problemas a nivel sistémico.

Algunos de los principales trastornos cardiovasculares descritos por la Organización Mundial de la Salud son:

- La cardiopatía coronaria
- El Accidente Cerebrovascular (ACV)
- La Cardiopatía hipertensiva
- La Cardiopatía isquémica
- El Infarto de miocardio
- Angina

En su mayoría las ECV tienen un origen multifactorial, siendo los principales factores de riesgo asociados una dieta desequilibrada, el sedentarismo, la obesidad, el consumo de tabaco o alcohol o niveles anormales de colesterol y triglicéridos, entre otros.

Para las personas con ECV o con alto riesgo cardiovascular (debido a la presencia de uno o más factores de riesgo), son fundamentales la detección precoz y el tratamiento temprano, por medio de servicios de orientación o la administración de fármacos, según corresponda.

### *1.2.1. PRESCRIPCIÓN HABITUAL PARA LA PREVENCIÓN Y TRATAMIENTO DE ECV*

A nivel individual, las intervenciones sanitarias de prevención de los primeros ataques cardíacos y accidentes cerebrovasculares, deben centrarse primordialmente en las personas que, si se tienen en cuenta todos los factores, presentan un riesgo cardiovascular medio a alto o en los individuos que presentan un solo factor de riesgo —por ejemplo, diabetes, hipertensión o hipercolesterolemia— con niveles superiores a los umbrales de tratamiento recomendados. La primera intervención (basada en un enfoque integral que tiene en cuenta todos los riesgos) es más rentable que la segunda y tiene el potencial de reducir sustancialmente los episodios cardiovasculares. Se trata de un enfoque viable dentro de los servicios de atención primaria en entornos de escasos recursos, que puede ser puesto en práctica incluso por trabajadores sanitarios que no son médicos.

Para la prevención secundaria de enfermedades cardiovasculares en pacientes con diagnóstico definitivo, por ejemplo, de diabetes, es necesario administrar tratamientos con los siguientes fármacos:

- ácido acetilsalicílico;
- betabloqueantes;
- inhibidores de la enzima convertidora de la angiotensina;

- estatinas.

La complejidad de las enfermedades de tipo cardiovascular conlleva, en muchos de los casos, la necesidad de combinar diferentes tratamientos a la vez. El correcto cumplimiento del tratamiento se asocia a resultados positivos para la salud del paciente mientras que la mala adherencia, aumenta la probabilidad de sufrir un nuevo evento cardiovascular lo cual puede implicar desde el empeoramiento de su enfermedad al fallecimiento del paciente.

## 1.3. Estudios de cohorte

Los estudios de cohorte son de tipo longitudinal, es decir, siguen la evolución de los pacientes en el tiempo. A diferencia de los ensayos clínicos aleatorizados donde el investigador asigna la exposición del sujeto a un determinado factor de manera controlada, en los estudios de cohorte los individuos participantes son elegidos de acuerdo con la exposición a este factor o a la predisposición a padecerlo. Una vez han sido clasificados, se recogen datos referentes a las variables de exposición y se sigue la evolución de la cohorte de manera que sea posible determinar si la aparición del evento de interés (o casos) difiere entre los pacientes expuestos y los no expuestos.

Los diferentes puntos en el tiempo en el cual se registran datos de los pacientes se conocen como *waves* (olas).

Figura 1 - Diagrama de un estudio de cohorte


Respecto al muestreo, algunos diseños de estudios de cohorte admiten que los pacientes sean incluidos en puntos alejados en el tiempo, siempre y cuando los datos que se registren sean previos al evento que es objeto de estudio. Esta es una de las principales razones por las cuales son un buen modelo para medir el riesgo de recurrencia de determinadas patologías.

Una parte importante de los estudios de cohorte es la correcta recogida de datos. Es imprescindible determinar si el paciente ha experimentado, o no, el evento que es objeto de estudio, si está en medio de un proceso diagnóstico o si ya ha estado expuesto al factor que se analiza. Es posible recoger los datos relativos a la exposición a partir de las historias clínicas de los pacientes, sus hábitos alimenticios y de vida, información relativa al área geográfica donde residen, etc. Por el contrario, la información relativa al evento de interés, suelen proceder tanto de la valoración de los investigadores en las visitar programadas, exámenes físicos, certificados de defunción, etc.

Dependiendo del tipo de evento que se desee estudiar, un diseño de tipo cohorte puede resultar sencillo y rápido (como es el caso de enfermedades provocada por exposiciones agudas, intensas y breves a determinados factores) o largo y costoso, ya que algunas patologías experimentan períodos prolongados de tiempo hasta llegar finalmente a manifestarse. En el caso de los resultados en casos de tipo agudo, suelen presentar una relación causa-efecto mucho más evidente que en casos de incubación prolongada. Un ejemplo exposición aguda, es la relación entre la ingesta de agua contaminada y las enfermedades de tipo diarreico. Por el contrario, un ejemplo de manifestación tardía, es el estudio de los efectos de la radiación nuclear (por accidente o fuga en una central) y determinados tipos de cáncer.

A continuación, se muestra un esquema de la distribución de individuos en un estudio de cohorte (en una tabla de contingencia 2x2), así como las principales tasas y estadísticos calculables:

Tabla 1. Esquema de la distribución de pacientes en un estudio de cohorte

| | **Evento** | **No-evento** | **Total** |
|---|---|---|---|
| **Expuestos** | A | B | A+B |
| **No-expuestos** | C | D | C+D |
| **Total** | A+C | B+D | A+B+C+D |

1. Tasa de incidencia:
   - Sobre pacientes expuestos: A/(A+B)
   - Sobre pacientes no expuestos: C/ (C+D)
2. Riesgo relativo: [Incidencia expuestos] / [Incidencia no expuestos]

La principal ventaja metodológica a la hora de optar por un diseño de tipo cohorte frente a otro diseño consiste en que los niveles del factor de estudio son evaluados durante el periodo de seguimiento de los pacientes, antes de que el evento de interés sea detectado. Así pues, es posible asumir que, la selección de los pacientes está ligada a el factor de riesgo y no a el estadio de la patología que es objeto de estudio. A continuación, se listan las principales ventajas e inconvenientes de este tipo de estudios.

Tabla 2. Resumen de las principales ventajas y desventajas de los estudios de cohorte

| Ventajas | Desventajas |
|---|---|
| 1. Estudiar factores de exposición poco comunes.<br>2. Establecer relaciones de causa-efecto, así como su secuencia temporal (entre la exposición y la manifestación del evento)<br>3. Poder examinar los efectos múltiples de una única exposición.<br>4. Tienden a limitar el sesgo en la averiguación de la exposición.<br>5. Calcular la tasa de incidencia del evento, así como su riesgo relativo en una determinada población | 1. El diseño es ineficiente para la evaluación de las enfermedades poco comunes ya que, a menudo, el tamaña muestral necesario es muy grande.<br>2. Pueden ser muy costosos y lentos.<br>3. La recogida de datos debe ser muy minuciosa.<br>4. La pérdida de seguimiento de los pacientes puede acarrear un sesgo importante, así como comprometer la validez del estudio. |

# Parte II

## 2. Datos faltantes en estudios de cohorte

Los estudios de cohorte suelen encontrarse ante la tesitura de contener datos faltantes (*missings*) en algunas variables. Esto implica que algunos datos de interés podrían no ser registrados en los individuos en las visitas programadas según el diseño del estudio. Esto puede venir dado por diversos motivos:  ya sea porque los individuos no son localizados, por eventos adversos, por razones administrativas, por fallo en los instrumentos de medida o simplemente porque los sujetos abandonan el estudio a partir de un cierto momento. Este último, es un caso especial de *missings* denominado *drop-out* (dejar o abandonar).

## 3. Procesos generadores de datos faltantes

Según Rubin (1976) se distinguen principalmente tres tipos de procesos generadores de *missings*:

- Datos perdidos completamente al azar (MCAR; *missing completely at random*)
- Datos perdidos al azar (MAR; *missing at random*)
- Datos perdidos no debidos al azar (MNAR; *missing not at random*)

### 3.1. Datos Perdidos Completamente al Azar o *Missing Completely at Random* (MCAR)

Hablamos de MCAR cuando la causa de la omisión de los datos no está relacionada con los mismos, es decir, las observaciones con datos perdidos son una muestra aleatoria del conjunto de observaciones. Puede definirse por la expresión:

$$\Pr(R_Z = 1|X, Z) = \Pr(R_Z = 1)$$

Donde Z representa la situación en la que hay una única variable con datos faltantes, X se corresponde con otro conjunto de variables que siempre es observado y RZ se trata de una variable indicadora (*dummy*) que adopta el valor de 1 si está ausente y 0 si es observada. Podemos traducir esta expresión como el hecho de que probabilidad de que Z falte no depende ni de las variables observadas (X) ni de los valores posiblemente ausentes de Z mismo.

Así pues, se considera que los datos perdidos son MCAR cuando las características de los sujetos con información son las mismas que las de los sujetos sin información. En este caso, la probabilidad de que un sujeto presente un valor faltante en una variable no depende ni de otras variables del cuestionario ni de los valores de la propia variable con valores perdidos.

Algunos ejemplos sencillos de MCAR son:

- Las personas que no nos proporcionan su salario tienen, en promedio, el mismo salario que las personas que nos lo proporcionan.
- Las características estadísticas (media, porcentajes) del resto de las variables son las misma para los sujetos que nos proporcionan su salario y para los que no lo proporcionan.

## 1.2. Datos perdidos al azar o *missing at random* (MAR)

Continuando con la asunción anterior de que Z es un caso con una única variable con datos faltantes y X el vector de variables siempre observadas, nuestra ecuación se convierte en:

$$\Pr(R_Z = 1|X, Z) = \Pr(R_Z = 1|X)$$

En este caso podríamos deducir que los *missings* contenidos den Z dependen de los datos registrados en X, pero no dependen de Z en sí mismos.

Así pues, la pérdida de datos es MAR cuando los sujetos con datos incompletos son significativamente diferentes de los que presentan datos completos en alguna variable, y el patrón de ausencia de datos puede ser predecible a partir de variables con datos observados completos. En este caso, la probabilidad de que se produzca la ausencia de una observación depende de otras variables, pero no de los valores de la variable con el valor ausente. Es imposible probar si la condición MAR es satisfecha y la razón es que dado que no conocemos la información faltante no podemos comparar los valores de aquellos sujetos que tienen información con los que no la tienen.

Un ejemplo de MAR sería la ausencia de valores en la variable sueldo si depende del estado civil, pero dentro de cada categoría, no estando la probabilidad de dicha ausencia relacionada con el sueldo en sí mismo.

## 1.3. Datos perdidos no debidos al azar o *missing not at random* (MNAR)

En el caso de que las asunciones de que los datos faltantes MCAR y MAR no pueden ser confirmadas, los datos son MNAR. En este caso, la pérdida de datos se corresponde con la probabilidad de que los datos perdidos sobre una variable Y dependa de los valores de dicha variable una vez que se han controlado el resto de las variables. Un ejemplo de este mecanismo sería la consideración de si los hogares de renta mayor son los que con menos probabilidad nos proporcionan el salario, una vez controladas el resto de las variables, entonces la pérdida de datos no es aleatoria ni ignorable.

# 2. Estrategias para el manejo de missings

El objetivo de cualquier análisis suele ser obtener estimaciones aquellos parámetros que describen alguna característica de la población que es objeto de estudio. Una vez obtenidas estas estimaciones, el investigador suele calcular algún indicador de la precisión de los estimadores. Finalmente, se llevan a cabo pruebas de contraste de hipótesis con la mayor potencia. La presencia de *missings* pueden alterar todas estas etapas debido a: sesgos en las estimaciones y disminución de su precisión, inducir a pérdida de la potencia de los test y alterar el nivel de significación.

Tratando de solucionar los problemas anteriormente planteados, se han propuesto en las últimas décadas diversos métodos para lidiar con los datos faltantes. Una primera aproximación al análisis, incorporando al análisis los datos faltantes, podría ser mediante alguno de los siguientes métodos:

- Análisis de datos completos (*listwise*)
- Análisis de datos disponibles (*pairwise*)
- Imputación por medias no condicionadas
- Imputación por medias condicionadas mediante métodos de regresión
- Máxima verosimilitud (MV)
- Imputación múltiple (IM)

## 2.1. Análisis de datos completos (*Listwise*)

En esta estrategia, aquellos casos que contienen datos faltantes son excluidos del análisis. El *listwise* es el método más habitualmente utilizado, llegando incluso a ser el método por defecto en determinados paquetes estadísticos como SPSS.

Al eliminar los casos no completos, se asume que éstos tienen las mismas características que los datos completos, y que la falta de respuesta se ha generado de manera aleatoria (premisa que habitualmente es falsa). Si la eliminación de casos no se acompaña con el ajuste apropiado, se obtendrán estimadores sesgados de los parámetros poblaciones lo que podría invalidar las conclusiones.

## 2.2. Análisis de datos disponibles (*Pairwise*)

El análisis de los datos disponibles constituye la alternativa más habitual al análisis por *listwise*.

Esta estrategia asume que los datos faltantes siguen una distribución MCAR. Se basa en la utilización de toda la información disponible sobre cada caso, eliminando aquellas observaciones que contienen *missings* y llevando a cabo los cálculos pertinentes con diferentes tamaños de muestra (limitando la comparación de resultados).

## 2.3. Imputación por medias no condicionadas

La sustitución de datos utilizando promedios es una metodología utilizada de manera habitual en el análisis de bases de datos y hay una extensa cantidad de literatura al respecto. Esta metodología asume que los datos faltantes siguen un patrón MCAR.

La imputación por medias no condicionadas trata de rellenar los registros vacíos con información de campos con información completa, de manera que los datos faltantes se reemplazan a partir de una selección aleatoria de valores observados.

Entre los principales efectos de aplicar esta estrategia está la afectación de la distribución de probabilidad de la variable imputada, la atenuación de la correlación de dicha variable respecto al resto de variables y la subestimación de la varianza. Esto es así puesto que, al incrementar el tamaño muestral de manera artificial, los estadísticos que definen su distribución se ven alterados.

## 2.4. Imputación por medias condicionadas mediante métodos de regresión

En el caso de que los datos faltantes sean MCAR, es posible utilizar modelos de regresión como estrategia para tratar de imputar información en una variable determinada (Y) a partir de covariables ($X_1$, $X_2$, …$X_n$) correlacionadas. Este método es especialmente útil con variables de tipo continuo.

Entre los principales inconvenientes de esta estrategia encontramos que, principalmente altera la relación de la variable Y con el resto de variables. Además, y al igual que en la imputación por medias, al incrementar la N, se altera la varianza de la distribución (la subestima).

## 2.5. Algoritmo esperanza-maximización (EM)

El EM se basa en encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables.

En este método, cada iteración consiste en un paso “E” y un paso “M”. El paso “E” encuentra la expectativa condicional de los datos "faltantes", dados los valores observados y las estimaciones actuales de los parámetros. Estas expectativas se sustituyen por los datos "perdidos". En el paso “M”, las estimaciones de máxima verosimilitud de los parámetros se calculan como si se hubieran rellenado los datos faltantes.

## 2.6. Imputación múltiple (IM)

El objetivo de la imputación múltiple es generar valores posibles para los valores perdidos, creando así varios conjuntos de datos "completos". Utiliza métodos de Monte-Carlo y sustituye los datos faltantes a partir de un número (m>1) de simulaciones. La metodología consta de varias etapas, y en cada simulación se analiza la matriz de datos completos a partir de métodos estadístico convencionales y posteriormente se combinan los resultados para generar estimadores robustos, su error estándar e intervalos de confianza.

# Parte III

## 3. Registro SORASE

Se incluyeron en el estudio un total de 1286 pacientes, que fueron clasificados en función del riesgo cardiovascular y seguidos durante un periodo total de 12 meses.

La depuración de la base de datos consistió en la recodificación de las variables de tipo cadena introducidas por los médicos, la agrupación de categorías similares, la resolución de queries y la comprobación de la hipótesis de normalidad de las variables.

Tabla 3. Clasificación de los pacientes según riesgo cardiovascular

| | Basal | | Mes 1 | | Mes 3 | | Mes 12 | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| Con antecedente de evento CV | 628 | 48,9% | 569 | 48,2% | 542 | 47,7% | 464 | 47,0% |
| Alto riesgo CV | 243 | 18,9% | 226 | 19,2% | 220 | 19,4% | 182 | 18,4% |
| Riesgo intermedio-moderado | 414 | 32,2% | 385 | 32,6% | 374 | 32,9% | 341 | 34,5% |
| Total | 1285 | 100,0% | 1180 | 100,0% | 1136 | 100,0% | 987 | 100,0% |

De entre las variables de respuesta contempladas en el estudio, analizaremos los relativas a la puntuación Morisky-Green relativa al cumplimiento del paciente respecto al tratamiento proporcionado *(polypill)*, ya que al combinar los diferentes tratamientos en una única píldora el cumplimiento debería ser mayor que en pacientes a los cuales se les recetan diferentes fármacos.

## 4. Causa de missings

Algunas de las razones por las cuales los pacientes discontinuaron el estudio fueron: razones económicas, abandono del estudio, reacciones adversas (dolor estomacal, irritación cutánea, etc.), fallecimiento del paciente, imposibilidad de re-contactar con el sujeto o el cambio de residencia.

En los tres grupos de riesgo, la disminución de pacientes a lo largo de las visitas de seguimiento de 1 mes, 3 meses y 12 meses fue similar.

# 5. Análisis descriptivo

Para cada paciente, se recogieron valores antropométricos y sociodemográficos en la visita basal.

Tabla 4. Descripción de variables

| Nombre | Descripción |
|---|---|
| gr_riesg | Grupo de riesgo de ECV |
| edad | Edad en años |
| sexo | Sexo del paciente (1=Femenino; 2=Masculino) |
| morgreen | Puntuación Morisky-Green obtenida por el paciente en la visita basal |
| morgreenc | Puntuación Morisky-Green (visita basal) clasificados como: 0=no cumplidor; 1=cumplidor |
| morgreen1 | Puntuación Morisky-Green obtenida por el paciente en la visita 1 mes |
| morgreenc1 | Puntuación Morisky-Green (visita 1 mes) clasificados como: 0=no cumplidor; 1=cumplidor |
| morgreen3 | Puntuación Morisky-Green obtenida por el paciente en la visita 3 meses |
| morgreenc3 | Puntuación Morisky-Green (visita 3 meses) clasificados como: 0=no cumplidor; 1=cumplidor |
| morgreen12 | Puntuación Morisky-Green obtenida por el paciente en la visita 12 meses |
| morgreenc12 | Puntuación Morisky-Green (visita 12 meses) clasificados como: 0=no cumplidor; 1=cumplidor |

Si llevamos a cabo un resumen *missings*, vemos que mientras que al inicio del estudio éstos representan menos del 5% de los registros, conforme avanzan las visitas, en las variables referentes al cumplimiento, los valores perdidos llegan a alcanzar el 25,7%, evidenciando la necesidad de un correcto tratamiento de dichos valores.

Tabla 5. Resumen de análisis de datos faltantes

| Estadísticos univariados | | | | | | | |
|---|---|---|---|---|---|---|---|
| | N | Media | Desviación estándar | Perdidos | | Número de extremos[a] | |
| | | | | Recuento | Porcentaje | Menor | Mayor |
| edad | 1286 | 57,4489 | 14,45310 | 0 | ,0 | 40 | 0 |
| gr_riesg | 1285 | | | 1 | ,1 | | |
| sexo | 1286 | | | 0 | ,0 | | |
| morgreen | 1232 | | | **54** | **4,2** | | |
| morgreenc | 1232 | | | **54** | **4,2** | | |
| morgreen1 | 1147 | | | **139** | **10,8** | | |
| morgreenc1 | 1147 | | | **139** | **10,8** | | |
| morgreen3 | 1106 | | | **180** | **14,0** | | |
| morgreenc3 | 1106 | | | **180** | **14,0** | | |
| morgreen12 | 955 | | | **331** | **25,7** | | |
| morgreenc12 | 955 | | | **331** | **25,7** | | |

A continuación, se muestra un resumen numérico de las variables mencionadas en la tabla anterior.

Tabla 6

| | | N | % | Media | DE | Mediana | Mínimo | Máximo |
|---|---|---|---|---|---|---|---|---|
| Grupo de riesgo CV | Con antecedente de evento CV | 628 | 48,9% | | | | | |
| | Alto riesgo CV | 243 | 18,9% | | | | | |
| | Riesgo intermedio-moderado | 414 | 32,2% | | | | | |
| Edad (años) | | | | 57,45 | 14,45 | 58,60 | 18,00 | 93,90 |
| Sexo | Femenino | 601 | 46,7% | | | | | |
| | Masculino | 685 | 53,3% | | | | | |
| Puntuación Morisky-Green (visita basal) | 0 | 114 | 9,3% | | | | | |
| | 1 | 128 | 10,4% | | | | | |
| | 2 | 175 | 14,2% | | | | | |
| | 3 | 225 | 18,3% | | | | | |
| | 4 | 590 | 47,9% | | | | | |
| Puntuacion Morisky-Green (visita basal): cumplidores | No cumplidores | 642 | 52,1% | | | | | |
| | Cumplidores | 590 | 47,9% | | | | | |
| Puntuación Morisky-Green (visita 1 mes) | 0 | 44 | 3,8% | | | | | |
| | 1 | 36 | 3,1% | | | | | |
| | 2 | 63 | 5,5% | | | | | |
| | 3 | 156 | 13,6% | | | | | |
| | 4 | 848 | 73,9% | | | | | |
| Puntuación Morisky-Green (visita 1 mes): cumplidores | No cumplidores | 299 | 26,1% | | | | | |
| | Cumplidores | 848 | 73,9% | | | | | |
| Puntuación Morisky-Green (visita 3 meses) | 0 | 43 | 3,9% | | | | | |
| | 1 | 39 | 3,5% | | | | | |
| | 2 | 85 | 7,7% | | | | | |
| | 3 | 133 | 12,0% | | | | | |
| | 4 | 806 | 72,9% | | | | | |
| Puntuación Morisky-Green (visita 3 meses): cumplidores | No cumplidores | 300 | 27,1% | | | | | |
| | Cumplidores | 806 | 72,9% | | | | | |
| Puntuación Morisky-Green (visita 12 meses) | 0 | 8 | 0,8% | | | | | |
| | 1 | 33 | 3,5% | | | | | |
| | 2 | 82 | 8,6% | | | | | |
| | 3 | 167 | 17,5% | | | | | |
| | 4 | 665 | 69,6% | | | | | |
| Puntuación Morisky-Green (visita 12 meses): cumplidores | No cumplidores | 290 | 30,4% | | | | | |
| | Cumplidores | 665 | 69,6% | | | | | |

# 6. Identificación de los missings

En primer lugar, testeamos la hipótesis de que nuestros datos sigan un mecanismo de tipo MCAR mediante el test de Little.

Tabla 7. Descriptiva del test de Little: edad, sexo, grupo de riesgo y cumplidores según MG

| Estadísticos univariados | | | | | | | |
|---|---|---|---|---|---|---|---|
| | N | Media | Desviación estándar | Perdidos | | Número de extremos[a] | |
| | | | | Recuento | Porcentaje | Menor | Mayor |
| edad | 1286 | 57,4489 | 14,45310 | 0 | ,0 | 40 | 0 |
| gr_riesg | 1285 | 1,8335 | ,88531 | 1 | ,1 | 0 | 0 |
| sexo | 1286 | 1,53 | ,499 | 0 | ,0 | 0 | 0 |
| morgreenc | 1232 | | | **54** | **4,2** | | |
| morgreenc1 | 1147 | | | **139** | **10,8** | | |
| morgreenc3 | 1106 | | | **180** | **14,0** | | |
| morgreenc12 | 955 | | | **331** | **25,7** | | |

En el análisis univariados descriptivo, vemos como el cuestionario de adherencia al tratamiento, a partir de la visita de 1 mes incrementan de manera importante la proporción de *missings* (>10,8%) representando el 25,7% en la visita de 12 meses.

Al llevar a cabo el test de Little, las variables introducidas en el modelo presentan una p=0,382, de manera que al ser p<0,05 podemos asumir que siguen una distribución MCAR.

# 7. Tratamiento de los missings

Una primera aproximación al análisis del dataset es el de Análisis por Intención de Tratar, en el cual todo paciente registrado en el estudio es analizado.

Tabla 8. Cumplidores por intención de tratar

| | | Cumplidores (visita basal) | | Cumplidores (visita 1 mes) | | Cumplidores (visita 3 meses) | | Cumplidores (visita 12 meses) | |
|---|---|---|---|---|---|---|---|---|---|
| | | N | % | N | % | N | % | N | % |
| Grupo de riesgo CV | Con antecedente de evento CV | 267 | 45,30% | 389 | 45,90% | 369 | 45,80% | 312 | 46,90% |
| | Alto riesgo CV | 110 | 18,60% | 161 | 19,00% | 144 | 17,90% | 113 | 17,00% |
| | Riesgo intermedio-moderado | 213 | 36,10% | 297 | 35,10% | 293 | 36,40% | 240 | 36,10% |
| Sexo | Femenino | 296 | 50,20% | 395 | 46,60% | 376 | 46,70% | 298 | 44,80% |
| | Masculino | 294 | 49,80% | 453 | 53,40% | 430 | 53,30% | 367 | 55,20% |

Vemos que al comparar los grupos de pacientes cumplidores según el cuestionario Morisky-Green la proporción de pacientes con antecedentes de ECV son los que presentan una mayor proporción de cumplimiento en la visita de 12 meses. El grupo de riesgo con un menor % de cumplimiento es el de alto riesgo CV. La proporción de pacientes cumplidores se incrementa en los 3 grupos de riesgo respecto a la visita basal.

Respecto al sexo del paciente, las mujeres presentan un menor % de cumplimiento respecto a los hombres a los 12 meses.

Si llevamos a cabo una prueba de $\chi^2$, obtenemos el siguiente resultado:

Tabla 9. Pruebas de $\chi^2$, por Intención de Tratar

| Pruebas de chi-cuadrado de Pearson | | | | | |
|---|---|---|---|---|---|
| | | Puntuación Morisky-Green (visita basal): cumplidores | Puntuación Morisky-Green (visita 1 mes): cumplidores | Puntuación Morisky-Green (visita 3 meses): cumplidores | Puntuación Morisky-Green (visita 12 meses): cumplidores |
| Grupo de riesgo CV | Chi-cuadrado | 9,732 | 7,683 | 13,470 | 2,276 |
| | gl | 2 | 2 | 2 | 2 |
| | Sig. | **,008*** | **,021*** | **,001*** | ,321 |
| Sexo | Chi-cuadrado | 5,564 | ,005 | ,784 | ,159 |
| | gl | 1 | 1 | 1 | 1 |
| | Sig. | **,018*** | ,942 | ,376 | ,690 |
| *. El estadístico de chi-cuadrado es significativo en el nivel ,05. | | | | | |

Otra aproximación es el análisis por protocolo, en la cual se analizan únicamente los casos completos.

Tabla 10. Cumplidores por protocolo

| | | Cumplidores (visita basal) | | Cumplidores (visita 1 mes) | | Cumplidores (visita 3 meses) | | Cumplidores (visita 12 meses) | |
|---|---|---|---|---|---|---|---|---|---|
| | | N | % | N | % | N | % | N | % |
| Grupo de riesgo CV | Con antecedente de evento CV | 202 | 44,5% | 310 | 44,1% | 310 | 44,7% | 308 | 47,0% |
| | Alto riesgo CV | 78 | 17,2% | 129 | 18,3% | 119 | 17,2% | 111 | 16,9% |
| | Riesgo intermedio-moderado | 174 | 38,3% | 264 | 37,6% | 264 | 38,1% | 237 | 36,1% |
| Sexo | Femenino | 220 | 48,5% | 327 | 46,5% | 324 | 46,8% | 295 | 45,0% |
| | Masculino | 234 | 51,5% | 376 | 53,5% | 369 | 53,2% | 361 | 55,0% |

En primer lugar, vemos como las N se han reducido con respecto al análisis por intención de tratar. El comportamiento de los datos es, en este caso, similar al del análisis previo.

Nuevamente, llevamos a cabo pruebas de $\chi^2$:

Tabla 11. Pruebas de $\chi^2$, por Protocolo

| **Pruebas de chi-cuadrado de Pearson** | | | | | |
|---|---|---|---|---|---|
| | | Puntuación Morisky-Green (visita basal): cumplidores | Puntuación Morisky-Green (visita 1 mes): cumplidores | Puntuación Morisky-Green (visita 3 meses): cumplidores | Puntuación Morisky-Green (visita 12 meses): cumplidores |
| Grupo de riesgo CV | Chi-cuadrado | 6,480 | 9,856 | 13,421 | 2,742 |
| | gl | 2 | 2 | 2 | 2 |
| | Sig. | **,039*** | **,007*** | **,001*** | ,254 |
| Sexo | Chi-cuadrado | 2,082 | ,007 | ,394 | ,196 |
| | gl | 1 | 1 | 1 | 1 |
| | Sig. | ,149 | ,935 | ,530 | ,658 |
| *. El estadístico de chi-cuadrado es significativo en el nivel ,05. | | | | | |

Obtenemos resultados similares a los del análisis por protocolo, no siendo en este caso el sexo significativo a nivel estadístico ($p>0.05$).

Una vez llevados a cabo el análisis del cumplimiento de los pacientes por intención de tratar (todos los pacientes incluidos en el estudio) y por protocolo (seleccionando aquellos lo completaron), procedemos a seleccionar qué método de imputación de datos faltantes es más conveniente utilizar en este dataset.

Creamos variables nuevas para el cumplimiento de cada visita, estableciendo que "0=*Missing*" y "1=Dato registrado". Comprobamos mediante pruebas de $\chi^2$ que, en el caso del cumplimiento, entre las visitas basal y la de 1 mes no hay asociación en los valores perdidos (p= 0,826), entre la de 1 mes

y la de 3 meses sí que la hay (p<0.0001) y la de 3 y 12 meses también (p<0.0001). No se observó asociación entre la visita basal y la de 12 meses (p=0.751).

Creamos un modelo de regresión siguiendo el GEE, con el cumplimiento a los 12 meses como variable dependiente, y el sexo, el grupo de riesgo y la edad como covariables, con la presencia de missings como respuesta.

Tabla 12. Modelo de GEE

| Variables en la ecuación | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | B | Error estándar | Wald | gl | Sig. | Exp(B) |
| Paso 0 | Constante | ,830 | ,070 | 139,083 | 1 | ,000 | 2,293 |

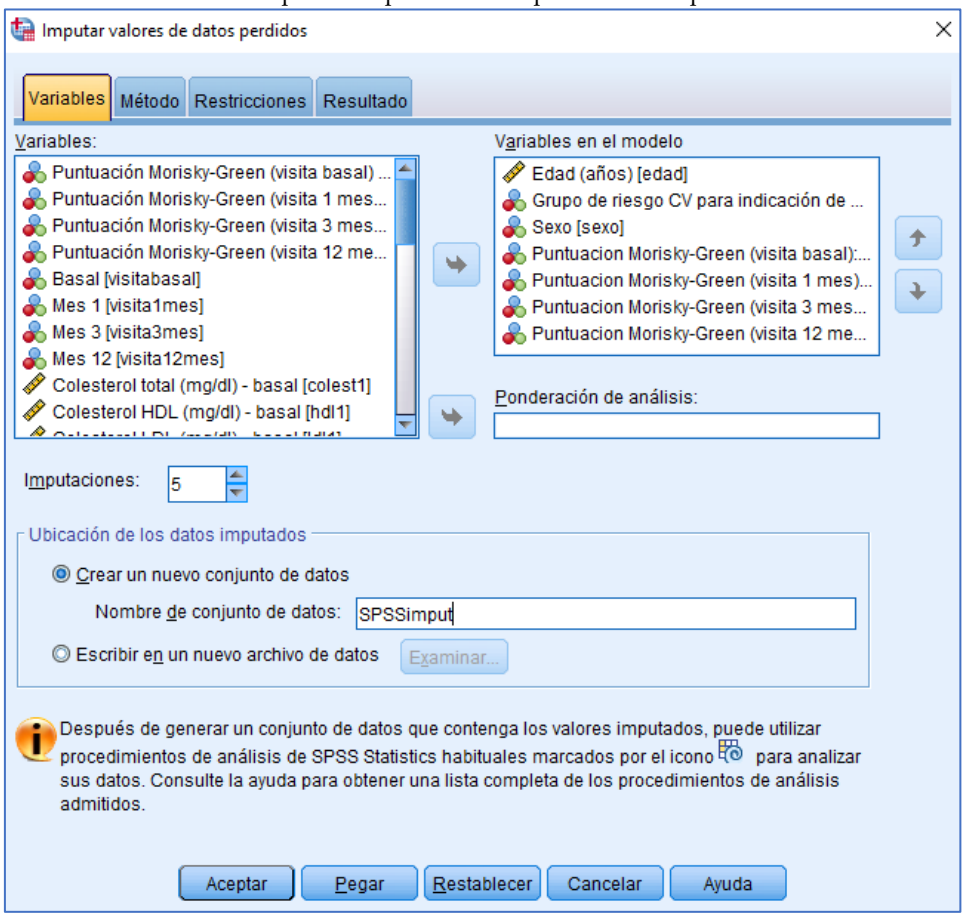
Observamos en la tabla 12 que no se incluyen en el modelo ninguna de las covariables introducidas.

De acuerdo a los resultados obtenidos en los test $\chi^2$ de y el GEE y teniendo en cuenta que el test de Little nos indicó que los datos eran MCAR, así pues, podemos proceder mediante un proceso de imputación múltiple y generar un modelo de regresión.

## 7.1. Análisis del dataset mediante el método de Imputación Múltiple

Tal y como hemos visto anteriormente, el objetivo de la imputación múltiple es generar valores posibles para los valores perdidos, creando así varios conjuntos de datos "completos". Los procedimientos analíticos que trabajan con conjuntos de datos de imputación múltiple producen resultados para cada conjunto de datos "completo", además de resultados combinados que estiman cuáles habrían sido los resultados si el conjunto de datos original no tuviera valores perdidos. En este caso, mediante el comando Analizar > Imputación múltiple > Imputar valores de datos perdidos. Especificamos que queremos generar 5 imputaciones.

Tabla 13. Captura del proceso de imputación múltiple en SPSS




Como resultado, generamos un dataset nuevo (SPSSimput) que contiene la variable "Imputation_", la cual consiste en los números de 0 a 5, referidos a la sesión de imputación en particular (Imputación = 0 se refiere al archivo de datos original).

Utilizando este nuevo dataset, generamos un modelo de regresión:

Tabla 14. Codificación de variables introducidas en la regresión.

| **Codificaciones de variables categóricas** | | | | |
|---|---|---|---|---|
| Número de imputación | | | Codificación de parámetro | |
| | | | -1 | -2 |
| 5 | Grupo de riesgo CV | Con antecedente de evento CV | 0,000 | 0,000 |
| | | Alto riesgo CV | 1,000 | 0,000 |
| | | Riesgo intermedio-moderado | 0,000 | 1,000 |
| | Sexo | Femenino | 0,000 | |
| | | Masculino | 1,000 | |
| | Puntuación Morisky-Green (visita 3 meses): cumplidores | No cumplidores | 0,000 | |
| | | Cumplidores | 1,000 | |
| | Puntuación Morisky-Green (visita 1 mes): cumplidores | No cumplidores | 0,000 | |
| | | Cumplidores | 1,000 | |
| | Puntuación Morisky-Green (visita basal): cumplidores | No cumplidores | 0,000 | |
| | | Cumplidores | 1,000 | |

En el proceso de codificación de las variables categóricas se crean dos variables nuevas: gr_riesg (1) y gr_riesg (2). El grupo de riesgo "Con antecedente de evento CV" ha sido tomaado por SPSS como categoría de referencia (tiene valores ceros en ambas), ya que es la que tiene una codificación absoluta más baja en la variable original, por lo que gr_riesg (1) es una dicotómica en la que el valor "1" es "Alto riesgo CV" y gr_riesg (2) es una dicotómica en la que el valor "1" es "Riesgo intermedio-moderado".

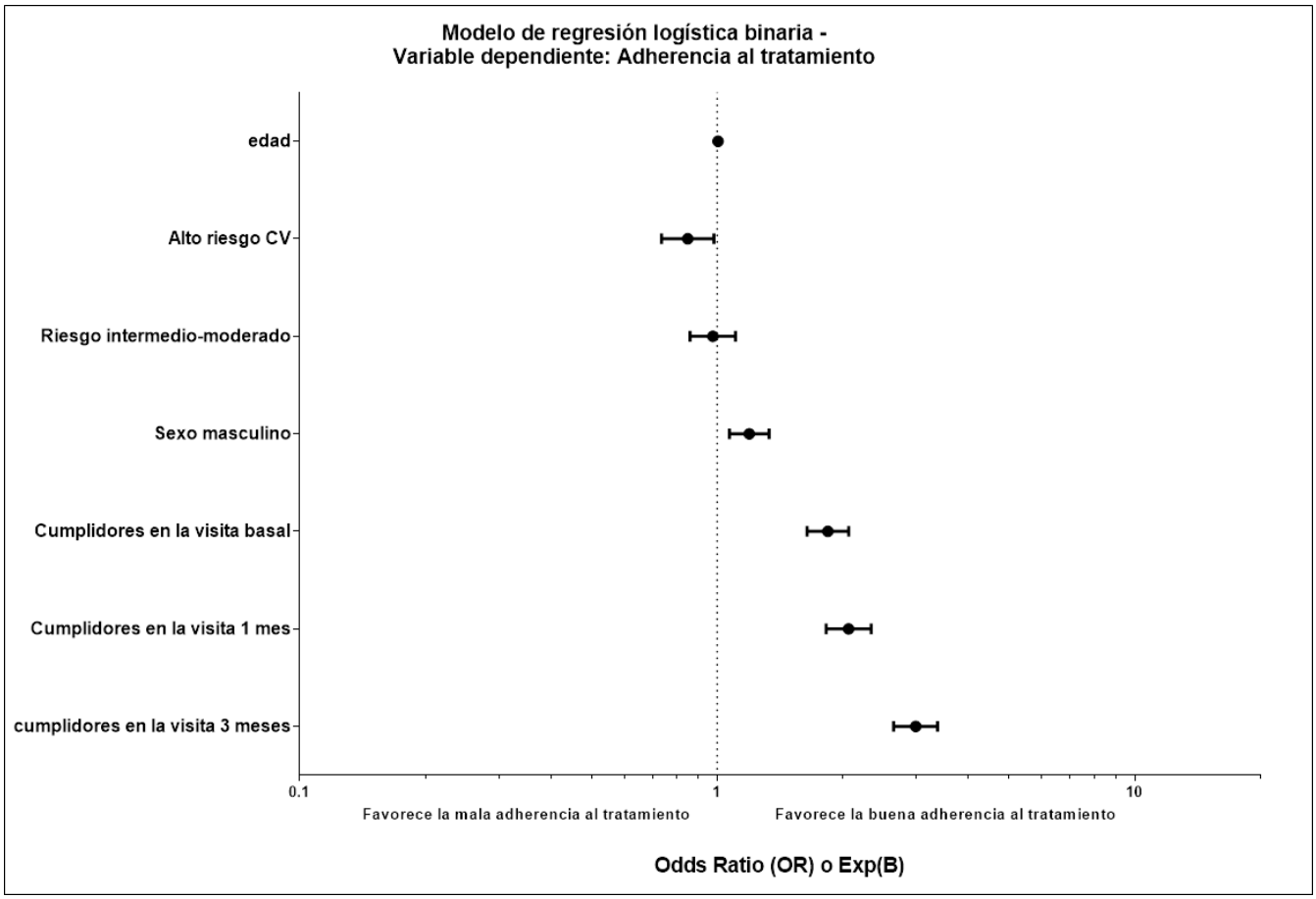
Tabla 15. Valores de regresión logística binaria. Variable de respuesta: Cumplimiento a los 12 meses

| **Variables en la ecuación** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Número de imputación | | | B | Error estándar | Wald | gl | Sig. | Exp(B) | 95% C.I. para EXP(B) | |
| | | | | | | | | | Inferior | Superior |
| 5 | Paso 1 | edad | 0,005 | 0,002 | 6,947 | 1 | **0,008** | 1,005 | 1,001 | 1,009 |
| | | gr_riesg | | | 5,041 | 2 | **0,08** | | | |
| | | gr_riesg(1) | -0,162 | 0,074 | 4,881 | 1 | **0,027** | 0,85 | 0,736 | 0,982 |
| | | gr_riesg(2) | -0,024 | 0,064 | 0,14 | 1 | 0,708 | 0,976 | 0,861 | 1,107 |
| | | sexo(1) | 0,177 | 0,056 | 10,083 | 1 | **0,001** | 1,193 | 1,07 | 1,331 |
| | | morgreenc(1) | 0,61 | 0,059 | 108,234 | 1 | **0,000** | 1,841 | 1,641 | 2,065 |
| | | morgreenc1(1) | 0,725 | 0,063 | 130,603 | 1 | **0,000** | 2,064 | 1,823 | 2,337 |
| | | morgreenc3(1) | 1,094 | 0,062 | 315,183 | 1 | **0,000** | 2,987 | 2,647 | 3,371 |
| | | Constante | -1,036 | 0,134 | 59,625 | 1 | **0,000** | 0,355 | | |

Vemos como la variable original de grupo de riesgo ("con antecedentes de evento CV") no tiene interpretación en la ecuación, está solo para indicarnos que de ella se han generado las dos *dummys*.

Según los resultados de la tabla 14, tanto la edad del paciente, como el grado de riesgo cardiovascular (alto riesgo CV) y la adherencia durante el estudio, influyen en la adherencia final al tratamiento ($p<0.05$ y $Wald<1.96$).

Figura 2. Representación gráfica de los resultados de la regresión logística



# Parte IV

## 8. Conclusiones

Durante el desarrollo de este Trabajo de Final de Máster, he tenido la oportunidad de trabajar con un dataset con una N de tamaño considerable (más de 1.000 pacientes en la visita basal) y de mejorar mis habilidades en el análisis de datos longitudinales, así como aprender sobre los diferentes métodos utilizados para manejar los datos faltantes es este tipo de estudios y aplicarlos mediante el paquete estadístico de SPSS. Además, dado el tamaño de la base de datos, así como la gran cantidad de variables registradas, me indujo a crear un script de Python que generara estadísticos descriptivos de manera automática de todas las variables con "rol=Entrada" en SPSS.

Respecto al procedimiento a la hora de llevar a cabo el análisis de datasets que contengan *missings*:

- En primer lugar, es conveniente llevar a cabo un análisis de *missings*, de manera que si el porcentaje de los desaparecidos es bajo (menos del 5%), podemos asumir que las inferencias no se verán afectadas por algún tipo de sesgo. En caso de que un porcentaje de datos faltantes superior al 5%, es necesario investigar la estructura de estos datos para confirmar si son MCAR, MAR o MNAR.
- Una alternativa a tratar de determinar la estructura de los datos consiste en analizar las diferentes técnicas de datos faltantes para manejarlos y evaluar si los diferentes enfoques son válidos.

Como recomendación para futuros análisis de estudios de tipo longitudinal o de cohorte, una vez depurada la base de datos, comprobar las posibles metodologías de análisis, ya que en función de cuál sea nuestra variable dependiente o de respuesta, así como la distribución y estructura de los datos debemos ser cuidadosos en la selección de la metodología más conveniente.

# 9. Anexos

## 9.1. Autorización de uso de base de datos

En Barcelona, a 05 de marzo de 2017

**REUNIDOS**

**FERRER INTERNACIONAL, S.A.**, con domicilio a efectos de este contrato en Av. Diagonal, 549, 5ª planta, 08029 Barcelona (en adelante denominada "FERRER"); con CIF A-08-041162

La Sra. Da. Mar Harmut Prats, con DNI 47258744-T y domicilio en C/Entença 60, 1º-4ª (de ahora en adelante el ESTUDIANTE DE MÁSTER) y,

El Dra. Da. Nuria Pérez Álvarez, con DNI 40562007N y domicilio en C/Espiga 11 3-2, 08027 Barcelona.

**MANIFIESTAN**

1. Que el ESTUDIANTE DE MÁSTER está llevando a cabo su TRABAJO DE FINAL DE MÁSTER sobre '*Tratamiento de los missings en estudios de cohorte*'.

2. Que el Dra. Da. Nuria Pérez Álvarez (de ahora en adelante el "Director de Trabajo de Final de Máster") dirige la Trabajo de Final de Máster del ESTUDIANTE DE MÁSTER sobre análisis de datos en presencia de datos faltantes, y que para obtener el Máster en Bioinformática y Bioestadística se requiere que el ESTUDIANTE DE MÁSTER, realice un proyecto en el que se aborde de manera teórica y se haga un análisis de datos práctico para ejemplificar la metodología aprendida solicitando el ESTUDIANTE DE MÁSTER utilizar información y/o documentación de FERRER.

3. FERRER posee información confidencial o de su propiedad (en adelante denominada como "INFORMACIÓN CONFIDENCIAL"), que incluye dicho sea a título enunciativo, pero no limitativo información técnica, científica, comercial y/o financiera, de forma oral o escrita (que incluye pero no se limita a documentación, dibujos, diseños, informes, correspondencia, formulas, datos, especificaciones) relativa a las actividades de FERRER y/o su Grupo de empresas.

4. Que es objetivo común de las partes firmantes preservar la confidencialidad de la información de FERRER utilizada por el ESTUDIANTE DE MÁSTER para la elaboración de su Trabajo de Final de Máster (de ahora en adelante, el "OBJETIVO") y con dicha finalidad las partes se reconocen capacidad legal para contratar y suscriben el presente Contrato con sujeción a los siguientes:

**PACTOS**

**Primero.**- El ESTUDIANTE DE MÁSTER se compromete a no divulgar o revelar a ningún tercero directa o indirectamente, total o parcialmente, la INFORMACIÓN CONFIDENCIAL que haya podido recibir de FERRER o que le pueda ser proporcionada por la misma o a la que de cualquier otro modo acceda y sea de la titularidad de FERRER o como resultado de la realización del OBJETIVO sin el previo consentimiento expreso y por escrito de FERRER. Asimismo, el ESTUDIANTE DE MÁSTER se compromete a no usarla para ninguna otra finalidad que no sea el cumplimiento del OBJETIVO.

Asimismo, el Director de Trabajo de Final de Máster se compromete a guardar absoluta confidencialidad respecto de aquella información que le sea comunicada por el ESTUDIANTE DE MÁSTER y, en particular, respecto de la INFORMACIÓN CONFIDENCIAL de FERRER.

**Segundo.**- Las obligaciones de confidencialidad del ESTUDIANTE DE MÁSTER y del Director de Trabajo de Final de Máster se establecen por un tiempo indefinido, que no podrá ser inferior a diez años desde la firma del presente contrato.

**Tercero.**- El ESTUDIANTE DE MÁSTER sólo revelará la INFORMACIÓN CONFIDENCIAL necesaria a los efectos de la comunicación, entrega y/o exposición pública de su Trabajo de Final de Máster y exclusivamente para la adquisición de la titulación de Máster en Bioinformática y Bioestadística y, en su caso, a los formadores que, encargados de dicha titulación, deban intervenir en el cumplimiento del OBJETIVO y sólo para este propósito. Estas personas deberán estar sujetas, asimismo, a obligación de confidencialidad.

**Cuarto.**- EL ESTUDIANTE DE MÁSTER y, en su caso, el Director de Trabajo de Final de Máster, se comprometen a devolver a FERRER la INFORMACIÓN CONFIDENCIAL así como cualquier soporte que la contenga, siempre que FERRER así se lo requiera y, en todo caso, cuando se haya cumplido el OBJETIVO, absteniéndose de guardarse copia de la misma.

**Quinto.**- En la ejecución del presente Contrato es posible que el ESTUDIANTE DE MÁSTER y/o el Director de Trabajo de Final de Máster puedan tener acceso a datos de carácter personal respecto de los cuáles FERRER sea responsable de acuerdo a lo previsto en la Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal (en adelante, la "Ley Orgánica 15/1999").

En este sentido, se obligan a respetar la confidencialidad de los datos a los cuales puedan tener acceso, y cumplir con todas las obligaciones de acuerdo con lo establecido en la Ley Orgánica 15/99 y su reglamento contenido en R.D. 1720/07 y/o cualesquiera otras normas que las sustituyan en el futuro.

**Sexto.**- Cualquier dato, información y/o resultados incluido, dicho sea a título enunciativo pero no limitativo, know-how, invenciones, mejoras, descubrimientos, materiales, resultados del trabajo, usos, métodos de trabajo o procedimientos de producción y otra propiedad intelectual y/o industrial realizados, obtenidos o desarrollados por el ESTUDIANTE DE MÁSTER utilizando parte o la totalidad de información y/o documentación o datos propiedad de FERRER (o proporcionados por ésta), serán propiedad exclusiva de FERRER quien podrá divulgarlos, explotarlos, utilizarlos y, en su caso, patentarlos de la manera que considere oportuna sin más limitaciones que las establecidas por la ley.

**Séptimo.**- Ningún dato y/o información y/o resultados obtenidos como consecuencia del OBJETIVO, podrán ser publicados ni total ni parcialmente sin el previo consentimiento

expreso y por escrito de FERRER en relación al contenido y/o momento de dicha presentación o publicación. El ESTUDIANTE DE MÁSTER se compromete a demorar durante el plazo que le requiera FERRER, dicha presentación o publicación.

**Octavo**.- Ninguna cláusula del presente acuerdo puede interpretarse como una cesión de derechos y/o licencia de cualquier tipo respecto de cualquier información de FERRER al ESTUDIANTE DE MÁSTER y/o al Director de TRABAJO DE FINAL DE MÁSTER.

**Noveno**.- El presente acuerdo se regirá de acuerdo con las leyes españolas, sometiéndose las partes expresamente a los Juzgados y Tribunales de Barcelona.

Y, en prueba de conformidad, firman las partes el presente acuerdo en el lugar y la fecha indicados "ut supra".

FERRER INTERNACIONAL S.A.

Da. Mar Harmut Prats

El ESTUDIANTE DE MÁSTER

Dra. Da. Nuria Pérez Álvarez

El Director de Trabajo Final de Máster

# 10. Glosario y resumen de abreviaturas

ECV – Enfermedad Cardiovascular

ACV – Accidente Cerebrovascular

MCAR – *Missing Completely at Random*

MAR – M*issing at Random*

LOCF – *Last observation carried forward*

# 11. Bibliografía

[1] P. D. Allison, *Missing Data*, no. n.{º} 136. SAGE Publications, 2002.

[2] E. Lazcano-Ponce, E. Fernández, E. Salazar-Martínez, and M. Hernández-Avila, "Estudios de cohorte. Metodología, sesgos y aplicación," *Salud Publica Mex.*, vol. 42, no. 3, pp. 230–241, Jun. 2000.

[3] Karahalios, L. Baglietto, J. B. Carlin, D. R. English, and J. A. Simpson, "A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures.," *BMC Med. Res. Methodol.*, vol. 12, p. 96, Jul. 2012.

[4] Jarrín, R. Geskus, S. Pérez-Hoyos, and J. del Amo, "Principales métodos de análisis en los estudios de cohortes de sujetos diagnosticados de infección por el virus de la inmunodeficiencia humana (VIH)," *Enferm. Infecc. Microbiol. Clin.*, vol. 28, no. 5, pp. 298–303, May 2010.

[5] "OMS | Enfermedades cardiovasculares," *WHO*, 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/es/. [Accessed: 15-Mar-2017].

[6] "Alianzas estratégicas para la innovación | Ferrer Corporate." [Online]. Available: http://www.ferrer.com/innovacion/alianzas-estrategicas-para-la-innovacion. [Accessed: 17-Mar-2017].

[7] "La polypill cardiovascular, desarrollada por la colaboración CNIC-Ferrer, aumenta la adhesión al tratamiento en esta patología - Atención Especializada - Elmedicointeractivo.com." [Online]. Available: http://www.elmedicointeractivo.com/articulo/atencion-especializada/polypill-cardiovascular-desarrollada-colaboracion-cnic-ferrer-aumenta-adhesion-tratamiento-patologia/20161115131644107170.html. [Accessed: 17-Mar-2017].

[8] M. Castro, C. Tutores, F. Gude Sampedro, A. Pérez, G. Máster, and T. Estadísticas, "IMPUTAC ON DE DATOS FALTANTES EN UN MODELO DE TIEMPO DE FALLO ACELERADO," 2014.

[9] I. Corporation, "IBM SPSS Missing Values 20."

[10] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art.," *Psychol. Methods*, vol. 7, no. 2, pp. 147–77, Jun. 2002.

[11] R. J. Little *et al.*, "The Prevention and Treatment of Missing Data in Clinical Trials," *N. Engl. J. Med.*, vol. 367, no. 14, pp. 1355–1360, Oct. 2012.

[12] C. K. Enders, *Applied Missing Data Analysis*. Guilford Press, 2010.

[13] "American Heart Association - Building healthier lives, free of cardiovascular diseases and stroke." [Online]. Available: http://www.heart.org/HEARTORG/. [Accessed: 18-May-2017].

[14] D. B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, p. 581, Dec. 1976.

[15] M. Szklo, "Population-based Cohort Studies," vol. 20, no. 1.

[16] "Qué es la enfermedad cardiovascular: MedlinePlus enciclopedia médica." [Online]. Available:

https://medlineplus.gov/spanish/ency/patientinstructions/000759.htm. [Accessed: 18-May-2017].

[17]J. M. Baena-Díez, M. Vidal-Solsona, A. O. Byram, I. González-Casafont, G. Ledesma-Ulloa, and N. Martí-Sans, "Epidemiología de las enfermedades cardiovasculares en atención primaria. Estudio Cohorte Zona Franca de Barcelona," *Rev. Española Cardiol.*, vol. 63, no. 11, pp. 1261–1269, Nov. 2010.

[18]L. A. Cupples, R. B. D'Agostino, K. Anderson, and W. B. Kannel, "Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study.," *Stat. Med.*, vol. 7, no. 1–2, pp. 205–22.

[19]W. Vach and M. Blettner, "Missing Data in Epidemiologic Studies," in *Encyclopedia of Biostatistics*, Chichester, UK: John Wiley & Sons, Ltd, 2005.

[20]J. Twisk and W. de Vente, "Attrition in longitudinal studies. How to deal with missing data.," *J. Clin. Epidemiol.*, vol. 55, no. 4, pp. 329–37, Apr. 2002.

[21]A. Karahalios, L. Baglietto, K. J. Lee, D. R. English, J. B. Carlin, and J. A. Simpson, "EMERGING THEMES IN EPIDEMIOLOGY The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study."

[22]J. Scheffer, "Dealing with Missing Data," *Res. Lett. Inf. Math. Sci*, vol. 3, pp. 153–160, 2002.