UCAM

UNIVERSIDAD CATÓLICA
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Programa de Doctorado en Ciencias de la Salud

Cátedra Internacional de Bioestadística y Big Data

Aplicación de métodos Big Data al análisis de datos biomédicos: Identificación de factores asociados con la presencia de datos incompletos y estudio de secuencias de eventos

Autor:
D. Juan José Piñero de Armas

Director:
Dr. D. David Prieto Merino

Murcia, Diciembre 2020

UCAM

UNIVERSIDAD CATÓLICA
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Programa de Doctorado en Ciencias de la Salud

Cátedra Internacional de Bioestadística y Big Data

Aplicación de métodos Big Data al análisis de datos biomédicos: Identificación de factores asociados con la presencia de datos incompletos y estudio de secuencias de eventos

Autor

D. Juan José Piñero de Armas

Director:

Dr. D. David Prieto Merino

Murcia, Diciembre 2020

UCAM

UNIVERSIDAD CATÓLICA
DE MURCIA

AUTHORIZATION OF THE DIRECTORS OF THE THESIS

FOR SUBMISSION:

Prof. Dr. David Prieto Merino, as Director of the doctoral thesis "Aplicación de métodos Big Data al análisis de datos biomédicos: Identificación de factores asociados con la presencia de datos incompletos y estudio de secuencias de eventos" by Mr. Juan José Piñero de Armas in the Ph.D. Program in Health Sciences, authorizes its submission as it fulfills the conditions necessary for its defense.

Sign, to comply with the Royal Decrees 99/2011, 1393/2007, 56/2005 and 778/98, in Murcia, December 11th, 2019.

Firmado por PRIETO MERINO ANTONIO DAVID - 51415275V el día 20/01/2021 con un certificado emitido por AC FNMT Usuarios

Dr. D David Prieto Merino.

A mi padre Rafael,

que desde el cielo me brinda luz y fuerzas.

A mi madre Candelaria,

por su sacrificio y cariño.

A mi novia Celia.

con quien he compartido los mejores momentos de mi vida,

A mi hermanita Nathalie.

AGRADECIMIENTOS

A todas las personas que hicieron posible este proyecto, muchas gracias por su apoyo y enseñanza. De forma muy especial, quiero dejar constancia de mi agradecimiento a mi director el Dr. David Prieto y al Dr. Luis Prieto, quienes me escogieron para realizar esta tesis y dedicaron su tiempo a transmitirme parte de su conocimiento.

A la Universidad Católica San Antonio de Murcia, y en particular, a la Facultad de Medicina por abrirme sus puertas y permitirme hacer esta Tesis y al Vicerrectorado de Investigación por todas sus atenciones, en especial a Silvia y Mariano, y por supuesto, a la Vicerrectora Dª Estrella Núñez.

A Dr. Juan Pablo Casas por acogerme para mi estancia investigadora en el Farr Institute of Health Informatics Research de Londres.

Y por encima de todo, gracias a mi familia, sin cuyo sacrificio y cariño no podría haber llegado hasta aquí. En especial a mi padre Rafael, que falleció el 28 de febrero de 2017, mientras realizaba la tesis, a mi madre Candelaria, a mi novia Celia, por estar siempre junto a mí y apoyarme en todo momento, y a mi hermana Nathalie.

Y por último, a mi amigo Virgilio por haber sido quien me animó a hacer el doctorado y a mi compañera María del Carmen.

ABSTRACT

The presence of missing data in biomedical databases can bias statiscal analyses and reduce their precision. Imputation methods can partially fix the problem, but rely on using proper models. We have used logistic regressions with random effects to identify what patient or medical center's variables are key factors to explain the missingness in other variables (weight and cigarettes) on large databases.

In order to deal with such a large amount of data, three different approaches were applied: 1) analysis with the complete dataset, 2) independent intra-centre analysis and pulling the results with a meta-analysis, 3) independent analysis in randomly partitioned blocks and pulling the results with a meta-analysis.

The most accurate results were obtained from the analysis of the whole database but it is extremely slow and often impossible to run in an average desktop computer because of memory limitations. Doing the analysis in randomized partitions provides accurate results and is much faster. By-center analysis does not require centers to share data but the results are not as accurate, it produced errors and it does not allow to perform a combined analysis of intra-center and inter-center variables in the same model.

We have applied this methodology to the database of the Area-7 public health service of Madrid, which contains 1400 variables collected over five years from a quarter of a million people.

The analysis of the whole dataset with random effects for individuals nested into centres showed that the odds ratio of missing data in WEIGHT is 0.809 for the covariate SEX (versus men), 2.184 for FOREIGNER (over nationals), 0.932 for AGE (per year), 0.854 for "Doctor Workload Pressure", 1.178 for "Nursing Workload Pressure" and 1.615 for YEAR (per year). All theses results had ($p < 10^{-16}$). These varaibles, included in the model as fixed effects, can explain 14.3% of the variability of the the odds of missingness in the variable WEIGHT. The factor Medical Center explains 8.9%, and the factor Person explains 57.8% of the variabilty.

For the variable CIGARETTES the analysis of the whole dataset showed that the odds ratio of missing data is 1.013 ($p = 0.057$) for the covariate AGE (per year), 1.082 for "Doctor Workload Pressure" ($p = 0.0009$), 0.836 ($p = 2.3{*}10^{-5}$) for "Nursing Workload Pressure" y 0.144 ($p < 10^{-16}$) for YEAR (per year). The model with additional variables showed that the odds ratio of missing data in CIGARETTES is 1.013 for the covariate WEIGHT ($p = 0.074$), and we saw that the variables SEX ($p = 0.18$) and FOREIGNER ($p = 0.30$) are not significant.

We have also explored different methods for the analysis and graphical representation of event sequences and applied them to the analysis of the medical events collected on the Area-database.

The simplest approaches just analyse separately the different types of events, modelling the time until their first occurrence, ignoring the following repetitions and assuming the occurrences are independent and follow a Poisson distribution. Markov models are also comonly applied to model the occurrence of consecutive events.

We have proposed a new approach to study sequences capable of analysing multiple different types of events, with recurrence and taking into account the time at risk and the transitions between consecutive and non-consecutive events. We focused our analysis on the shortest subsequences: pairs and triplets, for which we report the Incidence Rate (IR).

The results were graphically represented using connection networks, which helps to quickly visualize the sequences of events of any length. The inclusion of non-consecutive events in the analyses allowed us to discover some transitions that otherwise went unnoticed.

In those studies in which temporal information is approximate, it is not possible to know the date or exact order of occurrence of the events. We have developed a simple imputation method replacing the sequences by a weighted sum of fully ordered sequences generated by permuting the intra-annual events and using the frequency of the inter-annual pairs as a priori probability.

The analysis of the Area-7 data shows that the highest IRs correspond to pairs of events whose final event is TRASTLIPIDOS, HTA and GRIPE.

Key words— missing data, logistic regression, random effects, imputation, big data, sequence of events, connection networks

RESUMEN

La presencia de datos faltantes en las bases de datos biomédicas puede sesgar los análisis estadísticos y reducir su precisión. Los métodos de imputación de estos habitualmente denominados "*missing data*" corrigen parcialmente el problema pero necesitan modelos adecuados que relacionen su ocurrencia con el valor de las demás variables. Hemos utilizado regresiones logísticas con efectos aleatorios para identificar qué factores del paciente o centro médico están asociados con una mayor presencia de *missing data* en otras variables en bases de datos de gran tamaño.

Para poder analizar una cantidad tan grande de datos hemos aplicado tres enfoques diferentes: 1) análisis con el conjunto completo de los datos, 2) análisis independientes intra-centro y el posterior metaanálisis de sus coeficientes, 3) análisis independientes en particiones aleatorias de los datos y su posterior metaanálisis.

Los resultados más precisos se obtuvieron con el análisis simultáneo de toda la base de datos, pero es extremadamente lento y a menudo imposible de realizar en un ordenador medio debido a las limitaciones de memoria. El análisis de los datos particionados en bloques aleatorios arrojó resultados bastante precisos pero es mucho más rápido. El análisis por centros no requiere que los centros intercambien datos pero los resultados no son tan precisos, produjo errores durante los cálculos y no permite el análisis combinado de variables intra-centro e inter-centro en un mismo modelo.

Hemos aplicado esta metodología a la base de datos de atención primaria del Area-7 del servicio público de salud de Madrid, que contiene 1400 variables recopiladas a lo largo de cinco años de un cuarto de millón de personas. Hemos analizado la presencia de *missing data* en las variables peso y cigarrillos.

El análisis del conjunto completo de los datos con el modelo de efectos aleatorios para los individuos anidados a los centros mostró que la *odds ratio* de *missings* en el PESO es 0.809 para la covariable SEXO (frente a hombres), 2.184 para EXTRANJERO (frente a nacionales), 0.932 para EDAD (por año), 0.854 para PAMED, 1.178 para PAENF y 1.615 para YEAR (por año). Todos estos resultados se obtuvieron con ($p < 10^{-16}$). Estas variables, incluidas en el modelo como efectos fijos, explican el 14.3% de la variabilidad de las *odds* de tener *missing* en la variable PESO. El factor centro, explica otro 8.9% de la variabilidad, el factor individuo explica el 57.8%.

Para la variable CIGARRILLOS el análisis del conjunto completo de datos mostró que la la *odds ratio* de *missings* en CIGARRILLOS es 1.013 ($p = 0.057$) para la covariable EDAD, 1.082 para PAMED ($p = 0.0009$), 0.836 ($p = 2.3{\times}10^{-5}$) para PAENF y 0.144 ($p < 10^{-16}$) para YEAR. Los efectos fijos del modelo explican el 67% de la variabilidad de las *odds* de *missing* en la variable CIGARRILLOS. La variabilidad debida a los centros es casi cero. En el modelo con más variables podemos ver que la *odds ratio* de *missings* en CIGARRILLOS es 1.013 para PESO ($p = 0.074$) y comprobamos que las variable SEXO ($p = 0.18$) y EXTRANJERO ($p = 0.30$) no son significativas.

También hemos explorado diferentes métodos para el análisis y representación gráfica de secuencias y lo hemos aplicado al análisis de eventos médicos en bases de datos de gran tamaño. En la literatura este problema suele abordarse analizando por separado los diferentes tipos de evento, modelizando el tiempo hasta su primera ocurrencia, ignorando las repeticiones o asumiendo que son las ocurrencias son independientes y siguen una distribución de Poisson. También es común aplicar modelos de Markov a la ocurrencia de eventos consecutivos.

Proponemos un nuevo enfoque para estudiar secuencias capaz de analizar múltiples tipos de eventos diferentes, consecutivos y no consecutivos, descomponiéndolas en subsecuencias de menor tamaño y teniendo en cuenta el tiempo en riesgo, que reportamos utilizando la Tasa de Incidencia (IR).

Los resultados así obtenidos fueron representados gráficamente mediante redes de conexión, que han mostrado ser muy útiles para visualizar fácilmente las secuencias de eventos de cualquier longitud. La inclusión de eventos no consecutivos en los análisis permite descubrir transiciones que de otro modo pasan desapercibidas.

En aquellos estudios en los que la información temporal es aproximada no es posible conocer la fecha ni el orden exacto de ocurrencia de los eventos. Para solucionar este problema hemos desarrollado un método de imputación simple sustituyendo las secuencias por la suma ponderada de secuencias totalmente ordenadas obtenidas permutando los datos intraanuales y utilizando como probabilidad a priori la frecuencia de los pares interanuales.

Los métodos desarrollados han sido utilizados para analizar los eventos de la base de datos del Area-7 de Madrid, para los que hemos obtenido que Los IR más elevados corresponden a pares de eventos cuyo evento final es TRAST_LIPIDOS, HTA y GRIPE.

Palabras clave— datos faltantes, regresión logística, efectos aleatorios, imputación, big data, secuencias de eventos, redes de conexión

ÍNDICE GENERAL

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE CÓDIGOS

GLOSARIO DE TÉRMINOS

BIC	Criterio de información bayesiano.
EAP	Equipo de Atención Primaria o centro médico
glm	Modelo de regresión lineal generalizado.
Intercepto	Ordenada en el origen de los modelos de regresión.
IR	Incidence Rate. Tasa de Incidencia.
Odds	Probabilidad de ocurrencia de un evento dividida entre la probabilidad de que no ocurra.
Odds Ratio	Razón de momios o razón de oportunidades, cociente entre las odds para dos valores de una variable explicativa.
p	p-valor del coeficiente de la variable del modelo de regresión.
PAMED	Presión asistencial media del médico. Número de pacientes atendidos cada día.
PAENF	Presión asistencial media de enfermería.
POSMAS	Porcentaje de pacientes de 65 o más años de edad en el centro.

Capítulo 1

INTRODUCCIÓN

El incesante incremento de la cantidad de información y número de variables recopilados en las bases de datos biomédicas ha posibilitado la creación de modelos avanzados que posibilitan la realización de descubrimientos antes imposibles, pero la cada vez más compleja y cambiante estructura de los datos y la presencia de *missing data* y errores dificulta seriamente su procesamiento informático y análisis estadístico, haciendo necesario el desarrollo de nuevas técnicas[1].

Para realizar cualquier estudio es de vital importancia utilizar información que haya sido recopilada y estructurada adecuadamente y asegurar su calidad, [101, 100, 124, 126]. En esta tesis utilizamos métodos procedentes del campo del Big Data y de desarrollo propio para el procesado y análisis de datos biomédicos complejos y así estudiar dos problemas habituales en bioestadística: el análisis de *missing data* en bases de datos de muy gran tamaño, y el análisis de secuencias de eventos médicos.

1.1 IDENTIFICACIÓN DE FACTORES ASOCIADOS CON LA PRESENCIA DE *MISSING DATA*

La presencia de errores en las bases de datos médicas puede estar en parte relacionada con la calidad asistencial, con el tiempo que dedican los médicos a tareas administrativas y con el posterior tratamiento de la información [129, 4, 73, 155, 46, 130, 102, 11, 105, 3, 51, 130]

El análisis de múltiples bases de datos de investigación clínica reportó que las tasas de error detectadas oscilan entre el 2.3% y el 26.9%, [63]. Muchos de esos errores no son aleatorios y podrían sesgar gravemente a los resultados de los estudios realizados utilizando esos datos. El estudio [91] reportó la ocurrencia de 4 errores por cada 1000 prescripciones médicas.

[1] Métodos como las redes neuronales, sistemas difusos, algoritmos genéticos, tabú search, redes bayesianas, lasso o las redes de conexión, que permiten buscar patrones, realizar predicciones o automatizar la resolución de problemas muy grandes, pero pueden ser difíciles de interpretar y lentos.

Los *missing data*, también llamados datos faltantes o valores perdidos de una variable, son observaciones o registros que no tienen almacenado ningún valor o éste no es válido. Su presencia supone un problema común y puede influir en los resultados de los análisis sesgando las estimaciones, reduciendo su precisión y complicando los cálculos. Existen diferentes métodos, llamados de imputación, para sustituir esos valores perdidos o erróneos por estimaciones razonables que corrigen en parte el problema [16, 141, 85], pero necesitan de modelos adecuados que describan cómo se generan los *missing* y su relación con las demás variables.

Los *missing data* se suelen clasificar en tres tipos según el patrón de pérdida de los datos faltantes: 1) MCAR (Missing Completely At Random), la presencia de *missing data* en una variable es independiente del valor de ésta y del valor de las demás variables. 2) MAR (Missing At Random) es independiente de su valor, pero depende de otras variables. 3) NMAR (Not Missing At Random) depende de los valores de la propia variable.

En la primera parte de la tesis analizamos qué factores relacionados con el paciente o centro médico influyen en la aparición de *missing data MAR* en otras variables. Para ello utilizaremos modelos de regresión logística con efectos aleatorios, que serán aplicados a la base de datos del servicio público de salud del Area-7 de Madrid, que contiene información recogida durante siete años a casi un cuarto de millón de personas, [136, 137, 51].

La regresión logística es utilizada, [72], para el modelado de variables categóricas, como en nuestro caso, la presencia o ausencia de un valor. Los modelos de efectos aleatorios, por su parte, pueden ser utilizados para modelar variables que han sido medidas repetidamente en las mismas unidades de observación [13, 14, 164, 131], ya sea en función de tratamientos distintos, o en estudios longitudinales, en los que habitualmente las mediciones se repiten a lo largo del tiempo para una misma persona o lugar. Estos modelos permiten modelar adecuadamente la varianza y covarianza

dentro de cada individuo o centro y entre ellos [57, 71, 160, 19, 56, 13].

Este tipo de análisis necesitan de grandes cantidades de memoria, lo cual limita enormemente el tamaño máximo de los datos que pueden ser analizados. Podemos encontrar en la literatura especializada propuestas de algoritmos experimentales capaces de ajustar modelos de regresión con efectos aleatorios en bases de datos de gran tamaño [161, 145, 65, 159, 144, 27, 132, 134], pero sólo son aplicables a los modelos más sencillos y sin efectos anidados o utilizan métodos aproximados. A día de hoy desconocemos de la existencia de ningún software o librería que proporcione al usuario directamente este tipo de algoritmos, ya implementados. Debemos entonces explorar otras formas de abordar el problema.

Para salvar las limitaciones de memoria particionaremos los datos en bloques de tamaño reducido, en cada uno de los cuales ajustaremos un modelo de regresión independientemente. Utilizaremos técnicas de metaanálisis para combinar los coeficientes obtenidos en cada partición. Es habitual el empleo de estas técnicas para combinar los resultados obtenidos en diferentes estudios [23, 150, 44].

1.2 ANÁLISIS DE LAS SECUENCIAS DE EVENTOS

Otro problema habitual en los estudios médicos es el análisis de las secuencias temporales de los eventos en un mismo paciente. En la segunda parte de esta tesis proponemos diferentes métodos para analizar dichas secuencias a partir de la información recogida en bases de datos médicas.

El procedimiento más utilizado en bioestadística para el análisis de secuencias es la aplicación de modelos de riesgos proporcionales de Cox, que permiten modelizar el tiempo hasta la primera ocurrencia de un evento, ignorando sus repeticiones posteriores en un mismo paciente y considerando los eventos como independientes [45]. Pero las suposiciones del modelo de Cox son a menudo violadas, un mismo sujeto podría sufrir más de una vez el evento de interés, produciendo estimaciones

sesgadas.

En la literatura también podemos encontrar modelos estadísticos para analizar eventos recurrentes de tiempo-a-evento [5, 147], que sí permiten que los eventos se repitan en un mismo sujeto. En su forma más simple ignoran el orden de ocurrencia de los eventos o cuentan los eventos observados dentro de un período de tiempo dado (modelo Andersen-Gill) asumiendo que son independientes y siguen una distribución de Poisson [34, 74, 90, 94, 131]. Estos modelos no tienen en cuenta la ocurrencia de diferentes tipos de eventos. Un paciente puede sufrir sucesivamente varios tipos de eventos mutuamente excluyentes y ser considerado en una situación de riesgo competitivo, la ocurrencia de uno de ellos evitaría que cualquier otro evento suceda. En la mayoría de los estudios se llevan a cabo análisis independientes de cada tipo de evento, pero en realidad pueden estar correlacionados entre sí.

En presencia de riesgos competitivos, los métodos de supervivencia simples están sesgados porque asumen que todos los sujetos experimentarán el evento de interés si el período de seguimiento es lo suficientemente largo [17, 24]. En su lugar, se utilizan complejos modelos marginales bayesianos y semiparamétricos con la función de incidencia acumulativa y el hazard de la subdistribución, pero sus supuestos deben comprobarse cuidadosamente y no siempre se cumplen[58, 92, 37, 33, 38, 93, 104, 22].

Para el análisis de secuencias de eventos clínicos a menudo se utilizan modelos multiestado, que representan un número finito de estados de salud discretos y mutuamente excluyentes conectados por transiciones que expresan la probabilidad de que un paciente pase de un estado a otro [21, 80]. El modelo puede basarse en diferentes suposiciones:

- Modelos homogéneos en el tiempo: las tasas de transición son constantes en el tiempo. Los eventos fueron observados en puntos de tiempo preestablecidos y forman un proceso Markov de tiempo discreto.

- Modelos Markov de tiempo continuo: los datos observados son generados por una cadena Markov de tiempo continuo discretamente observada. Asume una distribución exponencial del tiempo de estadía para que la tasa de abandono de un estado no dependa del tiempo de ocupación. Las tasas de transición dependen únicamente del estado actual [21, 80, 86, 7].
- Modelos de Semi-Markov: las tasas de transición dependen no sólo del tiempo cronológico sino también del tiempo transcurrido en el estado actual, y la distribución del tiempo de estancia no se restringe sólo a la exponencial. [10, 35, 60, 148, 39].

Existen modelos Markov de orden superior en los que la probabilidad de un estado depende simultáneamente de varios estados consecutivos anteriores, pero las estimaciones de los parámetros son menos fiables [41, 113, 123]. Los modelos de Markov también se han utilizado en combinación con efectos aleatorios y análisis bayesianos [98, 125].

Otra alternativa comúnmente utilizada son los Modelos Ocultos de Markov (HMM) [18, 76, 75, 120, 122, 135, 143, 146, 154, 156, 158, 20, 9, 78, 118], desarrollados por Baum y Stratonovich, que asumen que los eventos observados discretamente se caracterizan por una variable latente subyacente con estados no observados, ocultos, que se mapean a cada estado de la enfermedad. [42, 115, 28, 110, 55, 83, 106, 43] hacen una revisión exhaustiva de los diferentes modelos de Markov utilizados en los estudios sanitarios y analizan los problemas que pueden surgir a partir de los mismos.

Se hace necesario un planteamiento general que incorpore simultáneamente los efectos de la repetición, la competencia, la dependencia temporal explícita, las covariables y el efecto de los eventos “distantes” no consecutivos [119].

En esta tesis proponemos un nuevo enfoque para el estudio de las secuencias de eventos que afectan a cada individuo, con el que podemos analizar simultáneamente

múltiples tipos de eventos teniendo en cuenta el tiempo en riesgo. Para ello exploramos diferentes modos de descomponer las secuencias originales en fragmentos más pequeños objeto de un posterior análisis. Consideramos las transiciones entre pares de eventos consecutivos, entre pares de eventos cualesquiera y subsecuencias de tres eventos, a los que llamaremos tripletes.

Hemos aplicado esta metodología para estudiar las secuencias de eventos recogidos en la base de datos del servicio público de salud del Area-7 de Madrid. En ella, para la mayoría de las variables, no se dispone de la fecha exacta de ocurrencia de los eventos, sólo del año. Por lo que se desconoce el orden exacto en el que tuvieron lugar los eventos interanuales, imposibilitando la reconstrucción completa de las secuencias. La opción de descartar los datos intraanuales nos haría perder poder estadístico y sesgaría los resultados. Para corregir este problema proponemos aplicar métodos de imputación de datos faltantes (*missing data*).

La *imputación múltiple* (IM) [50] se ha convertido en la forma estándar de tratar los datos faltantes en la mayoría de los análisis bioestadísticos múltivariantes [2] debido a su robustez, simplicidad matemática y que puede ser aplicada a todo tipo de variables siempre que el patrón de los *missings* sea *MAR*, no dependa del valor de la propia variable. Mediante un procedimiento iterativo sustituye en cada variable los *missings* por valores estimados utilizando modelos de regresión y simulaciones de Monte Carlo [95]. Existen dos enfoques principales para realizar la imputación múltiple [77]:

- La Modelización conjunta directa (JM), basada en la distribución posterior conjunta de las variables, suele asumir que las variables incompletas siguen una distribución normal multivariante [133]. En la práctica es complicada de utilizar.
- El FCS, especificación totalmente condicional o regresión secuencial mediante ecuaciones encadenadas (MICE), imputa los valores faltantes utilizando distribuciones condicionales univariantes para cada variable incompleta dadas todas

las demás, reutilizando cíclicamente los resultados de las iteraciones anteriores [32, 66]. Es más flexible y fácil de entender que la JM, y es ampliamente utilizado en bioestadística [16, 141, 29, 79, 139, 69].

Estos métodos, habitualmente utilizados con datos de sección transversal, también pueden aplicarse a datos longitudinales recogidos a intervalos regulares si consideramos las mediciones repetidas como variables diferentes [84, 121, 66, 128], pero sufren problemas de convergencia y colinealidad y fallan si los intervalos de tiempo son irregulares.

Para superar estas limitaciones se han desarrollado diferentes extensiones del enfoque JM: 1) Tratar todas las variables del modelo como continuas considerando las mediciones repetidas como variables distintas [67, 66, 133, 89], pero esta estrategia no funciona con datos binarios ni categóricos. 2) Utilizar una distribución multivariante conjunta [84], asumiendo que todas las variables incompletas son continuas con efectos aleatorios específicos del sujeto. 3) Tratar las variables discretas como normales latentes (LN) que siguen una distribución multivariante conjuntamente con las variables continuas [64]. Y extensiones del enfoque FCS: 4) Incluir las mediciones repetidas sólo dentro de bloques de tiempo preestablecidos [54]. 5) Utilizar un modelo multinivel para imputar los *missings* en cada variable dependiente del tiempo dada todas las demás, pasando iterativamente por los modelos de imputación univariante. 6) Permitir variaciones de error específicas del sujeto e imputar las variables binarias y categóricas como variables continuas [32]. 7) Imputar tanto las variables continuas como las binarias incompletas usando modelos con efectos aleatorios [53]. 8) Combinar el método MICE con el algoritmo Random Forest [30] como modelo condicional para la imputación, utilizando múltiples árboles de regresión para reducir el riesgo de sobreajuste. [138]

Para imputar *missings* en datos longitudinales con riesgos competitivos [58] se puede utilizar un modelo de riesgos proporcionales de Cox modelando

conjuntamente diferentes tipos de eventos [87], con un parámetro para el ratio de los riesgos entre tipos de eventos e incluyendo un término adicional para los evento de tipo desconocido [62, 59] o mal clasificados [127]. Una consecuencia indeseable de este enfoque es que no distingue entre *missings* de diferentes tipos de eventos. Algunos autores [68, 107, 96] proponen combinar la MI con el algoritmo de Random Forests utilizando pseudo-observaciones (valores obtenidos mediante remuestreo y estimadores de Kaplan-Meier) [6, 8] como variable de salida de un modelo Random Forest. Este enfoque también es usado con datos censurados por intervalos [48, 81, 109].

El problema del truncamiento en datos longitudinales [142, 151, 152, 31] es abordado mediante modelos de selección [70] y mezcla de patrones [61] o mediante MI [162]. Los riesgos competitivos con censura y truncamiento son estudiados en [82] con modelos latentes de tiempos de falla, asumiendo distribuciones de Weibull y utilizando bootstrap y Monte Carlo para calcular los estimadores de máxima verosimilitud. El mismo problema es estudiado en [25] utilizando estimaciones no paramétricas de la distribución conjunta de los tiempos de truncamiento y censura y de la máxima verosimilitud para modelar la curva de supervivencia, concluyendo que este enfoque, en general, puede no resultar válido.

La imputación de datos longitudinales también es tratado mediante modelos multiestado con tasas de transición dependientes del tiempo e interacciones [36, 97] y mediante modelos ocultos de Markov para datos categóricos[26].

Estos modelos de imputación se basan en asunciones difíciles de verificar y son computacionalmente intensivos, no son fácilmente aplicables a bases de datos de gran tamaño como la nuestra. La mayoría de los estudios que encontramos en la literatura con datos reales realizan los análisis utilizando sólo información de los pares de eventos consecutivos, por lo que no sirven para estudiar secuencias de mayor longitud teniendo en cuenta la influencia de eventos más distantes, que es uno de los

objetivos de esta tesis.

En nuestro enfoque utilizaremos un método aproximado de imputación para las variables que contienen fechas truncadas, generando todas las posibles permutaciones de los eventos interanuales para cada individuo y construyendo con ellas nuevas secuencias completas que pudieron ocurrir, compatibles con los datos originales. Asegurándonos así de que la secuencia real es tenida en cuenta, pero generando también secuencias ficticias que nunca ocurrieron. Este enfoque nos permite detectar secuencias que potencialmente han tenido lugar, incluidas las de mayor longitud, y que deberán ser objeto de un posterior estudio con datos de mayor calidad y modelos más complejos.

Testearemos la validez de este enfoque comparando, para las variables que lo permiten, las secuencias calculadas a partir de fechas completas con las calculadas al truncar previamente esas fechas y utilizar diferentes métodos de imputación. Tras realizar los análisis representamos los resultados en tablas y gráficamente mediante redes de conexión [15], que nos permiten visualizar rápidamente las secuencias de eventos de cualquier longitud. Es cada vez más habitual encontrar estudios en biomedicina que hacen uso de las redes de conexión para explorar datos y mostrar los resultados de los análisis. [153, 1].

Capítulo 2

OBJETIVOS

Los objetivos de esta tesis son:

A) Identificar qué factores están asociados con la presencia de *missing data* en otras variables en las bases de datos médicas.

   (1) Encontrar estrategias de análisis que posibiliten el calculo de los modelos de regresión cuando el tamaño de los datos es demasiado grande para la memoria del ordenador.

   (2) Identificar que factores relacionados con el paciente o el centro médico están asociados con la presencia de *missing data* en los datos del servicio público de salud del Area-7 de Madrid.

B) Estudiar las secuencias de eventos recogidas en bases de datos médicos de gran tamaño, incluso cuando la información temporal es incompleta.

   (3) Desarrollar un método sencillo con el que poder analizar secuencias de eventos médicos en presencia de recurrencia y competición.

   (4) Adaptar la metodología desarrollada para el análisis de las secuencias de eventos cuando sólo se dispone de información temporal parcial.

   (5) Explorar modos eficientes de visualizar las secuencias de eventos estudiadas.

   (6) Analizar las secuencias de eventos de la base de datos del servicio público de salud del Area-7 de Madrid.

Capítulo 3

METODOLOGÍA

3.1 IDENTIFICACIÓN DE FACTORES ASOCIADOS CON LA PRESENCIA DE *MISSING DATA*

3.1.1 Ámbito del estudio

El análisis de los *missing data* se realizó con la base de datos del servicio público de salud del Area-7 de Madrid[2], cuya delimitación mostramos en la Figura 3.1. Los datos contienen 1401 variables diferentes recogidas a 224 321 individuos a quienes se ha realizado un seguimiento desde 2006 hasta 2010. Esta información posibilitará el estudio de *missing data* y secuencias temporales de eventos médicos. La población diana es la población de Madrid y población española de similares características. El analista no ha realizado ningún diseño previo del experimento ni manipulado de ningún modo a los pacientes, simplemente ha utilizado la información procedente de todos los individuos contenidos en la base de datos para realizar análisis estadísticos.

[2] El Area-7 de Madrid está formada por 3 distritos sanitarios (Latina, Centro y Chamberí), 550 000 habitantes censados, 21 centros de salud, 2 hospitales (Clínico San Carlos y Fundación Jiménez Díaz), 3 centros de especialidades, 23 equipos atención primaria, 2 centros sanitarios especiales, 3 servicios de urgencias, 4 unidades de salud buco-dental y 5 unidades de fisioterapia. La demarcación y denominación de las diferentes áreas ha sufrido múltiples modificaciones en los últimos años.

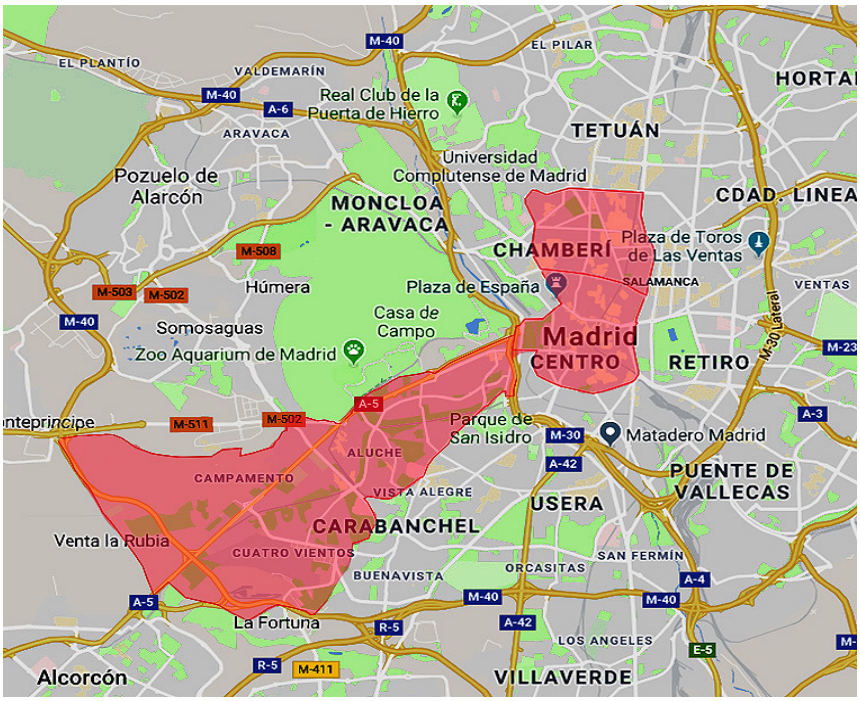


Figura 3.1: Delimitación geográfica del Area-7 del servicio público de salud de Madrid

3.1.2 Limpieza y reestructuración de los datos

Una fase habitualmente obviada en muchos estudios es la limpieza y preprocesamiento de los datos. En esta tesis queremos resaltar la importancia de este proceso por tratarse del más laborioso, y sin la cual no habría sido posible elegir las variables a estudiar ni continuar con los demás análisis. Se han realizado multitud de correcciones y transformaciones que merecen ser consideradas en detalle.

Los datos estudiados fueron obtenidos de una tabla en formato SPSS. Para poder limpiar y analizar los datos los hemos exportado a R, utilizando para ello el formato CSV ya que el intento de importar directamente el archivo .sav producía todo tipo de errores. La posibilidad de realizar todas las operaciones directamente con el programa SPSS fue descartada por ser extremadamente lento incluso para las operaciones más simples, llegando incluso a colgarse por su consumo excesivo de memoria. Hemos utilizado la versión 3.3.2 de R, con la ayuda de la librería data.table, que permite manejar grandes cantidades de datos eficazmente y la librerías lme4 y glmmTMB

para ajustar los modelos de regresión. También se ha utilizado Excel para crear una tabla con información resumida sobre cada variable y los procesos utilizados.

Los datos originales mostraban signos de haber sido manipulada por diferentes personas a lo largo del tiempo, han utilizado diferentes codificaciones para los datos. Por ejemplo, en algunas variables se indicaba la ausencia o presencia de una enfermedad para algunos años o personas mediante "0" y "1", mientras que para otros años o personas se utilizaba "S" y "N". Para poder realizar adecuadamente los análisis se han unificado los criterios.

Algunas de las variables contenidas en la tabla original sólo fueron tomadas para algunos pacientes o ciertos años, dejando vacío su valor para el resto. Es lo que se conoce como "*missing data*" o valores perdidos para esas variables. En otros casos se utilizaron deliberadamente valores de la variable claramente erróneos o fuera del rango de los posibles con la intención de indicar la presencia de *missing data*. El problema es que en muchos casos, incluso para una misma variable, el criterio utilizado fue cambiando: NA, NULL, 99999, fechas inexistentes, números negativos, etc. Además, la base de datos también contenía errores no deliberados, para la detección de los cuales ha sido necesario analizar las variables una a una. En nuestro estudio todos esos errores han sido considerados simplemente como *missing data* y han sido reemplazados por NA conforme a la sintaxis de R, tal como se detalla en el Apéndice B. No se ha pretendido valorar si los valores numéricos son ligeramente inferiores o superiores a lo habituales, sólo hemos considerado erróneos los registros que no contienen datos del tipo correcto o que son ilógicos.

Para poder trabajar adecuadamente se ha reestructurado la tabla. Los datos originales aparecían en formato "wide", es decir, las variable medidas diferentes años para un mismo individuo habían sido registrada como diferentes variables en una única fila y ocupando tantas columnas como años, por ejemplo, Peso_2006, Peso_2007, ..., Peso_2010. A partir de esa tabla hemos generado otra tabla en formato

"long" colapsando cada uno de esos conjuntos de variables en una única variable medida repetidamente, registrada en una sola columna y con tantas filas como años.

La Figura 3.2 muestra gráficamente el proceso con datos ficticios para la variable Peso. Podemos comprobar que existe una correspondencia biunívoca entre ambas representaciones. Además se ha creado una nueva columna que recoge el año. Lo mismo sucede con las demás variables anuales. Por otro lado, las variables que no se miden anualmente permanecen como estaban, ocupando una única columna. Nótese que la conversión puede generar nuevos *missings* en aquellas variables que sólo existiesen algunos años y no otros, por eso es necesario indicar previamente la posición de los *missings* originales, que son los que queremos estudiar, los posibles *missing* posteriores habrían sido generados artificialmente durante el postprocesado de la información y simplemente serían ignorados.

El nuevo formato de los datos permite hacer explicita la dependencia anual de las variables en nuestros modelos a la vez que posibilita la utilización de los datos de todos los años simultáneamente, tratando las variables como una misma y consiguiendo con ello mayor potencia estadística en los análisis. El Listado F.1 del Apéndice F muestra el código completo que hemos creado con este algoritmo.

En algunos casos la presencia de *missings* en los registros de una variable está estrictamente condicionada al valor de otra variable diferente. Diremos entonces que los *missings* de la primera están "vinculados" a la segunda. No es un fenómeno que suceda por azar sino que ha sido diseñado de ese modo. Por ejemplo, sólo consideramos *missing* la ausencia de dato en el número de cigarrillos que fuma un paciente si éste consta como fumador, y sólo consideramos *missing* la fecha de una intervención si realmente sucedió dicha intervención. Es decir, en las variables vinculadas sólo consideraremos *missings* los registros que realmente debían contener algún dato válido, el resto de valores faltantes o erróneos simplemente se ignoran y no se contabilizan para calcular las proporciones.

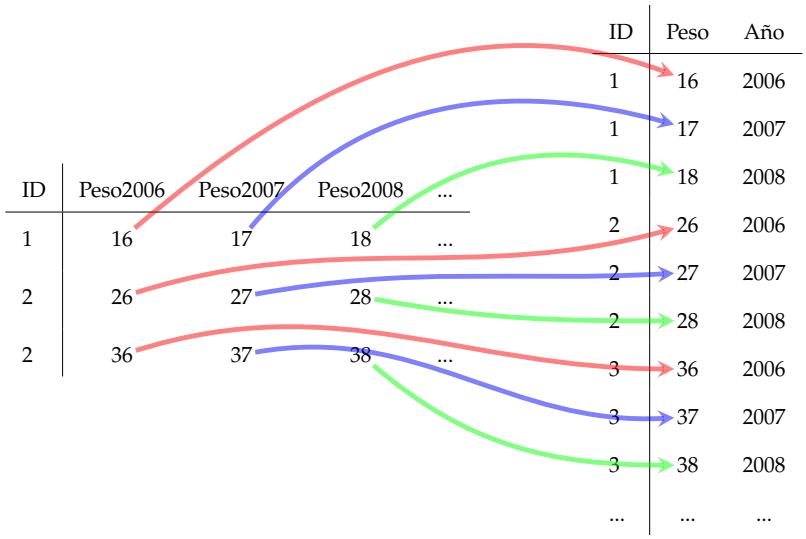


Figura 3.2: Transformación de los datos de formato Long a Wide. Ejemplo para una variable Peso tomada los años 2006, 2007 y 2008.

En el resto de variables, las no vinculadas a ninguna otra, consideramos *missing* cualquier valor faltante o error. La decisión de catalogar la variable como vinculada se ha tomado discrecionalmente, según el criterio del investigador. Podría haber otras variables vinculadas que se han pasado por alto. Para evitar ese problema se desarrolló un método automático para intentar identificarlas, comprobando en que variables todos los casos son *missing* cuando otra tiene valores *missing*, pero este método arrojó más de 6000 posibles combinaciones, demasiadas para ser verificadas por una persona, y que además podrían ser debidas al azar.

Se ha limpiado la tabla original de datos utilizando nombres válidos para las variables, eliminando las columnas duplicadas o sin ningún contenido valido o que no aportaban información adicional. Tras el proceso de reestructuración y limpieza de los datos hemos obtenido una tabla de 522 variables y 1 345 917 registros.

Nos hemos encontrado con dos grandes dificultades:

- La base de datos es demasiado grande para poder ser analizada en un ordenador promedio con los paquetes habituales de análisis estadístico, los cálculos son

muy extensos, y muchas veces ni siquiera pueden realizarse porque el proceso necesita más memoria RAM de la que dispone el ordenador, llegando incluso a hacer que el programa deje de funcionar.

- La base de datos contiene muchos errores y estos son de diferente tipo, lo que parece indicar que diferentes personas han manipulado la base de datos o han cambiado los criterios a lo largo del tiempo.

3.1.3 Métodos estadísticos

3.1.3.1 Generalidades y variables objeto de estudio

Queremos analizar la ocurrencia de *missing data* en las variables que puedan ser relevantes para la atención sanitaria. No estamos interesados en estudiar los valores concretos que adquieren estas variables sino en la presencia o ausencia de valores en sus registros, es decir, en la ocurrencia de *missings*, circunstancia que no tiene por qué estar relacionada con el valor de la propia variable, que además es desconocido para ese registro.

Modelamos entonces la ocurrencia de *missings* en una variable en función de los valores concretos que toman otras variables *predictoras* o covariables. La selección de estas covariables se ha realizado escogiendo manualmente aquellas consideradas importantes a juicio del analista y posteriormente minimizando el *criterio de información bayesiano* (BIC) mediante *backward stepwise selection*. Para lo cual se utilizó el conjunto completo de los datos. Estos modelos fueron reutilizados con los datos particionados por centro y con los particionados aleatoriamente descritos en los siguientes apartados. [3]

[3] Si cayésemos en la tentación de probar modelos con todas las posibles combinaciones de variables de nuestra base de datos obtendríamos resultados significativos para muchas de ellas, pero muchos sucederían por puro azar y deberíamos corregir los p-valores con métodos multitesting, lo cual a su vez afectaría negativamente a las variables que realmente sí producen un efecto significativo. Lo adecuado es preseleccionar previamente, con algún criterio lógico y la experiencia, unas pocas variables candidatas a ser predictoras y que pueden tener interés médico.

A priori pensamos que sería de interés el análisis de la presencia de *missings* en las siguientes variables *respuesta*: CIGARRILLOS, IMC, TIPO PROFESIONAL, TIPO DE USUARIO, COLESTEROL, EXTRANJERO, PESO, DIABETES, ALCOHOLISMO, PESO, FECHA DE NACIMIENTO, FUMADOR, TENSIÓN ARTERIAL, COLESTEROL, NÚMERO DE INTERVENCIONES MÉDICAS, NÚMERO DE PROTOCOLOS y PRESCRIPCIONES.

Como variables predictoras del paciente proponemos las siguientes: ID, YEAR, SEXO, EXTRANJERO, ALCOHOLISMO, TABACO, GRIPE, DIABETES, EDAD, PESO, RENTA, IMC, TOTAL DE FÁRMACOS DESDE LOS 65 AÑOS, TOTAL DE ANALITICAS, MFCONCENTRO (total de consultas concertadas por el paciente al médico en centro), ENFCONCENTRO (total de consultas concertadas por el paciente a enfermería en centro) y NÚMERO DE INGRESOS.

Como variables predictoras propias del centro: EAP (Equipo de Atención Primaria o centro médico), PAMED (presión asistencial del médico de familia, pacientes/día), PAENF (presión asistencial media de enfermería del EAP, pacientes/día) y POSMAS (porcentaje de pacientes con edad mayor o igual a 65 años del EAP).

No debemos confundir la presencia de *missings* en las variables predictoras con los *missings* en la variable respuesta que queremos estudiar, en nuestro caso PESO o CIGARRILLOS. Los primeros perjudican los análisis ya que obligan a eliminar la fila completa de datos, los segundos son el objeto de nuestro análisis y simplemente se convierten en unos y ceros de la variable respuesta en nuestros modelos logísticos.

Un primer intento de ajustar los modelos producía errores de convergencia que impedían proseguir con los análisis. Fue entonces necesario estandarizar todas las variables predictoras. Para la mayoría de las variables se utilizó automáticamente la transformación (variable-media)/desviación típica. Tres variables de interés en nuestro estudio fueron PAMED, PAENF y POSMAS se convirtieron entonces en (PAMED-32.176)/3.788, (PAENF-17.980)/2.380 y (POSMAS+0.069)/1.012

respectivamente. Otras variables se estandarizaron manualmente mediante un criterio que ofrecía resultados similares pero más fácil de interpretar: para el año se utilizó (YEAR-2006), para la edad (EDAD-50)/10, para el (PESO-70)/10. Nótese que la estandarización no afecta a los p-valores, sí a los coeficientes, pero a partir de ellos podemos calcular fácilmente los valores originales.

Se aplicaron análisis de regresión logística para obtener modelos que expliquen la proporción de *missings* en cada variable *respuesta* en función de las covariables *predictoras* [72]. Además, en caso necesario, se añadieron al modelo efectos aleatorios para así incorporar adecuadamente las medidas repetidas por individuo y/o centro médico y de ese modo tener en cuenta la variabilidad intrínseca de estos factores [13].

Incluimos el factor individuo (ID) cuando consideramos que la presencia de *missings* en los datos de una persona pudiera estar relacionada con ella, es decir, con factores intrínsecos a la persona no incluidos ya explícitamente en el modelo, como por ejemplo el parentesco con el médico, el atractivo físico, la religión, otras variables biológica o que algunos pacientes piden al médico la medición de más variables o se niegan a ofrecer datos o acuden sólo los viernes o el número de veces que esa persona es atendida. En caso contrario, si consideramos que los *missings* afectan a las personas independientemente de quienes sean, no incluimos el factor individuo en el modelo. No conocemos a priori cual de los dos criterios es el correcto, por lo que ejecutamos ambos modelos y los comparamos.

La inclusión en los modelos de una variable categórica como efecto fijo permite estimar un intercepto diferente para cada nivel de la variable, pero necesita tantos parámetros como niveles tenga esta. Si el número de niveles es elevado, como en nuestro caso ID, con $200\,000$ valores, los cálculos se complican, se reduce la potencia del modelo e incrementa el error estándar de los coeficientes estimados. En su lugar podemos incorporar la variable como efecto aleatorio, modelándola a través de una distribución con un número reducido de parámetros, habitualmente media y

varianza.

Para cuantificar la proporción de varianza explicada por los efectos fijos y por los efectos aleatorio utilizaremos las R^2 marginal y condicional definidas por Nakagawa y Schielzeth, [111].

3.1.3.2 Memoria y tiempo de cálculo

Las herramientas habitualmente utilizadas para realizar análisis estadísticos son poco adecuadas para lidiar con grandes cantidades de datos. La dificultad es incluso mayor cuando no sólo necesitamos transformarlos sino también realizar complejos ajustes estadísticos. El principal problema es que los algoritmos utilizados para realizar los ajustes utilizan grandes cantidades de memoria y son muy lentos, más cuanto mayor sea la base de datos.

Los algoritmos empleados para ajustar los modelos de regresión con efectos aleatorios son lentos y utilizan grandes cantidades de memoria, en el mejor de los casos, con los modelos lineales sin anidamiento, la memoria necesaria muestra una dependencia $O(nkp)$, donde n es el tamaño de los datos, p el número de variables y k el rango de la matriz de covarianzas, [163, 47].

Analizamos en detalle y con datos reales la memoria consumida y el tiempo necesarios para ejecutar diferentes modelos de regresión en función del número de filas (registros) de la base de datos y de las diferentes librerías utilizadas. Para ello, seleccionamos de la base de datos original aquellas filas que no contienen ningún *missing*[4] en las covariables de nuestros modelos. A continuación extraemos muestras de diferente tamaño y realizamos los ajustes, detallados en el Apartado 3.1.3.3, con cada una de esas muestras. El modelo completo utilizado ha sido $mPESO \sim SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + (1|EAP/ID)$, que

[4] En caso contrario, algunas muestras tendrán diferente número de filas, las que contienen *missings* son desechadas durante los análisis sin que el usuario se aperciba de ello, y la comparación no seria justa.

desarrollamos a lo largo de los próximos apartados.

Los análisis complejos de bases de datos de gran tamaño resultan extremadamente lentos e incluso imposibles de realizar en un ordenador "medio" con las librerías habituales de análisis estadístico, como lme4, dada la gran cantidad de memoria RAM necesaria. Para poder superar esta dificultad ideamos tres enfoques diferentes para poder analizar los datos, cuyos resultados compararemos:

a) Realizamos el análisis con el set completo de datos, pero engañando (forzando) a R para que utilice el propio disco duro "SSD" como si fuese memoria RAM, además utilizamos el paquete experimental glmmTMB.
b) Subdividimos aleatoriamente el conjunto de los datos en x trozos de menor tamaño. Ajustamos un modelo logístico sobre cada trozo por separado. Después unificamos todos los resultados mediante meta-análisis.
c) Realizamos el estudio por separado para cada centro en función de los factores que varían dentro de éste, tendremos 20x5 regresiones. Después unificamos los resultados mediante meta-análisis. Además haremos otra segunda regresión de los parámetros obtenidos dentro de cada centro en función de parámetros externos.

3.1.3.3 Análisis simultáneo del conjunto de los datos

El intento de analizar mediante un modelo de regresión con efectos aleatorios el conjunto completo de los datos utilizando los programas estadísticos habituales simplemente no funciona, produce un error o incluso hace que el programa se cuelgue por falta de memoria RAM. Añadirle más memoria al ordenador no es una solución escalable ya que la placa base está físicamente limitada. En nuestro caso, el ordenador dispone de 16GB y podría ser ampliado hasta un máximo de 32GB.

Hemos tenido que utilizar un paquete experimental, llamado glmmTMB, optimizado para trabajar con grandes cantidades de datos. Este paquete ha sido

creado a partir del motor de diferenciación automática Template Model Builder, se puede utilizar para ajustar modelos lineales generalizados con efectos mixtos. Este paquete asume que los efectos aleatorios son gaussianos en la escala del predictor lineal y los integra utilizando la aproximación de Laplace, más precisa que la quasi-verosimilitud penalizada pero menos que la cuadratura de Hermite utilizada por lme4 o que las cadenas de Markov.

Además, ha sido necesario engañar a R para que utilice el disco SSD para paginar los datos como si fuese la memoria RAM, haciéndole creer que dispone de 350GB de RAM. Para ello debemos añadir la instrucción *invisible(utils::memory.limit(350000))* al archivo .Rprofile, pero este "truco" sólo funciona con algunas librerías y a veces falla.

Los modelos logísticos con los que se ha ajustado el conjunto completo de datos han sido los siguientes:

Modelo simple, regresión logística con efectos fijos, sin tener en cuenta las medidas repetidas por individuo:

```
glm(mPESO ~ SEXO + YEAR + EDAD + PAMED + PAENF + EXTRANJERO + EAP,
        data=todos, family="binomial")
```

Modelo de efectos aleatorios con medidas repetidas por centro "EAP":

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
        (1|EAP), data=todos, family="binomial", REML=F)
```

Modelo de efectos aleatorios con medidas repetidas por individuo "ID":

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + EAP +
        (1|ID), data=todos, family="binomial", REML=F)
```

Modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro:

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
        (1|EAP/ID), data=todos, family="binomial", REML=F)
```

Hemos conseguido efectuar las regresiones pero el proceso es extremadamente lento dado el gran tamaño de los datos, los SSD tienen velocidades típicamente 100 veces inferiores a la de la RAM. De este modo cada regresión con el set completo de datos se puede demorar más de un día, y el proceso debe repetirse con cada combinación de parámetros que deseemos estudiar. En nuestro caso, hemos tenido que ir guardando los resultados parciales de cada proceso para no tener que repetir todos los cálculos anteriores en cada sesión.

Nótese además que parte del disco no está disponible porque es utilizado por archivos del usuario o por el propio sistema operativo. Esta solución no deja de ser un mero parche al problema. Podría suceder que la base de datos sea demasiado grande incluso para el disco SSD, en cuyo caso deberíamos realizar cálculo distribuido con herramientas como Spark y JuliaDB, mucho más complejas y que de momento no ofrecen la posibiliad de realizar modelos avanzados como los de efectos aleatorios, además de que son muy lentas.

3.1.3.4 Análisis de los datos particionados aleatoriamente

En este apartado proponemos otro enfoque para analizar los datos, que puede solventar parcialmente las dificultades para analizar bases de datos de gran tamaño. Subdividimos los datos en particiones escogidas aleatoriamente sin repetición. En nuestro caso 19 partes de igual tamaño. Elegimos este número de particiones para que coincida con el número de centros del apartado siguiente y obtener así resultados fáciles de comparar.

A continuación ajustamos separadamente un *minimodelo* logístico de efectos aleatorios con los datos de cada partición aleatoria. Los coeficientes obtenidos mediante estos minimodelos “intra-partición” explican para cada partición la

presencia de *missing data* en función de variables del paciente, de las variables internas del centro y del año. Los modelos aplicados son idénticos a los anteriormente utilizados con el conjunto completo de los datos, y su resultado debería arrojar coeficientes parecidos, sólo deberían cambiar los p-valores ya que cada partición dispone de menos datos.

En una segunda fase obtendremos una estimación de los coeficientes del modelo para el conjunto completo de datos, combinando los resultados de todas las particiones. Para ello realizaremos el meta-análisis de los coeficientes y errores estándar obtenidos en los minimodelos utilizando el paquete *metafor*, con la opción "FE", Fixed Effect, es decir, el modelo asume que todas las particiones proceden de una misma población homogénea y tienen las mismas características estadísticas, las diferencias entre ellas sólo serán debidas al azar. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

3.1.3.5 Análisis de los datos particionados por centro

En este apartado, en lugar de subdividir los datos aleatoriamente, los particionamos por centros médicos, obteniendo así 19 bloques de datos diferentes.[5] Con los datos de cada una de esas particiones por centro ajustamos separadamente un modelo logístico de efectos aleatorios, incluyendo sólo los factores que varían dentro de ese centro o en cada persona.

Los coeficientes obtenidos mediante estos minimodelos "intra-centro" explican para cada centro la presencia de *missing data* en función de variables del paciente, de las variables internas del centro y del año. Combinamos todos estos resultados realizando un meta-análisis de los coeficientes y sus errores estándar. Utilizamos para ello el paquete *metafor*, con la opción "ML", Maximum Likelihood, es decir, el modelo

[5] La base de datos original contiene información de 20 centros pero uno de ellos presenta demasiados errores y algunas de las variables en él recopiladas no han sido tomadas año tras año sino que se han repetido los valores artificialmente, por eso se ha tomado la decisión de eliminar ese centro.

asume que los resultados obtenidos para cada centro pueden variar más allá de lo que causaría el azar, debido a factores intrínsecos a cada centro. Es decir, en este apartado no se asume que las particiones provengan de una misma población homogénea. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

También realizamos un segundo set de regresiones para estudiar si los coeficientes obtenidos en cada modelo intra-centro también dependen de los parámetros que varían entre diferentes centros, y que no pudieron ser analizados simultáneamente en el paso previo porque la información era parcial. Sólo hemos considerado la variable inter-centro POSMAS.[6]

3.1.3.6 Modelos más complejos

En los apartados anteriores hemos utilizado modelos de complejidad media ya que la finalidad era simplemente comparar los resultados al particionar los datos de tres modos diferentes con un modelo funcional. La utilización de modelos más complejos, con interacciones o más variables, producía errores al ser aplicados a los datos fragmentados aleatoriamente y especialmente con los datos fragmentados por centros. En este nuevo apartado la finalidad es buscar el mejor modelo posible para el conjunto completo de los datos.

[6] Otra opción sería particionar los datos simultáneamente por centro y año, y realizar un ajuste dentro de cada uno de esos fragmentos. En este caso el modelo no tendría efectos aleatorios ya que en cada fragmento para cada persona sólo hay datos de un año, ya no hay medidas repetidas y además habríamos aislado el efecto de cada centro. Pero no podemos utilizar este método con nuestros datos porque muchas variables no se han tomado cada año en cada centro, simplemente se han replicado las medidas, por lo que producen errores al intentar utilizarlas en un modelo anual.

3.2 ANÁLISIS DE LAS SECUENCIAS DE EVENTOS

3.2.1 Ámbito del estudio.

Para el estudio de las secuencias de eventos hemos vuelto a utilizar la bases de datos del servicio público de salud del Area-7 de Madrid, que contiene 1401 variables diferentes recogidas a 224 321 individuos, a los que se ha realizado un seguimiento desde 2006 hasta 2010. A partir de los datos en formato long, Apartado 3.1.2, hemos extraído para nuestro análisis dos tipos de variables: aquellas que contienen fechas exactas (día/mes/año), y aquellas en las que sólo se recoge el año en el que ocurrió el evento (año). La mayoría de las variables de la base de datos pertenecen al segundo tipo, no recogen la fecha exacta del evento, que aparece truncada. Con lo que, además de tener una precisión de sólo un año, no es posible saber en que orden sucedieron los eventos interanuales de un mismo individuo.

Para evitar tener que descartar parte de la información y obtener resultados sesgados hemos desarrollado un método de imputación para las fechas de las variables truncadas. Testearemos la validez de este enfoque comparando los resultados obtenidos utilizando fechas exactas con los obtenidos al truncar manualmente esas mismas fechas. Y aplicaremos el mismo método a las variables que ya aparecen truncadas en la base de datos original.

Tras limpiar los datos, estudiamos las secuencias temporales de eventos, enfocando el análisis en las subsecuencias de menor longitud: los pares y tripletes. Ignoraremos el orden absoluto de los eventos ya que no estamos interesados en la posición exacta en la que tuvo lugar cada evento de ese individuo sino en sus diferencias relativas. Queremos saber por ejemplo si tras un ataque al corazón la persona falleció, no si ese ataque al corazón fue su quinto evento.

Cada individuo pudo haber sido diagnosticado del mismo tipo de evento repetidamente, lo cual complica los cálculos y posibilita diferentes modos de resumir

las secuencias: cuantas veces sucede un tipo de evento por cada vez que ha sucedido antes otro, o por cada vez que antes ha sucedido otro al menos una vez, o por cada vez que sucede otro posterior, etc. Nos centraremos entonces en una estadística fácil de interpretar y que no se presta a confusión: cuántas personas han sido diagnosticadas al menos una vez de una determinada secuencia de eventos.

También debemos tener en cuenta el tiempo transcurrido entre eventos. Un modo simple de hacerlo es utilizando la Tasa de Incidencia, en inglés Incidence Rate (IR), habitualmente definido como "número de veces que ocurre un evento divido por el tiempo de exposición al riesgo" [157]. En nuestro caso, lo definiremos como "número de veces que una persona ha sufrido una determinada secuencia de eventos al menos una vez, dividido por el tiempo de exposición al riesgo desde el penúltimo evento hasta el último o, en el caso de no existir último evento, hasta el final del estudio o fallecimiento.

Realizamos un primer análisis considerando sólo las subsecuencias de eventos consecutivos (Apartado 3.2.2). Posteriormente realizamos un segundo análisis considerando cualquier subsecuencia de eventos aunque no sean consecutivos (Apartado 3.2.3). Representamos gráficamente los resultados mediante un diagrama de nodos interconectados, llamado red de conexiones, en el que cada nodo representa un tipo de evento, y el grosor de las conexiones se calcula según el criterio que establezcamos, en nuestro caso el IR. El Listado F.7 del Apéndice F muestra el código utilizado para generar y dibujar las redes de conexiones para los tripletes y para asignar diferentes colores a las transiciones que pasan por un mismo modo. El análisis se ha realizado utilizando el programa R v3.4.1 y las librerías Diagrammer, VisNetwork, data.table, Lubridate y ggplot2.

3.2.2 Pares de eventos consecutivos

Comenzamos el análisis de las secuencias del modo más simple posible, considerando sólo las transiciones entre pares de eventos consecutivos, con los que

calculamos cuántas personas al menos una vez han sido diagnosticadas de un tipo de evento si inmediatamente antes fueron diagnosticadas de otro dado, ignorando los anteriores eventos. El resultado lo dividimos por el tiempo total en riesgo para ese par de eventos, obteniendo así su IR. A continuación detallamos el procedimiento de cálculo mediante un ejemplo con datos ficticios para dos personas. El Listado F.3 del Apéndice F muestra el código utilizado para analizar los datos reales.

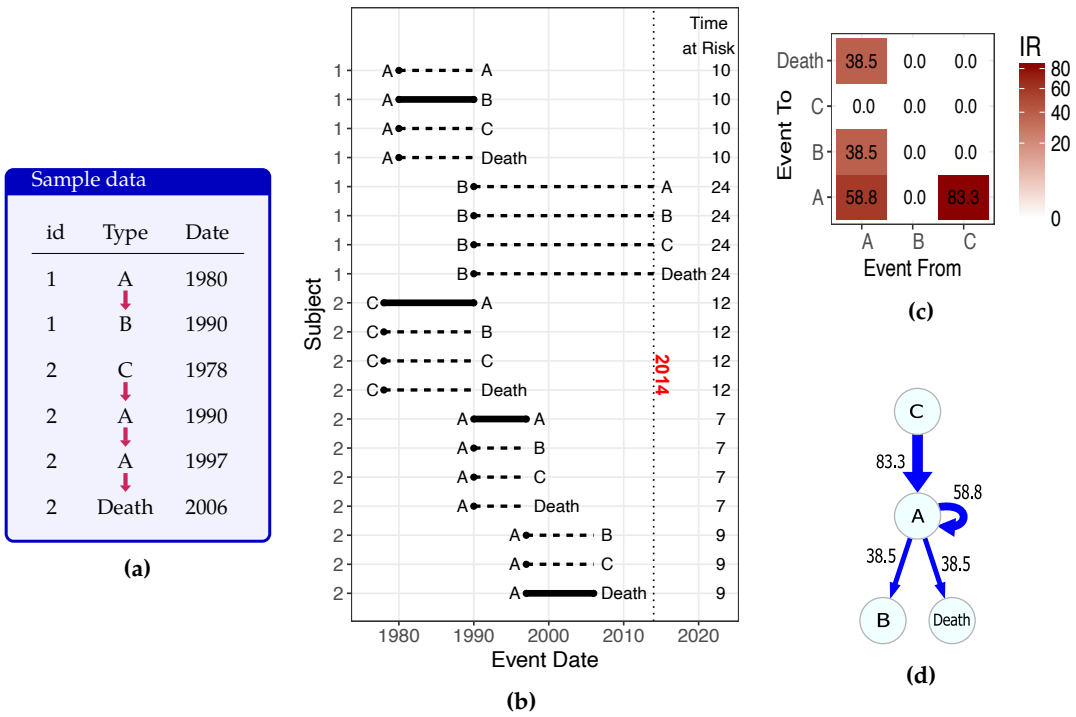


Figura 3.3: Procedimiento para calcular la Tasa de Incidencia (IR) de los pares de eventos consecutivos. a) Datos de ejemplo y transiciones generadas. b) Persona-año en riesgo para cada tipo de pares de eventos. Seleccionamos la primera ocurrencia de cada par y generamos todos los demás pares posibles. c) Tabla de IR para cada par (Eventos por cada 1000 Persona-Año). d) Red de conexiones de los IR.

- En la tabla de la Figura 3.3a podemos ver que la persona con id=1 sufre un evento A en 1980 y después otro evento B diez años más tarde. La persona id=2 sufre cuatro eventos: C, después A, después otro A y finalmente fallece. Nótese que una persona puede sufrir el mismo tipo de evento varias veces. El estudio finaliza en 2014. Las flechas rojas indican todas las transiciones (saltos) entre eventos consecutivos para cada persona, es decir el hecho de sufrir un evento e

inmediatamente después otro.

- Contamos el número de personas, N, que sufren (al menos una vez) cada par de tipos de eventos consecutivos. En Figura 3.3b observamos para la primera persona el par A→B, y para la segunda persona los pares C→A, A→A y A→Death. Por lo que cada uno de estos pares ha sido contabilizado en total en una persona. Tras la ocurrencia de un par la persona queda etiquetada como que ya lo ha sufrido y posteriores ocurrencias del mismo no se contabilizarían de nuevo. Señalamos los pares de eventos con línea oscura continua si realmente se ha registrado la ocurrencia de ambos eventos consecutivamente, y mediante línea discontinua si se ha registrado la ocurrencia del primer evento del par, quedando así la persona expuesta al segundo, pero no se ha llegado a registrar el segundo evento del par antes de que ocurriese otro evento diferente.
- Calculamos el tiempo en riesgo para la primera ocurrencia de cada par de tipos de eventos consecutivos y su exposición previa en cada persona, es decir, para cada par de tipos eventos consecutivos que sufre una persona calculamos el tiempo transcurrido desde la primera vez que sufre el primer tipo de evento del par estudiado hasta el evento siguiente si se correspondiese con el segundo evento del par estudiado. Si no ocurriese consecutivamente el segundo evento del par se irán sumando los tiempos de exposición a partir de las nuevas ocurrencias del primer tipo hasta que finalmente suceda consecutivamente el segundo evento del par buscado o hasta el fin del estudio o fallecimiento de esa persona. Una vez ha sucedido el par consecutivo, posteriores ocurrencias del mismo no computan para el cálculo del tiempo de exposición. En nuestro ejemplo la persona id=1 sufre un evento A, a partir del cual queda expuesto a sufrir cualquier otro evento A, B, C o Death hasta que realmente suceda uno de ellos. Diez años más tarde sufre B, resultando en una exposición a A, B, C y Death de diez años (tras haber sufrido A). Entonces habrá finalizado su exposición a futuros eventos a partir de A y comienza la exposición a nuevos eventos a partir de B. Puesto que la persona id=1 ya no sufre ningún otro evento

tras B simplemente queda expuesto a todos ellos hasta el final del estudio, es decir contabilizaremos 24 años para el tiempo de exposición de cada uno a partir de B. Para id=2 vemos como en 1978 sufre su primer evento, C. A partir del cual queda expuesto al riesgo de sufrir nuevos eventos de cualquier tipo hasta que en 1990 sufre un A. Por lo que tras C habrá estado expuesto a A, B, C y Death durante 12 años. Tras A volvemos a contar el tiempo de exposición, estará expuesto a A, B, C y Death durante 7 años, que es cuando sufre otro A. Por último, tras ese segundo A estará expuesto a B, C y Death (pero no a otro A porque este individuo ya quedó etiquetado antes como que ha sufrido el par A→A y no se le volvería a contabilizar el mismo par), y puesto que su siguiente y último evento es Death en 2006, la exposición a los tres eventos tras A habrá sido de 9 años.

- Sumamos los tiempos, para cada par de tipos de eventos, T, tanto los que sucedieron como los que sólo estuvieron expuestos. Obtenemos así la exposición total medida en "personas-año". Por ejemplo para el par A→A tenemos 10+7=17 personas-año.
- La tabla de la Figura 3.3c muestra el resultado del cálculo de la Tasa de Incidencia de cada tipo de evento tras otro dado, IR = N/T. Con ello tendremos el número de personas que sufren al menos una vez un determinado par de eventos consecutivos por cada persona-año de exposición desde el primer evento del par.

$$IR_{A\to A} = \frac{1}{10+7} = 0.0588 \qquad IR_{A\to B} = \frac{1}{10+7+9} = 0.0385$$

$$IR_{A\to Death} = \frac{1}{10+7+9} = 0.0385 \qquad IR_{C\to A} = \frac{1}{12} = 0.0833$$

- Vemos por ejemplo, como en una misma persona, el diagnóstico de C es seguido inmediatamente de un diagnóstico de A (al menos una vez) en 83.3 casos por cada 1000 personas-año de exposición tras C, o dicho de otro modo, en 83.3 personas sucede al menos una vez el par C→A por cada 1000 personas-año.

El diagnóstico de A es seguido inmediatamente de fallecimiento en 38.5 casos cada 1000 personas-año, es decir, 38.5 personas cada 1000 personas-año fallecen tras haber sufrido A. 58.8 personas son diagnosticadas (al menos una vez) de un evento A seguido de otro evento A por cada 1000 personas-año de exposición. Por último obtenemos que 38.5 personas son diagnosticadas (al menos una vez) de un evento A seguido un evento B por cada 1000 personas-año de exposición.

- La Figura 3.4d muestra la misma información pero utilizando una red de conexiones. Cuando apliquemos estos análisis a los datos reales, de gran tamaño y con multitud de relaciones, será necesario filtrar parte de la información resaltando sólo los saltos más relevantes.

Para calcular el Intervalo de Confianza de IR, tanto en este apartado como en todos los demás, hemos utilizado el método exacto de Poisson, donde t es el tiempo total de exposición al riesgo, α es 5% y N es el número de ocurrencias, que tal como lo hemos definido en nuestro caso es el número de individuos en los que ha tenido lugar al menos una vez ese par de eventos.

$$IC_{IR} = \left(\frac{\chi^2_{2N,\alpha/2}}{2t}, \frac{\chi^2_{2(N+1),1-\alpha/2}}{2t} \right) \tag{3.1}$$

3.2.3 Pares de eventos consecutivos como no consecutivos.

En este apartado analizamos todas las secuencias de pares de eventos, tanto consecutivos como no consecutivos, ignorando otros posibles eventos intermedios y posibilitando de ese modo la detección de transiciones que antes pasaban desapercibidas. Si una persona tuvo un ataque al corazón, después una gripe y por último una embolia, sólo podríamos detectar que tras el ataque al corazón en algún momento tuvo una embolia utilizando el análisis no consecutivo. Este enfoque no se ha hallado en otros artículos.

Utilizando los mismos datos de ejemplo del apartado anterior, pero considerando también los saltos no consecutivos, volvemos a calcular IR. Deberemos adaptar el

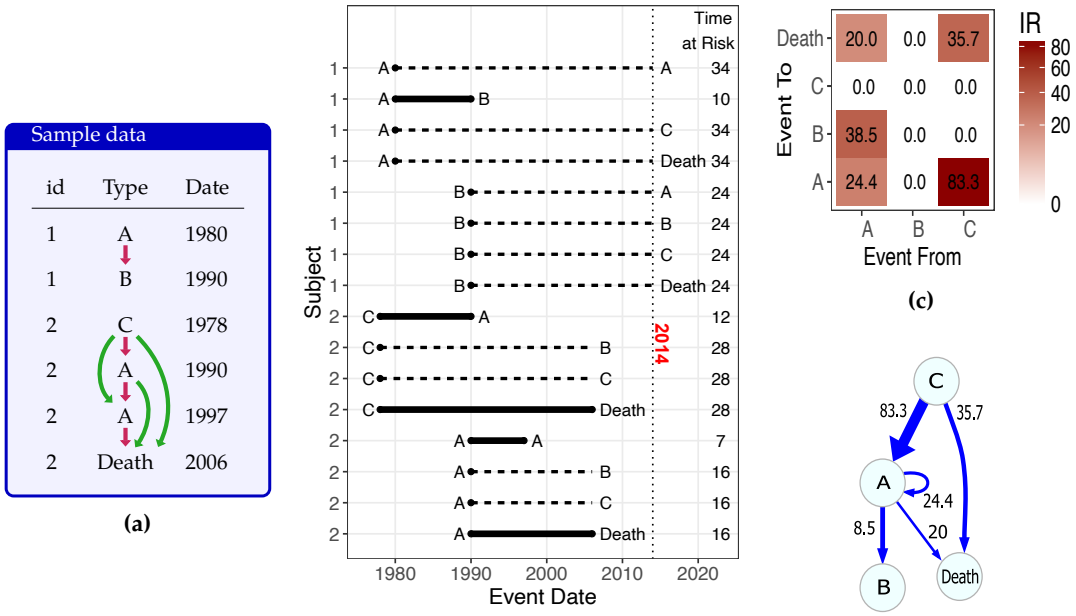


Figura 3.4: Procedimiento para calcular la Tasa de Incidencia (IR) de los pares de eventos de cualquier tipo, tanto consecutivos como no consecutivos. a) Datos de ejemplo y transiciones generadas. b) Persona-Año en riesgo. c) Tabla de IR para cada tipo de pares de eventos (Eventos por cada 1000 Persona-Año). d) Red de conexiones de los IR.

método de cálculo del tiempo de exposición al riesgo adecuadamente. El Listado F.2 del Apéndice F muestra el código utilizado para analizar los datos reales.

- Contamos el número de personas, N, que sufren al menos una vez cada par de tipos de eventos, tanto consecutivos (señaladas con flechas rojas en Figura 3.4a) como no consecutivos (señaladas con flechas verdes). En Figura 3.4b observamos para la primera persona el par A→B, y para la segunda persona los pares C→A, A→A, A→Death, C→Death. Por lo tanto, cada uno de esos pares ha sido contabilizado en una sola persona. Los pares C→A y A→Death aparecen una segunda vez en Figura 3.4a pero no los tenemos en cuenta para el cálculo de nuestro IR porque tal como lo hemos definido sólo nos interesa la primera ocurrencia de cada par de eventos, tras la cual, la persona queda ya etiquetada como que lo ha sufrido y posteriores ocurrencias de ese mismo par son ignoradas. Los pares de eventos se señalan con línea continua oscura si realmente se ha registrado su ocurrencia y mediante línea discontinua cuando hubo exposición al riesgo pero no sucedió el segundo evento del par.
- Calculamos el tiempo de exposición al riesgo para la primera ocurrencia de cada par de eventos en cada persona aunque no sean consecutivos, es decir, el tiempo transcurrido desde la primera vez que sufre un tipo de evento hasta la primera vez que sufre otro tipo de evento en cualquier momento aunque sucedan otros eventos entre ellos. En caso de no ocurrir nunca el segundo evento del par se calculará el tiempo hasta el fin del estudio o fallecimiento de esa persona. Posteriores ocurrencias de ese mismo par no computan para el cálculo del tiempo de exposición. Vemos por ejemplo como id=1 sufre un primer evento A, tras el cual queda expuesto a sufrir otros eventos A, B, C o Death. En 1990 sufre B, con lo que su exposición a B a partir de A es de 10 años. Además, sigue expuesto a otros A, C y Death hasta que finaliza el estudio, es decir 34 años a partir del primer A. A partir de B comienza su exposición a los eventos A, B, C y Death hasta que finaliza el estudio, es decir 24 años para cada uno.

- Continuamos con el proceso de cálculo del mismo modo que en el apartado anterior. Sumamos los tiempos para cada par de eventos, T, y calculamos con ellos IR=N/T como muestra la tabla de la Figura 3.4c.

$$IR_{A\to A} = \frac{1}{34+7} = 0.0244 \qquad IR_{A\to B} = \frac{1}{10+16} = 0.0385$$
$$IR_{A\to Death} = \frac{1}{34+16} = 0.0200 \qquad IR_{C\to A} = \frac{1}{12} = 0.0833$$
$$IR_{C\to Death} = \frac{1}{28} = 0.0357$$

- Multiplicamos el resultado por 1000 para facilitar su lectura. Vemos por ejemplo, como el diagnóstico de C es seguido en cualquier momento de un diagnóstico de A (al menos una vez) en 83.3 casos por cada 1000 personas-año. O como el diagnóstico de A es seguido en cualquier momento de fallecimiento en 20 casos cada 1000 personas-año, es decir, 20 personas cada 1000 personas-año fallecen si tuvieron en algún momento anterior al menos un A.
- La Figura 3.4d muestra la misma información pero utilizando una red de conexiones.

El análisis de transiciones no consecutivas nos ha permitido detectar la transición entre C y Fallecimiento, mediante el estudio de eventos consecutivo no se podía detectar.

3.2.4 Tripletes de eventos.

A continuación estudiamos las subsecuencias de tres eventos (tripletes) cualesquiera, tanto consecutivos como no consecutivos (lejanos). Este enfoque generaliza el anterior y puede ser útil para detectar relaciones más complejas en las que un tipo de evento no está asociado sólo a otro tipo dado sino a la combinación de dos.

Volvemos a utilizar los mismos datos pero esta vez analizamos las subsecuencias de tres tipos de eventos cualesquiera, tripletes, con las que calculamos IR. En este caso,

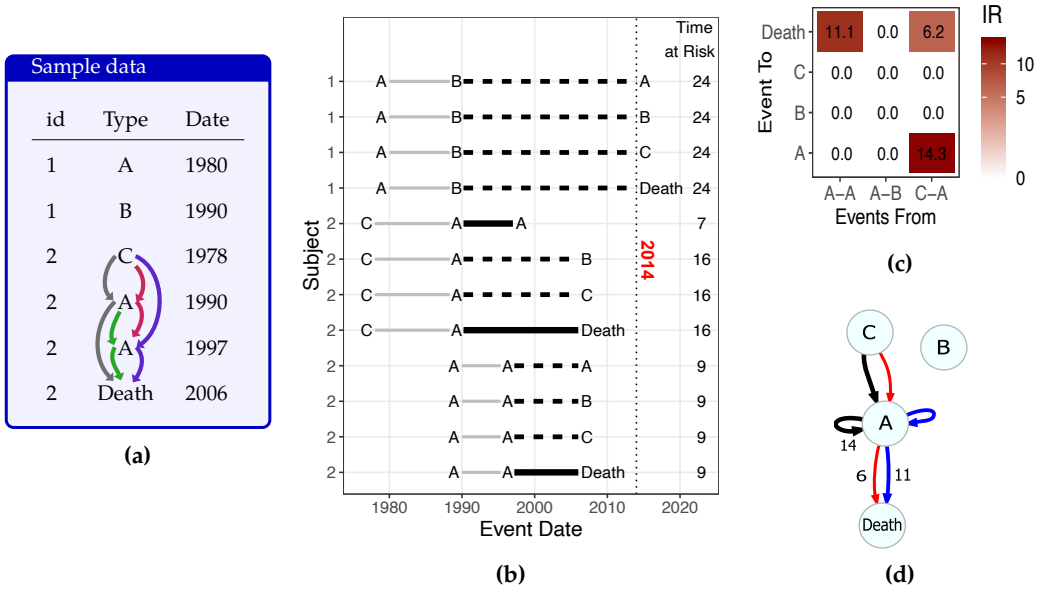


Figura 3.5: Procedimiento para calcular la Tasa de Incidencia (IR) para los tripletes de eventos. a) Datos de ejemplo y transiciones generadas. b) Persona-Año en riesgo para cada tipo de triplete de eventos. Seleccionamos la primera ocurrencia de cada triplete y generamos todos los demás pares posibles. c) Tabla de IR para cada triplete (Eventos por cada 1000 Persona-Año). d) Red de conexiones de los IR.

para calcular los tiempos de exposición al riesgo consideramos el tiempo transcurrido desde el segundo evento hasta el tercero de cada triplete. El Listado F.6 del Apéndice F muestra el código utilizado para analizar los datos reales.

- Contamos el número de personas, N, que sufren al menos una vez cada triplete de tipos de eventos, señalados en Figura 3.5a mediante flechas de colores. Vemos como la primera persona sólo tiene dos eventos, cantidad insuficiente para que tenga lugar ningún triplete. Para la segunda persona contabilizamos los tripletes C→A→Death, A→A→Death, C→A→A y C→A→Death. Cada uno de esos tripletes ha sido contabilizado en una sola persona. El triplete C→A→Death aparece dos veces en la misma persona pero sólo contamos su primera ocurrencia.
- En Figura 3.5b señalamos con línea continua gris tenue los pares de eventos iniciales que han sucedido en cada persona, con línea continua oscura el tiempo de exposición desde el segundo evento de cada par hasta el tercer evento de cada

triplete si realmente sucedió, y con línea discontinua el tiempo de exposición hasta un tercer evento si nunca llegó a suceder en nuestros datos pero la persona estuvo expuesta al riesgo de sufrirlo. Calculamos el tiempo de exposición al riesgo para la primera ocurrencia de cada triplete en cada persona aunque sus eventos no sean consecutivos, es decir, el tiempo transcurrido desde la primera vez que sufre un par de tipos de eventos hasta la primera vez que sufre otro tipo de evento en cualquier momento aunque sucedan otros eventos entre cualquiera de ellos. En caso de no ocurrir nunca el tercer evento del triplete se calculará el tiempo hasta el fin del estudio o fallecimiento de esa persona. Posteriores ocurrencias de ese mismo triplete no computan para el cálculo del tiempo de exposición. Vemos por ejemplo como id=2 sufre un evento C en 1978, luego un primer evento A en 1990, tras el cual queda expuesto a sufrir otros eventos A, B, C o Death. En 1997 sufre otro A,, con lo que su exposición a A a partir del par C→A es de 7 años. Además, sigue expuesto a otros B, C y Death hasta que fallece en 2006, es decir 16 años a partir del primer par C→A. Nótese que la fecha del primer evento no se utiliza para nada en el modelo simple que hemos creado.

- Continuamos con el proceso de cálculo de modo similar al apartado anterior. Sumamos los tiempos para cada triplete de eventos, T, y calculamos con ellos IR=N/T como muestra la tabla de la Figura 3.5c.

$$IR_{A\to A\to Death} = \frac{1}{9} = 0.1111 \qquad IR_{C\to A\to A} = \frac{1}{7} = 0.1429$$
$$IR_{C\to A\to Death} = \frac{1}{16} = 0.0625$$

- Multiplicamos el resultado por 100 para facilitar su lectura. Vemos por ejemplo, como el diagnóstico de un A seguido en cualquier momento de otro A es seguido en cualquier otro momento de un diagnóstico Death en 11.1 casos por cada 100 personas-año, es decir, 11.1 personas cada 100 personas-año fallecen si tuvieron en algún momento anterior al menos una vez el par A→A.

- La Figura 3.5d muestra la misma información pero utilizando una red de conexiones en la que se puede visualizar las trayectorias a lo largo de tres eventos.

3.2.5 Reconstrucción de las secuencias de eventos a partir de fechas truncadas

Para la mayoría de las variables contendidas en la base de datos no disponemos de la fecha exacta en la que tuvieron lugar los eventos, sólo se conoce el año, con lo que, además de tener poca precisión, no es posible saber en que orden tuvieron lugar los eventos interanuales y no podemos reconstruir la secuencia exacta para los individuos que sufrieron múltiples eventos durante un mismo año. Para evitar tener que descartar gran cantidad de información y obtener resultados sesgados hemos desarrollado un método de imputación que palía parcialmente el problema.

Sustituimos las secuencias en las que existe incertidumbre parcial o total en el orden de los eventos por la suma ponderada de un conjunto compatible de secuencias totalmente ordenadas obtenidas permutando los datos intraanuales, y a las que asignamos cierta probabilidad de ocurrencia. Nos aseguramos de este modo de que la secuencia que realmente sucedió está contenida entre las generadas, pero también estamos añadiendo también tiempo de exposición falso y generando secuencias que nunca sucedieron, incrementando falsamente su número. Aunque no conocemos la posición exacta de cada evento dentro del año sí que podemos calcular de modo aproximado los intervalos temporales entre eventos de diferentes años.

Ponderamos las secuencias de dos formas diferentes: a) asignando a todas las secuencias generadas la misma probabilidad de haber ocurrido realmente. b) asignando a las secuencias generadas probabilidades a partir de las frecuencias obtenidas previamente de los pares de eventos interanuales.

En la Figura 3.6 detallamos el procedimiento de generación de las secuencias

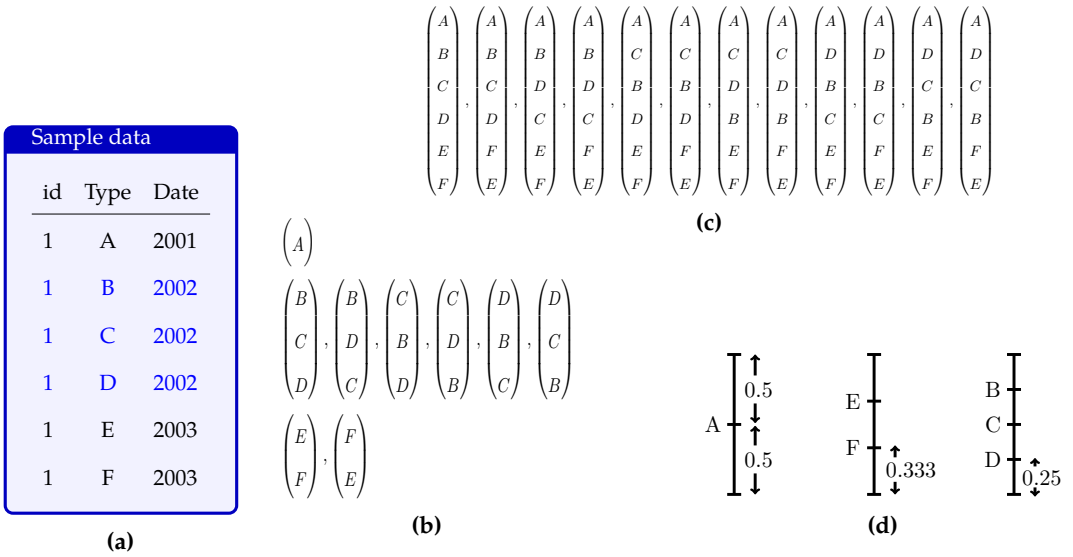


Figura 3.6: Procedimiento utilizado para la reconstrucción de las secuencias de eventos a partir de fechas truncadas: a) Datos de ejemplo para un individuo, en los que se desconoce el orden de los eventos interanuales. b) Permutaciones de los eventos dentro de cada año. c) Secuencias posibles de orden fijo obtenidas concatenando las permutaciones. El resultado final será la suma "ponderada" de todas ellas. d) Estimación de la fecha de los eventos ordenados según el número de eventos interanuales.

utilizamos datos ficticios para un individuo. La tabla de la Figura 3.6a contiene los datos de partida, en el año 2001 sufre un evento A, en el año 2002 sufre tres eventos, A, B, C, pero no sabemos en que orden, y en el año 2003 sufre otros dos, E y F. La Figura 3.6b muestra que para el año 2001, puesto que sólo hay un evento, sólo hay una posible subsecuencia, A. Para el año 2002 generamos las seis posibles permutaciones de los tres eventos, una de ellas será la que sucedió realmente. También generamos todas las posibles permutaciones para los dos eventos del año 2003. Finalmente concatenamos los resultados, Figura 3.6c, obteniendo un total de doce posibles secuencias compatibles con los datos originales. Nótese que el evento "Fallecimiento" sólo podría ocurrir en último lugar, descartaríamos las secuencias que lo contengan en otra posición.

En las secuencias generadas mediante este procedimiento el orden de los eventos queda perfectamente determinado, permitiéndonos estimar con mayor precisión la fecha en la que probablemente sucedieron, que a su vez permite estimar el tiempo

transcurrido entre los eventos interanuales, cosa que no se podía hacer con los datos originales.

Los años en los que ocurre un único evento el valor medio de la fecha de ocurrencia corresponde con la mitad del año, lo que en lenguaje común correspondería aproximadamente el 1 de julio. A los eventos únicos que suceden durante el año 2001 les asignaríamos la fecha 2001.5. Los años en los que suceden dos eventos la media de las fechas del primero de ellos se halla a un tercio del inicio del año. Y para los segundos eventos de un mismo año se halla a dos tercios del inicio. En el caso del año 2002 lo denotaríamos por 2002.333 y 2002.666 respectivamente. En general, en los años en los que ocurren *N* eventos el intervalo de tiempo transcurrido entre cada dos eventos consecutivos es de 1/(N+1), que puede ser fácilmente comprobado mediante simulaciones o matemáticamente. Por ejemplo, si en el año 2003 tuviésemos 3 eventos les asignaríamos las fechas 2003.25, 2003.5 y 2003.75. La Figura 3.6d resume gráficamente estos resultados, muestra como se subdivide el intervalo de un año según el número de eventos.

Aplicaremos este método a las variables que aparecen truncadas en la base de datos original, generando secuencias y fechas probables que servirán de entrada para los procedimientos desarrollados en los Apartados 3.2.2 a 3.2.4, donde calculábamos el IR de los pares y tripletes y ponderamos los resultados según el número de secuencias generadas en cada individuo. Testearemos la validez de este enfoque comparando los resultados obtenidos utilizando fechas exactas con los obtenidos al truncar manualmente esas fechas.

Para realizar los cálculos en el ordenador hemos reproducido el procedimiento descrito en la Figura 3.6, generando en paralelo todas las posibles secuencias para todos los individuos, calculando a partir de ellas todos los pares o tripletes y seleccionando sus primeras ocurrencias.

3.2.5.1 Ecuaciones para el cálculo simplificado del intervalo entre dos eventos.

También hemos implementando este método de imputación en un algoritmo alternativo para el análisis de secuencias de eventos consecutivos y no consecutivos, pero que sólo es aplicable al análisis de pares y tripletes y sólo si la ponderación de las secuencias generadas es equiprobable. En este caso realizamos los cálculos iterativamente, abortándolos al detectar la primera ocurrencia de cada par o triplete para cada individuo, y simplificando el proceso gracias a la utilización de resultados precalculados teóricamente.

Cuando los eventos del par analizado, A→B, ocurren en diferentes años calculamos el promedio de los intervalos de tiempo entre la primera ocurrencia de A y la primera ocurrencia de B como $\Delta_{tAB} = t_A - t_B$, donde t_A es el tiempo promedio de la primera ocurrencia del evento A, dada por $t_A = A\tilde{n}oA + \frac{1}{N_A+1}$, N_A es el número de ocurrencias de ese mismo tipo de evento durante ese año, y $AñoA$ es el año en el que sucede ese A. El resultado es independiente de la posible presencia de otros tipos de evento durante ese año. En el caso de que no pueda haber repetición de eventos durante un mismo año esta ecuación se reduce a $t_A = A\tilde{n}oA + \frac{1}{2}$. Análogamente el tiempo promedio de la ocurrencia del evento es B es $t_B = A\tilde{n}oB + \frac{1}{N_B+1}$

Cuando los eventos del par analizado, A→B, ocurren durante un mismo año el promedio de los intervalos de tiempo transcurridos entre la primera ocurrencia de A y la primera ocurrencia de un B posterior a A es dada por la Ecuación 3.2, donde los paréntesis interiores indican el número de combinaciones. Esta ecuación se reduce a $\frac{1}{3}$ cuando no es posible la repetición intraanual de eventos del mismo tipo. Además, la proporción de casos en los que B es posterior a A es $\frac{1}{2}$. Para la otra mitad de los casos calculamos la distancia promedio entre la primera ocurrencia de A y la siguiente ocurrencia de B otro año o hasta la fecha de fin de estudio, en este caso 2011.

$$t_{A\to B} = \frac{\left(\binom{N_a+N_b+1}{N_b+1} - N_a - 1\right)}{\left(\binom{N_a+N_b}{N_b} - 1\right)(1 + N_a + N_b)} \tag{3.2}$$

Estas ecuaciones son utilizadas para calcular la IR de los pares en el algoritmo que mostramos en el Código F.8 del Apéndice F, que tiene en cuenta simultáneamente todas las posibilidades. Para los tripletes volvemos a utilizar las mismas ecuaciones pero computando el tiempo transcurrido entre el segundo y tercer evento mediante un algoritmo recursivo más complejo que clasifica las subsecuencias en veinte subtipos y realiza cálculos parciales con ellos, Código F.9 del Apéndice F.

Capítulo 4

RESULTADOS

4.1 IDENTIFICACIÓN DE FACTORES ASOCIADOS CON LA PRESENCIA DE *MISSING DATA*

En esta sección mostramos los resultados de los análisis realizados para la identificación de los factores asociados con la presencia de *missing data* en otras variables utilizando la base de datos del Area-7 de Madrid. La Tabla 4.1 muestra un resumen de los modelos analizados en cada apartado de esta sección de la tesis.

Tabla 4.1: Índice de los modelos logísticos analizados en esta sección para la regresión de la *odds* de *missing* de las variables PESO o CIGARRILLOS en función de las demás covariables y de efectos aleatorios.

	Modelo $Miss \sim YEAR + EDAD + PAMED + PAENF + ...$	Junto	Particiones Aleatorias	Particiones Centros
PESO	+ SEXO + EXTRAN			Tabla 4.20
	+ SEXO + EXTRAN + EAP	Tabla 4.4	Tabla 4.12	
	+ SEXO + EXTRAN + (1 | EAP)	Tabla 4.5	Tabla 4.13	
	+ SEXO + EXTRAN + (1 | ID)	Tabla 4.6	Tabla 4.14	Tabla 4.21
	+ SEXO + EXTRAN + (1 | EAP/ID)	Tabla 4.7	Tabla 4.15	
CIGARRILOS	.			Tabla 4.20
	+ EAP	Tabla 4.8	Tabla 4.16	
	+ (1 | EAP)	Tabla 4.9	Tabla 4.17	
	+ (1 | ID)	Tabla 4.10	Tabla 4.18	Tabla 4.23
	+ (1 | EAP/ID)	Tabla 4.11	Tabla 4.19	

4.1.1 Diagnóstico de la memoria y el tiempo utilizados para el computo de los modelos

La Tabla 4.2 muestra, en escala doble logarítmica, la cantidad de memoria máxima necesaria para el ajuste de los datos para los principales modelos de regresión, librerías de R y modos de particionamiento en función del tamaño de la base de datos analizada (miles de filas). La Tabla 4.3 muestra el tiempo necesario (segundos) para calcular estos modelos.

Los modelos estudiados estudiados son de la forma $mPESO \sim SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + ...$, donde "..." representa las diferentes variantes: efectos aleatorios para los individuos (ID), efectos aleatorios para los centros médicos (EAP), efectos aleatorios para los individuos anidados a sus centros médicos (EAP/ID) y logístico simple incluyendo los centros como efectos fijos (EAP^*). Las librerías utilizadas son: glmmTMB, glm y lme4. Los métodos de particionado utilizados son: aleatorio, con 19 particiones aleatorias analizadas independientemente una tras otra y posterior metaanálisis. Junto, conjunto completo de los datos simultaneamente.

Algunas combinaciones de modelos y librerías no aparecen en este análisis porque el uso de memoria excedía las capacidades de nuestro ordenador o directamente provocaba el cuelgue del programa R.

Las mediciones para los modelos particionados se han realizado utilizando un sólo núcleo del ordenador ya que nuestra prioridad es minimizar la cantidad de memoria utilizada. El número de particiones aleatorias que hemos utilizado es 19, el mismo que en las demás secciones de la tesis.

Tabla 4.2: Memoria necesaria (MB) para el ajuste de los datos para los diferentes modelos[1] de regresión, librerías[2] y modos de particionamiento[3] en función del tamaño de la base de datos (Nº filas x1000). [1]Modelo de efectos aleatorios para los individuos (ID), efectos aleatorios para los centros médicos (EAP), efectos aleatorios para los individuos anidados a sus centros médicos (EAP/ID) y logístico simple incluyendo los centros como efectos fijos (EAP*). [2]Librería glmmTMB, librería base glm, librería lme4, particionado aleatorio con glmmTMB y metaanálisis (partition). [3]Ajuste de un modelo con el conjunto completo de los datos simultaneamente (Junto), ajuste secuencial de modelos independientes en 19 particiones (Parti).

		glmmTMB			lme4		glm
Partición	Tamaño	ID	EAP	EAP/ID	EAP	EAP/ID	EAP*
Junto	10000	2165	33	1146	128	348	46
Junto	20000	4252	36	2224	268	808	88
Junto	30000	6726	51	3207	339	1622	132
Junto	50000	11435	84	5988	620	8849	220
Junto	100000	24575	163	11036	1194		439
Junto	200000	50770	340	22185			881
Junto	400000	104220	641	48864			1755
Junto	985696	281240	1560	133030			4329
Partic	10000	97	3	88			
Partic	20000	152	4	132			
Partic	30000	214	5	193			
Partic	50000	377	7	299			
Partic	100000	814	10	509			
Partic	200000	1790	18	1039			
Partic	400000	3734	34	2019			
Partic	985696	10636	79	5054			

Tabla 4.3: Tiempo necesario (s) para el ajuste de los datos para los diferentes modelos[1] de regresión, librerías[2] y modos de particionamiento[3] en función del tamaño de la base de datos (Nº filas x1000). [1]Modelo de efectos aleatorios para los individuos (ID), efectos aleatorios para los centros médicos (EAP), efectos aleatorios para los individuos anidados a sus centros médicos (EAP/ID) y logístico simple incluyendo los centros como efectos fijos (EAP*). [2]Librería glmmTMB, librería base glm, librería lme4, particionado aleatorio con glmmTMB y metaanálisis (partition). [3]Ajuste de un modelo con el conjunto completo de los datos simultaneamente (Junto), ajuste secuencial de modelos independientes en 19 particiones (Parti).

		glmmTMB			lme4		glm
Partición	Tamaño	ID	EAP	EAP/ID	EAP	EAP/ID	EAP*
Junto	10000	28	11	20	5	19	1
Junto	20000	56	22	39	11	56	1
Junto	30000	92	32	58	15	136	1
Junto	50000	161	60	110	26	1118	1
Junto	100000	480	138	264	58		1
Junto	200000	4473	269	2158			2
Junto	400000	12186	615	7357			3
Junto	985696	44085	1808	24670			8
Partic	10000	30	16	30			
Partic	20000	44	22	41			
Partic	30000	68	33	73			
Partic	50000	124	57	113			
Partic	100000	312	121	237			
Partic	200000	744	263	506			
Partic	400000	1550	607	1101			
Partic	985696	4760	1550	3661			

4.1.2 Análisis de todos los datos conjuntamente

A continuación mostramos el resultado de ajustar el conjunto completo de los datos mediante diferentes modelos de regresión.

4.1.2.1 *Missing data* en la variable Peso

Tabla 4.4: Resultados del modelo logístico simple para la *odds* de *missing* en el PESO.

```
Call: glm(mPESO ~ SEXO + zYEO6 + zED50 + zPAMED + zPAENF + EXTRANJERO +
          EAP, family = "binomial")

Coefficients:
               Estimate  Std.Error  z-value Pr(>|z|)
  (Intercept) -0.040732   0.014270   -2.854  0.00431
  SEXO1       -0.066645   0.004529  -14.716  < 2e-16
  zYE06        0.211014   0.003108   67.892  < 2e-16
  zED50       -0.271519   0.001299 -208.958  < 2e-16
  zPAMED      -0.293332   0.006515  -45.026  < 2e-16
  zPAENF       0.181606   0.004767   38.093  < 2e-16
  EXTRANJERO1  0.300241   0.008406   35.716  < 2e-16
  EAP16070210  0.592201   0.021140   28.013  < 2e-16
  EAP16070310  0.159159   0.017794    8.945  < 2e-16
  EAP16070410  0.139682   0.018146    7.698 1.39e-14
  EAP16070610 -0.296017   0.018433  -16.059  < 2e-16
  EAP16071210 -0.638286   0.021428  -29.787  < 2e-16
  EAP16071310  0.948265   0.016764   56.566  < 2e-16
  EAP16071410 -0.865501   0.017304  -50.018  < 2e-16
  EAP16071510  1.031041   0.017355   59.408  < 2e-16
  EAP16071610  0.918529   0.016884   54.403  < 2e-16
  EAP16072110  0.496830   0.018805   26.420  < 2e-16
  EAP16072210  0.818194   0.018997   43.070  < 2e-16
  EAP16072410  0.668268   0.016251   41.120  < 2e-16
  EAP16072610  1.085774   0.016805   64.610  < 2e-16
  EAP16072810  0.295707   0.017506   16.892  < 2e-16
  EAP16073110 -0.533406   0.019035  -28.022  < 2e-16
  EAP16073210  0.418972   0.017593   23.814  < 2e-16
  EAP16073410  0.362170   0.020984   17.260  < 2e-16
  EAP16073610 -0.360641   0.017326  -20.815  < 2e-16

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1299004  on 985695  degrees of freedom
```

```
Residual deviance: 1183232  on 985671  degrees of freedom
AIC: 1183282   R2.Tjur 0.114
Number of Fisher Scoring iterations: 4
```

Tabla 4.5: Resultados del modelo de efectos aleatorios con medidas repetidas por centro para el PESO.

```
Family:  binomial  ( logit )
Formula: mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + (1|EAP)

     AIC      BIC   logLik  deviance  df.resid  marg.R2  cond.R2    ICC
 1183422  1183516  -591703   1183406    985688    0.102    0.185   0.09

Random effects:
  Groups Name        Variance Std.Dev.
  EAP  (Intercept)    0.3338   0.5777
  Number of obs: 985696, groups:  EAP 19

Fixed Effects:
               Estimate  Std.Error z-value Pr(>|z|)
  (Intercept)  0.234705   0.132700    1.77   0.0769
  SEXO1       -0.066648   0.004529  -14.72   <2e-16
  zYE06        0.211224   0.003106   68.00   <2e-16
  zED50       -0.271511   0.001299 -208.95   <2e-16
  zPAMED      -0.292663   0.006512  -44.94   <2e-16
  zPAENF       0.181639   0.004766   38.11   <2e-16
  EXTRANJERO1  0.300240   0.008406   35.72   <2e-16
```

Tabla 4.6: Resultados del modelo de efectos aleatorios con medidas repetidas por individuo para el PESO.

```
Family:  binomial  ( logit )
Formula: mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + EAP + (1|ID)

    AIC    BIC    logLik  deviance  df.resid  marg.R2 cond.R2   ICC
 887171  887477  -443559   887119    985670    0.211   0.820   0.77

Random effects:
  Groups Name        Variance Std.Dev.
  ID   (Intercept)     11.18    3.343
```

```
  Number of obs: 985696, groups:  ID, 201331

Fixed Effects:
               Estimate Std. Error z value Pr(>|z|)
  (Intercept)  0.141892   0.057074    2.49   0.0129
  SEXO1       -0.215475   0.018385  -11.72  < 2e-16
  zYE06        0.489766   0.004842  101.15  < 2e-16
  zED50       -0.731674   0.005716 -128.00  < 2e-16
  zPAMED      -0.609396   0.009861  -61.80  < 2e-16
  zPAENF       0.397242   0.007151   55.55  < 2e-16
  EXTRANJERO1  0.727136   0.032228   22.56  < 2e-16
  EAP16070210  1.537603   0.077463   19.85  < 2e-16
  EAP16070310  0.656999   0.071298    9.21  < 2e-16
  EAP16070410  0.546262   0.066324    8.24  < 2e-16
  EAP16070610 -0.544011   0.068048   -7.99  1.3e-15
  EAP16071210 -1.432313   0.077630  -18.45  < 2e-16
  EAP16071310  2.475922   0.068339   36.23  < 2e-16
  EAP16071410 -2.091644   0.069867  -29.94  < 2e-16
  EAP16071510  2.599268   0.068871   37.74  < 2e-16
  EAP16071610  2.333132   0.069688   33.48  < 2e-16
  EAP16072110  1.245185   0.068379   18.21  < 2e-16
  EAP16072210  2.043081   0.069469   29.41  < 2e-16
  EAP16072410  1.665251   0.063137   26.38  < 2e-16
  EAP16072610  2.714780   0.065223   41.62  < 2e-16
  EAP16072810  0.832684   0.070683   11.78  < 2e-16
  EAP16073110 -1.133322   0.071496  -15.85  < 2e-16
  EAP16073210  0.991614   0.065567   15.12  < 2e-16
  EAP16073410  0.808442   0.075975   10.64  < 2e-16
  EAP16073610 -0.962081   0.067022  -14.35  < 2e-16
```

Tabla 4.7: Resultados del modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro para el PESO.

```
Family:  binomial  ( logit )
Formula: mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + (1|EAP/ID)

    AIC      BIC   logLik  deviance  df.resid  marg.R2 cond.R2    ICC
 891958   892064  -445970    891940    985687    0.143   0.810   0.78

Random effects:
  Groups Name        Variance Std.Dev.
```

```
  ID:EAP (Intercept)   9.979    3.159
  EAP    (Intercept)   1.534    1.239
  Number of obs: 985696, groups:  ID 201331; EAP 19

Fixed Effects:
               Estimate  Std.Error z-value  Pr(>|z|)
  (Intercept)  0.795512   0.284654    2.79   0.0052
  SEXO1       -0.212108   0.017763  -11.94   <2e-16
  zYE06        0.479044   0.004778  100.25   <2e-16
  zED50       -0.703257   0.005490 -128.09   <2e-16
  zPAMED      -0.599718   0.009753  -61.49   <2e-16
  zPAENF       0.389932   0.007080   55.07   <2e-16
  EXTRANJERO1  0.780990   0.032282   24.19   <2e-16
```

4.1.2.2 *Missing data* en la variable Cigarrillos

Tabla 4.8: Resultados del modelo logístico simple para CIGARRILLOS.

```
Call: glm(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + EAP,
    family = "binomial", data = todos)

Coefficients:
               Estimate  Std.Error z-value Pr(>|z|)
  (Intercept) -2.23393    0.08077 -27.657  < 2e-16
  zYE06       -0.18760    0.01515 -12.381  < 2e-16
  zED50        0.06743    0.00757   8.908  < 2e-16
  zPAMED       0.08896    0.03285   2.708  0.00677
  zPAENF      -0.10412    0.02379  -4.376 1.21e-05
  EAP16070210  0.58260    0.11113   5.243 1.58e-07
  EAP16070310  0.65879    0.09820   6.709 1.96e-11
  EAP16070410  0.05715    0.10035   0.569  0.56904
  EAP16070610  0.15640    0.10266   1.524  0.12762
  EAP16071210  0.14129    0.11195   1.262  0.20691
  EAP16071310  0.72061    0.09026   7.984 1.42e-15
  EAP16071410 -0.28108    0.09273  -3.031  0.00244
  EAP16071510  0.84409    0.09304   9.072  < 2e-16
  EAP16071610  0.43839    0.09365   4.681 2.85e-06
  EAP16072110  0.77775    0.10138   7.672 1.70e-14
  EAP16072210  0.69410    0.10567   6.569 5.07e-11
  EAP16072410  0.60449    0.09087   6.652 2.89e-11
```

```
  EAP16072610  0.69778    0.09232   7.558 4.08e-14
  EAP16072810  0.31410    0.09616   3.266  0.00109
  EAP16073110 -0.09926    0.10592  -0.937  0.34870
  EAP16073210 -0.04086    0.09875  -0.414  0.67908
  EAP16073410  0.22670    0.11084   2.045  0.04083
  EAP16073610  0.01477    0.09620   0.154  0.87797

    Null deviance: 60672  on 100087  degrees of freedom
Residual deviance: 59073  on 100065  degrees of freedom
  (893866 observations deleted due to missingness)
AIC: 59119   R2.Tjur: 0.017
Number of Fisher Scoring iterations: 5
```

Tabla 4.9: Resultados del modelo de efectos aleatorios con medidas repetidas por centro para CIGARRILLOS.

```
Family: binomial  ( logit )
Formula: mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + (1|EAP)

     AIC      BIC   logLik deviance df.resid  marg.R2  cond.R2
 59176.5  59233.5 -29582.2  59164.5   100082    0.020    0.052

Random effects:
  Groups Name         Variance Std.Dev.
    EAP  (Intercept)   0.1108   0.3329
  Number of obs: 100088, groups:  EAP 19

Fixed effects:
               Estimate  Std.Error z-value  Pr(>|z|)
  (Intercept) -1.900682   0.081461 -23.332  < 2e-16
  zYE06       -0.182532   0.014502 -12.586  < 2e-16
  zED50        0.067822   0.007568   8.961  < 2e-16
  zPAMED       0.102874   0.031058   3.312 0.000925
  zPAENF      -0.102748   0.023382  -4.394 1.11e-05
```

Tabla 4.10: Modelo de efectos aleatorios con medidas repetidas por individuo para CIGARRILLOS.

```
Family: binomial  ( logit )
Formula: mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + EAP + (1|ID)

     AIC      BIC   logLik deviance df.resid  marg.R2  cond.R2
 14919.3  15147.7  -7435.7  14871.3   100064    0.006    0.997

Random effects:
  Groups Name        Variance Std.Dev.
   ID   (Intercept)     1127     33.58
  Number of obs: 100088, groups:  ID, 27758

Fixed effects:
               Estimate  Std.Error z-value Pr(>|z|)
  (Intercept) -11.76657    0.74511 -15.792  < 2e-16
  zYE06        -2.10819    0.10204 -20.661  < 2e-16
```

```
  zED50         0.09161    0.06892   1.329  0.18377
  zPAMED       -0.01124    0.17016  -0.066  0.94734
  zPAENF       -0.62160    0.12527  -4.962 6.98e-07
  EAP16070210   2.88155    1.01385   2.842  0.00448
  EAP16070310   0.46184    0.89882   0.514  0.60737
  EAP16070410  -0.14196    0.87415  -0.162  0.87099
  EAP16070610   0.17365    0.90276   0.192  0.84746
  EAP16071210  -0.14088    0.99439  -0.142  0.88733
  EAP16071310   0.48821    0.86116   0.567  0.57076
  EAP16071410  -0.19333    0.85442  -0.226  0.82099
  EAP16071510   1.38386    0.87318   1.585  0.11300
  EAP16071610   0.37621    0.89789   0.419  0.67522
  EAP16072110   1.34617    0.85994   1.565  0.11748
  EAP16072210   2.00632    0.94916   2.114  0.03453
  EAP16072410   1.54133    0.83089   1.855  0.06359
  EAP16072610   1.95276    0.84307   2.316  0.02055
  EAP16072810   1.45847    0.87749   1.662  0.09650
  EAP16073110   0.62204    0.92602   0.672  0.50176
  EAP16073210   0.55410    0.87244   0.635  0.52535
  EAP16073410   1.16169    0.96390   1.205  0.22813
  EAP16073610   0.91626    0.87098   1.052  0.29280
```

Tabla 4.11: Modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro para CIGARRILLOS.

```
Family: binomial  ( logit )
Formula: mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + (1|EAP/ID)

     AIC      BIC   logLik  deviance df.resid  marg.R2  cond.R2
 14914.5  14981.1  -7450.2   14900.5   100081    0.670       NA

Random effects:
  Groups Name            Variance    Std.Dev.
    ID:EAP (Intercept)  1.144e+03   3.383e+01
    EAP    (Intercept)  1.010e-08   1.005e-04
  Number of obs: 100088, groups: ID 27758; EAP 19

Fixed effects:
               Estimate  Std.Error z-value Pr(>|z|)
  (Intercept) -11.29288    0.14621  -77.24  < 2e-16
  zYE06        -1.94123    0.07864  -24.69  < 2e-16
```

```
zED50         0.12883    0.06778    1.90 0.057351
zPAMED        0.29915    0.09011    3.32 0.000901
zPAENF       -0.43933    0.10380   -4.23 2.31e-05
```

4.1.3 Análisis de los datos particionados aleatoriamente

4.1.3.1 *Missing data* en la variable Peso

Modelo logístico simple. A continuación mostramos los resultados de los análisis realizados con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios, para la los *missings* en la variable PESO.

```
glm(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + EAP,
data=partition, family="binomial")
```

La Tabla D.9 del Apéndice D resume los coeficientes y errores estándar obtenidos en dichos análisis intra-partición para cada variable. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones utilizando el paquete *metafor*, con la opción "FE", Tabla D.10 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.12 resumimos los coeficientes obtenidos en los meta-análisis de los coeficientes de las diferentes variables obtenidos en los modelos de regresión logística simple intra-partición para la *odds* de *missing data* en PESO.

Tabla 4.12: Resumen de los meta-análisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística simple para la *odds* de *missing data* en la variable PESO.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	-0.0407	0.01428	-2.85	0.00434
SEXO	-0.0667	0.00453	-14.72	5.12×10^{-49}
zYE06	0.2109	0.00311	67.82	0
zED50	-0.2716	0.00130	-209.00	0
zPAMED	-0.2937	0.00652	-45.06	0
zPAENF	0.1814	0.00477	38.03	0
EXTRANJERO	0.2996	0.00841	35.63	0

Modelo logístico con efectos aleatorios para los centros médicos. A continuación mostramos los resultados de los análisis de los *missings* en la variable PESO con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios sólo para los centros médicos, no para los individuos.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
        (1|EAP), data=partition, family="binomial", REML=F)
```

La Tabla D.33 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.34 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.13 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.13: Resumen de los Metaanálisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística con efectos aleatorios por centro médico para la *odds* de *missing data* en la variable PESO.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	0.2280	0.0308	7.40	1.34×10^{-13}
SEXO	-0.0667	0.0045	-14.73	4.19×10^{-49}
zYE06	0.2149	0.0031	69.83	0
zED50	-0.2714	0.0013	-208.9	0
zPAMED	-0.2809	0.0065	-43.45	0
zPAENF	0.1820	0.0047	38.38	0
EXTRANJERO	0.3005	0.0084	35.75	0

Modelo logístico con efectos aleatorios para los individuos. A continuación mostramos los resultados de los análisis realizados para la los *missings* en la variable PESO con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos, sin anidar. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
          EAP + (1|ID), data=partition, family="binomial", REML=F)
```

La Tabla D.17 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.18 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots. No incluimos en esta tabla los coeficientes particulares de los 19 centros EAP.

En la Tabla 4.14 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.14: Resumen de los meta-análisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística con efectos aleatorios de los individuos para la *odds* de *missing data* en la variable PESO.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	0.1412	0.0571	2.475	0.01332
SEXO	-0.2158	0.0184	-11.74	8.30×10^{-32}
zYE06	0.4896	0.0048	101.1	0
zED50	-0.7308	0.0057	-127.9	0
zPAMED	-0.6094	0.0099	-61.78	0
zPAENF	0.3972	0.0072	55.52	0
EXTRANJERO	0.7266	0.0322	22.55	0

Modelo logístico con efectos aleatorios para los individuos anidados a sus centros médicos. A continuación mostramos los resultados de los análisis realizados para la los *missing data* en la variable PESO con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos anidados a cada centro médico. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
   (1|EAP/ID), data=partition, family="binomial", REML=F)
```

La Tabla D.25 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.26 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.15 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.15: Resumen de los meta-análisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística con efectos aleatorios de los individuos anidados a cada centro para la *odds* de *missing data* en la variable PESO.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	0.7933	0.0669	11.87	1.74×10^{-32}
SEXO	-0.2129	0.0178	-11.96	5.54×10^{-33}
zYE06	0.4830	0.0047	101.5	0
zED50	-0.7032	0.0055	-127.9	0
zPAMED	-0.5871	0.0097	-60.40	0
zPAENF	0.3904	0.0071	55.26	0
EXTRANJERO	0.7842	0.0323	24.27	0

4.1.3.2 *Missing data* en la variable Cigarrillos

Modelo logístico simple. A continuación mostramos los resultados de los análisis de los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios.

```
glm(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + EAP, data=partition,
       family="binomial")
```

La Tabla D.41 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.42 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.16 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.16: Resumen de los meta-análisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística simple para la *odds* de *missing data* en la variable CIGARRILLOS.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	-2.2110	0.0819	-26.99	0
zYE06	-0.1915	0.0153	-12.52	6.05×10^{-36}
zED50	0.0675	0.0076	8.87	7.20×10^{-19}
zPAMED	0.0867	0.0331	2.62	0.00883
zPAENF	-0.1043	0.0241	-4.32	1.54×10^{-5}

Modelo logístico con efectos aleatorios para los centros médicos. A continuación mostramos los resultados de los análisis de los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios sólo para los centros médicos, no para los individuos.

```
glmmTMB(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + (1|EAP),
         data=partition, family="binomial", REML=F)
```

La Tabla D.59 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.60 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.17 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.17: Resumen de los Metaanálisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística con efectos aleatorios por centro médico para la *odds* de *missing data* en la variable CIGARRILLOS.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	-1.9575	0.0296	-66.23	0
zYE06	-0.1641	0.0113	-14.53	8.23×10^{-48}
zED50	0.0727	0.0076	9.58	9.36×10^{-22}
zPAMED	0.1664	0.0209	7.96	1.79×10^{-15}
zPAENF	-0.1113	0.0205	-5.42	5.82×10^{-8}

Modelo logístico con efectos aleatorios para los individuos. A continuación mostramos los resultados de los análisis de los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos, sin anidar. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mCIGARRILLOS~zYE06+zED50+zPAMED+zPAENF+EAP+(1|ID),
          data=partition, family="binomial", REML=F)
```

La Tabla D.47 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.48 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.18 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.18: Resumen de los meta-análisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística con efectos aleatorios de los individuos para la *odds* de *missing data* en la variable CIGARRILLOS.

	Estimate	Std.Error	z value	Pr(>|z|)
zYE06	-1.813	0.08105	-22.36	0
zED50	0.1288	0.07029	1.83	0.0668
zPAMED	0.2888	0.09501	3.04	0.0024
zPAENF	-0.4479	0.1091	-4.11	4.0×10^{-5}

Modelo logístico con efectos aleatorios para los individuos anidados a los centros. A continuación mostramos los resultados de los análisis de los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos anidados a cada centro médico. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mCIGARRILLOS~zYE06+zED50+zPAMED+zPAENF+(1|EAP/ID),
          data=partition, family="binomial", REML=F)
```

La Tabla D.53 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.54 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

En la Tabla 4.19 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.19: Resumen de los meta-análisis de los resultados obtenidos previamente en cada partición, en las que se ajustaron modelos de regresión logística con efectos aleatorios de los individuos anidados a cada centro para la *odds* de *missing data* en la variable CIGARRILLOS.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	-11.52	0.1584	-72.70	0
zYE06	-1.906	0.0806	-23.64	0
zED50	0.1246	0.0717	1.74	0.08214
zPAMED	0.2695	0.0958	2.81	0.00491
zPAENF	-0.4111	0.1109	-3.71	0.00021

4.1.4 Análisis de los datos particionados por centro

4.1.4.1 *Missing data* en la variable Peso

Modelo logístico simple. A continuación mostramos los resultados de los análisis de los *missings* en la variable PESO realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios.

```
glm(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO,
    data=centro, family="binomial")
```

La Tabla D.65 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todos los centros con la opción "ML", Tabla D.66 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, Figura D.49 y siguientes.

También realizamos un segundo set de regresiones, Tabla D.73 y siguientes, para estudiar si los coeficientes obtenidos en cada modelo intra-centro dependen de los parámetros que varían entre diferentes centros, y que no pudieron ser analizados simultáneamente en el paso previo porque la información era parcial. Sólo hemos

considerado la variable inter-centro POSMAS.

En la Tabla 4.20 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.20: Resumen de los meta-análisis de los resultados obtenidos previamente en cada centro médico, en las que se ajustaron modelos de regresión logística simple para la *odds* de *missing data* en la variable PESO.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	0.8401	0.4507	1.86	0.0623
SEXO	-0.0735	0.0168	-4.39	1.15×10^{-5}
zYE06	0.1239	0.1144	1.08	0.2788
zED50	-0.2738	0.0242	-11.33	8.56×10^{-30}
zPAMED	-0.6585	0.2198	-2.99	0.0027
zPAENF	0.2872	0.1276	2.25	0.0244
EXTRANJERO	0.2913	0.0382	7.63	2.32×10^{-14}

Modelo logístico con efectos aleatorios para los individuos. A continuación mostramos los resultados de los análisis de los *missings* en la variable PESO realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + (1|ID),
        data=centro, family="binomial", REML=F)
```

La Tabla D.80 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. Para cada uno de los coeficientes hemos calculado un meta-análisis combinando los resultados de todos los centros con la opción “ML”. La Tabla D.81 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, , Figura D.56 y siguientes.

La variable inter-centro POSMAS, porcentaje de pacientes del centro médico

con al menos 65 años, no puede ser introducida directamente en los modelos intra-centro por ser un parámetro anual fijo en cada uno, ha sido analizada en un proceso posterior, modelando cada uno de los coeficientes obtenidos en cada centro en función de POSMAS mediante regresiones lineales. De este modo, podemos testear si las variables intra-centro pudieran estar confundidas con variables externas. El restulado de estos modelos se expone en la Tabla D.88 y siguientes.

En la Tabla 4.21 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.21: Resumen de los meta-análisis de los resultados obtenidos previamente en cada centro médico, en las que se ajustaron modelos de regresión logística con efectos aleatorios de los individuos para la *odds* de *missing data* en la variable PESO.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	2.2700	1.0053	2.26	0.0239
SEXO	-0.2336	0.0447	-5.23	1.72×10^{-7}
zYE06	0.3696	0.2645	1.40	0.162
zED50	-0.7807	0.0781	-10.00	1.60×10^{-23}
zPAMED	-1.4735	0.4846	-3.04	0.0023
zPAENF	0.7565	0.3141	2.41	0.0160
EXTRANJERO	0.7442	0.1114	6.68	2.34×10^{-11}

4.1.4.2 *Missing data* en la variable Cigarrillos

Modelo logístico simple. A continuación mostramos los resultados de los análisis de los *missings* en la variable CIGARRILLOS realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios.

```
glm(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF,
       data=centro,family="binomial")
```

La Tabla D.95 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todos los centros con la opción "ML", Tabla D.96 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, Figura D.63 y siguientes.

En la Tabla 4.22 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.22: Resumen de los meta-análisis de los resultados obtenidos previamente en cada centro médico, en las que se ajustaron modelos de regresión logística simple para la *odds* de *missing data* en la variable CIGARRILLOS.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	-1.9150	0.1469	-13.03	7.96×10^{-39}
zYE06	-0.1870	0.0331	-5.65	1.57×10^{-8}
zED50	0.0610	0.0254	2.40	0.0163
zPAMED	0.0232	0.0491	0.47	0.6368
zPAENF	-0.1234	0.0409	-3.02	0.0026

Modelo logístico con efectos aleatorios para los individuos. A continuación mostramos los resultados de los análisis de los *missings* en la variable CIGARRILLOS realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos.

```
glmmTMB(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + (1|ID),
        data=centro, family="binomial", REML=F)
```

La Tabla D.101 del Apéndice D resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis

combinando los resultados de todos los centros con la opción "ML", Tabla D.102 y siguientes del Apéndice D. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, Figura D.68 y siguientes.

En la Tabla 4.23 resumimos los coeficientes obtenidos en el meta-análisis de las diferentes variables.

Tabla 4.23: Resumen de los meta-análisis de los resultados obtenidos previamente en cada centro médico, en las que se ajustaron modelos de regresión logística con efectos aleatorios de los individuos para la *odds* de *missing data* en la variable CIGARRILLOS.

	Estimate	Std.Error	z value	Pr(>|z|)
Inter	-13.0756	0.578	-22.62	0
zYE06	-6.0742	1.466	-4.14	3.42×10^{-5}
zED50	0.0927	0.0803	1.15	0.2485
zPAMED	-0.1465	0.2434	-0.60	0.5472
zPAENF	-0.5696	0.2653	-2.15	0.0318

4.1.5 Comparación de los resultados obtenidos con los diferentes modelos de regresión y métodos de particionado

A continuación, resumimos los principales resultados obtenidos para la ocurrencia de *missings* en función del modelo y método de particionado utilizados. En esta sección los resultados están referidos a las covariables originales de la base de datos, sin estandarizar.

La Tabla 4.24 muestra la *odds ratio* de ocurrencia de *missing* en la variable PESO y su intervalo de confianza del 95% para las diferentes variables explicativas, EXTRANJERO, SEXO (mujer frente a hombre), EDAD (años), PAENF, PAMED y YEAR (años), diferentes modelos de regresión (logística simple, efectos aleatorios por individuos, efectos aleatorios por centro médico y efectos aleatorios por individuo anidado a su centro) y diferentes métodos de particionado (conjunto completo, particionado aleatorio y particionado por centro médico). Nótese que los coeficientes

obtenidos en los modelos han sido exponenciados.

En la Tabla 4.24 también incluimos el Intercepto calculado con los diferentes modelos y el rango de valores que adquiere dicho intercepto asociado a la variabilidad entre centros, para lo cual, indicamos sus percentiles 2.5 y 97.5. Para el análisis con el conjunto completo de los datos este intervalo se corresponde aproximadamente con $e^{(Inter \pm 1.96 \cdot \sigma_{EAP})}$, donde utilizamos el Intercepto y la desviación típica de EAP obtenidos en el modelo. Para el análisis en particiones aleatorias se corresponde aproximadamente con $e^{\widehat{Inter} \pm 1.96 \cdot \sqrt{\widehat{var(EAP)}}}$, donde $\widehat{Inter}$ y $\widehat{var(EAP)}$ son respectivamente el promedio de los interceptos y el promedio de las varianzas de EAP obtenidos en cada partición. Para los análisis realizados en cada centro el intervalo equivale aproximadamente a $e^{\widehat{Inter} \pm 2.10 \cdot \sigma_{Inter}}$, donde $\widehat{Inter}$ y σ_{Inter} son respectivamente el promedio de los interceptos y el promedio de las varianzas de EAP obtenidos en cada partición..

Tabla 4.24: Comparación de los resultados para la ocurrencia de *missings* en la variable PESO en función del modelo[1] y método de particionado utilizados. *odds ratio* y su $IC_{95\%}$ para las diferentes variables explicativas originales: EXTRANJERO, SEXO (Mujer frente a hombre), EDAD (años), PAENF, PAMED y YEAR (años). [1]Logístico simple, efectos aleatorios por individuos, efectos aleatorios por centro, efectos aleatorios por individuo anidado a su centro. [2]Int.EAP muestra el 95% del rango de variabilidad del intercepto entre centros.

Variable explic.	Modelo[1]	Conjunto completo	Particiones Aleatorias	Por Centros
Int.EAP[2]	EAP	1.265 (0.408,3.924)	1.257 (0.406,3.892)	
Int.EAP[2]	ID			9.730 (7×10^{-4},1×10^{5})
Int.EAP[2]	EAP/ID	2.216 (0.195,25.13)	2.212 (0.197,24.80)	
EXTRA	GLM	1.350 (1.328,1.373)	1.349 (1.327,1.372)	1.338 (1.242,1.442)
EXTRA	ID	2.069 (1.942,2.204)	2.068 (1.942,2.203)	2.105 (1.692,2.618)
EXTRA	EAP	1.350 (1.328,1.373)	1.351 (1.328,1.373)	
EXTRA	EAP/ID	2.184 (2.050,2.326)	2.191 (2.056,2.334)	
SEXO	GLM	0.936 (0.927,0.944)	0.935 (0.927,0.944)	0.929 (0.899,0.960)
SEXO	ID	0.806 (0.778,0.836)	0.806 (0.777,0.835)	0.792 (0.725,0.864)
SEXO	EAP	0.936 (0.927,0.944)	0.935 (0.927,0.944)	
SEXO	EAP/ID	0.809 (0.781,0.838)	0.808 (0.781,0.837)	
EDAD	GLM	0.997 (0.997,0.997)	0.997 (0.997,0.997)	0.997 (0.997,0.998)
EDAD	ID	0.993 (0.993,0.993)	0.993 (0.993,0.993)	0.992 (0.991,0.994)
EDAD	EAP	0.997 (0.997,0.997)	0.997 (0.997,0.997)	
EDAD	EAP/ID	0.993 (0.993,0.993)	0.993 (0.993,0.993)	
PAENF	GLM	1.033 (1.031,1.034)	1.033 (1.031,1.034)	1.052 (1.007,1.099)
PAENF	ID	1.073 (1.070,1.075)	1.073 (1.070,1.075)	1.143 (1.025,1.274)
PAENF	EAP	1.033 (1.031,1.034)	1.033 (1.031,1.034)	
PAENF	EAP/ID	1.071 (1.069,1.074)	1.071 (1.069,1.074)	
PAMED	GLM	0.980 (0.979,0.981)	0.980 (0.979,0.981)	0.955 (0.927,0.984)
PAMED	ID	0.958 (0.957,0.960)	0.958 (0.957,0.960)	0.902 (0.845,0.964)
PAMED	EAP	0.980 (0.979,0.981)	0.981 (0.980,0.981)	
PAMED	EAP/ID	0.959 (0.958,0.960)	0.960 (0.959,0.961)	
YEAR	GLM	1.235 (1.227,1.242)	1.235 (1.227,1.242)	1.132 (0.905,1.416)
YEAR	ID	1.632 (1.617,1.647)	1.632 (1.616,1.647)	1.447 (0.862,2.430)
YEAR	EAP	1.235 (1.228,1.243)	1.240 (1.232,1.247)	
YEAR	EAP/ID	1.615 (1.599,1.630)	1.621 (1.606,1.636)	

La Tabla 4.25 muestra la *odds ratio* de *missings* en la variable CIGARRILLOS y

su intervalo de confianza del 95% para las diferentes variables explicativas, EDAD (años), PAENF, PAMED y YEAR (años), diferentes modelos de regresión (logística simple, efectos aleatorios por individuos, efectos aleatorios por centro médico y efectos aleatorios por individuo anidado a su centro) y diferentes métodos de particionado (conjunto completo, particionado aleatorio y particionado por centro médico). También incluimos el Intercepto calculado con los diferentes modelos y el rango de valores que adquiere dicho intercepto asociado a la variabilidad entre centros, para lo cual, indicamos sus percentiles 2.5 y 97.5.

Tabla 4.25: Comparación de los resultados para la ocurrencia de *missings* en la variable CIGARRILLOS en función del modelo[1] y método de particionado utilizados. *odds ratio* y su $IC_{95\%}$ para las diferentes variables explicativas originales: EDAD (años), PAENF, PAMED y YEAR (años). [1] Logístico simple, efectos aleatorios por individuos, efectos aleatorios por centro, efectos aleatorios por individuo anidado a su centro. [2]Int.EAP indica 95% del rango de variabilidad del intercepto entre centros.

Variable explic.	Modelo[1]	Conjunto completo	Particiones Aleatorias	Por Centros
Int.EAP[2]	EAP	0.149 (0.078,0.287)	0.141 (0.052,0.386)	
Int.EAP[2]	ID			5×10^{-6} (0.000,0.171)
Int.EAP[2]	EAP/ID	1.2461×10^{-5} (..459,..464)	6.20×10^{-6} (5.95,6.45)	
EDAD	GLM	1.007 (1.005,1.008)	1.007 (1.005,1.008)	1.006 (1.001,1.011)
EDAD	ID	1.009 (0.996,1.023)	1.013 (0.999,1.027)	1.009 (0.994,1.025)
EDAD	EAP	1.007 (1.005,1.008)	1.007 (1.006,1.009)	
EDAD	EAP/ID	1.013 (1.000,1.027)	1.013 (0.998,1.027)	
PAENF	GLM	0.957 (0.939,0.976)	0.957 (0.938,0.976)	0.949 (0.918,0.982)
PAENF	ID	0.770 (0.695,0.854)	0.828 (0.757,0.906)	0.787 (0.633,0.979)
PAENF	EAP	0.958 (0.939,0.976)	0.954 (0.938,0.971)	
PAENF	EAP/ID	0.831 (0.763,0.906)	0.841 (0.768,0.922)	
PAMED	GLM	1.024 (1.007,1.041)	1.023 (1.006,1.041)	1.006 (0.981,1.032)
PAMED	ID	0.997 (0.913,1.089)	1.079 (1.027,1.134)	0.962 (0.848,1.091)
PAMED	EAP	1.028 (1.011,1.044)	1.045 (1.034,1.056)	
PAMED	EAP/ID	1.082 (1.033,1.134)	1.074 (1.022,1.128)	
YEAR	GLM	0.829 (0.805,0.854)	0.826 (0.801,0.851)	0.829 (0.777,0.885)
YEAR	ID	0.121 (0.099,0.148)	0.163 (0.139,0.191)	0.002 (0.000,0.041)
YEAR	EAP	0.833 (0.810,0.857)	0.849 (0.830,0.868)	
YEAR	EAP/ID	0.144 (0.123,0.167)	0.149 (0.127,0.174)	

4.2 ANÁLISIS DE LAS SECUENCIAS DE EVENTOS

En esta sección mostramos los resultados de los análisis de las secuencias de eventos utilizando la base de datos del Area-7 de Madrid.

4.2.1 Descripción estadística de las variables.

Podemos clasificar las variables de la base datos del Area-7 de Madrid en dos grupos, aquellas que fueron recopiladas con su fecha completa, día/mes/año, y aquellas para las que sólo se ha registrado sí han tenido lugar o no un determinado año, de las que por lo tanto sólo se conoce el año de ocurrencia. A las primeras las llamaremos variables "completas" y a las segundas "truncadas". La mayoría de las variables de la base de datos son truncadas.

Realizamos primero el análisis estadístico de las variables "completas", pudiendo utilizar para ello las variables: *FECHA_TRASTLIPIDOS*, *FECHA_ARRITMIAS*, *FECHA_TABAQUISMO*, *FECHA_HTA*, *FECHA_INSF_CARD*, *FECHA_CARDIO_ISQ*, *FECHA_OBESIDAD*, *FECHA_DIABETES*, *FECHA_ACV*, *FECHA_ALCOHOLISMO*, *FECHA_VALVULOPATIAS*, *FECHA_ATERO_PER*. En ellas se hallan recogidos un total de $261\,850$ eventos pertenecientes a $128\,805$ personas.

Comenzamos analizando los eventos individualmente, sin tener en cuenta el orden absoluto en el que tuvieron lugar ni su relación con los demás eventos. En la Figura 4.1a podemos ver el número total de ocurrencias de cada tipo de evento. Comprobamos que el evento más frecuente es HTA ($72\,334$), seguido de TRASTLIPIDOS ($67\,837$) y OBESIDAD ($33\,208$), algunos de los cuales pueden pertenecer a la misma persona. La Figura 4.1b muestra cuantas personas sufren un número dado de eventos, vemos que el número de personas decrece con el número de eventos, $53\,880$ personas sufren un solo evento y $37\,699$ personas sufren dos, que pueden ser ocurrencias del mismo tipo.

La Figura 4.2 muestra el valor medio del intervalo de tiempo transcurrido entre

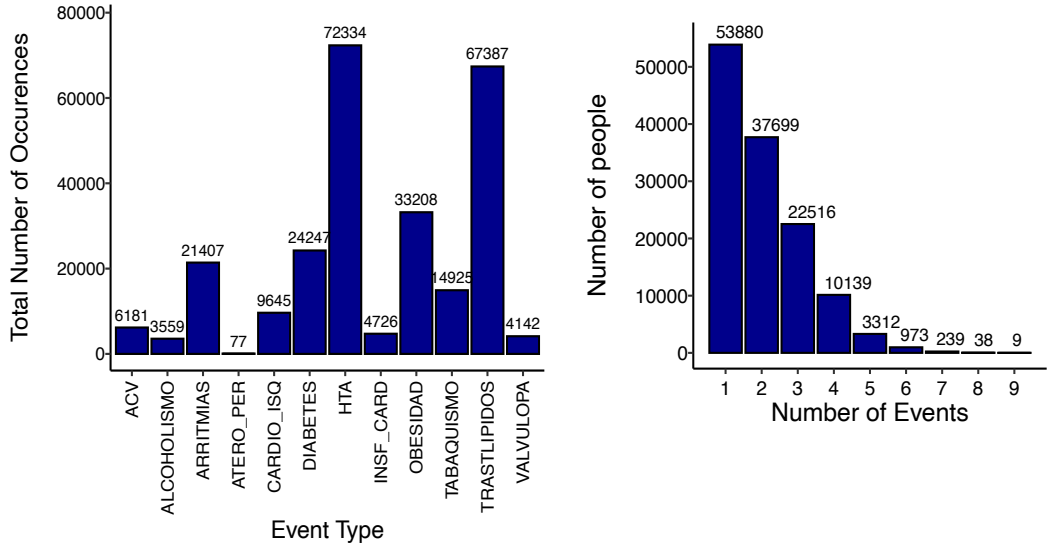


Figura 4.1: a) Total de eventos de cada tipo recogidos en las variables completas. b) Número de personas según la cantidad total de eventos que sufre.

pares de eventos de un mismo individuo: a) para eventos consecutivos, Tabla 4.2a, b) para eventos cualesquiera, tanto consecutivos como no consecutivos, Tabla 4.2b.

Representamos gráficamente la distribución de los intervalos de tiempo para los pares de eventos consecutivos: a) según el tipo de evento inicial de cada par, cualquiera que sea el evento final, Figura 4.3a. b) según el tipo de evento final, cualquiera que sea el evento inicial, Figura 4.3b. Vemos como, para la mayoría de los pares, este intervalo oscila entre varios meses y veinte años, siendo muchos más comunes los de 3 a 4 años y muy poco comunes los extremos. La proporción de pares de eventos en los que el primer evento es *TABAQUISMO* vuelve a aumentar pasados los veinte años, posiblemente porque tenemos registrado el tabaquismo desde edades muy tempranas. También podemos ver como el intervalo entre pares de eventos es menor cuando el primero de ellos es *ATERO_PERIFERICA*.

En las Figuras 4.4a y 4.4b volvemos a mostrar la distribución de los intervalos de tiempo pero considerando pares de eventos tanto consecutivos como no consecutivos.

En este caso observamos que, aunque en general los intervalos de tiempo son ligeramente más largos, los resultados son muy similares, volviendo a observar las mismas peculiaridades para *TABAQUISMO* y *ATERO_PERIFERICA*.

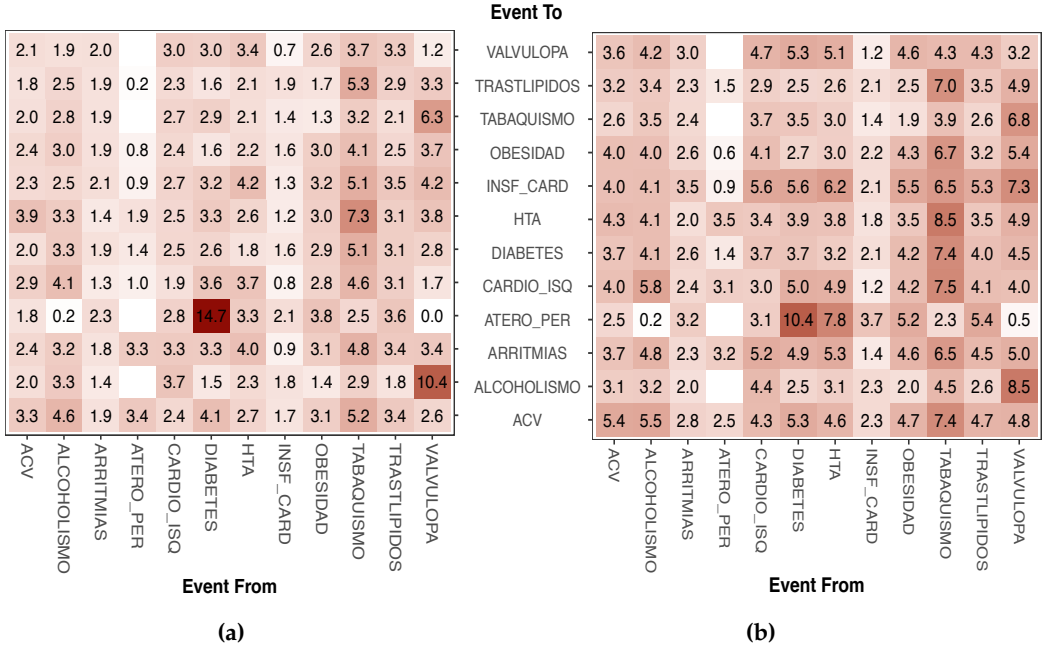


Figura 4.2: a) Tiempo medio transcurrido entre pares de eventos consecutivos (Años). b) Tiempo medio transcurrido entre cualquier par de eventos, tanto consecutivos como no consecutivos (Años).

La Figura 4.5a muestra la tasa de ocurrencia de cada tipo de evento por cada 10 000 personas a diferentes edades, es decir, el riesgo de sufrir cada tipo de evento. Podemos ver por ejemplo, como los dos tipos de evento más frecuentes entre las personas estudiadas son HTA y TRASTLIPIDOS, cuyos máximos se dan entre los 60 y 70 años. La Figura 4.5b muestra la proporción de veces que ocurre cada tipo de evento a diferentes edades sobre el total de ocurrencias de ese tipo de evento, es decir, la distribución de las edades en las que se sufre cada tipo de evento. Vemos por ejemplo, como el máximo riesgo, de sufrir ALCOHOLISMO se dan en torno a los 40 años, y el máximo riesgo de sufrir INSF_CARD a los 80 años.

A continuación mostramos los resultados del análisis descriptivo de las variables "truncadas" de la base de datos del Area-7 de Madrid para las que no se dispone de

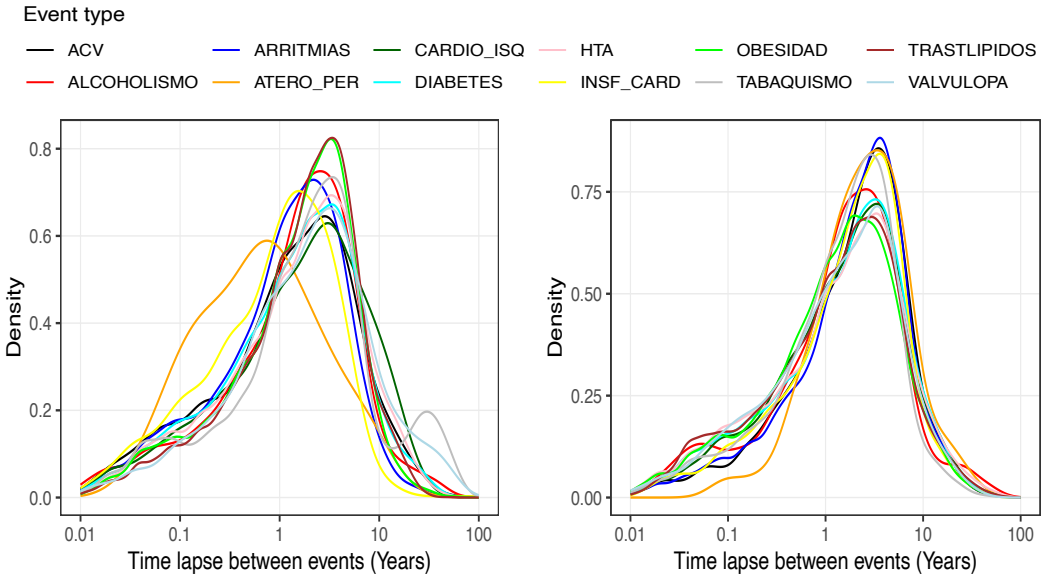


Figura 4.3: Distribución de los intervalos de tiempo entre pares de eventos consecutivos: a) Desde un tipo de evento dado hasta cualquier otro evento. b) Desde cualquier evento hasta un tipo de evento dado.

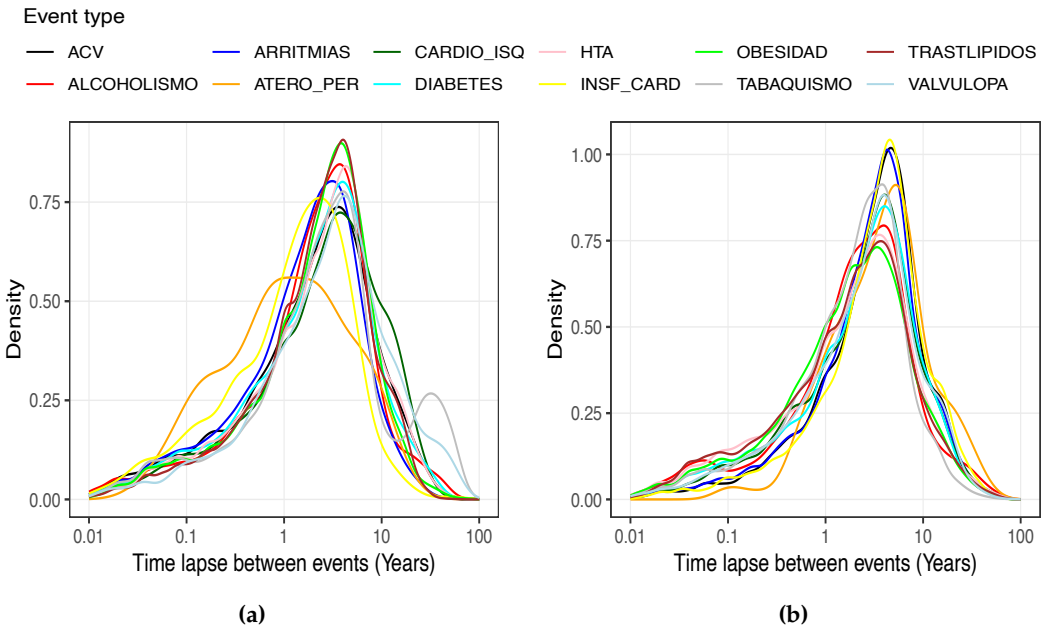
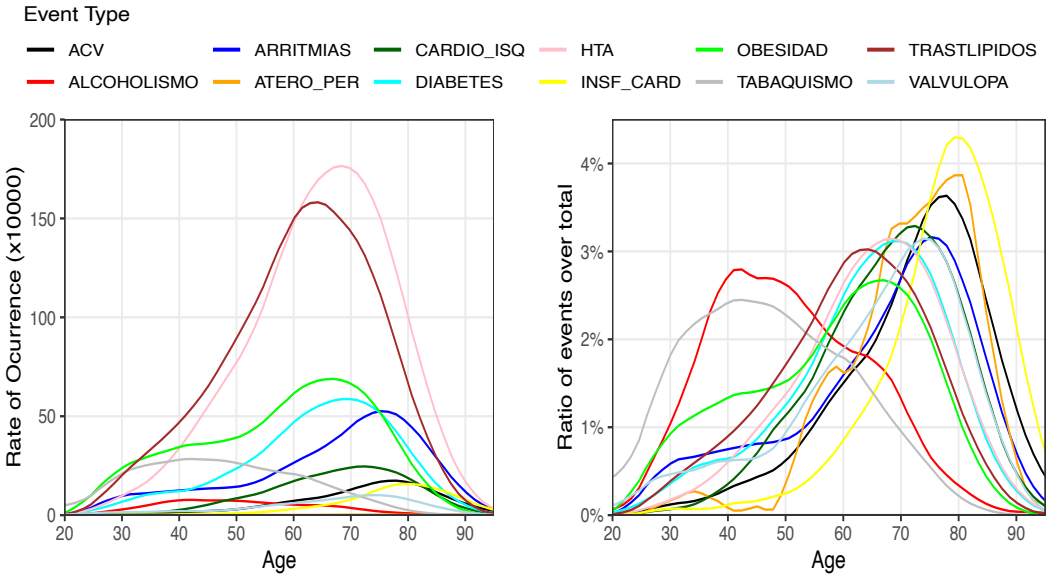


Figura 4.4: Distribución de los intervalos de tiempo entre pares de eventos tanto consecutivos como no consecutivos: a) Desde un tipo de evento dado hasta cualquier otro evento. b) Desde cualquier evento hasta un tipo de evento dado.



(a) **(b)**

Figura 4.5: a) Riesgo de sufrir un tipo de evento dado a diferentes edades: Ratio de ocurrencias de ese evento por cada 10000 personas-año. b) Distribución de edades de la gente que sufre cada tipo de evento. Proporción de eventos de un tipo dado a diferentes edades sobre el total de eventos de ese tipo.

la fecha completa sólo del año de ocurrencia de cada evento. Hemos descartado las variables que fueron creadas a posteriori a partir de otras preexistentes, las que contenían demasiados errores, algunas creadas para estudios con pequeños subgrupos de pacientes... y nos hemos centrado en las más interesantes desde un punto de vista médico. Obteniendo así 18 variables pertenecientes a 146 662 personas, que suman un total de 1 442 614 ocurrencias: *ACV, ALCOHOLISMO, ANEMIA, ARRITMIAS, ATERO_PERIFERICA, CARDIO_ISQUEMICA, COLESTEROL, DIABETES, FALLECIMIENTO, GRIPE, HTA, INSUFICIENCIA_CARDIACA, INSUF_RENAL, OBESIDAD, RETINOPATIA_DIABETICA, TABAQUISMO, TRASTLIPIDOS, VALVULOPATIAS*. Hemos creado una variable COLESTEROL categórica dicotomizando la variable continua recogida originalmente en la base de datos, considerando la ocurrencia de un evento cuando su valor es mayor de 200.

En el gráfico de barras de la Figura 4.6 podemos ver el número total de ocurrencias de cada tipo de evento. Comprobamos que los eventos más frecuente son HTA (323 391), TRASTLIPIDOS (282 410) y GRIPE (267 799).

Hemos tenido que prescindir de la variable fallecimiento, recogida en sólo 6831 personas, porque contiene graves errores: 1871 personas tienen registrados eventos clínicos en años posteriores a su propio fallecimiento. No podemos saber si la fecha es correcta para el resto de individuos. Todo parece indicar que la variable fallecimiento ha sido incorrectamente incorporada a posteriori en la base de datos. Asumiremos que el resto de variables sí que son correctas.

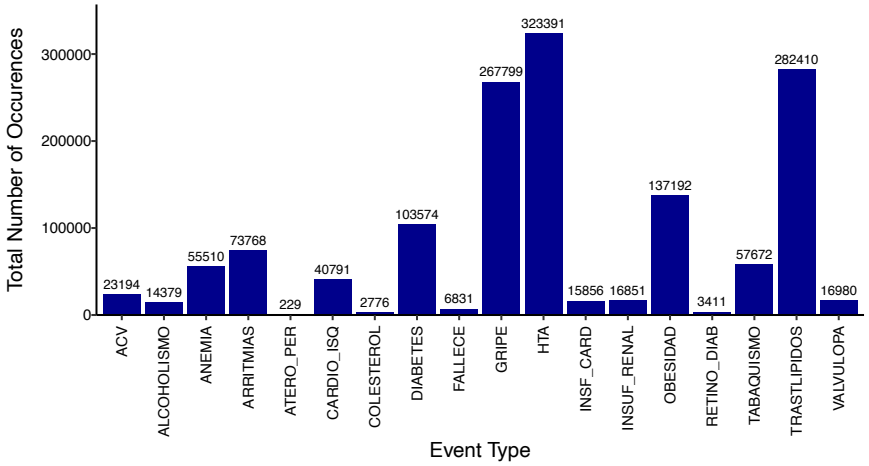


Figura 4.6: Total de eventos de cada tipo recogidos en las variables truncadas.

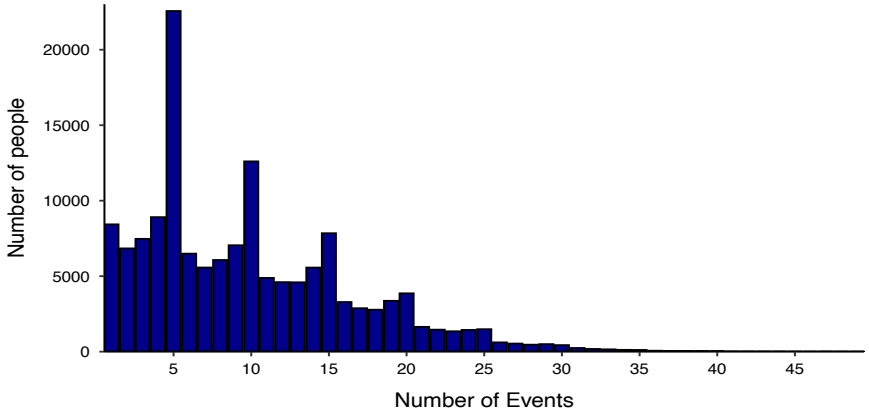


Figura 4.7: Número de personas según la cantidad total de eventos que sufre. Información de las variables truncadas.

La Figura 4.7 muestra cuantas personas sufren un número dado de eventos. Observamos un decrecimiento gradual en la frecuencia del número de eventos a medida que aumenta el número de ocurrencias, algunas personas sufren hasta 50 eventos. Además, el gráfico de

barras nos permite descubrir un patrón, es más frecuente la ocurrencia de un número de eventos múltiplo de 5, indicando que para muchos de los individuos recogidos en la base de datos la ocurrencia de un tipo de evento se repite todos los años.

4.2.2 Variables con fecha exacta.

4.2.2.1 Pares de eventos consecutivos.

La tabla de la Figura 4.8 muestra la tasa de incidencia de los diferentes pares de tipos de eventos consecutivos (número de ocurrencias por cada 1000 personas-año de exposición tras el primer evento). Vemos por ejemplo, como el diagnóstico de Diabetes es seguido inmediatamente del diagnóstico de HTA en la misma persona (al menos una vez) en 40.0 casos por cada 1000 personas-año de exposición tras la Diabetes. Dicho de otro modo, en 40.0 personas sucede al menos una vez el par Diabétes →HTA por cada 1000 personas-año. Comprobamos que el IR es elevado en casi todos los pares en los que el segundo evento es TRASTLIPIDOS o HTA. Comprobamos como el IR es bajo en casi todos los pares en los que el segundo evento es ATERO_PER o ALCOHOLISMO, y es elevado en los pares en los que el primero evento es ATERO_PER. En la Tabla E.1 del Apéndice E se muestra información adicional de los 24 pares con mayor IR que suceden al menos 5 veces: evento inicial, evento final, número de ocurrencias de ese par, tiempo de exposición al riesgo, IR y el intervalo de confianza de IR.

Podemos representar esta información gráficamente mediante redes de conexión, indicando cada tipo de evento de nuestra base de datos mediante un nodo y señalando las transiciones entre pares de eventos mediante flechas. La Figura 4.9 muestra la tasa de incidencia para los diferentes pares de eventos consecutivos por cada 1000 personas-año. Para evitar obtener una maraña ininteligible de conexiones hemos filtrado la información representando sólo los dos pares de mayor IR para cada nodo inicial. Vemos como el par HTA→TRASTLIPIDOS tiene una IR de 76 casos por cada 1000 personas-año, es decir, el diagnóstico HTA es seguido en la misma persona inmediatamente de un diagnóstico de TRASTLIPIDOS en 76 casos (al menos una vez) por cada 1000 personas-año. La Figura 4.10 también muestra mediante una red de conexiones la tasa de incidencia de los diferentes pares de eventos consecutivos pero representando sólo los dos pares de mayor IR para cada nodo final.

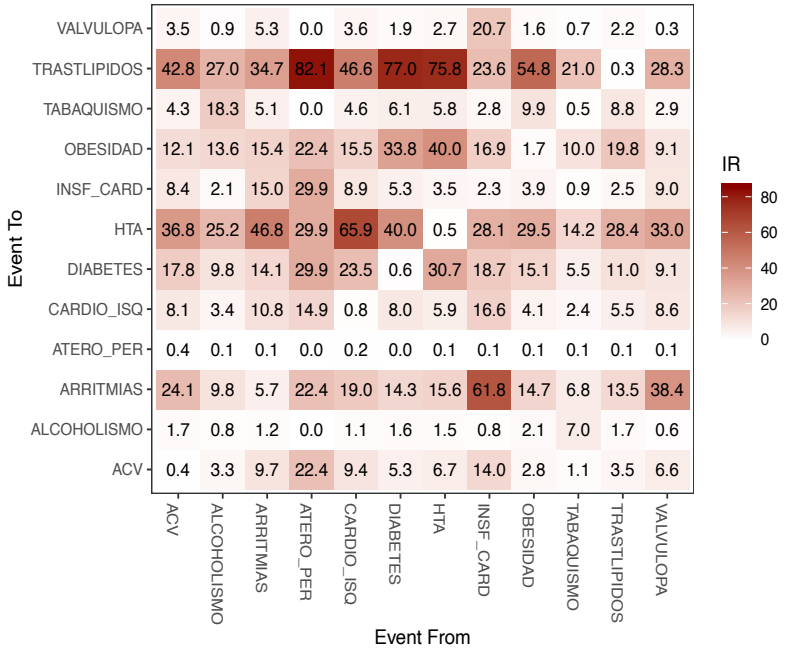


Figura 4.8: IR de los pares de eventos consecutivos (Ocurrencias por cada 1000 personas-año).

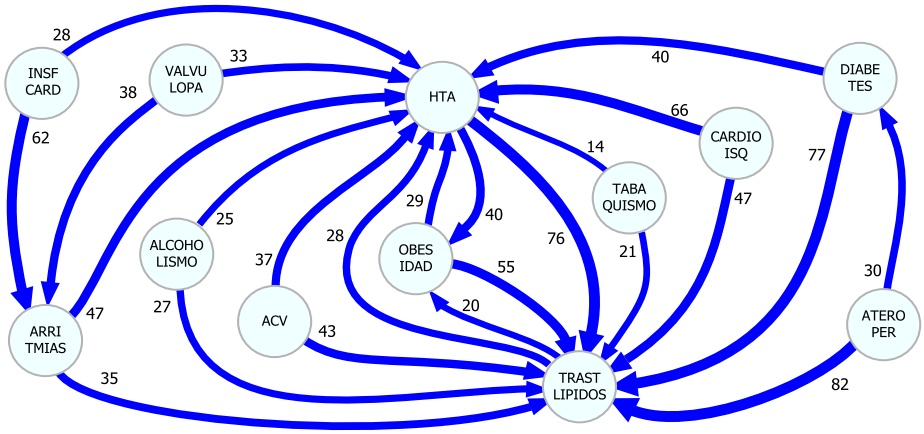


Figura 4.9: Red de conexiones de la IR de los pares de eventos consecutivos calculada con las fechas completas (Ocurrencias por cada 1000 personas-año). Sólo se ha representado los dos pares de mayor IR para cada nodo inicial.

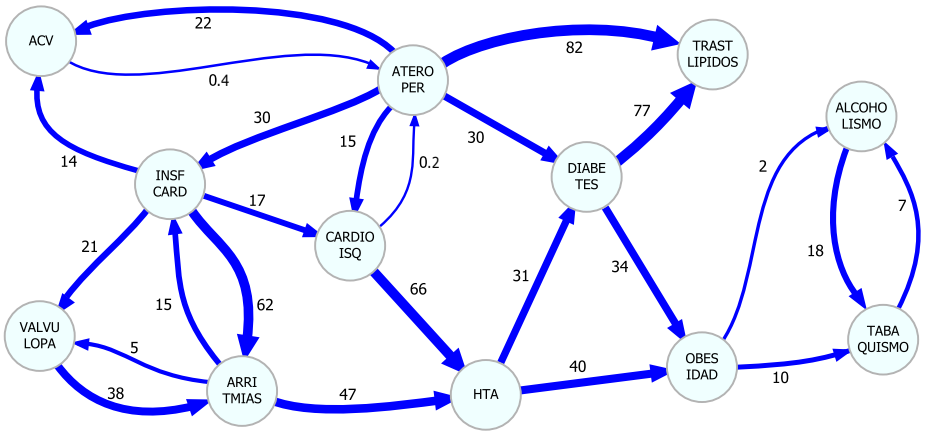


Figura 4.10: Red de conexiones de la IR de los pares de eventos consecutivos calculada con las fechas completas (Ocurrencias por cada 1000 personas-año). Sólo se ha representado los dos pares de mayor IR para cada nodo final.

4.2.2.2 Pares de eventos consecutivos y no consecutivos.

La tabla de Figura 4.11 muestra la tasa de incidencia para los diferentes pares de tipos de eventos, tanto consecutivos como no consecutivos, por cada 1000 personas-año. Vemos como el diagnóstico de Diabetes es seguido en cualquier momento del diagnóstico de HTA en la misma persona en 37.4 casos (al menos una vez) por cada 1000 personas-año. En la Tabla E.2 del Apéndice E se muestra información adicional de los 24 pares con mayor IR.

La Figura 4.12 muestra mediante una red de conexiones la tasa de incidencia para diferentes pares de eventos, tanto consecutivos como no consecutivos, por cada 1000 personas-año. Hemos filtrado la información representando sólo los dos pares de mayor IR para cada nodo inicial. Vemos como el par HTA→TRASTLIPIDOS tiene una IR de 78 casos por cada 1000 personas-año, es decir, el diagnóstico HTA es seguido en cualquier momento en la misma persona de un diagnóstico de TRASTLIPIDOS en 78 casos (al menos una vez) por cada 1000 personas-año. La Figura 4.13 también muestra la tasa de incidencia de los diferentes pares de eventos consecutivos y no consecutivos pero representando sólo los dos pares de mayor IR para cada nodo final.

4.2.2.3 Tripletes de eventos consecutivos y no consecutivos.

En este apartado no incluimos la tabla completa con la Tasa de Incidencia de todos los tripletes porque la información es demasiado extensa. Podemos encontrar una versión

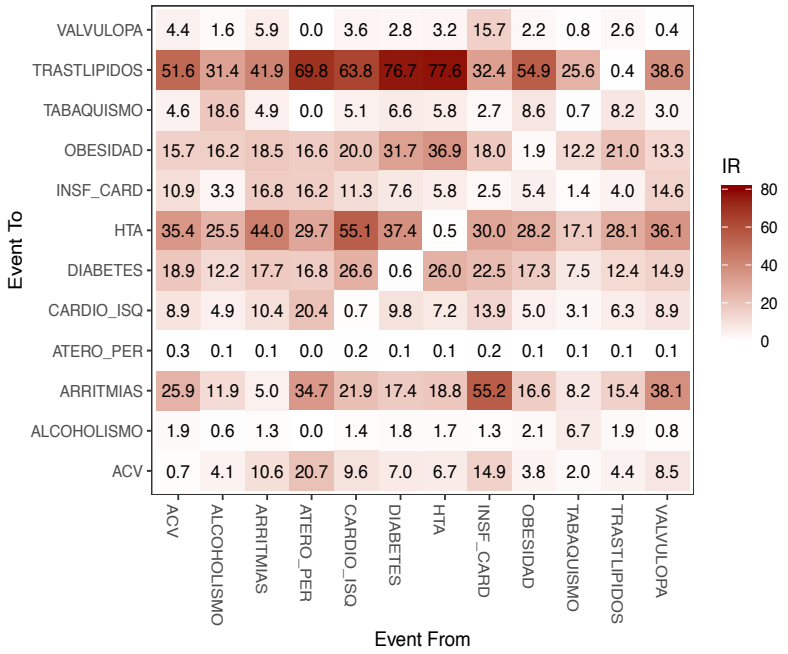


Figura 4.11: IR de los pares de eventos consecutivos y no consecutivos calculada con las fechas completas (Ocurrencias por cada 1000 personas-año).

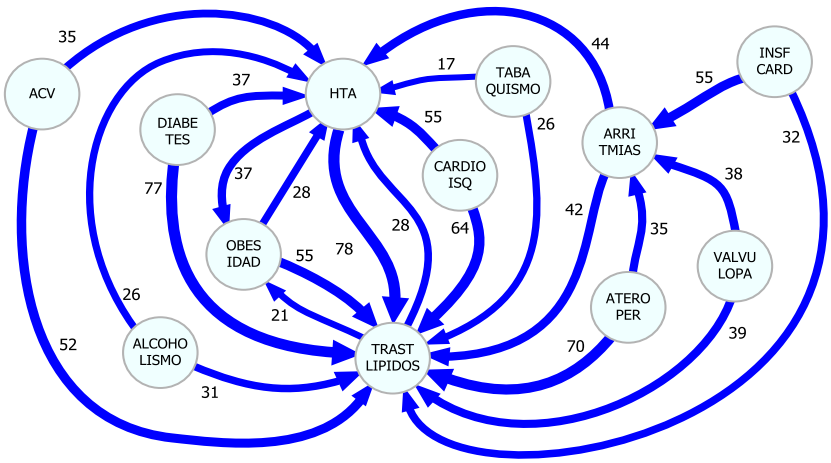


Figura 4.12: Red de conexiones de la IR de los pares de eventos consecutivos y no consecutivos calculada con las fechas completas (Ocurrencias por cada 1000 personas-año). Sólo se ha representado los dos pares de mayor IR para cada nodo inicial.

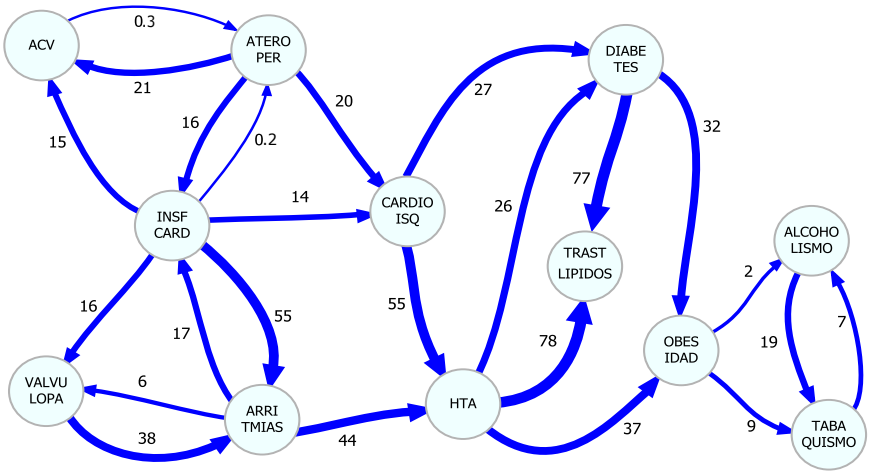


Figura 4.13: Red de conexiones de la IR de los pares de eventos consecutivos y no consecutivos calculada con las fechas completas (Ocurrencias por cada 1000 personas-año). Sólo se ha representado los dos pares de mayor IR para cada nodo final.

reducida, para los 24 tripletes con mayor IR, en la Tabla E.2 del Apéndice E.

La Figura 4.14 muestra mediante una red de conexiones la tasa de incidencia para los diferentes tripletes de eventos, tanto consecutivos como no consecutivos, por cada 1000 personas-año. En la gráfica se han representado todos los tipos de evento de nuestra base de datos, cada uno mediante un nodo. Mediante flechas de diferente color indicamos los tripletes de un mismo individuo. Para evitar obtener una maraña ininteligible de conexiones, ya que hay 1346 tripletes diferentes, debemos filtrar los resultados y quedarnos sólo con los que nos parezcan más importantes, en este caso hemos utilizado el siguiente criterio: para cada tipo de evento final mostramos sólo el triplete de mayor IR y con al menos 5 ocurrencias. Vemos como el triplete VALVULOPATÍA→CARDIO_ISQ→HTA (con flechas rojas) tiene una IR de 55 casos por cada 1000 personas-año, es decir, el diagnóstico VALVULOPATÍA es seguido en cualquier momento de un diagnóstico de CARDIO_ISQ y en cualquier momento posterior de HTA en la misma persona en 55 casos (al menos una vez en esa persona) por cada 1000 personas-año.

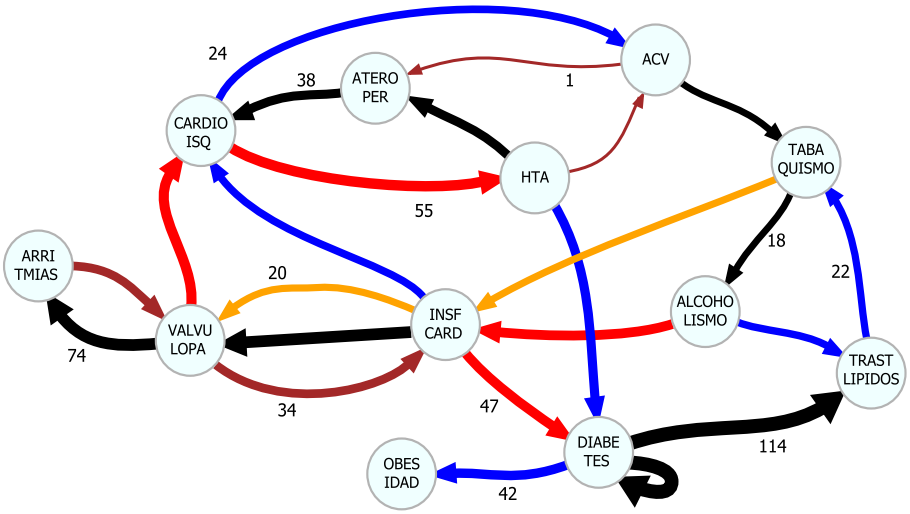


Figura 4.14: Red de conexiones de la IR de los tripletes de eventos consecutivos y no consecutivos calculada con las fechas completas (Ocurrencias por cada 1000 personas-año). Sólo se ha representado el triplete con mayor IR para cada nodo final.

4.2.3 Simulación de fechas truncadas a partir de las exactas.

4.2.3.1 Pares de eventos consecutivos y no consecutivos.

Con el fin de evaluar la metodología que hemos desarrollado para el análisis de secuencias cuando la información temporal es aproximada o incompleta repetimos los análisis anteriores con las mismas variables completas, pero forzando en ellas una menor precisión, para lo cual, truncamos manualmente las fechas originales, conservando sólo el año.

La tabla de la Figura 4.15a muestra la tasa de incidencia para los diferentes pares de tipos de eventos, tanto consecutivos como no consecutivos, por cada 1000 personas-año, calculada a partir de fechas simuladas truncando las fechas originales completas. Hemos imputado los datos asignando a todas las secuencias "candidatas" generadas la misma probabilidad de haber ocurrido realmente. Vemos como el diagnóstico de Diabetes es seguido en cualquier momento del diagnóstico de HTA en la misma persona en 58.8 casos (al menos una vez) por cada 1000 personas-año. En la Tabla E.4 del Apéndice E se muestra información adicional de los 24 pares con mayor IR que suceden.

La tabla de la Figura 4.15b muestra la misma información pero asignando a las secuencias generadas probabilidades a partir de las frecuencias obtenidas previamente de los pares de eventos interanuales. Los resultados obtenidos son prácticamente idénticos. En la Tabla E.5

se muestra información adicional de los 24 pares con mayor IR.

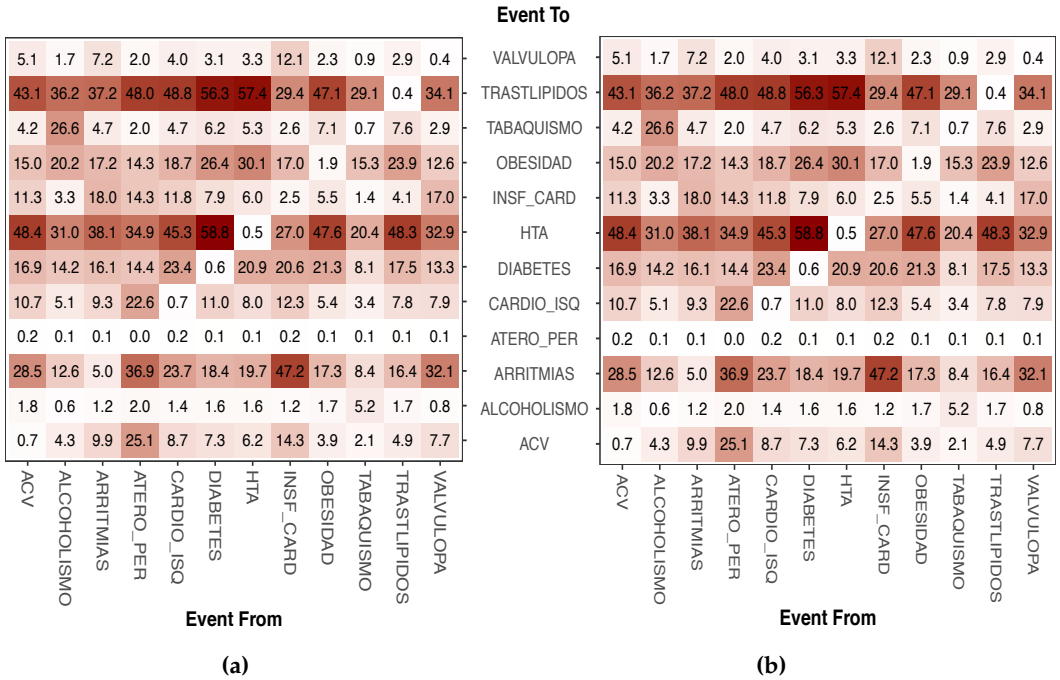


Figura 4.15: IR de los pares de eventos consecutivos y no consecutivos (Ocurrencias por cada 1000 personas-año). a) Método aproximado con imputación aleatoria. b) Método aproximado con imputación usando la probabilidad a priori interanual.

Para poder evaluar el efecto del truncamiento de las fechas independientemente del método de imputación volvemos a calcular el IR pero utilizando solamente los individuos que no han sufrido múltiples eventos intraanuales, es decir, aquellos que no tienen años repetidos, no existe incertidumbre en el orden de sus eventos y por lo tanto no aplicamos ningún método de imputación. Con la información de estos individuos comparamos los IR de los pares de eventos (tanto consecutivos como no consecutivos) calculadas a partir de fechas exactas, Figura 4.16a, con las IR calculadas a partir de fechas truncadas, Figura 4.16b. Vemos como el IR del par Diabetes→HTA calculado con las fechas exactas es 33.2, mientras que el resultado aproximado calculado mediante las fechas truncadas es 33.3 por cada 1000 personas-año. La comparación de estos resultados con los obtenidos utilizando la información de todos los individuos también nos permite evaluar el efecto de descartar la información intraanual.

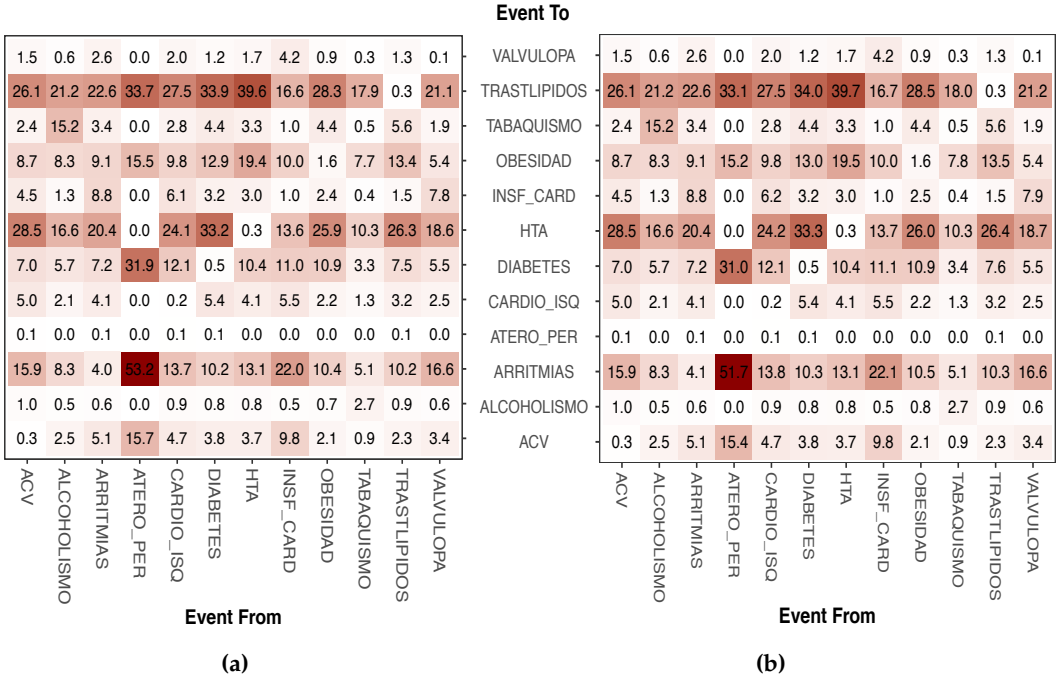


Figura 4.16: IR de los pares de eventos consecutivos y no consecutivos (Ocurrencias por cada 1000 personas-año) utilizando sólo información de las personas sin años repetidos. a) Con fechas exactas. b) Método aproximado con fechas truncadas.

4.2.3.2 Tripletes de eventos consecutivos y no consecutivos.

En este apartado no incluimos la tabla completa con la Tasa de Incidencia de todos los tripletes porque la información es demasiado extensa. Podemos encontrar una versión reducida, para los 24 tripletes con mayor IR, en la Tabla E.6 del Apéndice E. En ella se muestra la IR (por cada 1000 personas-año) de los diferentes tripletes de eventos, tanto consecutivos como no consecutivos, calculada a partir de fechas simuladas truncando las fechas originales completas y asignando a todas las secuencias posibles la misma probabilidad de haber ocurrido realmente.

La Tabla E.7 muestra la IR (por cada 1000 personas-año) de los diferentes tripletes de eventos, tanto consecutivos como no consecutivos, calculada a partir de fechas simuladas truncando las fechas originales completas, pero esta vez imputa los datos intraanuales utilizando las frecuencias obtenidas previamente de los pares de eventos interanuales. Vemos como el diagnóstico (al menos una vez) en una misma persona de DIABETES seguido en cualquier momento de otro diagnóstico DIABETES y luego de TRASTLIPIDOS tiene un IR

de 80.49 casos por cada 1000 personas-año, con un intervalo de confianza del 95% entre 47.7 a 127.2.

4.2.4 Variables con fecha truncada.

4.2.4.1 Pares de eventos consecutivos y no consecutivos.

En este apartado mostramos el resultado de calcular la Tasa de Incidencia de los diferentes pares de eventos que aparecen recogidos en la base de datos con la fecha truncada. Para realizar los análisis utilizamos el algoritmo alternativo Listado F.8, que hace uso de resultados teóricos precalculados en Apartado 3.2.5.1 para simplificar los cálculos y nos permite efectuarlos incluso cuando el número de eventos de cada individuo es muy elevado. El tiempo necesario para obtener los resultados ha sido de 8 horas.

La tabla de la Figura 4.17 muestra la tasa de incidencia para los diferentes pares de tipos de eventos, tanto consecutivos como no consecutivos, por cada 1000 personas-año que aparecen recogidos en la base de datos con la fecha truncada. Hemos imputado los datos asignando a todas las secuencias posibles la misma probabilidad de haber ocurrido realmente. Vemos como la mayoría de los pares en los que el segundo evento es HTA, GRIPE o TRASTLIPIDOS tienen valores de IR elevados. Esto se corresponde con una mayor frecuencia de esos tres tipos de eventos en la base de datos. Sin embargo, los pares cuyo primer evento es HTA, GRIPE o TRASTLIPIDOS no tienen en general un IR mayor que el resto de pares.

Observamos además que la mayoría de los pares en los que el evento final es igual al evento inicial, correspondientes a la diagonal de la tabla, tienen valores de IR cercanos a la unidad (1000 ocurrencias por cada 1000 personas-año) y mucho más altos que el resto de pares. Es decir, la mayoría de las ocurrencias de un evento son seguidas en algún momento por nuevas ocurrencias del mismo tipo de evento. La principal excepción es el par COLESTEROL→COLESTEROL, cuyo IR es de 112 ocurrencias por cada 1000 personas-año. Nótese el valor de IR puede ser superior a 1000 cuando ocurren múltiples eventos intraanuales[7], aunque esto no llega a suceder con ninguno de los pares que hemos analizado debido a que también existen muchos individuos en los que el número de ocurrencias es bajo o el tiempo de exposición alto.

[7] por ejemplo, si una persona sufriese el evento HTA en el año 2001 y los eventos HTA y ACV en el 2002, si sólo considerasemos a esa persona, el IR del par HTA→ACV sería de 1.7

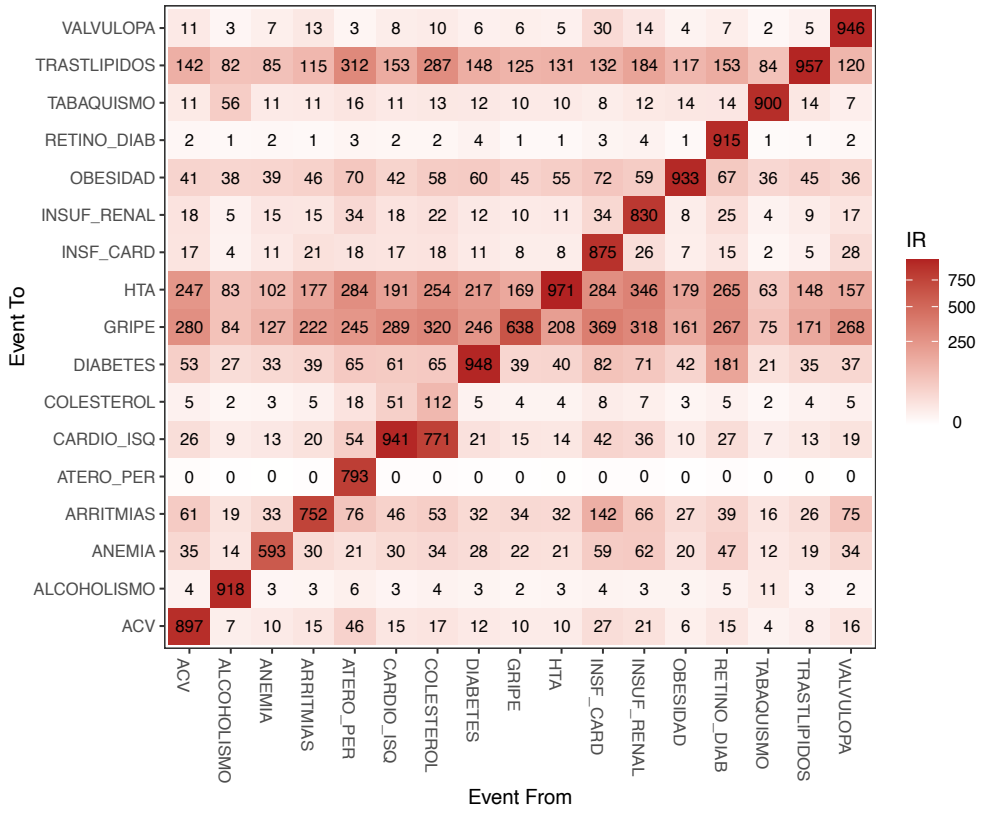


Figura 4.17: IR de los pares de eventos consecutivos y no consecutivos calculados a partir de las variables con fechas truncadas (Ocurrencias por cada 1000 personas-año).

La Figura 4.18 muestra la red de conexiones construida con la tasa de incidencia (por cada 1000 personas-año) de los diferentes pares de eventos, tanto consecutivos como no consecutivos. Hemos filtrado la información representando sólo los dos pares de mayor IR para cada nodo inicial y excluyendo aquellos en los que evento inicial y final son del mismo tipo. Vemos como en la mayoría de los casos los pares de eventos con mayor IR tienen como evento final HTA o GRIPE.

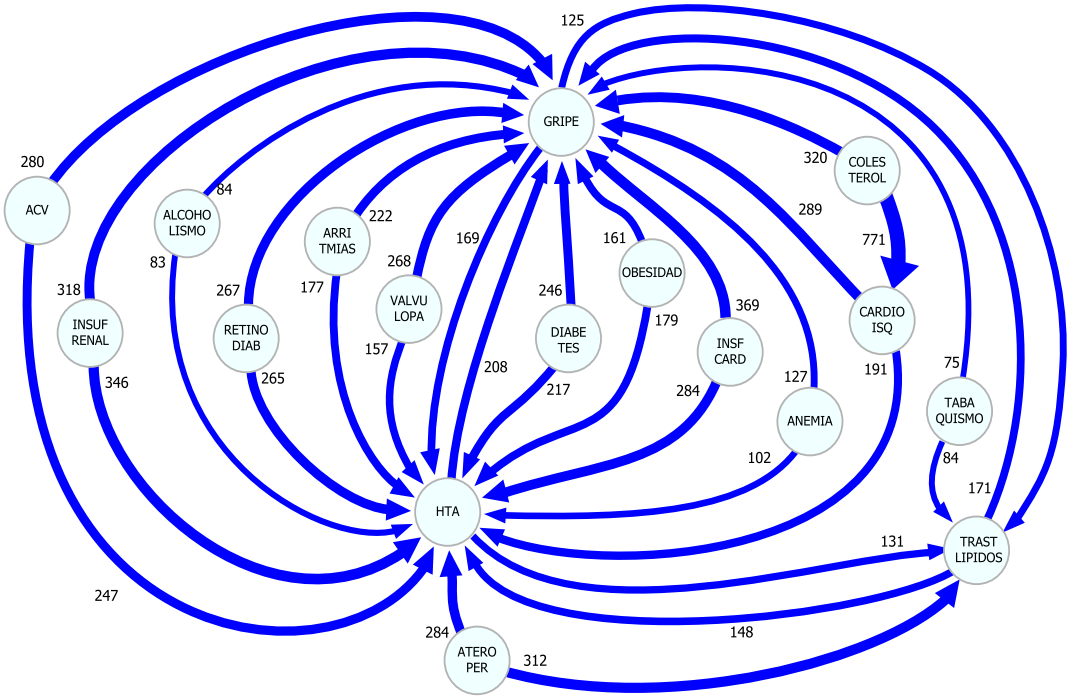


Figura 4.18: Red de conexiones de la IR de los pares de eventos consecutivos y no consecutivos (Ocurrencias por cada 1000 personas-año). Se han representado los dos pares de mayor IR para cada nodo inicial, excepto aquellos en los que nodo inicial y final coinciden.

La Figura 4.19 también muestra la tasa de incidencia de los diferentes pares de eventos consecutivos y no consecutivos pero representando sólo los dos pares de mayor IR para cada nodo final, de nuevo excluyendo aquellos en los que evento inicial y final son del mismo tipo. En este caso no observamos una dominancia tan clara de ningún tipo de evento, aunque podemos ver que en la mayoría de los casos los pares de eventos con mayor IR tienen como evento inicial ATERO_PER o INSF_CARD.

En la Tabla E.8 del Apéndice E se muestra información adicional de los 24 pares con mayor IR.

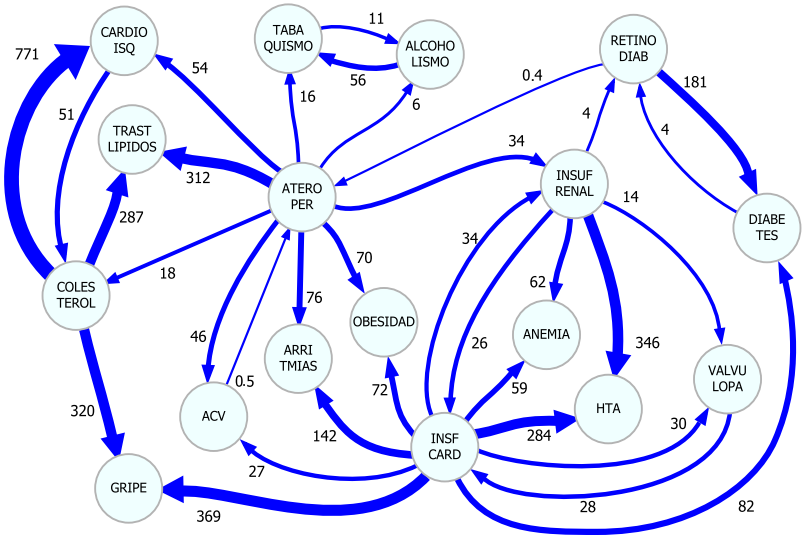


Figura 4.19: Red de conexiones de la IR de los pares de eventos consecutivos y no consecutivos (Ocurrencias por cada 1000 personas-año). Se han representado los dos pares de mayor IR para cada nodo final, excepto aquellos en los que nodo inicial y final coinciden.

4.2.4.2 Tripletes de eventos consecutivos y no consecutivos.

En este apartado mostramos el resultado de calcular la Tasa de Incidencia de los diferentes tripletes de eventos, tanto consecutivos como no consecutivos que aparecen recogidos en la base de datos con la fecha truncada. Hemos imputado los datos asignando a todas las secuencias posibles la misma probabilidad de haber ocurrido realmente. Utilizamos el algoritmo Listado F.9, que hace uso de resultados teóricos precalculados en Apartado 3.2.5.1 para simplificar los cálculos.

No incluimos la tabla completa con los valores de IR para todos los tripletes porque la información es demasiado extensa. Podemos encontrar una versión reducida, para los 24 tripletes con mayor IR, en la Tabla E.9 del Apéndice E.

La Figura 4.20 muestra la red de conexiones construida con la tasa de incidencia (por cada 1000 personas-año) de los diferentes tripletes de eventos, tanto consecutivos como no consecutivos que aparecen recogidos en la base de datos con la fecha truncada. Hemos filtrado la información representando sólo el triplete con mayor IR para cada nodo final excluyendo aquellos en los que evento inicial y final son del mismo tipo. Vemos que para la mayoría de los tripletes la IR es cercana a 1000 ocurrencias por cada 1000 personas-año. El mayor IR corresponde al triplete COLESTEROL→COLESTEROL→CARDIO_ISQ, con un valor de 1648

ocurrencias por cada 1000 personas-año. Nótese que puede haber valores de IR superiores a 1000 cuando existen múltiples ocurrencias intraanuales.

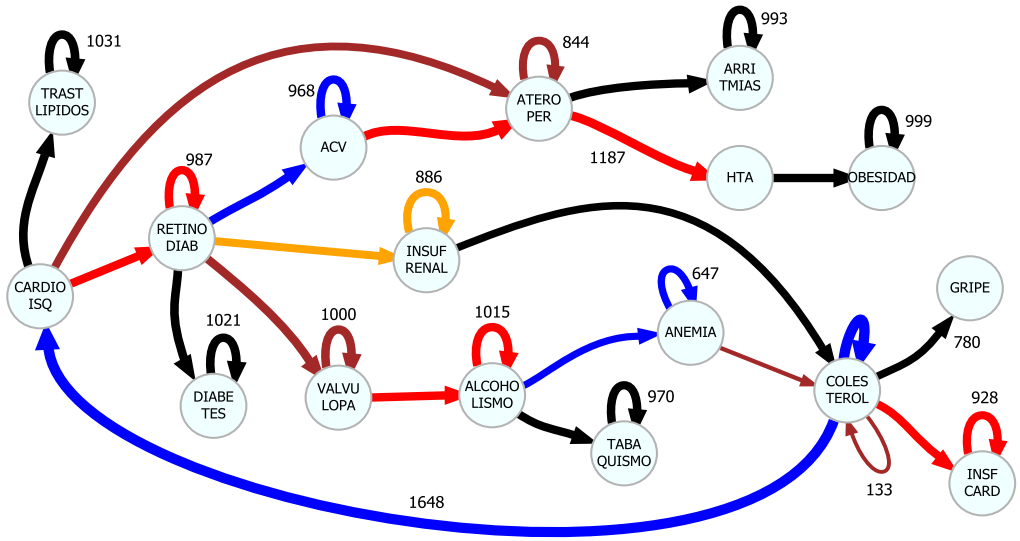


Figura 4.20: Red de conexiones de la IR de los tripletes de eventos consecutivos y no consecutivos (Ocurrencias por cada 1000 personas-año). Sólo se ha representado el triplete con mayor IR para cada nodo final. Se han excluido los tripletes cuyo evento inicial es del mismo tipo que el evento final.

4.2.5 Comparación de los resultados obtenidos con los diferentes métodos aplicados a las variables con fechas exactas.

Con el fin de cuantificar las discrepancias entre los resultados obtenidos con los diferentes métodos desarrollados proponemos calcular a) la media de las diferencias de los IR entre pares de eventos equivalentes $Dif(ResA, ResB) = \frac{\sum_{i=1}^{N}(IRB_i - IRA_i)}{N}$ y b) la media cuadrática de las diferencias de los IR. $Cuad(ResA, ResB) = \sqrt{\frac{\sum_{i=1}^{N}(IRB_i - IRA_i)^2}{N}}$. En la Tabla 4.26 resumimos esas discrepancias calculadas con las variables que aparecen con la fecha completa en la base de datos del Area-7 de Madrid. Todas las entradas de la tabla se refieren al análisis de pares de eventos tanto consecutivos como no consecutivos, excepto la entrada (*), en la que los cálculos se han realizado utilizando solamente los eventos consecutivos. Podemos ver los resultados obtenidos utilizando: fechas exactas vs fechas exactas pero sólo las personas sin múltiples eventos intraanuales, fechas exactas vs fechas truncadas con imputación simple, fechas exactas vs fechas truncadas con imputación utilizando probabilidad interanual, fechas exactas vs fechas exactas pero utilizando sólo eventos consecutivos, sólo personas sin múltiples eventos intraanuales vs fechas truncadas pero sólo las personas sin múltiples

eventos intraanuales, fechas truncadas e imputación simple vs fechas truncadas e imputación utilizando probabilidad interanual.

Tabla 4.26: Comparación de los resultados de los diferentes métodos utilizados para calcular la IR de los pares de eventos. a) Media de las diferencias. b) Media cuadrática de las diferencias.

Método1	Método2	$Dif^{(a)}$	$Cuad^{(b)}$
EXACTO	EXACTO SIN MULTIPLES	-6.37576	11.11141
EXACTO	APROX. SIMPLE	-0.23164	5.17811
EXACTO	APROX. CON PROB. INTERANUAL	-0.23118	5.17088
EXACTO	EXACTO SÓLO CONSECUTIVOS (*)	-0.76270	3.68109
EXACTO SIN MULTIPLES	APROX. SIN MULTIPLES	-0.00160	0.16433
APROX. SIMPLE	APROX. CON PROB. INTERANUAL	0.00046	0.01688

La presencia de algunas IR con valor nulo impide el cálculo de las variaciones relativas o porcentuales, que arrojarían valores infinitos. Por ello hemos optado por calcular las diferencias absolutas. La Tabla 4.27 resume las medias y las medias cuadráticas de las diferencias de las IR de los tripletes calculadas con los diferentes métodos desarrollados: fechas exactas vs fechas truncadas con imputación simple, fechas exactas vs fechas truncadas con imputación utilizando probabilidad interanual, fechas truncadas e imputación simple vs fechas truncadas e imputación utilizando probabilidad interanual.

Tabla 4.27: Comparación de los resultados de los diferentes métodos utilizados para calcular la IR de los tripletes de eventos. a) Media de las diferencias. b) Media cuadrática de las diferencias.

Método1	Método2	Dif	Cuad
EXACTO	APROX. SIMPLE	0.79805	14.89559
EXACTO	APROX. CON PROB. INTERANUAL	0.38507	14.58117
APROX. SIMPLE	APROX. CON PROB. INTERANUAL	-0.41655	5.42726

Capítulo 5

DISCUSIÓN

5.1 IDENTIFICACIÓN DE FACTORES ASOCIADOS CON LA PRESENCIA DE *MISSING DATA*

En la primera parte de esta tesis hemos explorado diferentes métodos para poder identificar eficientemente qué factores están relacionados con la presencia de *missing data* en las variables recogidas en bases de datos médicas de gran tamaño. Para ello utilizamos modelos de regresión logística con efectos aleatorios, y los hemos aplicado a la base de datos del servicio público de salud del Area-7 de Madrid, que contiene información recogida durante siete años a casi un cuarto de millón de personas.

Como paso previo antes de analizar en detalle los resultados estadísticos de cada modelo de regresión hemos llevado a cabo un estudio para cuantificar la memoria y tiempo necesarios para el calculo de cada modelo en función del tamaño de los datos, las librerías y el método de particionamiento utilizados.

Los análisis realizados muestran que la cantidad de memoria necesaria crece rápidamente con el tamaño de la base de datos estudiada, Tabla 4.2. Los modelos que menos memoria consumen durante el ajuste del conjunto completo de los datos son el modelo de efectos aleatorios por centros (1 | EAP), seguido del modelo de efectos fijos (EAP). A continuación aparece el modelo de efectos aleatorios para los individuos anidados a su centro médico (1 | EAP/ID), y por último, vemos que el modelo que más memoria consume es el de efectos aleatorios por individuos no anidadas a su centro (1 | ID).

La memoria utilizada por la librería lme4, considerada el estándar para ajustar modelos con efectos aleatorios en R, crece exponencialmente, llegando incluso a hacer que se cuelgue el programa cuando el número de filas analizadas es superior a 50000 en el modelo (1 | EAP/ID) o de 100000 en el modelo (1 | EAP). El modelo (1 | ID) producía errores con lme4 para tamaños incluso menores por lo que no se ha incluido en los resultados. Por ello, fue necesario buscar alternativas, encontrando finalmente la librería experimental glmmTMB, gracias a la cual ha sido posible analizar los modelos propuestos.

La función *glm* permite analizar el conjunto completo de los datos pero no tiene en cuenta los efectos aleatorios. Con ella hemos incluido la variable centro (EAP) como efecto fijo, obteniendo un intercepto para cada centro. No ha sido posible incluir la variable individuos

(ID) por tener un número demasiado elevado de valores.

El ajuste de modelos independientes en cada partición aleatoria de los datos y el posterior metaanálisis de los resultados utiliza unas 30 veces menos memoria que el análisis de todos los datos simultáneamente. Hemos fragmentado la base de datos en 19 particiones aleatorias, haciendo coincidir esta cantidad con el número de centros médicos para que la comparación de los resultados resulte más natural. En futuros estudios en los que solamente sea necesario analizar los datos en particiones ajustaremos su número en función del tamaño de los datos y la memoria disponible.

El tiempo necesario para realizar los ajustes crece rápidamente con el número de filas analizado, Tabla 4.3. El modelo más rápido de calcular es el modelo simple, a continuación los modelos con efectos aleatorios por centros (ignorando los individuos), luego los modelos con efectos aleatorios por personas anidadas dentro de cada centro y por último, los modelos que tienen en cuenta las medidas repetidas por personas pero sin anidar al centro, que resultan ser extremadamente lentos. En general, la librería glmmTMB resulta ser más rápida que lme4 excepto para el modelo con efectos aleatorios por centros, pero esta última sólo es utilizable cuando el número de filas analizadas no supera los 100000.

El análisis de los datos particionados ha mostrado ser, para el rango de tamaños estudiado, entre una y cinco veces más rápido que el análisis simultáneo del conjunto completo de los datos, obteniéndose las mayores diferencias cuando los datos originales no caben en memoria RAM.

Las mediciones para los modelos particionados se han realizado utilizando un sólo núcleo del ordenador ya que nuestra prioridad era minimizar la cantidad de memoria utilizada. El uso de más hilos (un máximo de cuatro en el nuestro ordenador) permite reducir el tiempo de cálculo pero multiplica la cantidad de memoria necesaria.

Los algoritmos utilizados para computar modelos de regresión con efectos aleatorios utilizan grandes cantidades de memoria y son muy lentos, más cuanto mayor sea el tamaño de los datos. Aunque la utilización de la librería glmmTMB hizo posible el cálculo de los modelos más complejos también fue necesario forzar a R para que utilizase el disco SSD como si fuese memoria RAM. Pero esta solución también es limitada, no es escalable y es muy lenta, algunos de los cálculos realizados duran en torno a un día.

Estos resultados preliminares justifican la necesidad de explorar nuevos métodos para abordar el análisis de los datos. Hemos propuesto particionarlos de dos formas diferentes: por centros médicos y aleatoriamente, realizando los análisis independientemente en cada bloque de datos, unificando posteriormente los resultados mediante un metaanálisis para cada variable. El particionamiento de los datos en 19 bloques nos ha permitido dividir la memoria necesaria para el ajuste de los modelos más complejos entre 5 y dividir el tiempo total de cálculo entre 15. La utilización de un número mayor de bloques, de menor tamaño, permitiría reducir aún más el tiempo y la memoria necesarios a costa de reducir la precisión de los resultados y generar más problemas de convergencia.

En el Apartado 4.1.2 se muestran en detalle los resultados obtenidos al analizar el conjunto completo de los datos. El modelo con efectos aleatorios para los individuos anidados a los centros, Tabla 4.7, muestra que la *odds ratio* de *missings* en el PESO es 0.8089 ($p < 10^{-16}$) para la covariable SEXO (frente a hombres), 2.184 ($p < 10^{-16}$) para EXTRANJERO (frente a nacionales), 0.932 ($p < 10^{-16}$) para EDAD (por año), 0.854 ($p < 10^{-16}$) para PAMED, 1.178 ($p < 10^{-16}$) para PAENF y 1.615 ($p < 10^{-16}$) para YEAR (por año). Nótese que tomamos como niveles de referencia para la regresión YEAR 2006, EDAD 50, PESO 70, PAMED 32, PAENF 18 y POSMAS 0. El cálculo de la R^2 marginal y condicional nos permite ver que los efectos fijos del modelo explican el 14.3% de la variabilidad de las *odds* de tener *missing* en la variable PESO. El factor centro, considerado efecto aleatorio, explica otro 8.9% de la variabilidad, mientras que el factor individuo explica el 57.8%.

Obtenemos resultados parecidos para los coeficientes de las variables, con diferencias inferiores al 2%, para el modelo con efectos aleatorios con individuos no anidados al centro médico, y los p-valores son casi idénticos. Las diferencias entre el modelo de individuos anidados a su centro y el modelo de individuos sin anidar pueden ser debidas a que el primero asume que los individuos pertenecen siempre al mismo centro médico, que tiene menos grados de libertad o al modo en que afectan otras variables de confusión.

El modelo con efectos aleatorios que además incluye POSMAS y los términos de interacción de PAMED y PAENF con YEAR muestra resultados similares y además nos indica que el efecto de PAMED disminuye ligeramente a lo largo de los años, mientras que PAENF apenas cambia. La *odds ratio* de *missings* en la variable PESO es 1.27 para POSMAS (por unidad que incrementa POSMAS), porcentaje de pacientes con edad mayor o igual a 65

años en ese centro médico.

Los resultados obtenidos en los modelos que no incluyen el factor individuo, ID, son muy similares entre sí, hasta el cuarto dígito, pero diferentes a los obtenidos cuando incluíamos ID. La inclusión de los efectos aleatorios para ID equivale a considerar que la presencia de *missings* en los datos de una persona puede estar relacionada con ella, es decir, con factores intrínsecos a la persona no incluidos ya explícitamente en el modelo, como por ejemplo, el parentesco con el médico, el atractivo físico, la religión, alguna variable biológica o que algunos pacientes piden al médico la medición de más variables o se niegan a ofrecer datos. El efecto de todos estos factores queda recogido en ID, mientras que en los modelos sin ID su efecto es absorbido por las demás variables y el error.

En el caso de la variable CIGARRILLOS, el modelo con efectos aleatorios para los individuos anidados a los centros con el conjunto completo de los datos, Tabla 4.11, muestra que la *odds ratio* de *missings* en CIGARRILLOS es 1.013 ($p = 0.057$) para la covariable EDAD, 1.082 para PAMED ($p = 0.00090$), 0.836 ($p = 2.3 * 10^{-5}$) para PAENF y 0.144 ($p < 10^{-16}$) para YEAR. Vemos que los efectos fijos del modelo explican el 67% de la variabilidad de las *odds* de tener *missing* en la variable CIGARRILLOS. La variabilidad debida a los centros es casi cero, es decir la toma de datos es casi igual en todos los centros. En el modelo con más variables, Tabla D.5, podemos ver que la *odds ratio* de *missings* en CIGARRILLOS es 1.013 para PESO ($p = 0.074$) y comprobamos que las variable SEXO ($p = 0.18$) y EXTRANJERO ($p = 0.30$) no son significativas.

El modelo con individuos no anidados por centros para CIGARRILLOS arroja resultados similares para YEAR, y resultados no similares aunque al menos del mismo signo para las variables EDAD y PAENF. Los resultados para PAMED son de signo contrario. Y de nuevo vemos que los modelos que no tienen en cuenta el factor individuo arrojan resultados diferentes a los anteriores pero muy similares entre sí.

En el Apartado 4.1.3 se muestran en detalle los resultados obtenidos al analizar independientemente los datos de cada partición aleatoria, tanto para el PESO, Tablas D.9, D.17, D.25 y D.33, como para los CIGARRILLOS, Tablas D.41, D.47, D.53 y D.59. Los resultados obtenidos en todas las particiones fueron combinados mediante meta-análisis, Tabla D.10 del Apéndice D y siguientes, permitiendo así estimar los coeficientes de cada variable para el

conjunto de los datos. Este proceso fue repetido con modelos de diferente tipo y complejidad. El resultado de los meta-análisis ha sido representado gráficamente mediante forest plots para cada variable, Figura D.1 y siguientes. Observamos como la variabilidad de los coeficientes calculados dentro de cada partición es similar a la variabilidad entre sus medias.

En el Apartado 4.1.4 se muestran en detalle los resultados obtenidos al analizar independientemente los datos de cada centro médico, tanto para el PESO, Tablas D.65 y D.80, como para los CIGARRILLOS, Tablas D.95 y D.101. Los resultados obtenidos en todos los centros son combinados mediante meta-análisis, Tabla D.66 y siguientes. Además, hemos representado gráficamente los meta-análisis de cada variable mediante forest plots, Figura D.49 y siguientes. En este caso observamos como la variabilidad dentro de cada centro es muy pequeña en comparación con la variabilidad entre las medias de los diferentes centros, existe una diferencia significativa entre los coeficientes calculados en diferentes centros, especialmente para las variables YEAR, PAMED, EDAD y PAENF.

Dada la imposibilidad de introducir la variable POSMAS en los ajustes intra-centro, ya que en la mayoría de los centros sólo adquiere dos valores diferentes, incluso en algunos sólo uno, realizamos a posteriori un segundo set de regresiones para estudiar si los coeficientes obtenidos en cada modelo intra-centro dependen del valor que POSMAS toma en los diferentes centros. Estos análisis no han arrojado ningún resultado significativo. Tablas D.73 a D.79 y D.88 a D.94.

Todos estos resultados han sido resumidos en tablas para comparar los diferentes modelos y métodos de particionado. En ellas, reportamos la *odds ratio* de ocurrencia de *missings* en las variables PESO y CIGARRILLOS para las distintas variables explicativas, además se incluye el $IC_{95\%}$ de las *odds ratio*. Consideramos como modelo de referencia el realizado con el conjunto completo de los datos y con efectos aleatorios para los individuos anidados a sus centros médicos.

Para la *odds ratio* de *missing* en PESO comprobamos que, Tabla 4.24, en general, los análisis de los datos particionados aleatoriamente y del conjunto completo de los datos arrojan valores muy similares, hasta el cuarto dígito. Las discrepancias son mayores, en el segundo o tercer dígito, en los análisis particionados por centro médico, especialmente para las variables PAMED y PAENF.

En el caso de la *odds* de *missing* en CIGARRILLOS volvemos a constatar, Tablas 4.25, que los cálculos realizados con el conjunto completo de los datos ofrecen resultados parecidos a los realizados en cada partición aleatoria por separado, hasta el tercer dígito. De nuevo observamos mayor discrepancia en los cálculos realizados mediante el particionado por centros médicos.

Las variables TABACO, GRIPE, RENTA_ZBS, IMC, TOTAL_FARMACOS_DESDE65_ANOS,MFDEMCENTRO, MFCONCENTRO, ENFCONCENTRO, ENFCONDOM, MFCONDOM, MFDOM, ENFDEMCENTRO, ENFDOM, NUM_INGRESOS, ALCOHOLISMO, TIPO_PROFESIONAL, TIPO_USU, DIABETES, DIAGNO_DIABETES, FRCV, CONSULTAS_MF, CONSULTAS_ENF, COLESTEROL, CARDIO_ISQ, TOTAL_ANALITICAS, también fueron incluidas en modelos previos pero no arrojaban coeficientes significativos y además perjudicaban al cálculo de las demás variables, tanto por el error que introducen en los modelos como por la reducción de filas útiles, ya que la mayoría de sus valores son *missing*.

5.2 ANÁLISIS DE LAS SECUENCIAS DE EVENTOS

En la segunda parte de la tesis hemos explorado nuevos métodos para el estudio y representación gráfica de las secuencias de eventos y los hemos utilizado para analizar las secuencias más comunes de eventos médicos en la base de datos del Area-7 de Madrid.

Hemos desglosado cada secuencia en subsecuencias de menor longitud: pares y tripletes, y hemos utilizado la Tasa de Incidencia (IR) para cuantificar su ocurrencia por unidad de tiempo de exposición. A partir de esas subsecuencias contamos el número de personas en las que cada par o triplete tuvo lugar al menos una vez, y lo dividimos entre el tiempo de exposición desde el penúltimo evento hasta el último evento, sin tener en cuenta el tiempo entre eventos anteriores ni nuevas ocurrencias de ese mismo par o triplete. Hemos considerado esta metodología más sencilla de interpretar que otras alternativas, que a su vez tienen sus propios inconvenientes. Habiendo desglosado las subsecuencias de eventos tal como se ha propuesto es inmediato que el investigador aplique cualquier otro criterio de interés.

Nuestro análisis amplía los encontrados en otros estudios al considerar no sólo las subsecuencias de eventos consecutivos sino también los no consecutivos. Esta generalización no sólo produce diferentes valores de IR sino que incluso permite revelar transiciones que

de otro modo pasarían desapercibidas, por ejemplo con el análisis de eventos consecutivos obtenemos un IR de 0 para el par DIABETES→ATERO_PER mientras que con el análisis no consecutivo obtenemos un IR de 0.1 (ocurrencias por cada 1000 personas-año), suficiente para poner de manifiesto la ocurrencia de dicho par. Para otros pares las diferencias entre ambos análisis no son tan drásticas pero si notables, por ejemplo el par ATERO_PER→ARRITMIAS produce un IR de 22.4 en el estudio de pares consecutivos y de 34.7 al considerar también los no consecutivos.

Esta discrepancia es debida a la presencia de otros eventos intercalados, que impiden detectar transiciones más lejanas si sólo se analizan eventos consecutivos. Cuanto más tiempo transcurre entre dos eventos, mayor será la probabilidad de que se produzcan otros eventos entre ellos. Nótese que nuestro análisis no consecutivo considera los eventos como no competitivos, es decir, la ocurrencia de un evento determinado no afecta a la presencia de otras subsecuencias ni modifica su IR.

Para comenzar el estudio de las secuencias de eventos de la base de datos del Area-7 de Madrid hemos analizado aquellos eventos que aparecen con la fecha exacta. Aunque sólo 12 de las variables [8] son de este tipo, contienen suficiente información para realizar un análisis preliminar de los datos, $261\,850$ ocurrencias de eventos pertenecientes a $128\,805$ personas. Además, utilizar esa misma información pero truncando previamente las fechas, nos ha permitido evaluar cómo varían los resultados presencia de fechas aproximadas.

La Figura 4.8 muestra la IR de los diferentes pares de tipos de eventos consecutivos. Vemos como, en general, los IR más altos corresponden a pares de eventos en los que el evento final es TRASTLIPIDOS o HTA. Especialmente elevados son los IR de los pares ATERO_PER→TRASTLIPIDOS (82.1 ocurrencias por cada 1000 personas-año), DIABETES→TRASTLIPIDOS (77.0), HTA→TRASTLIPIDOS (75.8), CARDIO_ISQ→HTA (65.9), INSF_CARD→ARRITMIAS (61.8) y OBESIDAD→TRASTLIPIDOS (54.8). El IR es muy bajo en casi todos los pares en los que el segundo evento es ATERO_PER o ALCOHOLISMO. Estos resultados se corresponden con los que cabría esperar de la frecuencia de los eventos individuales, Figura 4.1a, al menos para los eventos finales de cada par, no así para los iniciales.

[8] Trastornos lipídicos, Arritmias, Tabaquismo, HTA, Insuficiencia cardiaca, Cardiopatía Isquémica, Obesidad, Diabetes, ACV, Alcoholismo, Valvulopatias, Atero-periférico

La representación gráfica de los resultados mediante redes de conexión permite visualizar rápidamente la información más importante e identificar patrones, pero exige un filtrado previo de los datos que consideremos más importantes. Si mostramos los dos pares de mayor IR para cada nodo inicial, Figura 4.9 vemos que la mayoría de los pares consecutivos representados finalizan en HTA o TRASTLIPIDOS.

La Figura 4.11 muestra la IR de los diferentes pares de eventos tanto consecutivos como no consecutivos. De nuevo, en general, los IR más altos corresponden a pares de eventos en los que el evento final es TRASTLIPIDOS o HTA. Especialmente elevados son los IR de los pares HTA→TRASTLIPIDOS (77.6), DIABETES→TRASTLIPIDOS (76.7), ATERO_PER→TRASTLIPIDOS (69.8), INSF_CARD→ARRITMIAS (55.2), CARDIO_ISQ→HTA (55.1) y OBESIDAD→TRASTLIPIDOS (54.9). De nuevo el IR es muy bajo en casi todos los pares en los que el segundo evento es ATERO_PER o ALCOHOLISMO. Podemos visualizar las IR mediante redes de conexión en las Figuras 4.12 y 4.13.

También hemos analizado los tripletes de eventos, secuencias en las que la ocurrencia de un evento está condicionada por otros dos en un orden determinado. A partir del número de ocurrencias de cada triplete y del tiempo de exposición al evento final hemos calculado el IR para cada triplete, tanto consecutivos como no consecutivos. En la Tabla E.3 resumimos los resultados para los 24 tripletes con mayor IR.

Para visualizar fácilmente estos resultados hemos construido una red de conexiones, representando el triplete de mayor IR para cada tipo de evento final Figura 4.14. Hemos señalado con líneas de un mismo color las transiciones pertenecientes a un mismo triplete. Observamos que en la mayoría de los casos, los tripletes con mayor IR son aquellos en los que el último evento es TRASTLIPIDOS, como ya sucedía en el análisis de los pares de eventos, con un valor máximo de 114.39 ocurrencias por cada 1000 personas-año para el triplete DIABETES→DIABETES→TRASTLIPIDOS. Nótese que un triplete puede ser más frecuente que otros aunque no lo sean sus pares constituyentes o eventos individuales.

Con el fin de evaluar nuestra metodología en presencia de información temporal inexacta hemos vuelto a realizar los análisis utilizando los mismos datos pero truncando previamente las fechas, conservando sólo el año. De este modo, además de reducirse la precisión, provocamos otro efecto indeseado: la incertidumbre en el orden de los eventos que

tuvieron lugar durante un mismo año. Se hace necesario entonces utilizar algún método de imputación para poder proseguir con los análisis.

En esta tesis proponemos convertir, para cada persona, las secuencias de eventos de orden parcialmente indeterminado en la combinación de varias posibles secuencias de orden totalmente determinado obtenidas permutando los datos intraanuales. Cada una de esas nuevas secuencias tiene cierta probabilidad de haber ocurrido realmente. Realizamos los cálculos de dos modos: i) asumiendo que la probabilidad es igual para todas las secuencias, Figura 4.15a, y ii) asumiendo que la probabilidad es igual a la estimada a partir de los datos interanuales, Figura 4.15b.

Ambos métodos de imputación arrojan resultados prácticamente idénticos entre sí al ser aplicados a los datos truncados, Figuras 4.15a y 4.15b pero el método de imputación equiprobable es más sencillo y rápido. Sí que se observan diferencias apreciables entre los resultados obtenidos utilizando fechas exactas y los obtenidos utilizando fechas truncadas e imputación. Para cuantificar estas discrepancias hemos calculado el promedio y las medias cuadráticas de las diferencias de las IR arrojadas por los diferentes métodos de cálculo, Tabla 4.2.5.

También hemos probado otras formas de asignar el peso a las secuencias generadas, por ejemplo no calculando un IR parcial con los datos de un sólo individuo sino con el conjunto de todos, realizando varias iteraciones actualizando cada individuo con la mejor estimación del peso hasta ese momento. Los resultados han sido peores.

En promedio, para los pares de eventos, las IR calculadas mediante imputación de las fechas truncadas son ligeramente inferiores (-0.23 ocurrencias por cada 1000 personas-año) a las calculadas con fechas exactas, aunque sí existen discrepancias importantes en algunos pares concretos. Por ejemplo, la IR calculada imputando fechas truncadas es aproximadamente un 12% menor en los pares cuyo evento final es TRASTLIPIDOS, y un 20% mayor en los pares cuyo evento final es HTA.

Contrariamente a lo que sucede al estudiar el conjunto completo de los datos, el análisis de la información perteneciente exclusivamente a personas sin múltiples eventos intraanuales, Apartado 4.2.3.1 y Figura 4.16b, muestra muy pocas diferencias entre utilizar fechas exactas o truncadas. Ello nos permite concluir que la mayor parte de las discrepancias no son

provocadas por la imprecisión derivada del truncamiento sino por la ineficiencia del proceso de imputación, incapaz de estimar adecuadamente el orden real de las secuencias en los individuos que sufrieron múltiples eventos intraanuales.

También hemos evaluado cómo afecta el truncamiento de las fechas al análisis de los tripletes. En promedio, los métodos de imputación que hemos utilizado sobreestiman (0.80 ocurrencias por cada 1000 personas-año el simple y 0.39 el avanzado) la IR de los tripletes de eventos con respecto a los resultados obtenidos al utilizar fechas exactas, y producen mayor variabilidad (14.6) que cuando fueron utilizados con en el análisis de los pares de eventos (5.2), Tablas 4.26 y 4.27. Ello puede ser explicado por el reducido número de ocurrencias de cada triplete, en torno a diez veces inferior al de los pares. De nuevo, ambos métodos de imputación arrojan valores muy parecidos entre sí.

También hemos estudiado las variables que aparecen ya "truncadas" en la base de datos del Area-7 de Madrid, en las que sólo se recoge el año de ocurrencia de los eventos. Este conjunto de datos nos ha permitido incluir en nuestros análisis nuevos tipos de eventos que no aparecían anteriormente en las variables con fechas exactas (ANEMIA, COLESTEROL, GRIPE, INSUF_RENAL y RETINOPATIA_DIABETICA). Además, contiene un número de registros 5.5 veces superior, lo cual permite compensar parcialmente la menor precisión en los resultados causada por el truncamiento.

Las hemos analizado utilizando el método de imputación equiprobable, en el que se asigna la misma probabilidad de ocurrencia a todas las posibles secuencias de un individuo y a partir de ellas se calculan los pares y tripletes. Como hemos comprobado en apartados anteriores los resultados son similares a los obtenidos utilizando información a priori interanual.

Un primer intento de análisis de estos datos resultó fallido: Los algoritmos utilizados desarrollan todas las posibles secuencias en todos los individuos, permutando y combinando adecuadamente sus eventos, a partir de ellas calculan todos los pares o tripletes, eligen su primera ocurrencia y con ella calculan el IR. Esta metodología es rápida, R puede realizar muchas operaciones en paralelo, pero necesita cantidades ingentes de memoria. Esto no supuso un obstáculo para el análisis de las variables "exactas" pero sí que lo es ahora con las truncadas: tenemos 18 variables y cada individuo llega a sufrir hasta 49 ocurrencias de

eventos. Ha sido por tanto necesario rediseñar el algoritmo de cálculo, consiguiendo reducir la memoria necesaria (dividiendola entre cien), pero a costa de multiplicar por diez el tiempo de computación. En lugar de desarrollar todas las posibilidades y filtrar los resultados, el nuevo método realiza los cálculos uno a uno recursivamente, abortándolos al detectar la primera ocurrencia de cada par o triplete para cada individuo y utilizando resultados precalculados teóricamente.

En las variables truncadas, la mayoría de los pares en los que el segundo evento es HTA, GRIPE o TRASTLIPIDOS tienen valores de IR elevados, Figura 4.17. Esto se corresponde con una mayor frecuencia de esos tres tipos de eventos en la base de datos. Sin embargo, los pares cuyo primer evento es HTA, GRIPE o TRASTLIPIDOS no tienen en general un IR mayor que el resto de pares.

Observamos además que la mayoría de los pares en los que el evento final es igual al evento inicial tienen valores de IR cercanos a la unidad y mucho más altos que el resto de pares. La principal excepción es el par COLESTEROL→COLESTEROL, cuyo IR es de 112 ocurrencias por cada 1000 personas-año.

Las redes de conexiones construidas a partir del IR de los diferentes pares de eventos nos permiten representar gráficamente los resultados más significativos. En la mayoría de los casos los pares de eventos con mayor IR tienen como evento final HTA o GRIPE, Figura 4.18. En la mayoría de los casos los pares de eventos con mayor IR tienen como evento inicial ATERO_PER o INSF_CARD, Figura 4.19.

También hemos podido realizar con éxito el análisis de los tripletes de eventos a partir de los datos recogidos con la fecha truncada, Apartado 4.2.4.2. Para la mayoría de los tripletes representados la IR es cercana a 1000 ocurrencias por cada 1000 personas-año. El mayor IR corresponde al triplete COLESTEROL→COLESTEROL→CARDIO_ISQ, con un valor de 1648 ocurrencias por cada 1000 personas-año. De nuevo, consideramos especialmente útil el uso de redes de conexiones para visualizar los resultados de los diferentes tipos de tripletes de eventos, Figura 4.20.

El formato en el que se han registrado los eventos del Area-7, indicando si han ocurrido o no un determinado año, impone una restricción: cada año puede darse a lo sumo una ocurrencia de cada tipo de evento. Además, no sabemos en principio si la repetición de un

evento en años consecutivos para una misma persona es debida al acarreo de la información de un único evento o a que realmente ha sufrido nuevas ocurrencias de ese tipo de evento. Hemos comprobado, en la base de datos del Area-7 de Madrid, que no parece haberse realizado un acarreo automático de la información recogida para ninguno de los tipos de eventos estudiados, ya que para todos ellos se han identificado personas en las que la ocurrencia de un tipo de evento no es seguida por nuevas ocurrencias de ese evento en años posteriores.

Capítulo 6

CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

6.1 CONCLUSIONES

A lo largo de esta tesis hemos explorado y desarrollado diferentes métodos para el procesado y análisis de bases de datos biomédicas complejas que nos han permitido estudiar dos problemas habituales en bioestadística:

A) la identificación de factores asociados con la presencia de *missing data* en otras variables en las bases de datos médicas.

B) el análisis y representación gráfica de secuencias de eventos médicos con recurrencia y competición, incluso cuando sólo se disponga de información temporal parcial.

Consideramos que se han cumplido satisfactoriamente los objetivos propuestos para este trabajo.

6.1.1 IDENTIFICACIÓN DE FACTORES ASOCIADOS CON LA PRESENCIA DE *MISSING DATA*

1) Con el fin de poder calcular los modelos estadísticos cuando el tamaño de los datos es demasiado grande para la memoria disponible hemos estudiado diferentes estrategias de particionamiento del análisis. Hemos dividido los datos en bloques de menor tamaño, ajustado los modelos de regresión independientemente en cada uno de ellos y combinado los resultados mediante metaanálisis.

   (a) El análisis del conjunto completo de los datos es el más preciso pero extremadamente lento y no es escalable, la memoria utilizada desborda rápidamente las capacidades del ordenador.

   (b) El análisis de los datos particionados aleatoriamente es escalable, rápido y arroja resultados muy parecidos a los obtenidos en el análisis del conjunto completo de los datos.

   (c) El análisis de los datos particionados por centro es igual de rápido que el del aleatorio pero menos preciso. Para algunas variables los resultados difieren considerablemente de los obtenidos con el conjunto completo de los datos. Esta estrategia es mucho más susceptible de producir errores durante los cálculos e impide el análisis combinado de variables intra-centro e inter-centro en un mismo modelo. Por otro lado, permite realizar los cálculos independientemente en cada centro sin necesidad

de intercambiar información entre ellos ni mantener una gigantesca base de datos centralizada.

(d) Consideramos por tanto que la mejor opción para calcular modelos de regresión logística con efectos aleatorios en bases de datos de gran tamaño es realizar los cálculos en particiones aleatorias. La menor precisión provocada por el particionamiento es compensada por la gran cantidad de datos disponibles, que permiten seguir detectando efectos significativos en las variables. Aunque podríamos tener problemas con variables categóricas cuyos niveles sean muy poco frecuentes. Debemos elegir el número de particiones en función del tamaño de los datos y la memoria disponible, teniendo en cuenta qué las particiones pequeñas producen resultados menos precisos pero más rápido.

2) El análisis de la base de datos del servicio público de salud del Área-7 de Madrid mediante modelos de regresión logística con efectos aleatorios ha mostrado que la presencia de *missing data* en algunas variables médicas está asociada con el valor que adquieren otros factores relacionados con el paciente y el centro médico. Observamos que la probabilidad de tener *missing data* en la variable PESO es mayor en hombres, jóvenes, extranjeros, cuanto mayor sea la presión asistencial de enfermería y menor sea la presión asistencial del médico y ha ido aumentando cada año. El factor centro explica un 8.9% de la variabilidad. La probabilidad de tener *missing data* en la variable CIGARRILLOS es mayor en ancianos y cuanto mayor sea la presión asistencial del médico y menor la presión asistencial de enfermería y ha ido disminuyendo cada año. En este caso la variabilidad entre centros es prácticamente nula.

6.1.2 ANÁLISIS DE LAS SECUENCIAS DE EVENTOS

3) La descomposición de las secuencias de eventos en múltiples subsecuencias de menor tamaño ha mostrado ser un procedimiento útil para identificar fácilmente las subsecuencias más relevantes. Nuestro análisis, en el que hemos calculado la Tasa de Incidencia (IR) de pares y tripletes de eventos, amplía otros estudios al considerar no sólo los eventos consecutivos sino también los no consecutivos, lo cual nos permite descubrir transiciones que de otro modo pasan desapercibidas.

4) Para poder analizar las variables con información temporal incompleta hemos desarrollado dos métodos de imputación permutando los eventos intraanuales: asignando

probabilidades iguales a cada nueva secuencia o de acuerdo a las frecuencias observadas interanuales. Ambos métodos de imputación arrojan resultados prácticamente idénticos pero el método de imputación equiprobable es más sencillo y rápido y permite simplificar el algoritmo de análisis.

5) Las redes de conexiones han mostrado ser una herramienta útil para visualizar fácilmente los resultados de los análisis, especialmente para las secuencias más largas. Para ello puede es necesario decidir un criterio para filtrar las transiciones que consideremos más relevantes, en nuestro caso según la IR. Debemos ser precavidos ya que aunque los pares excluidos tengan menor IR podrían ser importantes por otros motivos. Nótese que no podemos concatenar los resultados de los pares de eventos para inferir los resultados de los tripletes, ya que estaríamos mezclando erróneamente información de diferentes personas. Nuestro enfoque es diferente al de Markov porque este último asume que cada transición es independiente de las otras.

6) Los algoritmos desarrollados nos han permitido analizar las secuencias de eventos de la base de datos del Area-7 de Madrid, detectando los pares y tripletes de eventos de mayor IR para cada tipo de evento inicial o final.

   (a) Aunque en algunos casos se observan diferencias entre los resultados obtenidos utilizando fechas completas y con fechas truncadas, los métodos desarrollados son en general prometedores y nos han permitido realizar los análisis con todos los datos evitando tener que descartar gran cantidad de información y obtener así resultados sesgados. Los resultados se corresponden con los que cabría esperar de la frecuencia de los eventos individuales, al menos para los eventos finales de cada par, no así para los iniciales.

   (b) El análisis de los eventos que aparecen con la fecha completa ha mostrado que en general, los pares de eventos con mayor IR son aquellos en los que el evento final es Trastorno Lipídico o HTA. Especialmente elevados son las IR de los pares ATERO_PER→TRASTLIPIDOS, DIABETES→TRASTLIPIDOS, HTA→TRASTLIPIDOS, CARDIO_ISQ→HTA, INSF_CARD→ARRITMIAS y OBESIDAD→TRASTLIPIDOS. El IR es muy bajo en casi todos los pares en los que el segundo evento es ATERO_PER o ALCOHOLISMO.

   (c) En la mayoría de los casos, los tripletes con mayor IR son aquellos en los que el último evento es TRASTLIPIDOS.

(d) El análisis de los eventos que aparecen con la fecha truncada ha permitido incluir nuevas variables en el análisis y ha mostrado que la mayoría de los pares en los que el segundo evento es HTA, GRIPE o TRASTLIPIDOS tienen valores de IR elevados.

(e) Las IR calculadas mediante imputación de las fechas truncadas son, en general, ligeramente inferiores a las calculadas con fechas exactas. En el caso de los tripletes, en promedio, los métodos de imputación sobreestiman la IR y producen mayor variabilidad que con los pares de eventos.

6.2 LIMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

1. Con el fin de superar las limitaciones de memoria el procesado de los datos ha requerido de la fragmentación de los datos originales, procesado por separado de cada fragmento, volcado de los resultados intermedios en el disco y su posterior recombinación. Pero este modo de proceder sólo es aplicable a tareas que pueden ser fácilmente ejecutadas en paralelo.

2. A pesar de la optimización de los análisis, el tamaño de los datos estudiados está en el límite de lo que puede ser manejado adecuadamente en un ordenador medio con las librerías estadísticas comunes, que intentan cargar simultáneamente toda la información en memoria, esto incluye R con las librerías lme4 o glmmTMB, SPSS, Stata, Python con numpy y statsmodels, y Julia con la librería MixedModels.jl.

3. Para futuros estudios aconsejamos la utilización de programas especializados en el tratamiento y visualización de grandes cantidades de datos, como Tableau o Qlikview o almacenar la información en bases de datos como MonetDB o TileDB. También sería recomendable, si se dispone de los recursos necesarios, utilizar plataformas de cálculo distribuido como Spark con la librería Photon-ML. La desventaja de estas tecnologías es que, aunque facilitan la manipulación de los datos y la obtención de estadísticas básicas, son más complicadas, muchas de sus librerías son experimentales y en general no permiten ajustar modelos complejos como los utilizados en nuestro estudio.

4. Es importante concienciar a médicos y analistas para que los datos sean adecuadamente recopilados y estructurados en origen, registrando las fechas y otras variables con

precisión y utilizando criterios homogéneos. Todo ello permitiría simplificar los análisis, evitar errores triviales y obtener resultados más precisos sin necesidad de utilizar complejos métodos estadísticos, [49].

5. En estudios de mayor tamaño, en los que no sea viable realizar un análisis previo con el conjunto completo de los datos, puede ser necesario realizar la selección de las covariables independientemente en cada partición, lo que podría generar modelos diferentes o incluso incompatibles en cada una. Es necesario desarrollar métodos capaces de combinar modelos que difieran no sólo en el valor de los coeficientes sino también en su estructura.

6. Dado el gran número de covariables, los métodos de regresión regularizada, como Lasso o ElasticNet, podrían ofrecer mejores resultados para seleccionar los modelos con la combinación óptima de variables minimizando los problemas de sobreajuste, pero son computacionalmente exigentes y también necesitaríamos aplicarlos en cada partición y combinar sus resultados mediante un metaanálisis.

7. Para poder analizar las secuencias de eventos con información temporal incompleta hemos imputado los datos intra-anuales utilizando la distribución de los intervalos de tiempo inter-anuales, pero la información disponible está sesgada, no es posible reconstruirla eficientemente más allá de la zona de truncamiento. Para futuros estudios aconsejamos modelar por separado los intervalos de tiempo y el número de ocurrencias de cada par de eventos mediante distribuciones truncadas (de Poisson, Exponencial,...) o modelar directamente la IR mediante alguna distribución genérica como la Beta truncada.

8. También sería útil la utilización de bootstrap y modelos bayesianos para estimar los parámetros de los modelos y calcular sus intervalos de credibilidad. En la práctica, el uso de modelos de elevada complejidad y el remuestreo puede resultar inviable en bases de datos de gran tamaño. En la medida de los posible, debemos intentar sustituir procedimientos complejos de cálculo por ecuaciones precalculadas obtenidas mediante razonamientos teóricos y utilizar aproximaciones validas.

9. Otra línea de investigación importante sería combinar la información proveniente de diferentes conjuntos de datos o con diferentes formatos, por ejemplo la obtenida de las variables truncadas y la de las variables sin truncar.

10. Nuestros análisis se han realizado considerando las secuencias como eventos ordenados. También podrían realizarse ignorando el orden de parte de los eventos. Podríamos por ejemplo calcular la IR de la transición a un cierto tipo de evento si la persona sufrió previamente otros dos eventos sin importar cuál de ellos ocurrió primero, que por otro lado, sería otra alternativa para el análisis de las secuencias de orden parcialmente indeterminado.

11. Otra línea de trabajo interesante para el cálculo de IR en tripletes o secuencias de mayor tamaño sería encontrar el modo de incluir en los resultados la información de los intervalos temporales entre los primeros eventos de las secuencias y considerar la competitividad de los eventos incluso cuando no son consecutivos.

12. Sugerimos utilizar bibliotecas como Visnetwork que permiten tratar los gráficos de forma interactiva seleccionando y filtrando fácilmente con el ratón las rutas y nodos más relevantes para el analista. Las redes de conexión también pueden ser utilizadas para realizar un análisis de redes diferenciales, [149, 99], estudiar cómo cambian las relaciones entre los nodos ante la variación de algún factor como puede ser la edad o el sexo o investigar correlaciones complejas.

Capítulo 7

BIBLIOGRAFÍA

BIBLIOGRAFÍA

[1] Ravindra K. Ahuja, Thomas L. Magnanti y James B. Orlin. *Network flows: theory, algorithms, and applications*. 1993.

[2] Paul D. Allison. «Multiple Imputation for Missing Data. A Cautionary Tale». 28 (2000).

[3] Raquel Alonso y Luis Llanes. «Control de calidad en la gestión de las reclamaciones de los usuarios del área sanitaria 10 de atención especializada de la comunidad de Madrid (2000-2005)». *Revista de Calidad Asistencial* 24.2 (2009).

[4] Elske Ammenwerth y H-P Spötl. «The time needed for clinical documentation versus direct patient care». *Methods of information in medicine* 48.01 (2009).

[5] Leila DAF Amorim y Jianwen Cai. «Modelling recurrent events: a tutorial for analysis in epidemiology». *International journal of epidemiology* 44.1 (2015).

[6] P. K. Andersen. «Generalised linear models for correlated pseudo-observations, with applications to multi-state models». 90 (2003).

[7] Per Kragh Andersen y Niels Keiding. «Multi-state models for event history analysis». *Statistical methods in medical research* 11.2 (2002).

[8] Per Kragh Andersen y Maja Pohar Perme. «Pseudo-observations in survival analysis». 19 (2010).

[9] H Ansari, M A Mansournia, S Izadi, M Zeinali, M Mahmoodi y K Holakouie-Naieni. «Predicting CCHF incidence and its related factors using time-

series analysis in the southeast of Iran: comparison of SARIMA and Markov switching models.» *Epidemiology and infection* 143 (4 mar. de 2015).

[10] Hilary Aralis y Ron Brookmeyer. «A stochastic estimation procedure for intermittently-observed semi-Markov multistate models with back transitions.» *Statistical methods in medical research* (ene. de 2017).

[11] JM Aranaz y C Moya. «Seguridad del paciente y calidad asistencial». *Rev Calid Asist* 26.6 (2011).

[12] Juan Jose Piñero de Armas. «Analysis of Sequences of Cardiovascular Events». Congreso Internacional "XVII Conferencia Española y VII Encuentro Iberoamericano de Biometría CEB-EIB 2019". Valencia, jun. de 2019.

[13] R. Harald Baayen, Douglas J. Davidson y Douglas M. Bates. «Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items». *Journal of memory and language* 59.4 (2008).

[14] Paul B Baltes y John R Nesselroade. *History and rationale of longitudinal research*. Academic Press, 1979.

[15] Albert-Laszlo Barabasi y Zoltan N Oltvai. «Network biology: understanding the cell's functional organization». *Nature reviews genetics* 5.2 (2004).

[16] J Barnard y X L Meng. «Applications of multiple imputation in medical studies: from AIDS to NHANES.» *Statistical methods in medical research* 8 (1 mar. de 1999).

[17] Adrian Barnett y Nick Graves. «Competing risks models and time-dependent covariates.» *Critical care (London, England)* 12 (2 2008).

[18] Francesco Bartolucci y Alessio Farcomeni. «A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates.» *Biometrics* 71 (1 mar. de 2015).

[19] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker y col. «Lme4: Linear Mixed-Effects Models Using Eigen and S4». *R package version* 1.7 (2014).

[20] Leonard E. Baum y Ted Petrie. «Statistical Inference for Probabilistic Functions of Finite State Markov Chains». 37 (1966).

[21] J. Robert Beck y Stephen G. Pauker. «The Markov Process in Medical Prognosis». 3 (1983).

[22] Aurélien Belot, Laurent Remontet, Guy Launoy, Valérie Jooste y Roch Giorgi. «Competing risk models to estimate the excess mortality and the first recurrent-event hazards.» *BMC medical research methodology* 11 (mayo de 2011).

[23] C. S. Berkey, D. C. Hoaglin, F. Mosteller y G. A. Colditz. «A random-effects regression model for meta-analysis». 14 (1995).

[24] Jan Beyersmann y Martin Schumacher. «Time-dependent covariates in the proportional subdistribution hazards model for competing risks.» *Biostatistics (Oxford, England)* 9 (4 oct. de 2008).

[25] Warren B. Bilker y Mei-Cheng Wang. «Bootstrapping left truncated and right censored data». *Communications in Statistics - Simulation and Computation* 26.1 (1997).

[26] Laura Boeschoten, Danila Filipponi y Roberta Varriale. «Combining Multiple Imputation and Hidden Markov Modeling to Obtain Consistent Estimates of Employment Status» (ene. de 2020).

[27] Cristian Bologa, Vernon Shane Pankratz, Mark L Unruh, Maria Eleni Roumelioti, Vallabh Shah, Saeed Kamran Shaffi, Soraya Arzhan, John Cook y Christos Argyropoulos. «Generalized Mixed Modeling in Massive Electronic Health Record Databases: what is a healthy serum potassium?» *arXiv preprint arXiv:1910.08179* (2019).

[28] Marco Bonetti, Raffaella Piccarreta y Gaia Salford. «Parametric and nonparametric analysis of life courses: an application to family formation patterns.» *Demography* 50 (3 jun. de 2013).

[29] Mark Bounthavong, Jonathan H Watanabe y Kevin M Sullivan. «Approach to addressing missing data for electronic medical records and pharmacy claims data research.» *Pharmacotherapy* 35 (4 abr. de 2015).

[30] Leo Breiman. «Random forests». *Machine learning* 45.1 (2001).

[31] Stef van Buuren. *Flexible Imputation of Missing Data, Second Edition*. 2018.

[32] Stef van Buuren y Karin Groothuis-Oudshoorn. «mice: Multivariate Imputation by Chained Equations in R». 45 (2011).

[33] Jianwen Cai y Douglas E Schaubel. «Marginal means/rates models for multiple type recurrent event data.» *Lifetime data analysis* 10 (2 jun. de 2004).

[34] Qing Cai, Mei-Cheng Wang y Kwun Chuen Gary Chan. «Joint modeling of longitudinal, recurrent events and failure time data for survivor's population.» *Biometrics* 73 (4 dic. de 2017).

[35] Qi Cao, Erik Buskens, Talitha Feenstra, Tiny Jaarsma, Hans Hillege y Douwe Postmus. «Continuous-Time Semi-Markov Models in Health Economic Decision Making: An Illustrative Example in Heart Failure Disease Management.» *Medical decision making : an international journal of the Society for Medical Decision Making* 36 (1 ene. de 2016).

[36] Baojiang Chen. «Statistical Methods for Multi-State Analysis of Incomplete Longitudinal Data» (2008).

[37] Bingshu E Chen, Richard J Cook, Jerald F Lawless y Min Zhan. «Statistical methods for multivariate interval-censored recurrent events.» *Statistics in medicine* 24 (5 mar. de 2005).

[38] Bingshu Eric Chen y Richard J Cook. «Tests for multivariate recurrent events in the presence of a terminal event.» *Biostatistics (Oxford, England)* 5 (1 ene. de 2004).

[39] Tomoaki Chiba, Hideitsu Hino, Shotaro Akaho y Noboru Murata. «Time-Varying Transition Probability Matrix Estimation and Its Application to Brand Share Analysis.» *PloS one* 12 (1 2017).

[40] Minkyoung Choi, Marco Mesa-Frias, Eveline Nüesch, James Hargreaves, David Prieto-Merino, Ann Bowling, G Davey Smith, Shah Ebrahim, Caroline E Dale y Juan P Casas. «Social capital, mortality, cardiovascular events and cancer: a systematic review of prospective studies». *International journal of epidemiology* 43.6 (2014).

[41] Germinal Cocho, Pedro Miramontes, Ricardo Mansilla y Wentian Li. «Bacterial genomes lacking long-range correlations may not be modeled by low-order Markov chains: the role of mixing statistics and frame shift of neighboring genes.» *Computational biology and chemistry* 53 Pt A (dic. de 2014).

[42] D Commenges. «Multi-state models in epidemiology.» *Lifetime data analysis* 5 (4 dic. de 1999).

[43] Richard J. Cook y Jerald F. Lawless. «Statistical Issues in Modeling Chronic Disease in Cohort Studies». 6 (2014).

[44] Harris Cooper, Larry V Hedges y Jeffrey C Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2019.

[45] D. R. Cox y PAWL Lewis. «The statistical analysis of series of events». *The Annals of Mathematical Statistics* (1966).

[46] Caitlin M Cusack, George Hripcsak, Meryl Bloomrosen, S Trent Rosenbloom, Charlotte A Weaver, Adam Wright, David K Vawdrey, Jim Walker y Lena Mamykina. «The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting». *Journal of the American Medical Informatics Association* 20.1 (2013).

[47] Gregory Darnell, Stoyan Georgiev, Sayan Mukherjee y Barbara E Engelhardt. «Adaptive randomized dimension reduction on massive data». *The Journal of Machine Learning Research* 18.1 (2017).

[48] Gipeum Do y Yang-Jin Kim. «Analysis of interval censored competing risk data with missing causes of failure using pseudo values approach». *Journal of Statistical Computation and Simulation* 87.4 (2017).

[49] Justin Doods, Florence Botteri, Martin Dugas y Fleur Fritz. «A European inventory of common electronic health record data elements for clinical trial feasibility». *Trials* 15.1 (2014).

[50] Bradley Efron. «Missing Data, Imputation, and the Bootstrap». *Journal of the American Statistical Association* 89.426 (1994).

[51] S Garrido Elustondo, E García Esquina, I Viúdez Jiménez, C López Gómez, E Más Cebrián y M Ballarín Bardají. «Estudio de la Calidad de Vida Profesional en trabajadores de Atención Primaria del Área 7 de la Comunidad de Madrid». *Revista de calidad asistencial* 25.6 (2010).

[52] JR Emberson, PH Whincup, DA Lawlor, D Montaner y S Ebrahim. «Coronary heart disease prevention in clinical practice: are patients with diabetes special? Evidence from two studies of older men and women». *Heart* 91.4 (2005).

[53] Craig K. Enders, Brian T. Keller y Roy Levy. «A fully conditional specification approach to multilevel imputation of categorical and continuous variables.» 23 (2018).

[54] Craig K. Enders, Stephen A. Mistler y Brian T. Keller. «Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation.» 21 (2015).

[55] Daniel M Faissol, Paul M Griffin y Julie L Swann. «Bias in Markov models of disease.» *Mathematical biosciences* 220 (2 ago. de 2009).

[56] Julian J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Vol. 124. CRC press, 2016.

[57] S. Fieuws, Geert Verbeke y G. Molenberghs. «Random-effects models for multivariate repeated measures.» *Statistical methods in medical research* 16 (5 oct. de 2007).

[58] Jason P. Fine y Robert J. Gray. «A Proportional Hazards Model for the Subdistribution of a Competing Risk». 94 (1999).

[59] Alexis Gabadinho, Gilbert Ritschard, Matthias Studer y Nicolas S Müller. «Mining sequence data in R with the TraMineR package: A user's guide». *Geneva: Department of Econometrics and Laboratory of Demography, University of Geneva* (2009).

[60] Florence Gillaizeau, Etienne Dantan, Magali Giral y Yohann Foucher. «A multistate additive relative survival semi-Markov model.» *Statistical methods in medical research* 26 (4 ago. de 2017).

[61] Robert J. Glynn, Nan M. Laird y Donald B. Rubin. *Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse*. 1986.

[62] Els Goetghebeur y Louise Ryan. «Analysis of competing risks survival data when some failure types are missing». *Biometrika* 82.4 (1995).

[63] Saveli I Goldberg, Andrzej Niemierko y Alexander Turchin. «Analysis of data errors in clinical research databases». *AMIA annual symposium proceedings*. Vol. 2008. American Medical Informatics Association. 2008.

[64] Harvey Goldstein, James Carpenter, Michael G Kenward y Kate A Levin. «Multilevel models with multivariate mixed response types». *Statistical Modelling* 9.3 (2009).

[65] Guangbao Guo. «Taylor quasi-likelihood for limited generalized linear models». *Journal of Applied Statistics* (2020).

[66] Brendan Halpin. «Multiple imputation for categorical time series». *The Stata Journal* 16.3 (2016).

[67] Brendan Halpin. «Multiple imputation for life-course sequence data» (2012).

[68] Seungbong Han, Adin-Cristian Andrei y Kam-Wah Tsui. «Multiple imputation for competing risks survival data via pseudo-observations». 25 (2018).

[69] Panteha Hayati Rezvan, Katherine J Lee y Julie A Simpson. «The rise of multiple imputation: a review of the reporting and implementation of the method in medical research.» *BMC medical research methodology* 15 (abr. de 2015).

[70] James J Heckman. «The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models». *Annals of economic and social measurement, volume 5, number 4*. NBER, 1976.

[71] Donald Hedeker. «A Mixed-Effects Multinomial Logistic Regression Model». *Statistics in medicine* 22.9 (2003).

[72] David W. Hosmer Jr, Stanley Lemeshow y Rodney X. Sturdivant. *Applied Logistic Regression*. Vol. 398. John Wiley & Sons, 2013.

[73] George Hripcsak, David K Vawdrey, Matthew R Fred y Susan B Bostwick. «Use of electronic clinical documentation: time spent and team interactions». *Journal of the American Medical Informatics Association* 18.2 (2011).

[74] Chiung-Yu Huang, Mei-Cheng Wang y Ying Zhang. «Analysing panel count data with informative observation times». 93 (2006).

[75] Qi Huang, Dwayne Cohen, Sandra Komarzynski, Xiao-Mei Li, Pasquale Innominato, Francis Lévi y Bärbel Finkenstädt. «Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data.» *Journal of the Royal Society, Interface* 15 (139 feb. de 2018).

[76] Zhengxing Huang, Wei Dong, Fei Wang y Huilong Duan. «Medical Inpatient Journey Modeling and Clustering: A Bayesian Hidden Markov Model Based Approach.» *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2015 (2015).

[77] Md Hamidul Huque, John B. Carlin, Julie A. Simpson y Katherine J. Lee. «A comparison of multiple imputation methods for missing data in longitudinal studies.» *BMC medical research methodology* 18 (1 dic. de 2018). epublish.

[78] Lurdes Y T Inoue, Ruth Etzioni, Christopher Morrell y Peter Müller. «Modeling disease progression with longitudinal markers». *Journal of the American Statistical Association* 103.481 (2008).

[79] Amalia Karahalios, Laura Baglietto, John B Carlin, Dallas R English y Julie A Simpson. «A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures.» *BMC medical research methodology* 12 (jul. de 2012).

[80] Richard Kay. «A Markov model for analysing cancer markers and disease states in survival studies». *Biometrics* (1986).

[81] Lois G Kim, Claire Carson, Debbie A Lawlor y Shah Ebrahim. «Geographical variation in cardiovascular incidence: results from the British Women's Heart and Health Study». *BMC public health* 10.1 (2010).

[82] Debasis Kundu, Debanjan Mitra y Ayon Ganguly. «Analysis of left truncated and right censored competing risks data». *Computational Statistics & Data Analysis* 108 (2017).

[83] Kajsa Kvist, Per Kragh Andersen, Jules Angst y Lars Vedel Kessing. «Event dependent sampling of recurrent events.» *Lifetime data analysis* 16 (4 oct. de 2010).

[84] Joseph L. y Recai M. Yucel. «Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values». *Journal of Computational and Graphical Statistics* 11.2 (2002).

[85] N. M. Laird. «Missing data in longitudinal studies.» *Statistics in medicine* 7 (1-2 1988). ppublish.

[86] Jane M. Lange y Vladimir N. Minin. «Fitting and interpreting continuous-time latent Markov models for panel data». 32 (2013).

[87] Bryan Lau y Catherine Lesko. «Missingness in the Setting of Competing Risks: from Missing Values to Missing Potential Outcomes». 5 (2018).

[88] DA Lawlor, C Bedford, M Taylor y S Ebrahim. «Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study». *Journal of Epidemiology & Community Health* 57.2 (2003).

[89] K. J. Lee y J. B. Carlin. «Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation». 171 (2010).

[90] Keunbaik Lee y Michael J. Daniels. «A Class of Markov Models for Longitudinal Ordinal Data». 63 (2007).

[91] Timothy S Lesar, Laurie Briceland y Daniel S Stein. «Factors related to errors in medication prescribing». *Jama* 277.4 (1997).

[92] Jianing Li, Thomas H. Scheike y Mei-Jie Zhang. «Checking Fine and Gray subdistribution hazards model with cumulative sums of residuals». 21 (2015).

[93] Li-An Lin, Sheng Luo, Bingshu E Chen y Barry R Davis. «Bayesian analysis of multi-type recurrent events and dependent termination with nonparametric covariate functions.» *Statistical methods in medical research* 26 (6 dic. de 2017).

[94] D. Lin. «Nonparametric estimation of the gap time distribution for serial events with censored data». 86 (1999).

[95] Roderick J. A. Little y Donald B. Rubin. *Statistical Analysis with Missing Data. Little/Statistical Analysis with Missing Data*. 2002.

[96] Brent R. Logan, Mei-Jie Zhang y John P. Klein. «Marginal Models for Clustered Time-to-Event Data with Competing Risks Using Pseudovalues». *Biometrics* 67.1 (2011).

[97] Wenjie Lou. «MULTI-STATE MODELS WITH MISSING COVARIATES». *Theses and Dissertations–Statistics. 16.* (2016).

[98] Junsheng Ma, Wenyaw Chan, Chu-Lin Tsai, Momiao Xiong y Barbara C Tilley. «Analysis of transtheoretical model of health behavioral changes in a nutrition intervention study–a continuous time Markov chain model with Bayesian approach.» *Statistics in medicine* 34 (27 nov. de 2015).

[99] David Macleod. «Differential networks (and other statistical issues) for the analysis of metabolomic data». Tesis doct. London School of Hygiene & Tropical Medicine, 2017.

[100] Jonathan I Maletic y Andrian Marcus. «Data cleansing: A prelude to knowledge discovery». *Data mining and knowledge discovery handbook*. Springer, 2009.

[101] Jonathan I Maletic y Andrian Marcus. «Data Cleansing: Beyond Integrity Analysis.» *Iq*. Citeseer. 2000.

[102] Emilio A Martínez Marco y Jesús Aranaz Andrés. «¿Existe relación entre el reingreso hospitalario y la calidad asistencial?» *Revista de Calidad Asistencial* 17.2 (2002).

[103] Margaret May, Debbie A. Lawlor, Peter Brindle, Rita Patel y Shah Ebrahim. «Cardiovascular disease risk assessment in older women: can we improve on

Framingham? British Women's Heart and Health prospective cohort study». *Heart* 92.10 (2006).

[104] Yassin Mazroui, Simone Mathoulin-Pélissier, Gaetan Macgrogan, Véronique Brouste y Virginie Rondeau. «Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data.» *Biometrical journal. Biometrische Zeitschrift* 55 (6 nov. de 2013).

[105] José M Mena Mateo, Luis Sánchez Perruca, Juan Cárdenas Valladolid y col. «Implantación y evaluación informática de un programa de mejora de la calidad asistencial en el Área 4 de Atención Primaria de Madrid». *Rev. calid. asist* (2006).

[106] Julia Mikolai y Mark Lyons-Amos. «Longitudinal methods for life course research: A comparison of sequence analysis, latent class growth models, and multi-state event history models for studying partnership transitions». *Longitudinal and Life Course Studies* (2017).

[107] Ulla B Mogensen y Thomas A Gerds. «A random forest approach for competing risks based on pseudo-values». *Statistics in medicine* 32.18 (2013).

[108] Silvia Montoro-García, María Pilar Zafrilla-Rentero, Francisco Miguel Celdrán-de Haro, Juan José Piñero-de Armas, Fidel Toldrá, Luis Tejada-Portero y José Abellán-Alemán. «Effects of dry-cured ham rich in bioactive peptides on cardiovascular health: A randomized controlled trial». *Journal of Functional Foods* 38 (2017).

[109] Margarita Moreno-Betancur y Aurélien Latouche. «Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values». *Statistics in Medicine* 32.18 (2013).

[110] Lawrence H Moulton y Mlchael J Dibley. «Multivariate time-to-event models for studies of recurrent childhood diseases.» *International journal of epidemiology* 26.6 (1997).

[111] Shinichi Nakagawa y Holger Schielzeth. «A general and simple method for obtaining R2 from generalized linear mixed-effects models». *Methods in ecology and evolution* 4.2 (2013).

[112] Melanie Nichols, Nick Townsend, Peter Scarborough y Mike Rayner. «Trends in age-specific coronary heart disease mortality in the European Union over three decades: 1980–2009». *European Heart Journal* 34.39 (14 de oct. de 2013).

[113] João Nicolau. «A New Model for Multivariate Markov Chains. Multivariate Markov chains». 41 (2014).

[114] Eveline Nüesch, Perel Pablo, Caroline E Dale, David Prieto-Merino, Meena Kumari, Ann Bowling, Shah Ebrahim y Juan P Casas. «Incident disability in older adults: prediction models based on two British prospective cohort studies.» *Age and ageing* 44 (2 mar. de 2015).

[115] Elena Olariu, Kevin K Cadwell, Elizabeth Hancock, David Trueman y Helene Chevrou-Severac. «Current recommendations on the estimation of transition probabilities in Markov cohort models for use in health care decision-making: a targeted literature review.» *ClinicoEconomics and outcomes research : CEOR* 9 (2017).

[116] Andrés Parrilla-Almansa, Carlos Alberto González-Bermúdez, Silvia Sánchez-Sánchez, Luis Meseguer-Olmo, Carlos Manuel Martínez-Cáceres, Francisco Martínez-Martínez, José Luis Calvo-Guirado, Juan José Piñero de Armas, Juan Manuel Aragoneses, Nuria García-Carrillo y col. «Intraosteal Behavior of Porous Scaffolds: The mCT Raw-Data Analysis as a Tool for Better Understanding». *Symmetry* 11.4 (2019).

[117] Neil Pearce y Harvey Checkoway. «A simple computer program for generating person-time data in cohort studies involving time-related factors». *American journal of epidemiology* 125.6 (1987).

[118] Edsel A Peña. «Dynamic Modelling and Statistical Analysis of Event Times.» *Statistical science : a review journal of the Institute of Mathematical Statistics* 21 (4 nov. de 2006).

[119] Edsel A. Peña y Myles Hollander. *Models for Recurrent Events in Reliability and Survival Analysis*. 2004.

[120] Tim Pfeiffer, Nicolai Heinze, Robert Frysch, Leon Y Deouell, Mircea A Schoenfeld, Robert T Knight y Georg Rose. «Extracting duration information in a picture category decoding task using hidden Markov Models.» *Journal of neural engineering* 13 (2 abr. de 2016).

[121] Matthew Powney, Paula Williamson, Jamie Kirkham y Ruwanthi Kolamunnage-Dona. «A review of the handling of missing longitudinal outcome data in clinical trials.» *Trials* 15 (jun. de 2014).

[122] Ali Rafei, Einollah Pasha y Roohangiz Jamshidi Orak. «A warning threshold for monitoring tuberculosis surveillance data: an alternative to hidden Markov model.» *Tropical medicine & international health : TM & IH* 20 (7 jul. de 2015).

[123] Adrian E Raftery. «A model for high-order Markov chains». *Journal of the Royal Statistical Society. Series B (Methodological)* (1985).

[124] Erhard Rahm y Hong Hai Do. «Data cleaning: Problems and current approaches». *IEEE Data Eng. Bull.* 23.4 (2000).

[125] Stephen L Rathbun y Saul Shiffman. «Mixed effects models for recurrent events data with partially observed time-varying covariates: Ecological momentary assessment of smoking.» *Biometrics* 72 (1 mar. de 2016).

[126] Patrick Rockenschaub, Vincent Nguyen, Robert W Aldridge, Dionisio Acosta, Juan Miguel García-Gómez y Carlos Sáez. «Data-driven discovery of changes in clinical code usage over time: a case-study on changes in cardiovascular disease recording in two English electronic health records databases (2001–2015)». *BMJ open* 10.2 (2020).

[127] Bart Van Rompaye, Shabbar Jaffar y Els Goetghebeur. «Estimation With Cox Models. Cause-specific Survival Analysis With Misclassified Cause of Failure». 23 (2012).

[128] Patrick Royston. «Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables». *The Stata Journal* 9.3 (2009).

[129] Gordon D Rubenfeld. «Using computerized medical databases to measure and to improve the quality of intensive care». *Journal of critical care* 19.4 (2004).

[130] R Sáchez González, Rocio Álvarez Nido y Soledad Lorenzo Borda. «Calidad de vida profesional de los trabajadores de atención primaria del Área 10 de Madrid». *Medifam* 13.4 (2003).

[131] Issaka Sagara, Roch Giorgi, Ogobara K Doumbo, Renaud Piarroux y Jean Gaudart. «Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data.» *Malaria journal* 13 (jul. de 2014).

[132] Ankan Saha y Dhruv Arya. «Generalized mixed effect models for personalizing job search». *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2017.

[133] JL Schafer. «Multiple Imputation, o prime Statistical Methods in Medical Research 8» (1999).

[134] Jürg Schelldorfer, Peter Bühlmann y SARA VAN DE GEER. «Estimation for high-dimensional linear mixed-effects models using ℓ1-penalization». *Scandinavian Journal of Statistics* 38.2 (2011).

[135] Tunny Sebastian, Visalakshi Jeyaseelan, Lakshmanan Jeyaseelan, Shalini Anandan, Sebastian George y Shrikant I Bangdiwala. «Decoding and modelling of time series count data using Poisson hidden Markov model and Markov ordinal logistic regression models.» *Statistical methods in medical research* (ene. de 2018).

[136] Secretaría General del Servicio Madrileño de Salud. *Memoria anual de actividad de la gerencia asistencial de atención primaria año 2018*. Ed. por Servicio Madrileño de Salud. Secretaría General del Servicio Madrileño de Salud. Sep. de 2019. URL: http://www.madrid.org/bvirtual/BVCM020342.pdf.

[137] Secretaría General del Servicio Madrileño de Salud. *Memoria anual de actividad del servicio madrileño de salud año 2018*. Ed. por Servicio Madrileño de Salud.

Secretaría General del Servicio Madrileño de Salud. Mayo de 2019. URL: http://www.madrid.org/bvirtual/BVCM020283.pdf.

[138] Anoop D. Shah, Jonathan W. Bartlett, James Carpenter, Owen Nicholas y Harry Hemingway. «Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study.» *American journal of epidemiology* 179 (6 mar. de 2014). ppublish.

[139] Gowri Shivasabesan, Biswadev Mitra y Gerard M O'Reilly. «Missing data in trauma registries: A systematic review.» *Injury* (mar. de 2018).

[140] Gunnar Steineck, Philip H Kass y Anders Ahlbom. «A comprehensive clinical epidemiological theory based on the concept of the source person-time and four distinct study stages». *Acta Oncologica* 37.1 (1998).

[141] Jonathan AC Sterner, Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood y James R. Carpenter. «Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls». *BMJ. British medical journal* 339.7713 (2009).

[142] Sundarraman Subramanian. «Multiple imputations and the missing censoring indicator model». 102 (2011).

[143] M J Sweeting, V T Farewell y D De Angelis. «Multi-state Markov models for disease progression in the presence of informative examination times: an application to hepatitis C.» *Statistics in medicine* 29 (11 mayo de 2010).

[144] Zilong Tan. «Approximate Inference for High-Dimensional Latent Variable Models». Tesis doct. Duke University, 2018.

[145] Zilong Tan, Kimberly Roche, Xiang Zhou y Sayan Mukherjee. «Scalable algorithms for learning high-dimensional linear mixed models». *arXiv preprint arXiv:1803.04431* (2018).

[146] Howard H Z Thom, Christopher H Jackson, Daniel Commenges y Linda D Sharples. «State selection in Markov models for panel data with application to psoriatic arthritis.» *Statistics in medicine* 34 (16 jul. de 2015).

[147] Simon Thompson, Stephen Kaptoge, Ian White, Angela Wood, Philip Perry, John Danesh y Emerging Risk Factors Collaboration. «Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies». *International journal of epidemiology* 39.5 (2010).

[148] Andrew C Titman y Linda D Sharples. «Semi-Markov models with phase-type sojourn distributions.» *Biometrics* 66 (3 sep. de 2010).

[149] Beatriz Valcárcel, Peter Würtz, Nafisa-Katrin Seich al Basatena, Taru Tukiainen, Antti J Kangas, Pasi Soininen, Marjo-Riitta Järvelin, Mika Ala-Korpela, Timothy M Ebbels y Maria de Iorio. «A differential network approach to exploring differences between biological states: an application to prediabetes». *PLoS One* 6.9 (2011).

[150] Wolfgang Viechtbauer. «Conducting meta-analyses in R with the metafor package». *Journal of Statistical Software* 36.3 (2010).

[151] Qihua Wang y Gregg E. Dinse. «Linear regression analysis of survival data with missing censoring indicators». 17 (2011).

[152] Qihua Wang, Gregg E. Dinse y Chunling Liu. «Hazard function estimation with cause-of-death data missing at random». 64 (2012).

[153] Stanley Wasserman y Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

[154] Andrew Whalen, Gregor Gorjanc, Roger Ros-Freixedes y John M Hickey. «Assessment of the performance of hidden Markov models for imputation in animal breeding.» *Genetics, selection, evolution : GSE* 50 (1 sep. de 2018).

[155] Jesse O Wrenn, Daniel M Stein, Suzanne Bakken y Peter D Stetson. «Quantifying clinical narrative redundancy in an electronic health record». *Journal of the American Medical Informatics Association* 17.1 (2010).

[156] Amy Ming-Fang Yen y Hsiu-Hsi Chen. «Bayesian measurement-error-driven hidden Markov regression model for calibrating the effect of covariates on multistate outcomes: Application to androgenetic alopecia.» *Statistics in medicine* 37 (21 sep. de 2018).

[157] T Kue Young. *Population health: concepts and methods*. Oxford University Press, USA, 2005.

[158] Ali Zare, Mahmood Mahmoodi, Kazem Mohammad, Hojjat Zeraati, Mostafa Hosseini y Kourosh Holakouie Naieni. «Assessing misdiagnosis of relapse in patients with gastric cancer in Iran cancer institute based on a hidden Markov multi-state model.» *Asian Pacific journal of cancer prevention : APJCP* 15 (9 2014).

[159] Ningshan Zhang, Kyle Schmaus, Patrick O Perry y col. «Fitting a deeply nested hierarchical model to a large book review dataset using a moment-based estimator». *The Annals of Applied Statistics* 13.4 (2019).

[160] Ting Zhang y Wei Biao Wu. «Inference of time-varying regression models». 40 (2012).

[161] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang y Deepak Agarwal. «Glmix: Generalized linear mixed models for large-scale response prediction». *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

[162] Yue Zhao, Amy H. Herring, Haibo Zhou, Mirza W. Ali y Gary G. Koch. «A Multiple Imputation Method for Sensitivity Analyses of Time-to-Event Data with Possibly Informative Censoring». 24 (2014).

[163] Xiang Zhou. «A unified framework for variance component estimation with summary statistics in genome-wide association studies». *The annals of applied statistics* 11.4 (2017).

[164] Eric R. Ziegel, P. Diggle, K. Liang y S. Zeger. «Analysis of Longitudinal Data». 37 (1995).

Capítulo 8

APÉNDICES

Apéndice A

MENCIÓN INTERNACIONAL EN EL TÍTULO DE DOCTOR

Con el objetivo de cumplir con el requisitos especificados en el Real Decreto 99/2011 para la obtención de la Mención Internacional en el Título de Doctor, a la que opta la presente tesis doctoral, a continuación se incluye un resumen de la tesis en inglés.

A.1 INTRODUCTION

The ongoing increase in the amount of information and the number of variables collected every day by the biomedical activity has made it possible to create advanced models and offer a huge potential to boost medical and epidemiological research and enables discoveries that were impossible before, but its complexity and changing structure come with a set of challenging problems.

One of them is the occurrence of *missing data*, registers that do not contain valid values, that seriously hinders its computer processing and can skew statistical analyses and reduce their accuracy. Imputation methods make it possible to fill these gaps with reasonable estimates that partially correct the problem but rely on using proper models. In order to use them, it is necessary to know the type of *missing data*: related to the value of the variable,

conditioned to other variables or completely at random [16, 141, 85].

In order to carry out any study it is crucial to use information that has been properly collected and structured and to ensure its quality, [101, 100, 124, 126]. The presence of errors in medical databases may be partially related to the quality of care, to the time doctors spend on administrative tasks and to the subsequent processing of the information, [129, 4, 73, 155, 46, 130, 102, 11, 105, 3, 51, 130].

The analysis of several clinical research databases reported error rates from 2.3 to 26.9%, [63]. Many of those errors were not random and could seriously bias the results of studies performed using these data. The [91] study reported a rate of 4 errors per 1000 prescriptions.

In the first part of this thesis we analyzed which factors related to the patient or medical center are related with the occurrence of *missing data* (MAR) in other variables. In order to accomplish this, logistic regression models with random effects will be used, which will be applied to the Area-7 public health service database of Madrid, which contains information about 1400 variables collected over five years from nearly a quarter of a million people..

Logistic regression is used to model categorical variables, as in our case, the presence or absence of a value [72]. Random effects models, on the other hand, can be used to model variables that have been measured repeatedly over time for the same person or place [13, 14, 164, 131] and [57, 71, 160, 19, 56, 13].

In order to properly combine the results obtained in different studies, it is common to use meta-analysis techniques. We will use these techniques to improve the accuracy of the coefficients obtained by adjusting the data partitioned into small blocks, and thus be able to carry out analyses when the data do not fit in the computer memory [23, 150, 44].

Another common problem in medical studies is the analysis of sequences of medical events. In the second part of this thesis we explore different methods for the analysis and graphical represention of these sequences using the information collected in medical databases.

The simplest approaches to analysing sequences just modelize the time to the first occurrence of an event ignoring subsequent repetitions or counts the events observed within

a given time period usually ignoring its subsequent repetitions and assuming they are independent and follow a Poisson distribution [45, 5, 147, 34, 74, 90, 94, 131].

In order to take into account the occurrence of different event types Markov models are often used, which represent a finite number of discrete and mutually exclusive health states connected by transitions that represent the probability of a patient moving from one state to another, but this models ignore the effect of non-consecutive events [21, 80, 86, 7, 10, 35, 60, 148, 39].

In the presence of competitive risks, simple survival methods are biased, [17, 24], and complex Bayesian and semi-parametric marginal models with the cumulative incidence function and the hazard of sub-distribution are used instead, but their assumptions must be carefully tested and are not always satisfied, [58, 92, 37, 33, 38, 93, 104, 22].

There are higher order Markov models in which the probability of a state depends simultaneously on several previous consecutive states, but parameter estimates are less reliable [41, 113, 123]. Markov models have also been used in combination with random effects and Bayesian analysis [98, 125].

Another commonly used alternative are the Hidden Markov Models (HMM) [18, 76, 75, 120, 122, 135, 143, 146, 154, 156, 158, 20, 9, 78, 118], which assume that discreetly observed events are characterized by an underlying latent variable with unobserved, hidden states that map to each disease state [42, 115, 28, 110, 55, 83, 106, 43] make an exhaustive review of the different Markov models used in health studies and analyze the problems that may arise from them.

A more general approach is needed to simultaneously incorporate repetition, competition, explicit time dependence, covariates and the effect of non-consecutive 'distant' events [119]. We propose a new method to studying sequences able to deal with multiple types of recurrent events and taking into account the time at risk. We show that considering both consecutive and non-consecutive events in our analyses not only produce different IR values but also allows us to reveal transitions that otherwise went unnoticed.

We have applied this methodology to study the events collected in the database of the Area-7 public health service of Madrid. For most variables in this database the event date

is unknown because only the year has been recorded. Therefore, we can not recover the exact order in which the inter-annual events took place. We propose to apply *missing data* imputation methods to deal with date truncation.

Multiple imputation methods(MI), [50, 2, 95, 133, 32, 66, 16, 141, 29, 79, 139, 69], usually used with cross-sectional data, can also be applied to longitudinal data collected at regular intervals if we consider repeated measurements as different variables, but suffer from convergence problems and collinearity if missingness depends on the value of the variable itself. Furthermore, they need to be extended to be used with irregular time intervals or competitive risks, [58, 87, 62, 59, 127] with joint multivariate distributions and random effects [84], using Random Forest methods [30, 68, 107, 96], or by using more complex models [67, 66, 133, 89, 54, 32, 53].

Censoring [48, 109, 81] and truncation in longitudinal data [142, 151, 152, 31] are usually addressed by means of selection models [70] and pattern mixing [61] or Multiple Imputation [162]. Competitive risks with censorship and truncation are studied in [82] with latent failure time models, assuming Weibull distributions and using bootstrap and Monte Carlo to calculate maximum likelihood estimators. The same problem is studied in [25] using non-parametric estimates of the joint distribution of truncation and censoring times and maximum likelihood to model the survival curve, concluding that this approach, in general, may not be valid.

The imputation of longitudinal data is also treated by multi-state models with time-dependent transition rates and interactions [36, 97], and by hidden Markov models for categorical data [26].

These imputation methods are based on hard-to-verify assumptions, computationally intensive and not easily applicable to large databases like ours. Most of the studies we can find in the literature working with real data perform the analyses using only information from the pairs of consecutive events, thus they are not useful to study longer sequences taking into account the influence of more distant events, which is one of our objectives.

Our approach will impute truncated dates by generating all possible permutations of the inter-annual events for each person and creating new sequences that could happen, compatible with the original data. We will test the validity of this approach by comparing

the results calculated from complete dates with those calculated with manually truncated dates and using different imputation methods. Then we will show the results in tables and graphs by means of connection networks [15], which allow us to quickly visualize sequences of events of any length [153, 1].

THE AIMS OF THIS THESIS ARE:

A) To identify what patient or medical center variables are associated with the occurrence of missing data in other variables in medical databases.
   (1) To explore different tools and methods to calculate regression models when the size of the data is too large for the computer's memory.
   (2) To identify which factors related to the patient or the medical center are associated with the occurrence of missing data in the database of the public health service of the Area-7 in Madrid.

B) To study the sequences of events collected in large medical databases, even when the time information is incomplete
   (3) To develop a simple method to analyze sequences of medical events in the presence of recurrence and competition.
   (4) To adapt the methodology developed for the study of sequences to be able to analyse variables with incomplete time information.
   (5) To explore efficient ways to visualize the sequences of events studied.
   (6) To analyse the sequence of events in the Area-7 database.

A.2 METHODS

A.2.1 IDENTIFICATION OF FACTORS ASSOCIATED WITH THE OCCURRENCE OF MISSING DATA IN OTHER VARIABLES

Before proceeding with the analysis itself, it has been necessary to carry out a complex and time-consuming cleansing process of the data, which contained multitude of errors. This step was required to explore and select what variables would be used in our models. We also studied the variables to discern real missing data from those cases that simply do not make sense and do not have to be taken into account, such as the the number of cigarettes that a non-smoking person smokes. The data was reshaped from wide to long format, piling up different year's measurements of the same variables. Eventually obtaining a new database of

522 variables and $1\,345\,917$ rows.

The variables finally chosen to study the occurrence of missing data were PESO (Weight) and CIGARRILLOS (Number of Cigarettes). And the covariates included in our models were: YEAR, EAP (medical center code), PAMED (doctor workload pressure, patients/day), PAENF (nursing workload pressure, patients/day) POSMAS (percentage of patients >= 65 years old), SEXO (gender), and EXTRANJERO (Foreigner).

We built mixed multilevel logistic models to analyse the factors determinant of missingness in other variables of interest in the database. The data have been fitted with the logistic model $mPESO \sim SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + (1|EAP/ID)$ and also with other alternatives, including interactions and additional variables.

The complexity and huge amount of data made necessary to explore three different approaches to deal with this problem:
1) Analysing simultaneously the complete dataset.
2) Doing separate analyses in each medical center and pulling the results with a meta-analysis.
3) Doing separate analyses partitions randomly generated from the database and pulling the results with a meta-analysis.

The calculations were performed using the program R v3.3, and the libraries data.table, lme4, glmmTMB and metafor. In addition, it has been necessary to force R to use the SSD disk as if it were RAM. Predictor variables have been standardized to avoid convergence errors.

A.2.2 ANALYSIS OF SEQUENCES OF EVENTS

The study of the sequences of events was carried out using again the database of the Area-7 public health service of Madrid. This dabase contains 1401 different variables collected from $224\,321$ individuals, who have been tracked from 2006 to 2010. The data have had to be transformed from Wide to Long format and contain two types of variables: i) variables recorded with exact dates "dd/mm/yy", and ii) variables recorded with truncated dates, just the year. The variables we are going to analyze are *ACV, ALCOHOLISMO, ANEMIA, ARRITMIAS, ATERO_PERIFERICA, CARDIO_ISQUEMICA, COLESTEROL, DIABETES, FALLECIMIENTO, GRIPE, HTA, INSUFICIENCIA_CARDIACA, INSUF_RENAL, OBESIDAD,*

RETINOPATIA_DIABETICA, TABAQUISMO, TRASTLIPIDOS, VALVULOPATIAS.

We studied the sequences of events, focusing the analysis on the shorter subsequences: pairs and triplets. In order to quantify how many people have been diagnosed at least once from a given sequence and also take into account the time elapsed between the events we used the Incidence Rate (IR), [157], number of occurrences divided by the time of exposure. For our purposes we define it as the "number of people that underwent a given sequence of events at least once, divided by the exposure time from the penultimate event to the last one or until the study's end or decease" [117, 140]. We performed the calculation through two different procedures: by considering only the subsequences of consecutive events (Section 3.2.3) and by considering any subsequence of events whether or not consecutive. The exact procedure is shown with an example at Sections 3.2.2 to 3.2.4 and the code used to perform the computation at Listings F.2, F.3 and F.6.

The date produces an uncertainty in the exact order in which inter-annual events take place, making necessary to develop an imputation procedure, Section 3.2.5. We replaced the sequences with partially unordered events by the weighted sum of a compatible set of fully ordered sequences obtained by permuting the intra-annual data, which we assigned a certain probability of occurrence. Even though we do not know the exact position of each event within a year, we can calculate approximately the time intervals between events occurring in different years.

The new sequences were weighted in two different ways: a) with equal probability of having actually occurred. b) using the frequencies previously calculated from the inter-annual events. Section 3.2.5 We simplified the computation by combining this brute-froce procedure with pre-calculated theoretical formulas, Section 3.2.5.1.

Then we graphically displayed these results by means of a diagram of interconnected nodes, known as connection network, in which each node represents an event type, and the thickness of the connections is calculated according to an established criterion, in our case the IR value. The analysis has been performed using the program R v3.4.1 and the libraries Diagrammer, VisNetwork, data.table, Lubridate and ggplot2.

A.3 RESULTS AND DISCUSSION

A.3.1 IDENTIFICATION OF FACTORS ASSOCIATED WITH THE OCCURRENCE OF MISSING DATA IN OTHER VARIABLES

The analyses carried out with the Area-7 database showed that the amount of memory needed to adjust regression models grows rapidly with the size of the data, Table 4.2. The models that use the least memory are the model with random effects for centers and the model with simple fixed effects. The model that consumes the most memory is the model with random effects for people not nested in centers.

We also found that the lme4 library, considered the standard to fit models with random-effects, uses huge amounts of memory, even making it unfeasible to analyse the whole dataset simultaneously, the program simply crashes with more than $50\,000$ rows. It was therefore necessary to look for alternatives, eventually finding the experimental library glmmTMB.

In Table 4.3 we see how the time required to compute the models grows exponentially with the number of rows analysed. In this case we see how the fastest model is the simple model, followed by the models with random effects for centers, then the models with random effects for people nested in centers, and finally, the models with random effects for people but without nesting them in centers, which turns out to be extremely slow.

The algorithms used to compute random-effects regression models use large amounts of memory and are very slow. Although the use of the glmmTMB library made it possible to calculate the most complex models, it was also necessary to force R to use the SSD unit as if it were RAM memory. But this solution is also limited, not scalable and very slow.

Our findings justify the need to explore new methods for analysing the data. In this thesis we proposed to split them in two different ways: 1) by medical centers, and 2) randomly. And then to carry out the analyses independently in each block of data, subsequently pullying and unifying the results by means of a meta-analysis. In the Tables 4.2 and 4.3 we have also included the results of the analysis of partitioned (in 19 blocks) data and see that this procedure divides by 5 the need for memory and by 15 the computation time.

The Section 4.1.2 shows in detail the results obtained by analysing the complete dataset.

The random-effects model for individuals nested in centers, Table 4.7, shows that the odds ratio of missing data in PESO is 0.8089 for the covariate SEXO (versus men), 2.184 for EXTRANJERO (over nationals), 0.932 for EDAD (per year), 0.854 for PAMED, 1.178 for PAENF and 1.615 for YEAR (per year). All these results have ($p < 10^{-16}$). We use as reference levels YEAR 2006, EDAD 50, PESO 70, PAMED 32, PAENF 18 and POSMAS 0.

We get almost the same results, with differences smaller than 2%, with the random-effects model with ID not nested to centers, and the p-values are almost identical too. The random-effects model that includes the POSMAS variable and the interactions of PAMED and PAENF with YEAR shows similar results and indicates that the effect of PAMED decreases slightly over the years, while PAENF hardly changes. The odds ratio of missing data in PESO is 1.27 for the covariate POSMAS (per unit increment), the percentage of patients with age greater than or equal to 65 years in that medical center.

The different models without the ID factor yield results very similar to each other, up to the fourth digit, irrespective to the exact variables and interactions included, but different from those we get when also including the ID. The inclusion of random effects for ID is equivalent to considering that the occurrence of missings is related to factors intrinsic to the individual not already included explicitly in the model, such as his kinship with the doctor, physical attractiveness, religion, some biological variable or that some patients ask the doctor to measure more variables or refuse to offer information. The effect of all these factors is captured in ID, while in non ID models their effect is absorbed by the other variables and the error.

In Table 4.24 we summarise the results obtained for PESO using the different models when analysing the complete set of data.

For the variable CIGARRILLOS, again with the full dataset, the random-effects model for ID nested in centers, Table 4.11, shows that the odds ratio of missing data in CIGARRILLOS is 1.013 ($p = 0.057$) for the covariate EDAD (per year), 1.082 for PAMED ($p = 0.00090$), 0.836 ($p = 2.3 * 10^{-5}$) for PAENF y 0.144 ($p < 10^{-16}$) for YEAR (per year). The model with additional variables, Table D.5, shows that the odds ratio of missing data in CIGARRILLOS is 1.013 for the covariate PESO ($p = 0.074$), and we see that the variables SEXO ($p = 0.18$) and EXTRANJERO ($p = 0.30$) are not significant.

The model with individuals not nested in centers produces similar results for the variable YEAR, and different results but at least of the same sign for the variables EDAD and PAENF. The results for PAMED are of opposite sign. Again, we see that models not taking into account the ID factor give different results from the previous ones but very similar to each other. In the Table 4.24 we summarize the results obtained for CIGARRILLOS using the different models when analysing the complete dataset.

The Section 4.1.3 shows in detail the results obtained by independently analysing the data in each random partition, both for PESO, Tables D.9, D.17, D.25 and D.33, and for CIGARRILLOS, Tables D.41, D.47, D.53 and D.59. The results obtained in all the partitions are combined by using meta-analyses, Table D.10 and onwards in Appendix D, which make it possible to estimate the coefficients of each variable for the dataset. This process was repeated with different models. In addition, we have graphically displayed the meta-analyses of each variable using forest plots, Figure D.1 and onwards. We see that the variability of the computed coefficients within each partition is similar to the variability between their means.

The Section 4.1.4 shows in detail the results obtained by independently analysing the data from each medical center, both for PESO, Tables D.65 and D.80, and CIGARRILLOS, Tables D.95 and D.101. The results obtained in all centers are combined with a meta-analysis, Table D.66, Figure D.49 and onwards. In this case we can see that the variability within each center is very small compared to the variability between the means of the different centers, there is a significant difference between the coefficients calculated in different centers, especially for the variables YEAR, PAMED, AGE and PAENF.

Given the impossibility of introducing the POSMAS variable in the intra-center adjustments, since in most centers it only acquires one or two different values, we perform a second set of regressions to study whether the coefficients obtained in each intra-center model depend on the inter-center variation of POSMAS. These analyses have not yielded any significant results. Tables D.73 to D.79 and Tables D.88 to D.94.

All these results have been summarized in tables to compare the different models and partitioning methods. We have reported the value of "z" instead of the p-value because in most cases the latter is too small and is automatically rounded to zero, making it impossible to make further comparisons.

For the odds of missing in PESO we can see, Table 4.24, that the coefficients of the covariates yield very similar values, up to the fourth digit, when all the data are analysed together and when analysed independently in each partition. The z's are also almost identical. However, in the analyses partitioned by medical center, the discrepancies are, in general, greater, around 7% for the SEX and AGE covariates, and much greater for the other variables.

For the odds of missing data in CIGARRILLOS we see again, Table 4.25, that the calculations made with the complete set of data offer results similar to those made in each partition separately, but now just up to the second digit. Calculations made partitioning by medical centers do not give such similar values and the z's calculated in this way are less extreme.

A.3.2 ANALYSIS OF SEQUENCES OF EVENTS

Preliminary analysis of the data. In the second part of the thesis we explored new methods for the study and graphic representation of sequences of events and used them to analyse the most common sequences of cardiovascular events in the Madrid Area-7 database.

This database contains 12 variables recorded with complete dates, dd/mm/yy, which hold information on the occurrence of 261 850 events in 128 805 people. Figures 4.1a and 4.1b show the total number of occurrences of each event type and the number of people that have suffered a given number of events. We see that the most frequent event is HTA (72 334), followed by TRASTLIPIDOS (67 837) and OBESITY (33 208).

Figures 4.2a, 4.2b, 4.3a, 4.3b, 4.4a and 4.4b show the average value and distribution of the time intervals between pairs of events on the same individual for consecutive and non-consecutive events according to the initial or final event type. Figure 4.5a shows the risk of suffering each type of event at different ages. Figure 4.5b shows the distribution of the number of people that has suffered each type of event.

The Area-7 database also contains 18 variables recorded with truncated dates, which hold information on the occurrence of 1 442 614 events in 146 662 people. Figures 4.6 and 4.7 show the total number of occurrences of each event type and the number of people that have suffered a given number of events. We had to discard the death variable, collected from only 6831 people, because it contains serious errors: 1871 people have clinical events recorded after

their own death. We see that the most frequent event are HTA (323 391), TRASTLIPIDOS (282 410) and GRIPE (267 799). We can see a progressive decrease in the frequency of events as the number of occurrences increases, some people suffer up to 50 events. In addition, the barplot allows us to discover a pattern, multiple of 5 occurrences are more frequent than others.

Variables with exact dates. Figure 4.8 shows the IR of the different pairs of consecutive event types. In general, the highest IRs correspond to pairs of events in which the final event is TRASTLIPIDOS or HTA. Especially high are the IRs of the ATERO_PER→TRASTLIPIDOS pairs (82.1 occurrences per 1000 person-years), DIABETES→TRASTLIPIDOS (77.0), HTA→TRASTLIPIDOS (75.8), CARDIO_ISQ→HTA (65.9), INSF_CARD→ARRITMIAS (61.8) and OBESIDAD→TRASTLIPIDOS (54.8). The IR is very low in almost all the pairs in which the second event is ATERO_PER or ALCOHOLISMO. These results agree with what is expected from the frequency of single events, Figure 4.1a, at least for the final ones.

The representation of the results by means of connection networks allows us to quickly visualize the most important information and identify patterns, but requires a previous filtering of the most relevant data. If we show the two pairs with the highest IR for each initial node, Figure 4.9 we see that most of the consecutive pairs represented end up in HTA or TRAST_LIPID.

The IR of the different pairs of events, both consecutive and non-consecutive, is shown in Figure 4.11. Again, in general, the highest IRs correspond to the pairs in which the final event is TRASTLIPIDOS or HTA. Particularly high are the IRs of the HTA→TRASTLIPIDOS (77.6), DIABETES→TRASTLIPIDOS (76.7), ATERO_PER→TRASTLIPIDOS (69.8), INSF_CARD→ARRITMIAS (55.2), CARDIO_ISQ→HTA (55.1) and OBESIDAD→TRASTLIPIDOS (54.9). Again, the IR is very low in almost all the pairs in which the second event is ATERO_PER o ALCOHOLISMO. Figures 4.12 and 4.13 displays the IR values by means of connection networks
.

We have also analysed the triplets of events, sequences in which the occurrence of one event is conditioned by two others in a given order. From the number of occurrences of each triplet and the time of exposure to the final event, we have calculated the IR for each triplet, both consecutive and non-consecutive. Table E.3 summarises the results for the 24 triplets

with the highest IR.

Figure 4.14 shows the connection network with the highest IR triplet for each type of final event. We have highlighted the transitions belonging to the same triplet with lines of the same colour. In most cases, the triplets with the highest IRs are those in which the last event is TRASTLIPIDOS, with a maximum value of 114.39 occurrences per 1000 person-years for the triplet DIABETES→DIABETES→TRASTLIPIDOS.

In order to evaluate our methodology in the presence of inexact temporal information, we have performed the analyses again with the same data but now truncating the dates, keeping only the year. In this way, apart from reducing the accuracy, we get another undesirable effect: the uncertainty in the order of the intra-annual events. It is then necessary to use some method of imputation. We propose to convert, for each person, the sequences of events of partially undetermined order into the combination of several possible sequences of totally determined order generated by permuting the intra-annual data. Each of these new sequences has a certain probability of having actually occurred. We perform the calculations in two ways: i) assuming that the probability is equal for all sequences, Figure 4.15a, and ii) assuming that the probability is equal to the estimated inter-annual frequencies, Figure 4.15b.

Both imputation methods produce nearly identical results when applied to our manually truncated data, Figures 4.15a and 4.15b, but the equiprobable imputation method is simpler and faster. On the other hand, some differences can be noticed when comparing the results obtained using exact dates with the those obtained by imputing truncated dates. To quantify these discrepancies, we have calculated the average and quadratic averages of the IR differences produced by the different methods of calculation, 4.2.5.

On average, the IRs calculated by imputing truncated dates are slightly lower (-0.23 occurrences per 1000 person-years) than those calculated with exact dates, although there are important discrepancies in some specific pairs. For example, the IR calculated by imputing the truncated date is approximately 12% lower for pairs whose final event is TRASTLIPIDOS, and 20% higher for pairs whose final event is HTA.

As opposed to the results obtained by studying the full dataset, analysing the information pertaining exclusively to individuals without multiple intra-annual events, Section 4.2.3.1 and Figure 4.16b, shows very little difference between using exact or truncated dates. This

leads us to conclude that most of the discrepancies are not caused by the imprecision derived from truncation but by the inefficiency of the imputation process, which is unable to properly estimate the real order of the sequences.

We have also evaluated how the truncation of the dates affects the analysis of the triplets. On average, it overestimates (0.80 occurrences per 1000 person-years for the simple and 0.39 for the advanced) the IR of the triplets of events with respect to the results obtained when using exact dates, and it produces greater variability (14.6) than when we analysed the pairs of events (5.2), Tables 4.26 and 4.27. This can be explained by the reduced number of occurrences of each triplet, around ten times smaller than the pairs. Again, both imputation methods produce very similar values.

Variables with truncated dates. We studied the "truncated" variables from the Madrid Area-7 database, which only contain the year of occurrence of the events. This dataset has allowed us to include in our analyses new types of events, (ANEMIA, CHOLESTEROL, FLU, INSUF_RENAL and RETINOPATHY_DIABETICS). In addition, it contains 5.5 times more records, which partially compensates for the lower precision in the results caused by truncation. We analysed them using the equiprobable imputation method, in which the same probability of occurrence is assigned to all possible sequences of an individual and from these the pairs and triplets are calculated. Again, the results are similar to those obtained using a priori inter-annual information.

Given the large number of occurrences a person can suffer, the method used in the previous section is not applicable. It has then been necessary to redesign the computation algorithm, managing to reduce the necessary memory by performing the calculations one by one recursively, aborting them on the detection of the first occurrence of each pair or triplet for each individual and using pre-calculated theoretic results.

Most pairs where the *second* event is HTA, GRIPE or TRASTLIPIDOS have high IR values, Figure 4.17. This corresponds to a higher frequency of these three types of events in the database. However, the pairs whose *first* event is HTA, GRIPE or TRASTLIPIDOS, in general, don not have higher IRs. Most pairs where the final event is equal to the initial event have IR values close to one and much higher than the others. The main exception is COLESTEROL→COLESTEROL, whose IR is 112 occurrences per 1000 person-years.

The connections networks built from the IR of the different pairs of events allow us to graphically represent the most significant results. In most cases, the pairs of events with the highest IR have as their final event HTA or GRIPE, Figure 4.18. In most cases, the pairs of events with the highest IR have as their initial event ATERO_PER or INSF_CARD, Figure 4.19.

We have also been able to successfully analyse the triplets of events from the data collected with the truncated date, Section 4.2.4.2. For most of the triplets represented the IR is close to 1000 occurrences per 1000 person-years. The highest IR corresponds to COLESTEROL→COLESTEROL→CARDIO_ISQ, with a value of 1648 occurrences per 1000 person-years. Again, we consider the use of networks to display the results of the different types of triplets of events to be particularly useful, Figure 4.20.

A.4 CONCLUSIONS AND FUTURE WORK

Throughout this thesis we have explored and developed different methods to analyse complex biomedical databases that have allowed us to study two common problems in biostatistics:

A) the identification of factors associated with the occurence of missing data in large databases, and
B) the analysis of sequences of events in the presence of recurrence and competition, even when temporal information is incomplete.

We consider that the objectives proposed for this work have been successfully achieved.

A.4.1 IDENTIFICATION OF FACTORS ASSOCIATED WITH THE OCCURRENCE OF MISSING DATA IN OTHER VARIABLES

1) In order to be able to calculate the statistical models when the size of the data is too big for the available memory we have studied different strategies to partition the analysis. We have split the data into smaller blocks, adjusted the regression models independently in each of them and combined the results using meta-analysis.
   (a) The simultaneous analysis of the entire dataset produces the most accurate results but is extremely slow and not scalable.

(b) The analysis of randomly partitioned blocks of data is considerably faster, scalable and offers very similar result to those obtained in the analysis of the complete dataset.

(c) The analysis of the data separately at each center is fast but less accurate. For some variables the results differ considerably from those obtained with the complete dataset. This strategy does not allow to perform a combined analysis of intra-center and inter-center variables in the same model and is more likely to produce errors. On the other hand, it allows the calculations to be made independently in each center without requiring to share information between them nor maintaining a huge database.

(d) The idea of performing data analysis at randomly partitioned blocks has been shown to be very efficient to analyse large datasets. The lower precision of these methods is usually compensated by the large amount of data. We must choose the number of partitions according to the size of the data and the available memory, taking into account that small partitions produce less accurate but faster results.

2) The analysis of the database of the Area-7 public health service of Madrid by using logistic regression models with random effects has shown that the presence of missing data in some variables is associated with the value that other factors related to the patient and the medical center acquire. We observe that the probability of having missing data in the variable WEIGHT is greater in men, young people, foreigners, the greater the nursing workload pressure is, the lower the doctor workload pressure is, and increases over the years. The center factor explains 8.9% of the variability. The probability of having missing data in the CIGARETTE variable is higher in the elderly and the higher the doctor's care pressure and the lower the nursing care pressure and has been decreasing each year. In this case the variability between centers is nearly zero.

A.4.2 ANALYSIS OF SEQUENCES OF EVENTS

3) Breaking down each sequence into smaller subsequences has been shown to be useful to easily identify the most relevant sequences, which can then be further studied by the doctor. Our analysis, in which we calculated the IR of the pairs and triplets, extends those found in other studies by considering not only the subsets of consecutive events but also non-consecutive ones. This generalisation allows us to discover transitions that would otherwise go unnoticed.

4) To be able to analyse the variables with incomplete ttime information we have developed two methods of imputation by permuting intra-annual events: assigning equal probabilities to each new sequence or according to the observed inter-annual frequencies. Both imputation methods give almost identical results but the equiprobable imputation is simpler and faster and allows to get a simplified analysis algorithm.

5) Connection networks have shown to be a useful tool to easily visualize the results, especially for longer sequences. This may require deciding how to filter out the transitions that we consider most relevant, in our case according to the IR.

6) The algorithms developed have allowed us to analyse the sequences of events in the Area-7 database, detecting the pairs and triplets of events with the highest IR for each type of initial or final event.

   (a) Although in some cases there are differences between the results obtained using complete and truncated dates, the methods developed are promising and have allowed us to perform the analyses using all the data, avoiding to discard large amounts of information.

   (b) The analysis of the events with the full date has shown that, in general, the pairs of events with the highest IR are those in which the final event is LIPIDICDISORDER or HBP. Particularly high are the IR of the pairs ATERO_PER→LIPIDICDISORDER , DIABETES→LIPIDICDISORDER , HBP→LIPIDICDISORDER, CARDIO_ISQ→HBP, HEART_FAILUER→ARRHYTMIAS and OBESITY→LIPIDICDISORDER. The IR is very low in most pairs in which the second event is ATERO_PER or ALCOHOLISM.

   (c) In most cases, the highest IR triplets are those in which where the last event is LIPIDICDISORDER.

   (d) The analysis of the events that appear with the truncated date has allowed to include new variables in the analysis, and has shown that most of the pairs in which the second event is HTA, FLU o LIPIDICDISORDER have high IR values.

   (e) The IR calculated by imputing truncated dates is, in general, slightly lower than those calculated with exact dates. In the case of triplets, on average, imputation methods overestimate the IR and produce greater variability than with event pairs.

A.4.3 LIMITATIONS AND FUTURE WORK

1. In order to overcome memory limitations we had to fragmentate the original data on the disk, process each fragment separately and merge it back. But this procedure is not applicable to all kind of tasks.
2. Despite the optimization of the analyses, the size of the studied data is at the limit of what can be properly handled on an average computer with the common statistical libraries, which attempt to load all the information into memory. This includes R with the lme4 or glmmTMB libraries, SPSS, Stata, Python with Numpy and Statsmodels, and Julia with the MixedModels.jl library.
3. For future studies, we recommend the use of specialized programs for the processing and visualization of large amounts of data, such as Tableau or Qlikview, or to store the information in databases such as MonetDB or TileDB. It would also be advisable, if the enough resources are available, to use distributed calculation platforms such as Spark with the Photon-ML library. The disadvantage of these technologies is that, although they facilitate the manipulation of the data and the obtaining of basic statistics, they are more complicated, many of their libraries are experimental and in general they do not allow to adjust complex models like those used in our study.
4. We consider it very important to make physicians and analysts aware of the importance of properly collecting and structuring the data from its source, recording the dates and other variables accurately and using an homogeneous criteria. This would allow to simplify the analyses, avoid trivial errors and obtain more accurate results without having to use complex statistical methods, [49].
5. In larger studies it could be necessary to perform the selection of covariates independently in each partition, which could generate different or even incompatible models in each one. It is necessary to develop methods capable of combining models that differ not only in the value of the coefficients but also in their structure.
6. Given the large number of covariates, regularized regression methods, such as Lasso or ElasticNet, could offer better results to select the models with the optimal combination of variables minimizing overfit problems, but they are computationally demanding.
7. In order to be able to analyze the sequences of events with incomplete temporal information we have imputed the intra-annual data using the distribution of the

inter-annual intervals, but the available information is biased, it is not possible to reconstruct it efficiently beyond the truncation zone. For future studies we advise to model separately the time intervals and the number of occurrences of each pair of events using truncated distributions (Poisson, Exponential,...) or to model directly the IR using some generic distribution such as truncated Beta.

8. It would also be useful to use bootstrap and Bayesian models to estimate the parameters of the models and calculate their credibility intervals. In practice, the use of highly complex models and resampling may not be feasible in large databases. As far as possible, we should try to replace complex calculation with pre-calculated equations and use valid approximations.

9. It would also be important to combine information from different datasets or with different formats, for example the obtained from truncated and untruncated variables.

10. Another line of work for the calculation of IR in triplets or larger sequences would be to find a way to include in the results the information of the time intervals between the first events of the sequences and to consider the competitiveness of the events even when they are not consecutive. Or also ignoring the order of some events.

11. We suggest using libraries such as Visnetwork that allow you to edit the charts interactively by easily selecting and filtering with the mouse the most relevant paths and nodes for the analyst. Connection networks can also be used to perform a differential network analysis, [149, 99], to study how the relationships between the nodes change with the variation of some factor such as age or gender, or to investigate complex correlations.

Apéndice B

RESUMEN DEL PROCESO DE LIMPIEZA Y TRANSFORMACIÓN DE LOS DATOS

Resumen del proceso de limpieza de los datos utilizados para el análisis de *missing data*:

- Exportamos datos desde SPSS a formato csv. De otro modo produce errores.
- Exportamos las definiciones, tipos de variable y comentarios desde SPSS al archivo de Excel sutituir.xlsx, que iremos actualizando con más información. Entre los comentarios se encuentras indicaciones explicitas de valores fuera de rango considerados *missings*. Muchos otros valores no habíann sido especificados.
- Cargamos el archivo csv con la función fread de la libreria data.table de R. Probamos otras librerías, pero ésta es la más rápida.
- En las columnas que contienen fecha sustituimos los ceros por NA y eliminamos los 10/14/1582, valor que aparece indicando fecha errónea en muchos casos.
- Sustituimos los valores fuera de rango indicados en sustituir.xlsx por NA.

- Eliminamos las columnas llenas de NULL y las completamente vacías.
- Eliminamos las columnas duplicadas, idénticas.
- Eliminamos las columnas que contienen fechas que pueden ser fácilmente calculadas a partir de otras columnas, por ejemplo sumando un año.
- Guardamos el resultado del proceso de limpieza preliminar en el archivo limpio.csv
- Comprimimos el archivo para que ocupe menos espacio en disco.
- Reescribimos adecuadamente el nombre de algunas variables fecha con el número al final del nombre para seguir siempre el mismo criterio. Por ejemplo tto_2007_
- Eliminamos las columnas que contienen el texto "OK" en su nombre. No son datos sino comprobaciones posteriores de la base de datos, añadidas por algún informático o analista.
- Modificamos el nombre de las columnas que contienen un número expresando año pero aparecen una única vez.
- Abreviamos algunos nombres de columnas que se repiten muchas veces: TOTAL_DERIVACIONES es sustituido por TD. PROTOCOLO por P. PRESCRIPCION por PRE.
- Transformamos los datos, contenidos en el archivo limpio.csv, de formato wide a formato long. Para ello necesitamos subdividir el archivo original en trozos y utilizando un bucle con el que ir operando sobre cada trozo. De este modo evitamos los problemas de memoria que impiden operar sobre todo el archivo de una sola vez. El resultado de cada trozo se va concatenando sucesivamente al final del anterior. Es un proceso en el que el resultado de cada trozo es independiente del resto.
- A partir de aquí usamos los datos en formato long.
- Comprobamos el resultado es correcto y que todas las filas empiezan por ID.
- Corregimos las fechas para que no haya ceros ni errores. preseleccionar columnas que contengan alguna fecha, y buscar en ellas cualquier cosa que no sea NA o fecha.
- Eliminamos las columnas que contienen la letra "X".
- Eliminamos las columnas que contienen 99999.
- Eliminamos los decimales del PESO.
- Eliminamos acentos y sustituimos la letra ñ por n.
- Seleccionamos en el Excel las variables que van a ser utilizadas como predictoras, outcome o ambas cosas. Dejamos fuera las variables OK y las repetidas y las de tipo fecha_situacion.
- Creamos una tabla resumiendo los valores de cada variable para poder clasificarlas.

- Unificamos Consultas_enf_10 con Consultas_enf cuando year=0 o 10 para cada ID y eliminamos el nombre antiguo.
- Unificamos todos los year=2000 con year=2010
- Buscamos manualmente posibles variables vinculadas, en las que la presencia de un *missing* no deba ser tenida en cuenta como tal ya que realmente no debía existir ningún valor en ese registro.
- Además, realizamos una búsqueda automática para intentar localizar más vinculaciones entre variables.
- Clasificamos las variables en subgrupos para reducir el número de posibles vinculaciones a explorar. Por ejemplo todas las relacionadas con el tabaco o todas las relacionadas con la diabetes.
- Añadimos una columna indicando que valor de cada variable representa un valor nulo: NO, 0,...
- Sustituimos momentáneamente todos los valores de la base de datos por 0, 1 o 2, según indiquen un valor nulo, un valor válido o la ausencia de valor (*missing*). Al pasar luego a formato long se pueden generar nuevos *missing*, pero no los consideramos *missing*, los ignoramos por ser fruto del post-procesado.
- Generamos tablas cruzadas de variables dos a dos con 0, 1, 2, NA.
- Abandonamos la idea de buscar ligaduras automáticamente porque se obtenían más de 6000 posibles. Demasiadas para comprobarlas. Nos limitaremos a la clasificación manual.
- Entonces los *missings* serán considerados *missing* siempre que no hayan sido generados en el post-procesado ni estén vinculados a otra variable que indique que no deba existir valor. Creamos dos bases de datos para su posterior análisis.
- En la primera base de datos, los *missings* en una variable vinculada a otra variable (padre) no se considerarán *missing* si la variable padre también es *missing*, 0 o NO, simplemente se ignorará.
- En la segunda base de datos, los *missings* en una variable vinculada a otra variable (padre) sí se considerarán *missing* aunque la variable padre también sea *missing*. Sólo se ignorará el *missing* si la variable padre es 0 o NO.
- Para las variables no ligadas sustituimos los *missings* por 1 y los demás valores por 0.
- Además, debemos realizar cambios especiales a las variables de los electrocardiogramas.

Si FECHA_xx_ELECTROCARD==xx, TOTAL_ELECTROCARD==yy

xx > yy o yy==NA entonces xx=NA

xx <= yy y xx==NA entonces xx=1

xx <= yy y xx!=NA entonces xx=0

- Sustituimos los valores erróneos de la variable PAIS_NAC por NA.
- Una vez hemos decidido que valores de *missing* son considerados *missing* y cuales no generamos una base de datos binaria de las variables outcome que queremos analizar.
- Eliminamos el año 2011 porque sólo contiene unas pocas variables a posteriori que no interesan y además perjudicaba todo el estudio añadiendo *missings* que no son *missings* reales e incrementando el tamaño de la base de datos.
- Eliminamos los datos de los años ya fallecidos, ignorar tras sustituir.

Resumen del proceso de limpieza de los datos utilizados para el análisis de las secuencias de eventos:

- Escogemos las variables de interés.
- Transformamos la base de datos a formato long.
- Eliminamos *missings* y duplicados.
- La lectura de las fechas resulta errónea porque se han mezclado formatos. Todas las variables fecha del año 2006 y las variables CARDIO_ISQ, DIABETES y ACV del año 2010 aparecen como m/d/y. Las convertimos al formato d/m/y, como el resto.
- Creamos archivo de datos con los eventos que aparecen con la fecha exacta y lo ordenamos.
- Creamos archivo de datos con fechas truncadas para los eventos definidos anualmente y lo ordenamos.

Apéndice C

TABLAS DE CONTINGENCIA PARA LA SELECCIÓN DE VARIABLES EN LOS MODELOS DE *MISSING DATA*

A continuación mostramos las tablas de contingencia calculadas para algunas variables de interés. Para cada variable se han sustituido los valores 0, "N" y otros con significado nulo por un 0. Cada *missing* o valor erróneo ha sido sustituido por un 1. El resto de valores, los correctos, han sido susituidos por un 2. A partir de esos nuevos valores 0, 1 y 2 hemos calculado las tablas de contingencia para cada par de variables. Por ejemplo, en la tabla ARRITMIAS VS FECHA_ARRITMIAS podemos comprobar como existen 73 768 casos en los que la variable ARRITMIAS contiene un valor correcto y su FECHA también, y 1 047 832 casos en los que la variable ARRITMIAS consta explícitamente como nula, porque no hubo arritmia, y la FECHA de esa arritmia es un *missing*, ya que no había ninguna fecha que tomar.

Con estos resultados podemos comprobar que variables se corresponden exactamente con otras, que variables parecen no tener ninguna relación. Y lo que más interesante, podemos

investigar que variables parecen tener sus valores *missing* completamente condicionados los *missing* o nulos de otra. Por ejemplo comprobamos que en los casos en los que TABACO=0, es decir no fumador, la variable CIGARRILLOS se registró como 0 en 112 002 casos pero en otros 900 964 simplemente se dejó sin anotar. Pero claramente eso no es un error ni lo consideraremos *missing* para nuestro análisis ya que en esos casos ese dato no era necesario.

EXTRANJERO

PAIS_NAC	0	1	2
0	0	0	0
1	2724	42	252
2	1200786	10890	131226

INSFCARD

FECHA INSFCARD	0	1	2
0	0	0	0
1	1105744	0	0
2	0	0	15856

PRO_INSFCARD

FECHA INSFCARD	0	1	2
0	0	0	0
1	1105219	0	525
2	12036	0	3820

VALVULOPAT

FECHAVAL VULOPAT	0	1	2
0	0	0	0
1	1104620	0	0
2	0	0	16980

ARRITMIAS

FECHA ARRITMIAS	0	1	2
0	0	0	0
1	1047832	0	0
2	0	0	73768

CARDIOISQ

FECHA CARDIOISQ	0	1	2
0	0	0	0
1	1080809	0	0
2	0	0	40791

INCIDHTA

YEAR DIAGHTA	0	1	2
0	0	0	0
1	992318	0	11482
2	112994	0	4806

DIAGACV

YEAR DIAGACV	0	1	2
0	0	0	0
1	875036	0	0
2	19784	0	2460

ATEROPERI

FECHA ATEROPERI	0	1	2
0	0	0	0
1	1121371	0	0
2	0	0	229

INCID_DIABET

PRO DIABET	0	1	2
0	836121	0	1991
1	0	0	0
2	54949	0	4219

TRAST_LIP

FECHA TRAST_LIP	0	1	2
0	0	0	0
1	839190	0	0
2	0	0	282410

TABAQUIS

FECHA TABAQUIS	0	1	2
0	0	0	0
1	1063928	0	0
2	0	0	57672

ALCOHOLIS

FECHA ALCOHOLIS	0	1	2
0	0	0	0
1	1107221	0	0
2	0	0	14379

PRO_ALCOHOL

FECHA ALCOHOLIS	0	1	2
0	0	0	0
1	1102705	0	4516
2	12329	0	2050

PRO_ALCOHOL

ALCOHOLIS	0	1	2
0	1102705	0	4516
1	0	0	0
2	12329	0	2050

TABACO

CIGAR	0	1	2
0	112002	0	0
1	900964	0	9687
2	0	0	98947

INR_RANGO

INR CONTROL	0	1	2
0	4483	0	3076
1	788	882150	1176
2	0	0	5607

HDL

PRO HIPERCOL	0	1	2
0	0	1051891	2704
1	0	0	0
2	0	65962	1043

DIAG_TRAST_LIP

YEAR_DIAG TRAST_LIP	0	1	2
0	0	0	0
1	1010010	0	0
2	103917	0	7673

DIAG_DIABET

YEAR_DIAG DIABET	0	1	2
0	0	0	0
1	1081870	0	0
2	37512	0	2218

CVD

ING_CVD	0	1	2
0	1309512	0	17298
1	0	0	0
2	0	0	19110

DIAG_OBESID

YEAR_DIAG OBESID	0	1	2
0	0	0	0
1	1071185	0	0
2	46911	0	3504

DIAG_CARDIOISQ

YEAR_DIAG CARDIOISQ	0	1	2
0	0	0	0
1	646281	0	0
2	24959	0	1720

HTA

FECHA_HTA	0	1	2
0	0	0	0
1	798209	0	0
2	0	0	323391

ACV

FECHA_ACV	0	1	2
0	0	0	0
1	1098406	0	0
2	0	0	23194

DIABET

FECHA DIABET	0	1	2
0	0	0	0
1	1018026	0	0
2	0	0	103574

OBESID

FECHA OBESID	0	1	2
0	0	0	0
1	984408	0	0
2	0	0	137192

PESO

PRO_OBESID	0	1	2
0	0	682629	351648
1	0	0	0
2	0	20720	66603

DIAS_REINGRESO

REINGRESO	0	1	2
0	0	24450	0
1	0	1314450	0
2	0	0	7020

DIAG_CIE9M RECO1CMBD

FECHA.1 CMBD	0	1	2
0	0	0	0
1	0	1314450	0
2	0	0	31470

Apéndice D

DETALLES DE LOS MODELOS DE REGRESIÓN Y DEL METAANÁLISIS DE LOS COEFICIENTES

D.1 Análisis de todos los datos conjuntamente

Tabla D.1: Índice de modelos logísticos alternativos para la regresión de la *odds* de *Missing* en las variables PESO o CIGARRILLOS en función de las demás covariables y de efectos aleatorios con el conjunto completo de los datos.

	Modelos logísticos $Miss \sim SEXO + YEAR + EDAD + PAMED + PAENF + ...$	Tabla
PESO	$+TABACO + GRIPE + RENTA\_ZBS + IMC +$ $POSMAS + ENFCONCENTRO + TOT\_FARMAC65 +$ $MFDEMCENTRO + (1\|EAP/ID)$	Tabla D.2
	$+PAMED : YEAR + PAENF : YEAR + POSMAS +$ $EXTRAN + (1\|EAP/ID)$	Tabla D.3
	$+PAMED : YEAR + PAENF : YEAR + zPOSMAS +$ $EXTRAN + (1\|EAP)$	Tabla D.4
CIGARRILOS	$+zPE70 + EXTRANJERO + (1\|EAP/ID)$	Tabla D.5
	$+zPE70 + EAP + zPAMED : zYE06 + zPAENF : zYE06$ $+EXTRAN$	Tabla D.6
	$+zPE70 + zPAMED : zYE06 + zPAENF : zYE06$ $+EXTRAN + (1\|EAP/ID)$	Tabla D.7
	$+zPE70 + zPAMED : zYE06 + zPAENF : zYE06$ $+EXTRAN + (1\|EAP)$	Tabla D.8

D.1.1 *Missing* data en la variable Peso

En este apartado incluimos los resultados obtenidos con modelos de regresión alternativos para el PESO utilizando el conjunto completo de los datos:

Tabla D.2: Resultados del modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro para el PESO con variables adicionales.

```
Family: binomial  ( logit )
Formula: mPESO ~ SEXO + TABACO + GRIPE + zYE06 + zED50 + zRENTA_ZBS + zIMC +
  zPAMED + zPAENF + zPOSMAS + zENFCONCENTRO + zTOTAL_FARMACOS_DESDE65_ANOS +
  zMFDEMCENTRO + (1|EAP/ID)
```

```
     AIC      BIC   logLik deviance df.resid
 26224.4  26357.0 -13096.2  26192.4    29246

Random effects:
 Groups Name        Variance Std.Dev.
   ID:EAP (Intercept) 4.961    2.227
     EAP    (Intercept) 2.252    1.501
  Number of obs: 29262, groups:  ID 6877; EAP 19

Fixed Effects:
                              Estimate  Std.Error  z-value   Pr(>|z|)
  (Intercept)                 -5.10910    0.64718   -7.894   2.92e-15
  SEXO                         0.34551    0.07151    4.832   1.35e-06
  TABACO                       0.24123    0.09030    2.671   0.007556
  GRIPE                        0.44314    0.05432    8.158   3.41e-16
  zYE06                        1.33264    0.05777   23.070    < 2e-16
  zED50                       -0.05404    0.44387   -0.122   0.903094
  zRENTA_ZBS                  -0.68664    0.32779   -2.095   0.036192
  zIMC                        -0.28205    0.03560   -7.923   2.33e-15
  zPAMED                      -1.85082    0.06706  -27.599    < 2e-16
  zPAENF                       0.85394    0.04487   19.030    < 2e-16
  zPOSMAS                     -0.39334    0.11336   -3.470   0.000521
  zENFCONCENTRO                0.05364    0.02339    2.293   0.021844
  zTOTAL_FARMACOS_DESDE65_ANOS 0.15709    0.03554    4.420   9.87e-06
  zMFDEMCENTRO                -0.01605    0.03002   -0.534   0.593030
```

Tabla D.3: Resultados del modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro con interacciones y POSMAS para el PESO.

```
Family: binomial  ( logit )
Formula: mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + zPAMED:zYE06 +
  zPAENF:zYE06 + zPOSMAS + EXTRANJERO + (1|EAP/ID)

      AIC       BIC    logLik  deviance  df.resid
 890763.0  890904.6 -445369.5  890739.0    985684

Random effects:
  Groups Name        Variance Std.Dev.
    ID:EAP (Intercept) 10.051   3.17
    EAP    (Intercept)  1.612   1.27
```

```
  Number of obs: 985696, groups:  ID 201331; EAP 19

Fixed Effects:
               Estimate  Std.Error z-value Pr(>|z|)
  (Intercept)   0.868873   0.291792    2.98   0.0029
  SEXO         -0.212288   0.017824  -11.91   <2e-16
  zYE06         0.442009   0.005029   87.90   <2e-16
  zED50        -0.705001   0.005511 -127.92   <2e-16
  zPAMED       -0.496806   0.010991  -45.20   <2e-16
  zPAENF        0.321428   0.010531   30.52   <2e-16
  zPOSMAS       0.241153   0.015664   15.40   <2e-16
  EXTRANJERO    0.783591   0.032382   24.20   <2e-16
  zYE06:zPAMED -0.094786   0.003215  -29.48   <2e-16
  zYE06:zPAENF  0.041435   0.003391   12.22   <2e-16
```

Tabla D.4: Resultados del modelo de efectos aleatorios con medidas repetidas por centros con interacciones y POSMAS para el PESO.

```
Family: binomial  ( logit )
Formula: mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + zPAMED:zYE06 +
    zPAENF:zYE06 + zPOSMAS + EXTRANJERO + (1|EAP)

      AIC       BIC    logLik  deviance  df.resid
1182900.9 1183030.7 -591439.4 1182878.9    985685

Random effects:
  Groups Name        Variance Std.Dev.
    EAP    (Intercept) 0.3484   0.5903
  Number of obs: 985696, groups:  EAP 19

Fixed effects:
               Estimate  Std.Error z-value  Pr(>|z|)
  (Intercept)   0.270301   0.135612    1.99   0.0462
  SEXO1        -0.066632   0.004530  -14.71   <2e-16
  zYE06         0.193679   0.003273   59.18   <2e-16
  zED50        -0.271645   0.001300 -208.99   <2e-16
  zPAMED       -0.247484   0.007258  -34.10   <2e-16
  zPAENF        0.146554   0.007056   20.77   <2e-16
  zPOSMAS       0.113737   0.010299   11.04   <2e-16
  EXTRANJERO1   0.300485   0.008411   35.73   <2e-16
  zYE06:zPAMED -0.041278   0.002127  -19.41   <2e-16
```

```
zYE06:zPAENF  0.020320   0.002262    8.98   <2e-16
```

D.1.2 *Missing data* en la variable Cigarrillos

Modelos de regresión alternativos para los CIGARRILLOS utilizando el conjunto completo de los datos:

Tabla D.5: Resultados del modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro y las variables SEXO y EXTRANJERO para CIGARRILLOS.

```
Family: binomial  ( logit )
Formula: mCIGARRILLOS ~ SEXO + zYE06 + zED50 + zPE70 + zPAMED + zPAENF +
           EXTRANJERO + (1|EAP/ID)

     AIC      BIC   logLik deviance df.resid
  8392.4   8482.4  -4186.2   8372.4    59993

Random effects:
  Groups Name        Variance  Std.Dev.
    ID:EAP (Intercept) 1.136e+03 3.370e+01
    EAP    (Intercept) 2.955e-09 5.436e-05
  Number of obs: 60003, groups:  ID 19555; EAP 19

Fixed effects:
               Estimate  Std.Error z-value Pr(>|z|)
  (Intercept) -11.91092    0.26067  -45.69   <2e-16
  SEXO1         0.37651    0.28399    1.33   0.1849
  zYE06        -1.85199    0.11430  -16.20   <2e-16
  zED50         0.20532    0.08956    2.29   0.0219
  zPE70         0.13247    0.07422    1.78   0.0743
  zPAMED        0.21145    0.11761    1.80   0.0722
  zPAENF       -0.34471    0.15145   -2.28   0.0228
  EXTRANJERO1   0.55140    0.52705    1.05   0.2955
```

Tabla D.6: Resultados del modelo de regresión simple con interacciones para CIGARRILLOS.

```
Call: glm(mCIGARRILLOS ~ SEXO + zYE06 + zED50 + zPE70 + zPAMED + zPAENF +
     EAP + zPAMED:zYE06 + zPAENF:zYE06 + EXTRANJERO,
      family = "binomial", data = todos)

Coefficients:
                Estimate  Std.Error z-value Pr(>|z|)
  (Intercept)  -2.516249   0.108421 -23.208  < 2e-16
  SEXO1         0.193366   0.034708   5.571 2.53e-08
  zYE06        -0.196010   0.021517  -9.109  < 2e-16
  zED50         0.138502   0.010984  12.610  < 2e-16
  zPE70         0.057458   0.009340   6.152 7.67e-10
```

```
  zPAMED        0.154269   0.047195   3.269 0.001080
  zPAENF       -0.182827   0.048168  -3.796 0.000147
  EXTRANJERO1   0.376187   0.062919   5.979 2.25e-09
  zYE06:zPAMED -0.079321   0.015412  -5.147 2.65e-07
  zYE06:zPAENF  0.061833   0.016299   3.794 0.000148
  EAP16070210   0.683709   0.157443   4.343 1.41e-05
  EAP16070310   0.453117   0.129193   3.507 0.000453
  EAP16070410   0.054100   0.133242   0.406 0.684722
  EAP16070610   0.097016   0.133786   0.725 0.468354
  EAP16071210   0.140263   0.147567   0.951 0.341855
  EAP16071310   0.410538   0.127876   3.210 0.001325
  EAP16071410  -0.226976   0.119426  -1.901 0.057360
  EAP16071510   0.482632   0.139294   3.465 0.000531
  EAP16071610   0.268848   0.130625   2.058 0.039573
  EAP16072110   0.623916   0.134381   4.643 3.44e-06
  EAP16072210   0.623300   0.145427   4.286 1.82e-05
  EAP16072410   0.562545   0.121432   4.633 3.61e-06
  EAP16072610   0.493149   0.130732   3.772 0.000162
  EAP16072810   0.084041   0.131641   0.638 0.523206
  EAP16073110  -0.256670   0.140895  -1.822 0.068499
  EAP16073210  -0.002781   0.130889  -0.021 0.983049
  EAP16073410   0.211709   0.150129   1.410 0.158485
  EAP16073610   0.090659   0.126297   0.718 0.472866

    Null deviance: 32062  on 60002  degrees of freedom
Residual deviance: 31212  on 59975  degrees of freedom
  (933951 observations deleted due to missingness)
AIC: 31268
```

Tabla D.7: Resultados del modelo de efectos aleatorios con medidas repetidas por centros e individuos anidados a cada centro e interaciones para CIGARRILLOS.

```
Family: binomial  ( logit )
Formula: mCIGARRILLOS ~ SEXO + zYE06 + zED50 + zPE70 + zPAMED + zPAENF +
    zPAMED:zYE06 + zPAENF:zYE06 + EXTRANJERO + (1 | EAP/ID)

     AIC      BIC   logLik deviance df.resid
  8391.4   8499.4  -4183.7   8367.4    59991

Random effects:
  Groups Name        Variance  Std.Dev.
```

```
    ID:EAP (Intercept) 1.133e+03 3.366e+01
    EAP    (Intercept) 1.803e-08 1.343e-04
  Number of obs: 60003, groups:  ID 19555; EAP 19

Fixed effects:
                 Estimate  Std.Error z-value Pr(>|z|)
  (Intercept)  -11.85275    0.26214  -45.22  < 2e-16
  SEXO1          0.37607    0.28415    1.32  0.18568
  zYE06         -1.88776    0.11973  -15.77  < 2e-16
  zED50          0.20311    0.08962    2.27  0.02343
  zPE70          0.13033    0.07448    1.75  0.08016
  zPAMED         0.32602    0.13152    2.48  0.01318
  zPAENF        -0.50063    0.18412   -2.72  0.00655
  EXTRANJERO1    0.54676    0.52742    1.04  0.29990
  zYE06:zPAMED  -0.21390    0.09647   -2.22  0.02661
  zYE06:zPAENF   0.13291    0.08651    1.54  0.12446
```

Tabla D.8: Resultados del modelo de efectos aleatorios con medidas repetidas por centros e interaciones para CIGARRILLOS.

```
Family: binomial  ( logit )
Formula: mCIGARRILLOS ~ SEXO + zYE06 + zED50 + zPE70 + zPAMED + zPAENF +
    zPAMED:zYE06 + zPAENF:zYE06 + EXTRANJERO + (1|EAP)

     AIC      BIC   logLik deviance df.resid
 31304.2  31403.2 -15641.1  31282.2    59992

Random effects:
  Groups Name        Variance Std.Dev.
    EAP    (Intercept) 0.06514  0.2552
  Number of obs: 60003, groups:  EAP 19

Fixed effects:
                Estimate  Std.Error z-value Pr(>|z|)
  (Intercept)  -2.293301   0.072371  -31.69  < 2e-16
  SEXO1         0.194420   0.034694    5.60 2.10e-08
  zYE06        -0.181014   0.019279   -9.39  < 2e-16
  zED50         0.140046   0.010983   12.75  < 2e-16
  zPE70         0.058096   0.009338    6.22 4.93e-10
  zPAMED        0.188327   0.042415    4.44 8.99e-06
  zPAENF       -0.176990   0.046250   -3.83 0.000130
```

```
EXTRANJERO1    0.378224   0.062881    6.01 1.80e-09
zYE06:zPAMED -0.076353   0.015200   -5.02 5.08e-07
zYE06:zPAENF  0.061670   0.016104    3.83 0.000128
```

D.2 Análisis de los datos particionados aleatoriamente

D.2.1 *Missing data* en la variable Peso

D.2.1.1 Modelo logístico simple.

Resultados de los análisis realizados con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios, para la los *missings* en la variable PESO.

```
glm(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + EAP,
data=partition, family="binomial")
```

La Tabla D.9 resume los coeficientes y errores estándar obtenidos en dichos análisis intra-partición para cada variable. A partir de esos datos, para cada uno de los coeficientes hemos realizado un metaanálisis combinando los resultados de todas las particiones utilizando el paquete *metafor*, con la opción "FE", Tabla D.10 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.9: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable PESO realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística simple.

Part	Inter	SEXO	YE06	ED50	PAMED	PAENF	EXTRA	std.err Inter	std.err SEXO	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF	std.err EXTRA
16	0.005	-0.066	0.233	-0.273	-0.333	0.234	0.320	0.062	0.020	0.014	0.006	0.028	0.021	0.037
11	-0.070	-0.072	0.195	-0.283	-0.318	0.136	0.272	0.062	0.020	0.014	0.006	0.028	0.021	0.036
13	0.040	-0.061	0.204	-0.274	-0.305	0.172	0.323	0.063	0.020	0.014	0.006	0.029	0.021	0.037
1	-0.011	-0.082	0.222	-0.267	-0.273	0.201	0.286	0.063	0.020	0.014	0.006	0.028	0.021	0.037
3	-0.067	-0.022	0.216	-0.268	-0.269	0.166	0.292	0.062	0.020	0.014	0.006	0.028	0.021	0.036
7	-0.046	-0.059	0.228	-0.273	-0.265	0.198	0.312	0.062	0.020	0.014	0.006	0.028	0.021	0.036
6	-0.122	-0.083	0.212	-0.265	-0.288	0.164	0.318	0.062	0.020	0.014	0.006	0.029	0.021	0.037
9	-0.005	-0.058	0.205	-0.275	-0.300	0.195	0.332	0.062	0.020	0.014	0.006	0.028	0.021	0.037
19	0.017	-0.053	0.188	-0.272	-0.308	0.167	0.291	0.062	0.020	0.014	0.006	0.028	0.021	0.036
2	-0.062	-0.067	0.231	-0.276	-0.243	0.185	0.242	0.063	0.020	0.014	0.006	0.028	0.021	0.036
17	-0.066	-0.049	0.208	-0.267	-0.296	0.156	0.324	0.063	0.020	0.014	0.006	0.028	0.021	0.037
10	-0.024	-0.053	0.209	-0.275	-0.292	0.175	0.328	0.062	0.020	0.014	0.006	0.028	0.021	0.037
8	-0.121	-0.080	0.215	-0.266	-0.290	0.195	0.381	0.062	0.020	0.014	0.006	0.029	0.021	0.037
15	-0.144	-0.047	0.213	-0.268	-0.292	0.200	0.288	0.063	0.020	0.014	0.006	0.028	0.021	0.037
14	-0.042	-0.127	0.194	-0.265	-0.311	0.165	0.231	0.062	0.020	0.014	0.006	0.028	0.021	0.036
4	0.024	-0.069	0.201	-0.276	-0.330	0.206	0.250	0.064	0.020	0.014	0.006	0.029	0.021	0.036
5	0.056	-0.084	0.211	-0.273	-0.265	0.175	0.322	0.063	0.020	0.013	0.006	0.028	0.021	0.037
12	-0.116	-0.049	0.215	-0.274	-0.310	0.178	0.211	0.062	0.020	0.014	0.006	0.028	0.021	0.036
18	-0.016	-0.087	0.207	-0.272	-0.296	0.180	0.382	0.061	0.020	0.014	0.006	0.029	0.021	0.037

Tabla D.10: Metaanálisis de los Interceptos de las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$Inter,sei=resuG7$std.Inter, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 27.5366   15.5282  -53.0732  -52.1288  -52.8379

Test for Heterogeneity:
  Q(df = 18) = 15.5282, p-val = 0.6254

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.0407  0.0143  -2.8524  0.0043  -0.0687  -0.0127
```

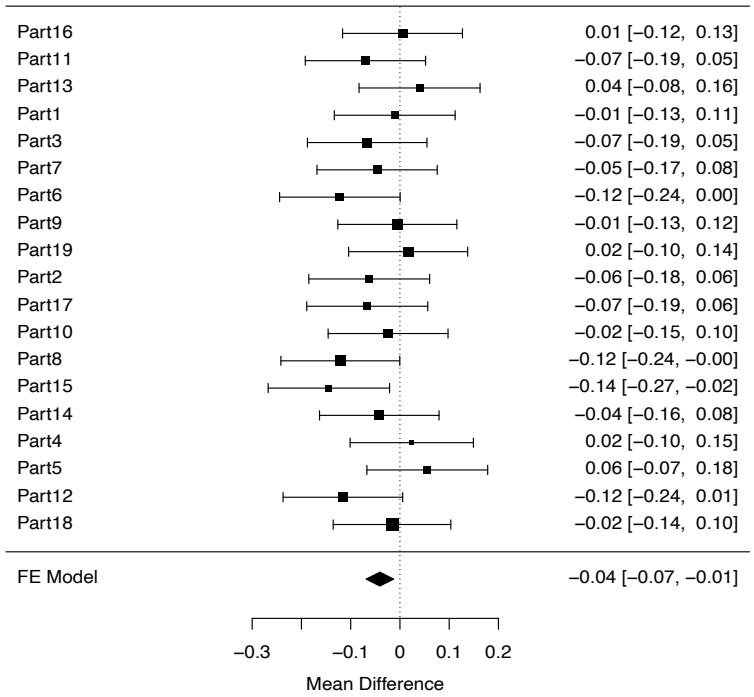


Figura D.1: Forest plot de los Interceptos de las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

Tabla D.11: Metaanálisis de los coeficientes de SEXO obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$SEXO,sei=resuG7$std.SEXO, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 46.0290   22.1583  -90.0579  -89.1135  -89.8226

Test for Heterogeneity:
  Q(df = 18) = 22.1583, p-val = 0.2250

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.0667  0.0045  -14.7156  <.0001  -0.0756  -0.0578
```

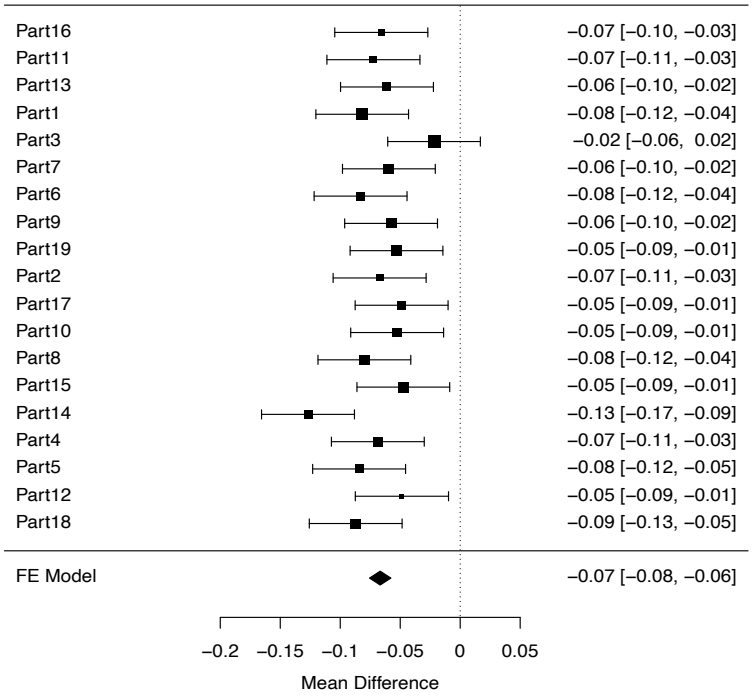


Figura D.2: Forest plot de los coeficientes de SEXO obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

Tabla D.12: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$zYE06,sei=resuG7$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik   deviance        AIC        BIC       AICc
  57.0890    14.3400  -112.1781  -111.2336  -111.9428

Test for Heterogeneity:
  Q(df = 18) = 14.3400, p-val = 0.7067

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.2109  0.0031  67.8213  <.0001  0.2048  0.2170
```

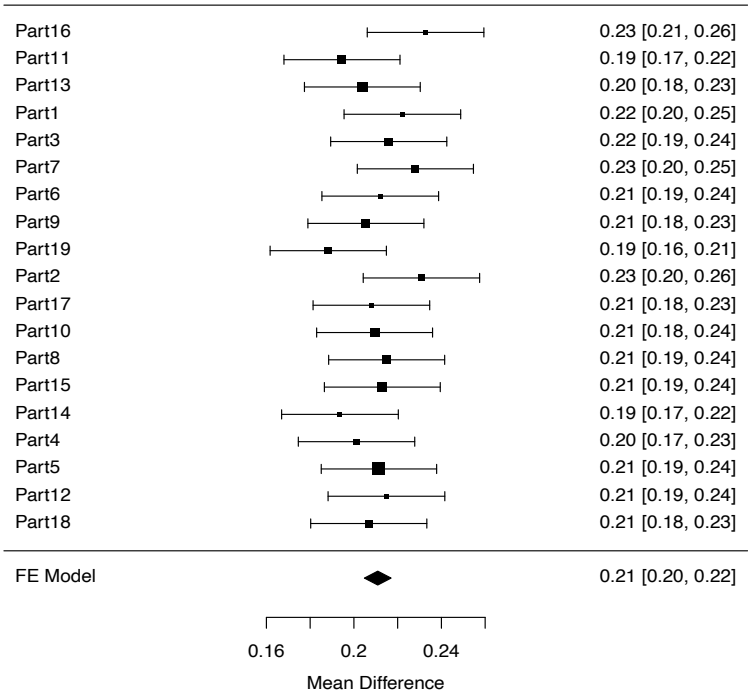


Figura D.3: Forest plot de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

Tabla D.13: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$zED50,sei=resuG7$std.zED50, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik   deviance        AIC        BIC       AICc
  74.8924    11.8756  -147.7849  -146.8404  -147.5496

Test for Heterogeneity:
  Q(df = 18) = 11.8756, p-val = 0.8536

Model Results:
  estimate      se       zval    pval    ci.lb    ci.ub
   -0.2716  0.0013  -208.9560  <.0001  -0.2742  -0.2691
```

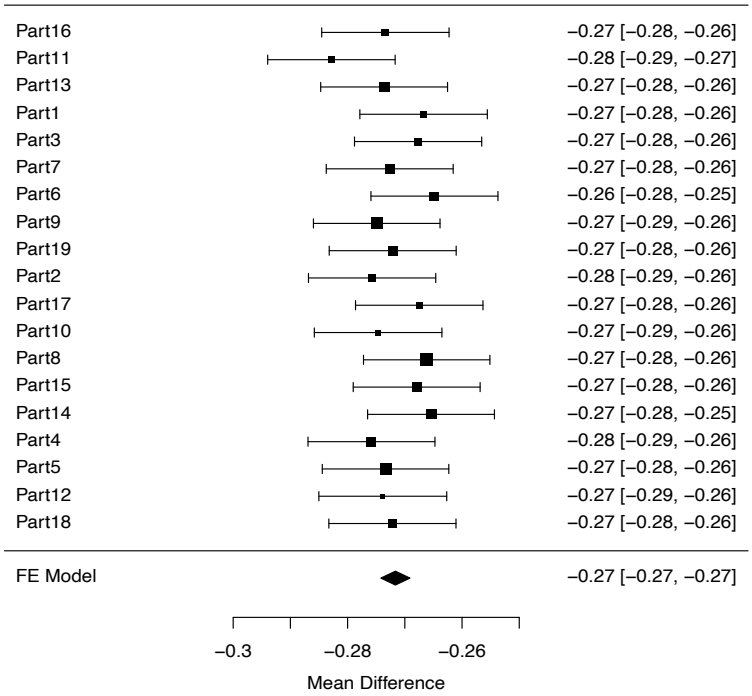


Figura D.4: Forest plot de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

Tabla D.14: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$zPAMED,sei=resuG7$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 44.1562   12.0822  -86.3123  -85.3679  -86.0770

Test for Heterogeneity:
  Q(df = 18) = 12.0822, p-val = 0.8430

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.2937  0.0065  -45.0552  <.0001  -0.3065  -0.2809
```

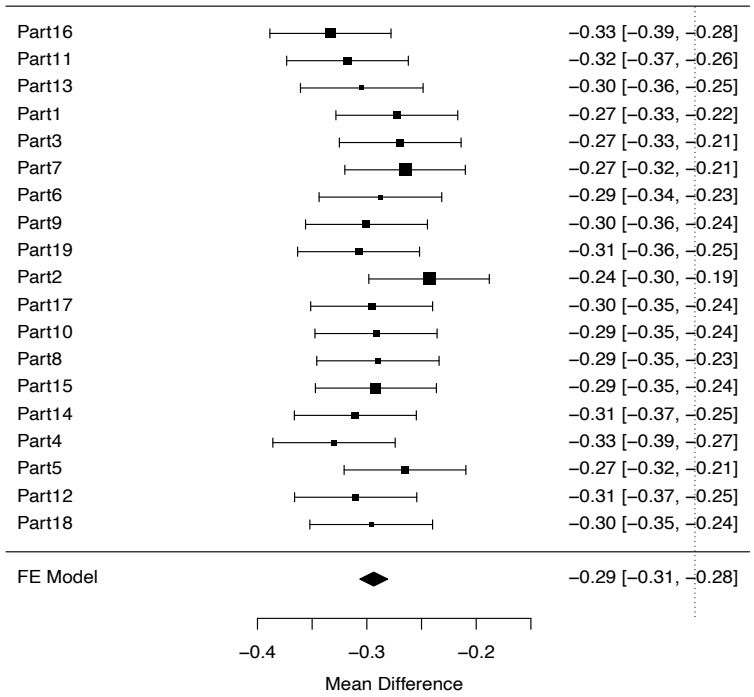


Figura D.5: Forest plot de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

Tabla D.15: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$zPAENF,sei=resuG7$std.zPAENF, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 46.1460   19.9681  -90.2920  -89.3476  -90.0567

Test for Heterogeneity:
  Q(df = 18) = 19.9681, p-val = 0.3346

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.1814  0.0048  38.0344  <.0001  0.1721  0.1908
```

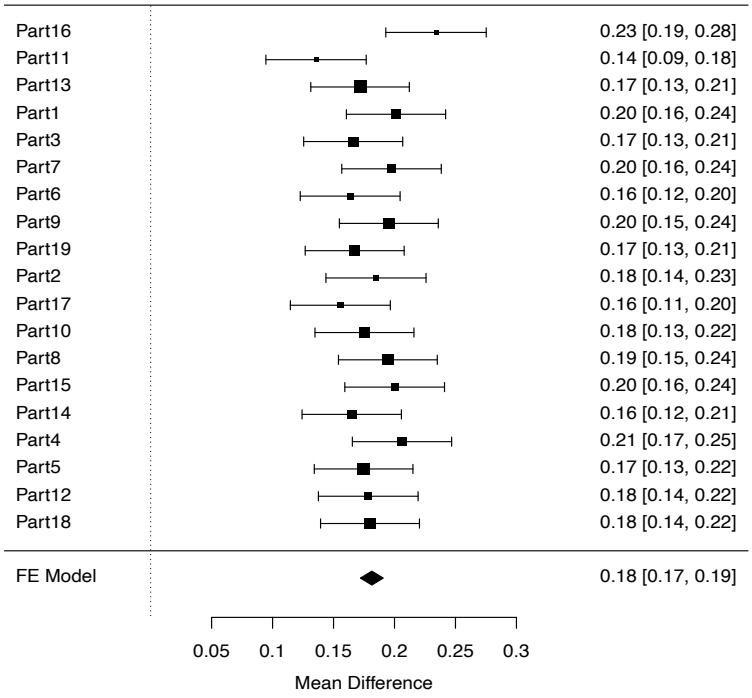


Figura D.6: Forest plot de los coeficientes de PAENF obtenidos en las regresiones simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

Tabla D.16: Metaanálisis de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

```
rma(resuG7$EXTRANJERO,sei=resuG7$std.EXTRANJERO, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 31.3791   27.9469  -60.7582  -59.8137  -60.5229

Test for Heterogeneity:
  Q(df = 18) = 27.9469, p-val = 0.0629

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.2996  0.0084  35.6263  <.0001  0.2832  0.3161
```

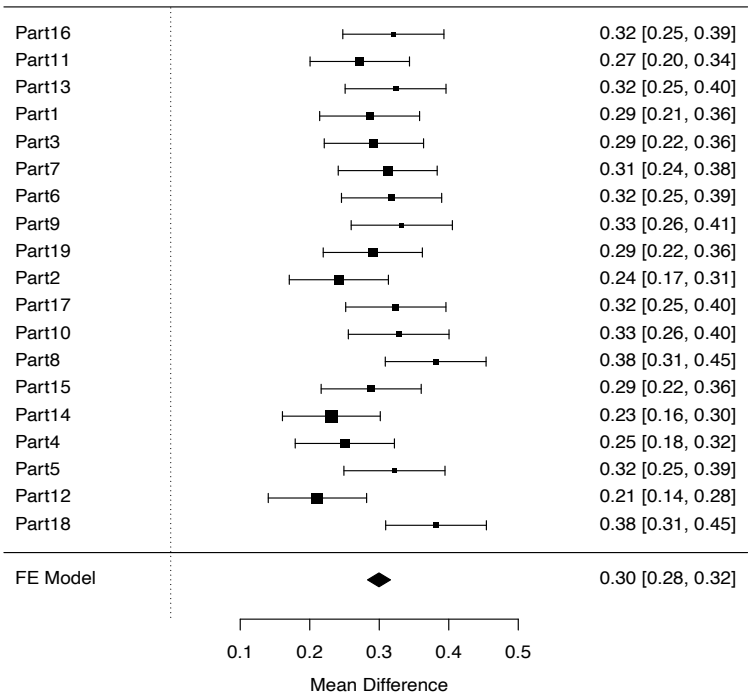


Figura D.7: Forest plot de los coeficientes de EXTRANJERO obtenidos en las regresiones simples intra-partición-aleatoria para la *odds* de *missings* en PESO.

D.2.1.2 Modelo logístico con efectos aleatorios para los individuos.

Resultados de los análisis realizados para la los *missings* en la variable PESO con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios (medidas repetidas) para los individuos, **sin** anidar a los centros. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
          (1|ID), data=partition, family="binomial", REML=F)
```

La Tabla D.17 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.18 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.17: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable PESO realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los individuos.

Part	Inter	SEXO	YE06	ED50	PAMED	PAENF	EXTRA	std.err Inter	std.err SEXO	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF	std.err EXTRA
5	0.547	-0.222	0.502	-0.736	-0.605	0.443	0.686	0.250	0.079	0.021	0.025	0.042	0.031	0.139
15	0.353	-0.285	0.506	-0.719	-0.548	0.388	0.689	0.241	0.080	0.021	0.025	0.043	0.031	0.140
16	0.274	-0.237	0.449	-0.726	-0.705	0.377	0.907	0.239	0.080	0.021	0.025	0.043	0.031	0.141
9	0.434	-0.284	0.494	-0.698	-0.616	0.397	0.701	0.246	0.079	0.021	0.025	0.043	0.031	0.142
12	0.107	-0.350	0.527	-0.716	-0.676	0.454	0.853	0.255	0.081	0.021	0.025	0.044	0.031	0.141
2	0.033	-0.133	0.498	-0.693	-0.571	0.408	0.822	0.253	0.079	0.021	0.025	0.043	0.031	0.140
19	0.136	-0.319	0.553	-0.749	-0.594	0.456	0.818	0.249	0.079	0.021	0.025	0.042	0.031	0.142
8	0.262	-0.241	0.461	-0.744	-0.604	0.354	0.712	0.250	0.082	0.021	0.025	0.044	0.032	0.145
13	0.072	-0.159	0.466	-0.709	-0.563	0.388	0.762	0.238	0.077	0.021	0.024	0.042	0.031	0.134
14	0.239	-0.352	0.492	-0.695	-0.578	0.397	0.868	0.241	0.082	0.021	0.025	0.043	0.031	0.140
11	0.158	-0.252	0.460	-0.725	-0.638	0.373	0.793	0.247	0.080	0.021	0.025	0.043	0.031	0.139
6	-0.139	-0.061	0.476	-0.733	-0.565	0.375	0.598	0.253	0.079	0.021	0.025	0.043	0.031	0.139
1	0.013	-0.044	0.497	-0.754	-0.565	0.385	0.687	0.230	0.079	0.021	0.025	0.043	0.031	0.140
18	-0.190	-0.152	0.451	-0.759	-0.677	0.339	0.556	0.259	0.078	0.021	0.024	0.042	0.031	0.137
4	-0.162	-0.270	0.490	-0.748	-0.720	0.472	0.506	0.255	0.080	0.021	0.025	0.043	0.031	0.140
17	0.295	-0.189	0.487	-0.745	-0.622	0.360	0.643	0.248	0.080	0.021	0.025	0.043	0.031	0.138
10	0.093	-0.228	0.508	-0.750	-0.588	0.391	0.713	0.267	0.084	0.021	0.026	0.044	0.032	0.147
3	-0.135	-0.131	0.494	-0.753	-0.542	0.375	0.804	0.244	0.081	0.021	0.025	0.043	0.031	0.139
7	0.258	-0.213	0.496	-0.738	-0.607	0.414	0.704	0.272	0.084	0.021	0.026	0.043	0.032	0.148

Tabla D.18: Metaanálisis de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$Inter,sei=resuG8$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
   2.8009   12.3028   -3.6018   -2.6574   -3.3665

Test for Heterogeneity:
  Q(df = 18) = 12.3028, p-val = 0.8312

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.1412  0.0571  2.4751  0.0133  0.0294  0.2530
```

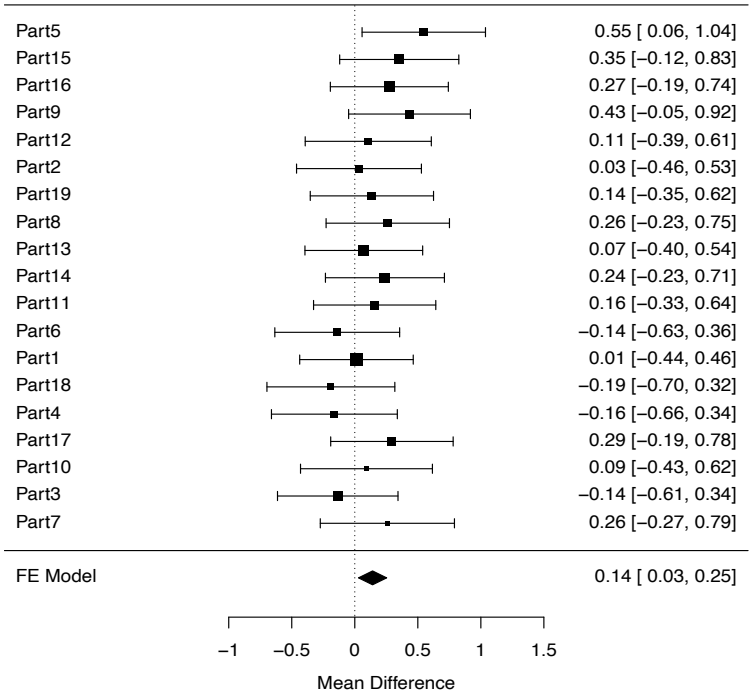


Figura D.8: Forest plot de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.19: Metaanálisis de los coeficientes de SEXO obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$SEXO,sei=resuG8$std.SEXO, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 19.6788   21.6149  -37.3577  -36.4133  -37.1224

Test for Heterogeneity:
  Q(df = 18) = 21.6149, p-val = 0.2495

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.2158  0.0184  -11.7363  <.0001  -0.2518  -0.1798
```

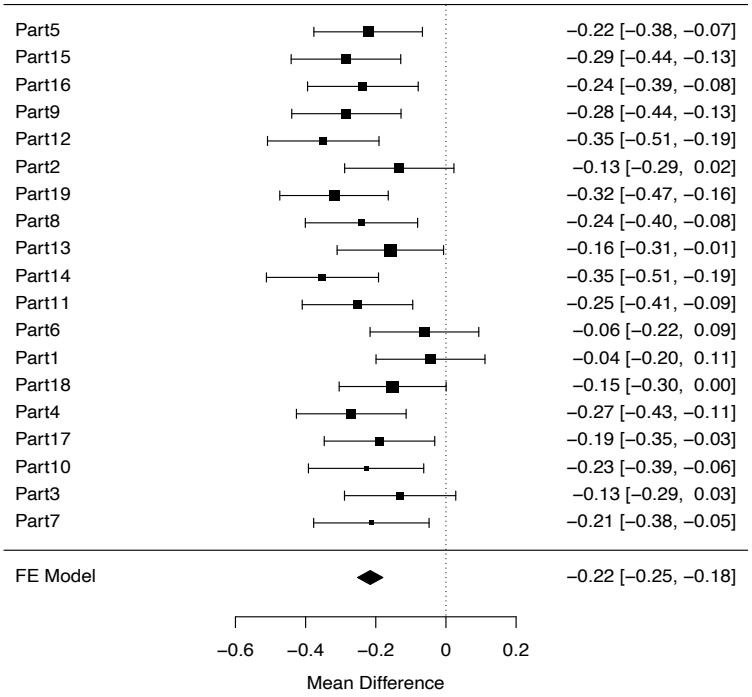


Figura D.9: Forest plot de los coeficientes de SEXO en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.20: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$zYE06,sei=resuG8$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 42.2770   27.1228  -82.5539  -81.6095  -82.3186

Test for Heterogeneity:
  Q(df = 18) = 27.1228, p-val = 0.0767

Model Results:
  estimate      se      zval    pval   ci.lb   ci.ub
    0.4896  0.0048  101.0943  <.0001  0.4802  0.4991
```

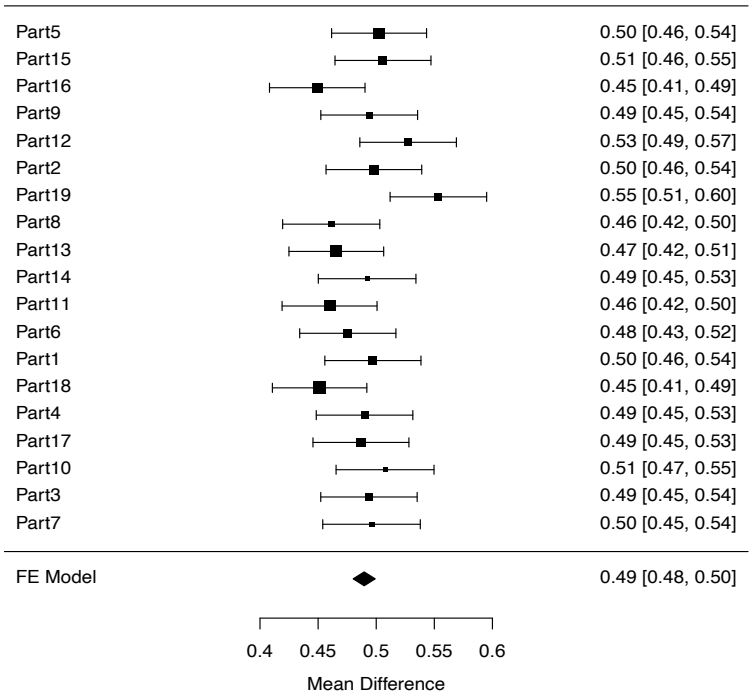


Figura D.10: Forest plot de los coeficientes de YEAR en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.21: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$zED50,sei=resuG8$std.zED50, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 46.2322   12.9273  -90.4645  -89.5201  -90.2292

Test for Heterogeneity:
  Q(df = 18) = 12.9273, p-val = 0.7959

Model Results:
  estimate      se       zval    pval    ci.lb    ci.ub
   -0.7308  0.0057  -127.9093  <.0001  -0.7420  -0.7196
```

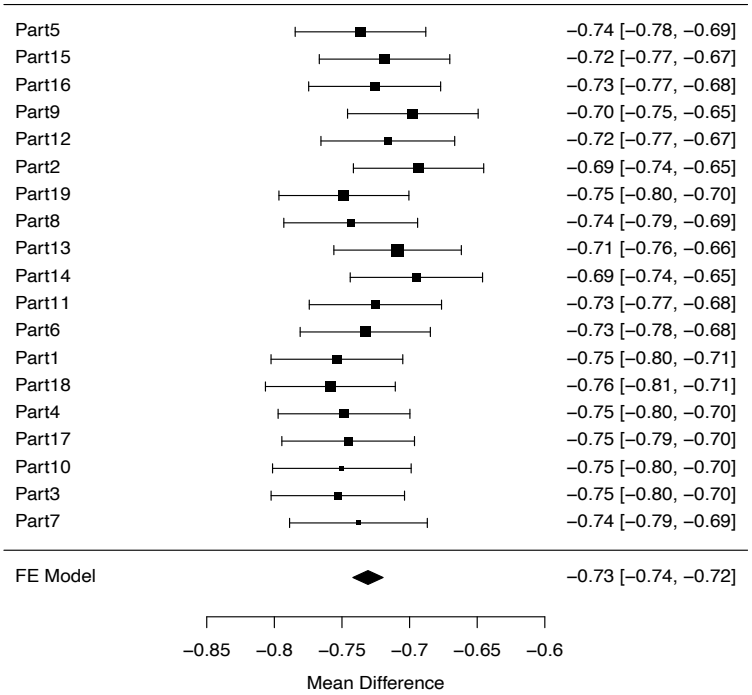


Figura D.11: Forest plot de los coeficientes de EDAD en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.22: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$zPAMED,sei=resuG8$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 29.0452   26.5589  -56.0905  -55.1461  -55.8552

Test for Heterogeneity:
  Q(df = 18) = 26.5589, p-val = 0.0876

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.6094  0.0099  -61.7838  <.0001  -0.6287  -0.5901
```

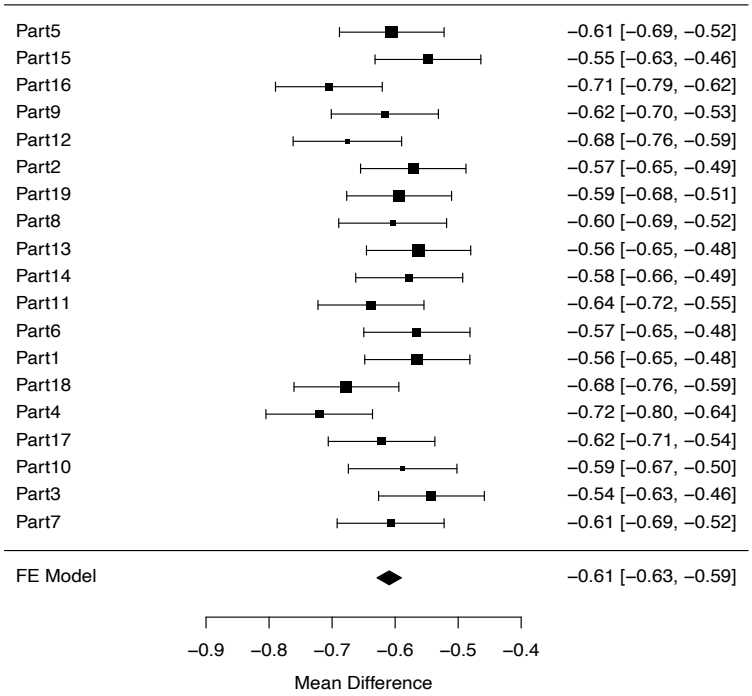


Figura D.12: Forest plot de los coeficientes de PAMED en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.23: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$zPAENF,sei=resuG8$std.zPAENF, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 36.2823   24.2903  -70.5646  -69.6201  -70.3293

Test for Heterogeneity:
  Q(df = 18) = 24.2903, p-val = 0.1457

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.3972  0.0072  55.5231  <.0001  0.3832  0.4112
```

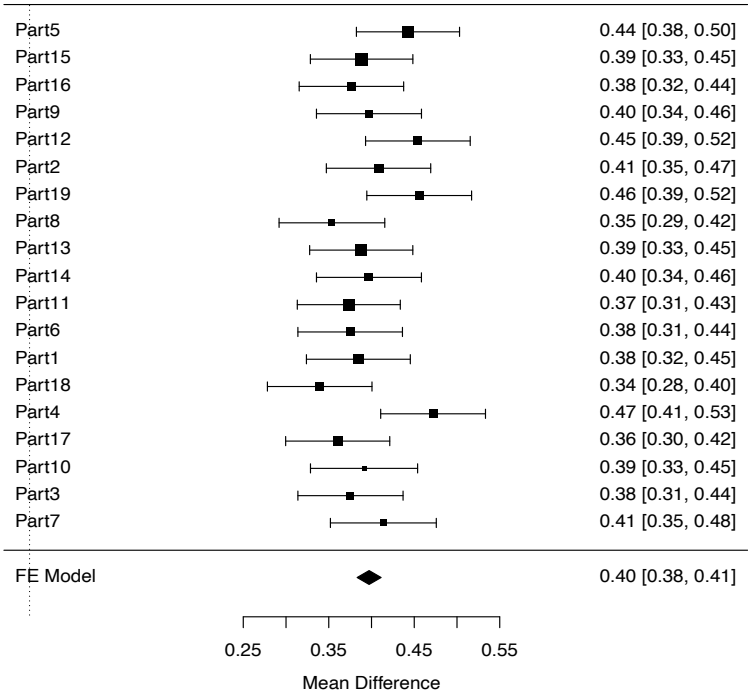


Figura D.13: Forest plot de los coeficientes de PAENF en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.24: Metaanálisis de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuG8$EXTRANJERO,sei=resuG8$std.EXTRANJERO, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 14.5675   10.5083  -27.1349  -26.1905  -26.8996

Test for Heterogeneity:
  Q(df = 18) = 10.5083, p-val = 0.9140

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.7266  0.0322  22.5456  <.0001  0.6635  0.7898
```

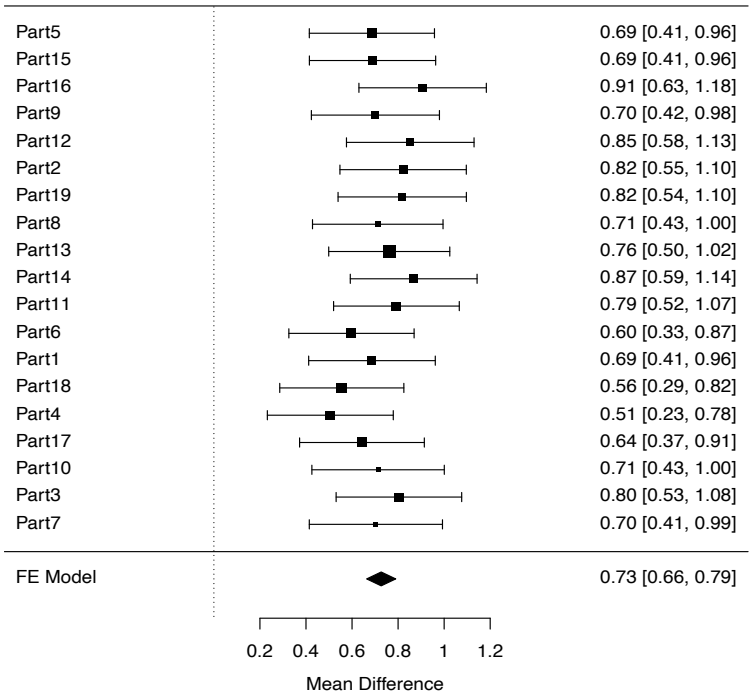


Figura D.14: Forest plot de los coeficientes de EXTRANJERO en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

D.2.1.3 Modelo logístico con efectos aleatorios para los individuos anidados a sus centros médicos.

Resultados de los análisis realizados para la los *missing data* en la variable PESO con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos anidados a cada centro médico. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
   (1|EAP/ID), data=partition, family="binomial", REML=F)
```

La Tabla D.25 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.26 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.26: Metaanálisis de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$Inter,sei=resuG6$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
   5.4763    0.9144   -8.9526   -8.0082   -8.7173

Test for Heterogeneity:
  Q(df = 18) = 0.9144, p-val = 1.0000

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.7933    0.0668  11.8677  <.0001  0.6623  0.9243
```

Tabla D.25: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable PESO realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los individuos anidados al centro médico.

Part	EAP var	Inter	SEXO	YE06	ED50	PAMED	PAENF	EXTRA	std.err Inter	std.err SEXO	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF	std.err EXTRA
9	1.574	0.858	-0.269	0.482	-0.711	-0.544	0.366	0.947	0.296	0.079	0.021	0.025	0.044	0.031	0.141
17	1.416	0.861	-0.329	0.473	-0.659	-0.616	0.385	0.819	0.282	0.075	0.020	0.023	0.041	0.030	0.136
16	1.495	0.735	-0.147	0.474	-0.742	-0.600	0.346	0.677	0.301	0.078	0.021	0.025	0.042	0.031	0.145
3	1.425	0.648	-0.109	0.475	-0.678	-0.616	0.390	0.645	0.292	0.077	0.021	0.024	0.042	0.031	0.139
19	1.205	0.843	-0.214	0.490	-0.722	-0.587	0.387	0.737	0.303	0.079	0.021	0.024	0.043	0.031	0.146
12	1.594	0.750	-0.312	0.528	-0.685	-0.535	0.445	1.067	0.270	0.078	0.021	0.024	0.042	0.031	0.139
14	1.534	0.814	-0.309	0.499	-0.692	-0.533	0.432	0.606	0.272	0.075	0.021	0.023	0.042	0.030	0.137
5	1.472	0.753	-0.115	0.475	-0.728	-0.593	0.384	1.080	0.290	0.078	0.021	0.024	0.042	0.031	0.139
18	1.576	0.835	-0.287	0.492	-0.734	-0.627	0.415	0.840	0.309	0.079	0.021	0.025	0.043	0.031	0.143
8	1.620	0.896	-0.168	0.445	-0.701	-0.553	0.306	0.702	0.285	0.077	0.021	0.024	0.042	0.031	0.141
6	1.704	0.896	-0.327	0.458	-0.720	-0.584	0.364	0.797	0.301	0.079	0.021	0.024	0.042	0.031	0.140
7	1.561	0.729	-0.156	0.506	-0.709	-0.625	0.399	0.680	0.296	0.079	0.021	0.024	0.043	0.031	0.143
1	1.576	0.754	-0.165	0.483	-0.691	-0.602	0.408	0.792	0.292	0.078	0.021	0.024	0.042	0.031	0.142
2	1.318	0.745	-0.140	0.497	-0.710	-0.499	0.376	0.796	0.263	0.077	0.021	0.024	0.042	0.031	0.141
10	1.456	0.766	-0.114	0.491	-0.722	-0.629	0.427	0.933	0.307	0.079	0.021	0.024	0.043	0.031	0.148
13	1.628	0.857	-0.246	0.462	-0.675	-0.623	0.412	0.728	0.303	0.077	0.021	0.024	0.042	0.030	0.141
4	1.788	0.766	-0.187	0.486	-0.708	-0.528	0.387	0.709	0.297	0.076	0.020	0.024	0.042	0.030	0.140
11	1.344	0.761	-0.165	0.469	-0.732	-0.634	0.375	0.644	0.302	0.078	0.021	0.024	0.043	0.030	0.141
15	1.597	0.819	-0.280	0.495	-0.657	-0.631	0.414	0.702	0.289	0.077	0.021	0.023	0.042	0.031	0.137

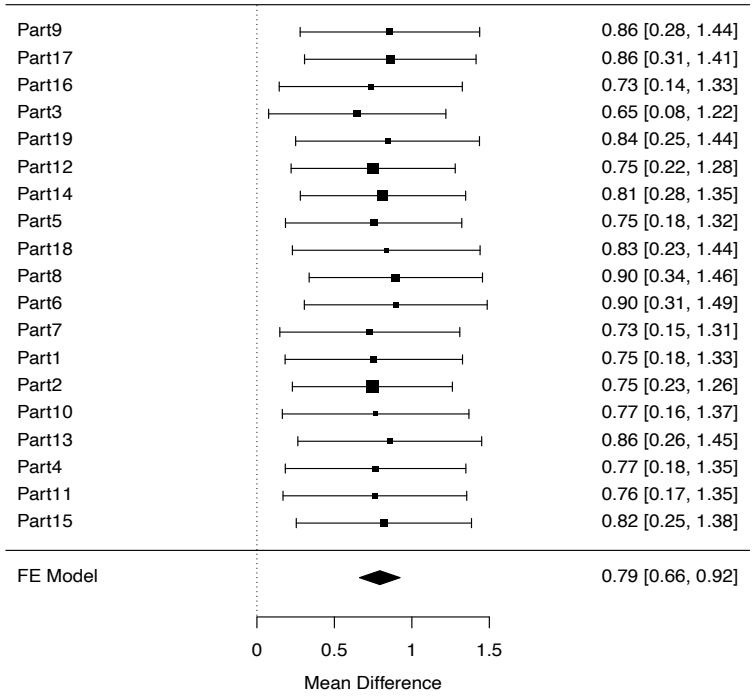


Figura D.15: Forest plot de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

Tabla D.27: Metaanálisis de los coeficientes de SEXO obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$SEXO,sei=resuG6$std.SEXO, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  21.9176   18.3924  -41.8352  -40.8907  -41.5999

Test for Heterogeneity:
  Q(df = 18) = 18.3924, p-val = 0.4301

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.2129  0.0178  -11.9632  <.0001  -0.2477  -0.1780
```

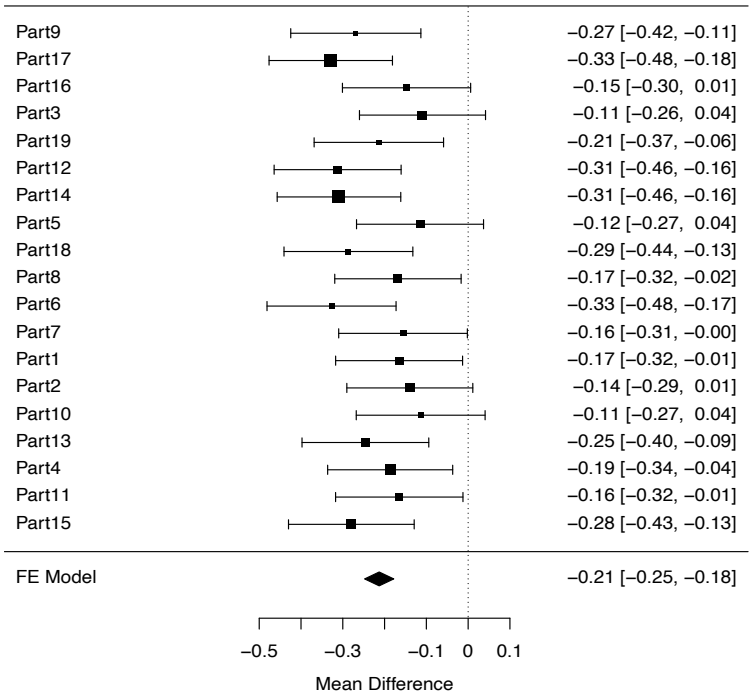


Figura D.16: Forest plot de los coeficientes de SEXO de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

Tabla D.28: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$zYE06,sei=resuG6$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  48.8487   14.6457  -95.6973  -94.7529  -95.4620

Test for Heterogeneity:
  Q(df = 18) = 14.6457, p-val = 0.6861

Model Results:
  estimate      se      zval    pval   ci.lb   ci.ub
    0.4830  0.0048  101.4948  <.0001  0.4737  0.4924
```

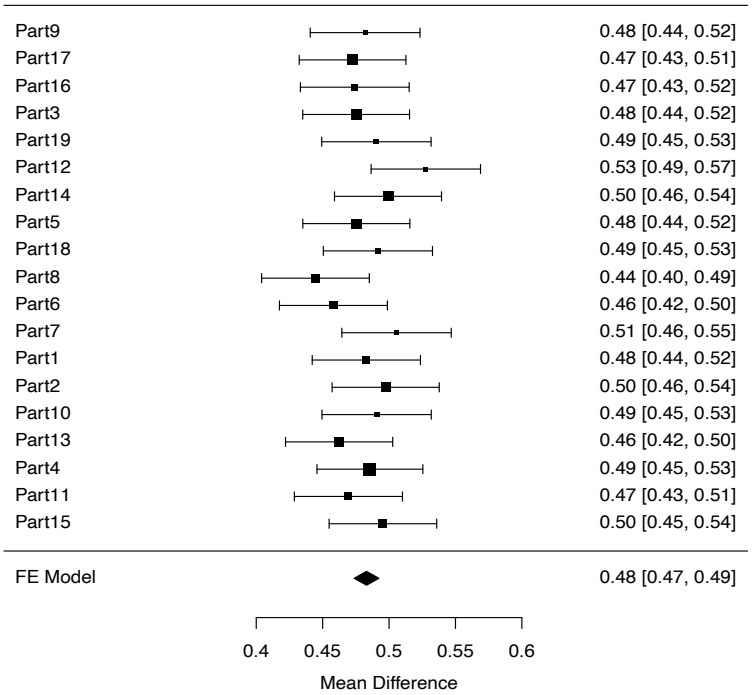


Figura D.17: Forest plot de los coeficientes de YEAR de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

Tabla D.29: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$zED50,sei=resuG6$std.zED50, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  43.5206   19.7975  -85.0411  -84.0967  -84.8058

Test for Heterogeneity:
  Q(df = 18) = 19.7975, p-val = 0.3443

Model Results:
  estimate      se       zval    pval    ci.lb    ci.ub
   -0.7032  0.0055  -127.8602  <.0001  -0.7139  -0.6924
```

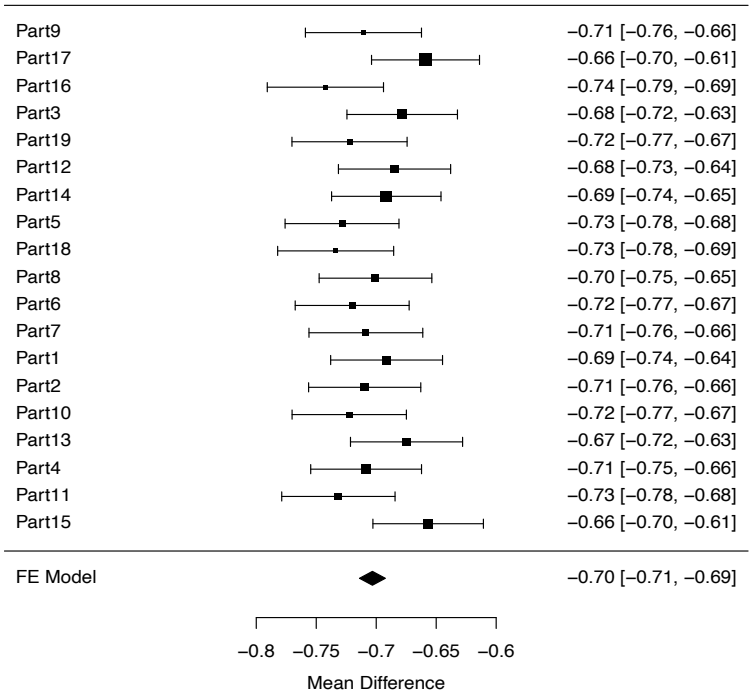


Figura D.18: Forest plot de los coeficientes de EDAD de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

Tabla D.30: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$zPAMED,sei=resuG6$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  33.6086   17.9876  -65.2172  -64.2728  -64.9819

Test for Heterogeneity:
  Q(df = 18) = 17.9876, p-val = 0.4565

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.5871  0.0097  -60.3977  <.0001  -0.6061  -0.5680
```

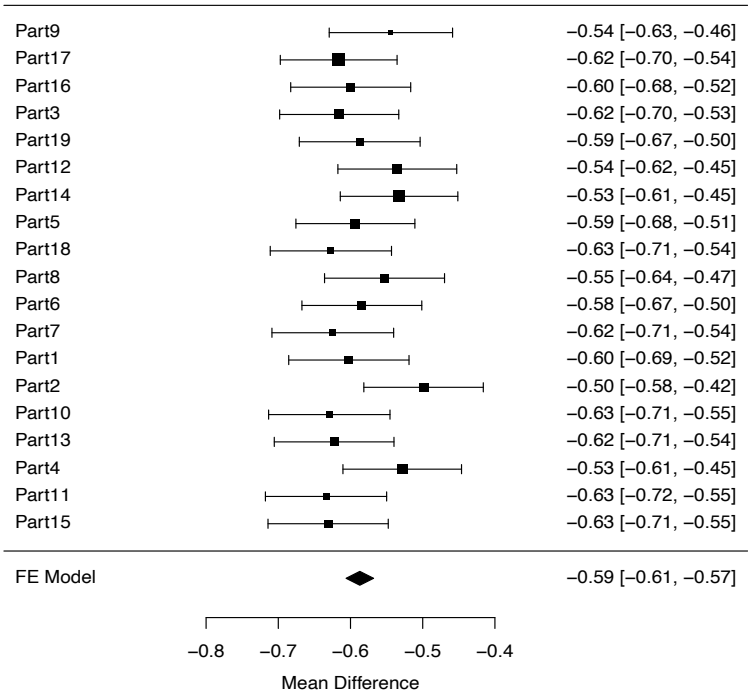


Figura D.19: Forest plot de los coeficientes de PAMED de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

Tabla D.31: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$zPAENF,sei=resuG6$std.zPAENF, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  38.6642   20.0040  -75.3284  -74.3840  -75.0931

Test for Heterogeneity:
  Q(df = 18) = 20.0040, p-val = 0.3326

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.3904  0.0071  55.2614  <.0001  0.3766  0.4043
```

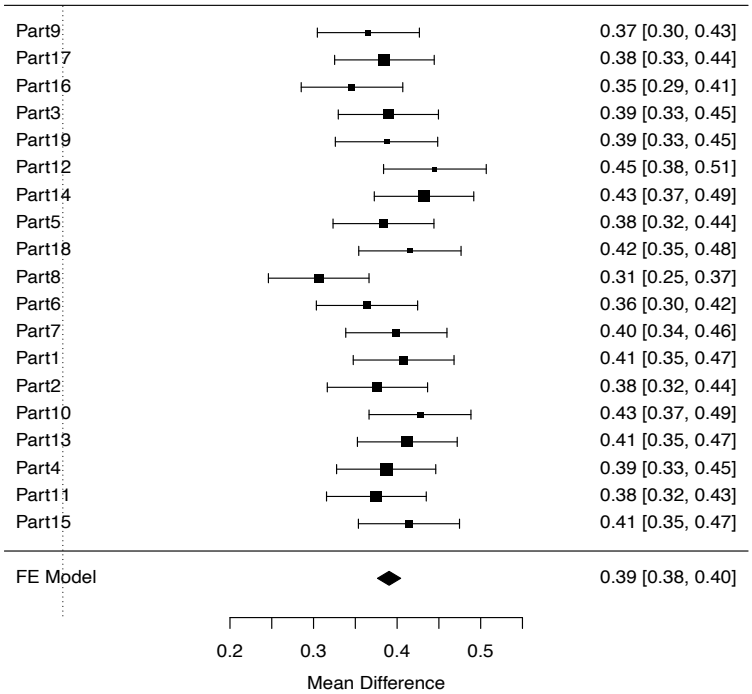


Figura D.20: Forest plot de los coeficientes de PAENF de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

Tabla D.32: Metaanálisis de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG6$EXTRANJERO,sei=resuG6$std.EXTRANJERO, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  11.1461   17.2477  -20.2923  -19.3478  -20.0570

Test for Heterogeneity:
  Q(df = 18) = 17.2477, p-val = 0.5061

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.7842  0.0323  24.2651  <.0001  0.7209  0.8475
```

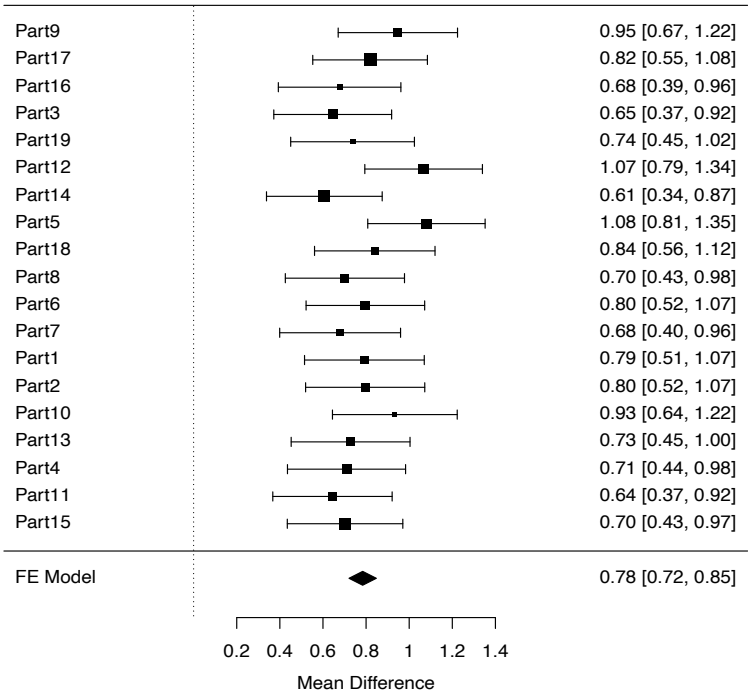


Figura D.21: Forest plot de los coeficientes de EXTRANJERO de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en PESO.

D.2.1.4 Modelo logístico con efectos aleatorios para los centros médicos.

Resultados de los análisis de los *missings* en la variable PESO con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios de los centros médicos, no para los individuos.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO +
        (1|EAP), data=partition, family="binomial", REML=F)
```

La Tabla D.33 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.34 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.33: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable PESO realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los centros médicos.

Part	EAP var	Inter	SEXO	YE06	ED50	PAMED	PAENF	EXTRA	std.err Inter	std.err SEXO	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF	std.err EXTRA
6	0.336	0.229	-0.063	0.208	-0.273	-0.257	0.170	0.300	0.132	0.020	0.013	0.006	0.028	0.021	0.037
12	0.315	0.248	-0.107	0.218	-0.271	-0.268	0.183	0.312	0.131	0.020	0.013	0.006	0.028	0.021	0.037
15	0.329	0.229	-0.059	0.213	-0.266	-0.275	0.172	0.312	0.133	0.020	0.013	0.006	0.028	0.021	0.037
17	0.318	0.179	-0.061	0.240	-0.279	-0.256	0.223	0.326	0.130	0.020	0.013	0.006	0.028	0.021	0.037
3	0.282	0.240	-0.078	0.214	-0.272	-0.289	0.167	0.276	0.135	0.020	0.013	0.006	0.028	0.021	0.036
16	0.357	0.186	-0.051	0.223	-0.263	-0.247	0.171	0.339	0.129	0.020	0.013	0.006	0.028	0.021	0.037
13	0.335	0.236	-0.060	0.206	-0.278	-0.267	0.152	0.337	0.139	0.020	0.013	0.006	0.028	0.021	0.037
19	0.319	0.232	-0.087	0.223	-0.273	-0.253	0.203	0.304	0.131	0.020	0.013	0.006	0.028	0.021	0.037
2	0.345	0.224	-0.056	0.217	-0.274	-0.306	0.204	0.305	0.135	0.020	0.013	0.006	0.028	0.021	0.037
8	0.343	0.243	-0.048	0.202	-0.266	-0.265	0.176	0.273	0.132	0.020	0.013	0.006	0.028	0.021	0.037
4	0.357	0.197	-0.052	0.230	-0.283	-0.290	0.194	0.282	0.136	0.020	0.013	0.006	0.028	0.021	0.037
7	0.337	0.248	-0.080	0.212	-0.268	-0.278	0.173	0.321	0.136	0.020	0.013	0.006	0.028	0.021	0.037
18	0.349	0.200	-0.044	0.226	-0.269	-0.310	0.211	0.260	0.139	0.020	0.013	0.006	0.028	0.021	0.037
5	0.293	0.241	-0.075	0.218	-0.271	-0.274	0.183	0.263	0.137	0.020	0.013	0.006	0.028	0.021	0.037
11	0.323	0.310	-0.092	0.185	-0.272	-0.326	0.166	0.289	0.139	0.020	0.013	0.006	0.028	0.021	0.037
9	0.350	0.252	-0.060	0.200	-0.265	-0.302	0.173	0.305	0.135	0.020	0.013	0.006	0.029	0.021	0.037
10	0.382	0.267	-0.101	0.200	-0.272	-0.317	0.165	0.289	0.140	0.020	0.013	0.006	0.028	0.021	0.036
1	0.302	0.201	-0.043	0.220	-0.268	-0.278	0.183	0.299	0.133	0.020	0.013	0.006	0.028	0.021	0.037
14	0.342	0.185	-0.051	0.230	-0.273	-0.281	0.189	0.321	0.133	0.020	0.013	0.006	0.028	0.021	0.037

Tabla D.34: Metaanálisis de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$Inter,sei=resuG5$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
  20.1665    1.0204  -38.3329  -37.3885  -38.0976

Test for Heterogeneity:
  Q(df = 18) = 1.0204, p-val = 1.0000

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.2280  0.0308  7.4022  <.0001  0.1677  0.2884
```

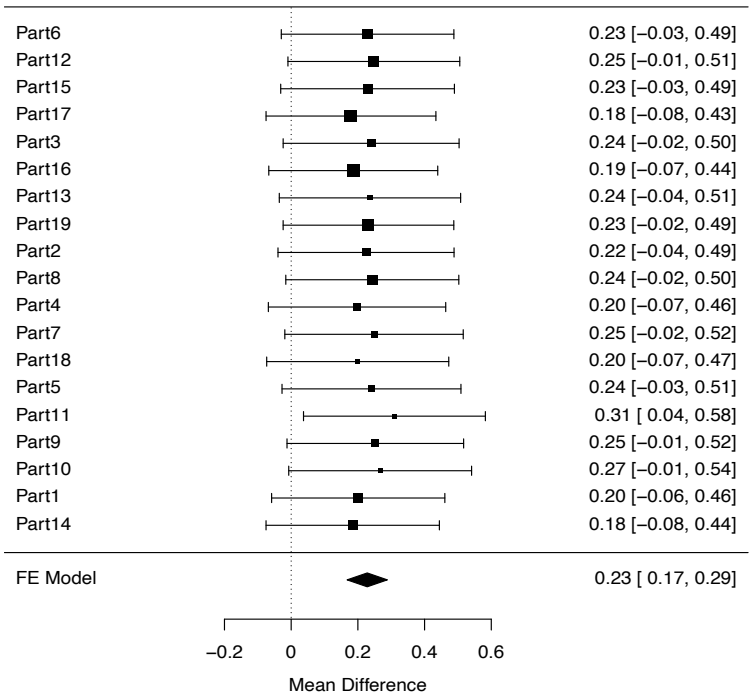


Figura D.22: Forest plot de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

Tabla D.35: Metaanálisis de los coeficientes de SEXO obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$SEXO,sei=resuG5$std.SEXO, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 48.4744   17.2757  -94.9488  -94.0044  -94.7135

Test for Heterogeneity:
  Q(df = 18) = 17.2757, p-val = 0.5042

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.0667  0.0045  -14.7292  <.0001  -0.0756  -0.0578
```

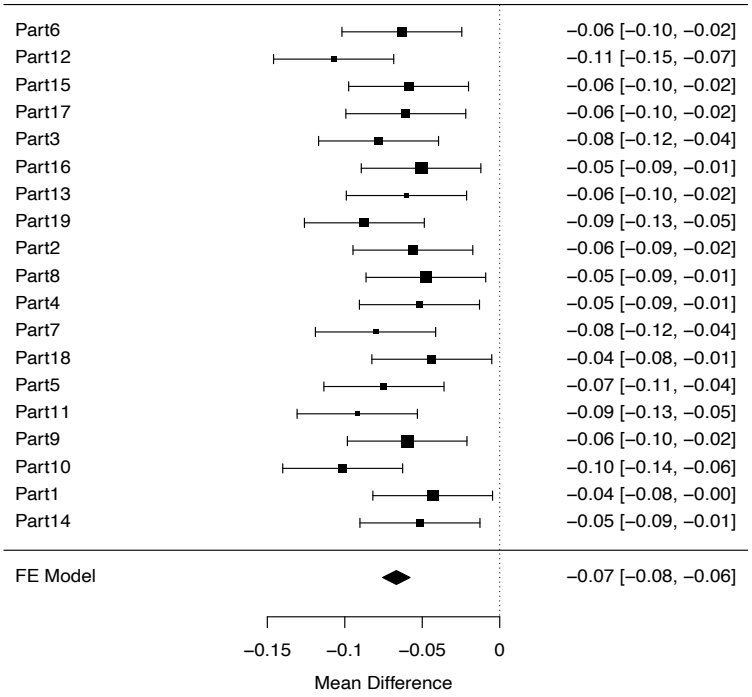


Figura D.23: Forest plot de los coeficientes de SEXO de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en PESO.

Tabla D.36: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$zYE06,sei=resuG5$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik   deviance        AIC        BIC       AICc
  56.0472    16.8122  -110.0944  -109.1500  -109.8591

Test for Heterogeneity:
  Q(df = 18) = 16.8122, p-val = 0.5360

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.2149  0.0031  69.8263  <.0001  0.2089  0.2210
```

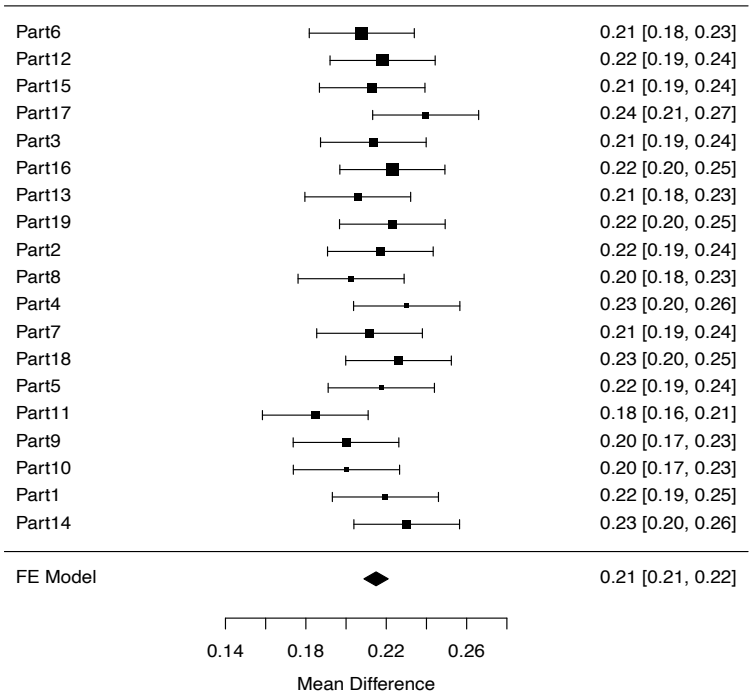


Figura D.24: Forest plot de los coeficientes de YEAR de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en PESO.

Tabla D.37: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$zED50,sei=resuG5$std.zED50, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
   logLik   deviance        AIC        BIC       AICc
  73.9674    13.7341  -145.9348  -144.9904  -145.6996

Test for Heterogeneity:
  Q(df = 18) = 13.7341, p-val = 0.7463

Model Results:
  estimate      se       zval    pval    ci.lb    ci.ub
   -0.2714  0.0013  -208.8537  <.0001  -0.2740  -0.2689
```

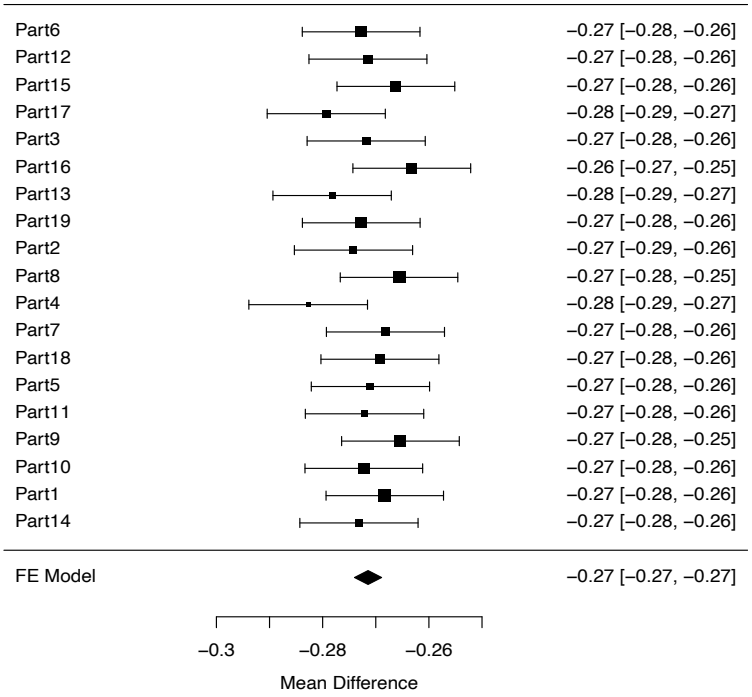


Figura D.25: Forest plot de los coeficientes de EDAD de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en PESO.

Tabla D.38: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$zPAMED,sei=resuG5$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 44.5490   11.6059  -87.0981  -86.1537  -86.8628

Test for Heterogeneity:
  Q(df = 18) = 11.6059, p-val = 0.8669

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.2809  0.0065  -43.4512  <.0001  -0.2936  -0.2683
```

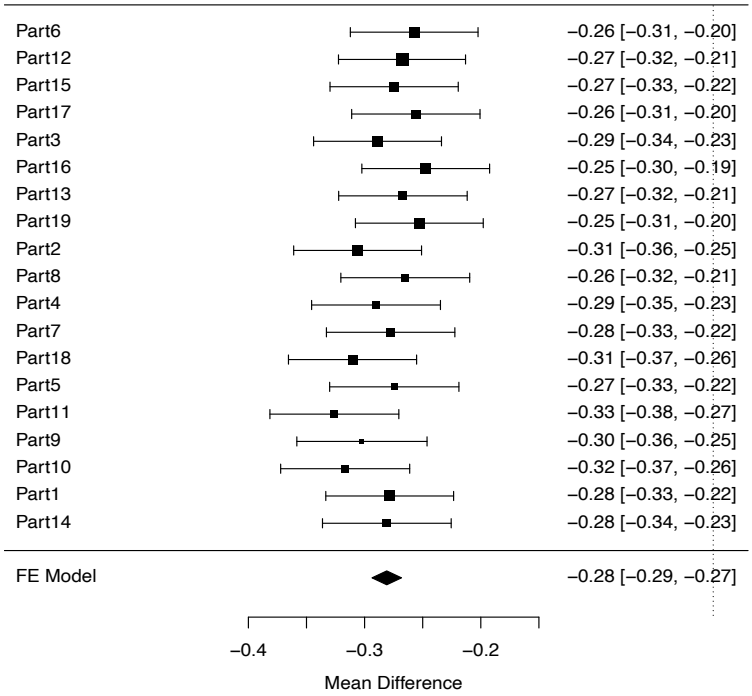


Figura D.26: Forest plot de los coeficientes de PAMED de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en PESO.

Tabla D.39: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$zPAENF,sei=resuG5$std.zPAENF, method="FE", measure="MD")


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 49.2899   13.9005  -96.5798  -95.6354  -96.3445

Test for Heterogeneity:
  Q(df = 18) = 13.9005, p-val = 0.7356

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.1820  0.0047  38.3831  <.0001  0.1727  0.1913
```

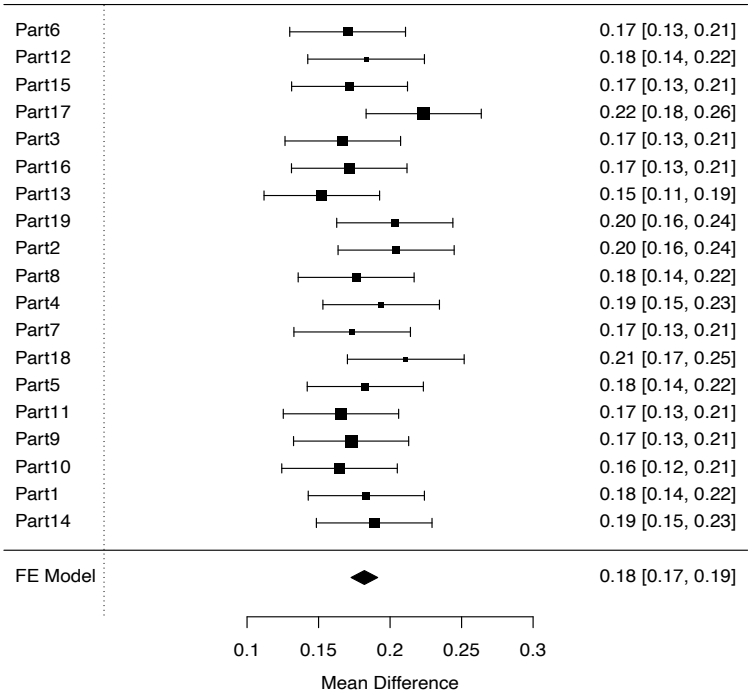


Figura D.27: Forest plot de los coeficientes de PAENF de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en PESO.

Tabla D.40: Metaanálisis de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en PESO.

```
rma(resuG5$EXTRANJERO,sei=resuG5$std.EXTRANJERO, method="FE", measure="MD"))

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 41.7915    7.1399  -81.5830  -80.6385  -81.3477

Test for Heterogeneity:
  Q(df = 18) = 7.1399, p-val = 0.9889

Model Results:
  estimate      se     zval    pval   ci.lb   ci.ub
    0.3005  0.0084  35.7473  <.0001  0.2841  0.3170
```

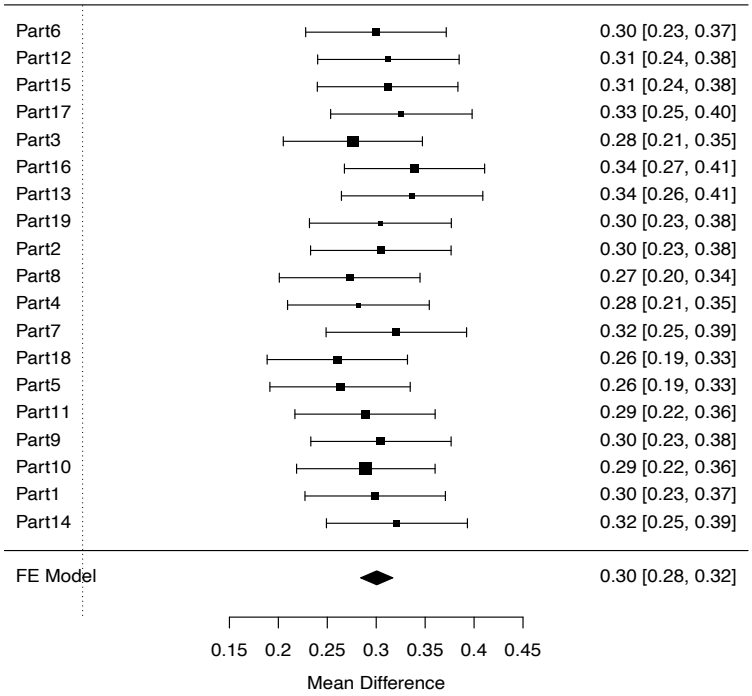


Figura D.28: Forest plot de los coeficientes de EXTRANJERO de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en PESO.

D.2.2 *Missing data* en la variable Cigarrillos

D.2.2.1 Modelo logístico simple.

Resultados de los análisis de los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios.

```
glm(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + EAP, data=partition,
       family="binomial")
```

La Tabla D.41 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.42 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.42: Metaanálisis de los Interceptos obtenidos en cada partición, en las que se ajustaron regresiones logísticas simples para la *odds* de *missing data* en CIGARRILLOS.

```
rma(resuH3$Inter,sei=resuH3$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 -4.7785   12.6172   11.5571   12.5015   11.7924

Test for Heterogeneity:
  Q(df = 18) = 12.6172, p-val = 0.8138

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -2.2114  0.0819  -26.9938  <.0001  -2.3720  -2.0508
```

Tabla D.41: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable CIGARRILLOS realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística simple.

Part	Inter	YE06	ED50	PAMED	PAENF	std.err Inter	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF
4	-2.314	-0.229	0.063	0.177	-0.264	0.354	0.067	0.033	0.146	0.107
16	-2.024	-0.232	0.093	0.094	-0.147	0.340	0.063	0.031	0.135	0.096
9	-2.144	-0.189	0.069	0.045	-0.022	0.351	0.066	0.034	0.154	0.106
1	-2.359	-0.169	0.098	0.131	-0.205	0.354	0.067	0.033	0.147	0.104
19	-2.700	-0.247	0.069	0.068	-0.163	0.434	0.068	0.034	0.142	0.112
15	-3.595	-0.164	0.034	0.024	0.102	0.721	0.066	0.033	0.142	0.104
7	-1.900	-0.131	0.106	0.227	-0.071	0.310	0.070	0.035	0.149	0.110
18	-1.943	-0.211	0.025	0.181	-0.304	0.302	0.069	0.034	0.144	0.112
14	-2.116	-0.168	0.032	0.194	-0.125	0.326	0.068	0.033	0.143	0.105
13	-1.949	-0.161	0.095	0.061	-0.123	0.307	0.070	0.033	0.149	0.106
5	-1.958	-0.368	0.065	-0.253	-0.315	0.333	0.069	0.033	0.149	0.106
8	-2.107	-0.289	0.065	0.065	-0.353	0.332	0.067	0.033	0.150	0.112
3	-2.265	-0.039	0.050	0.262	0.202	0.345	0.065	0.033	0.137	0.097
2	-2.309	-0.232	0.055	0.080	-0.190	0.364	0.064	0.033	0.146	0.097
11	-2.517	-0.144	0.092	0.094	0.108	0.401	0.067	0.033	0.145	0.107
10	-2.628	-0.180	0.004	0.145	-0.038	0.433	0.064	0.033	0.138	0.109
6	-2.185	-0.043	0.099	0.048	0.207	0.347	0.065	0.033	0.143	0.105
17	-2.900	-0.269	0.096	-0.147	-0.097	0.468	0.065	0.033	0.144	0.102
12	-2.188	-0.189	0.074	0.112	-0.268	0.331	0.068	0.033	0.142	0.107

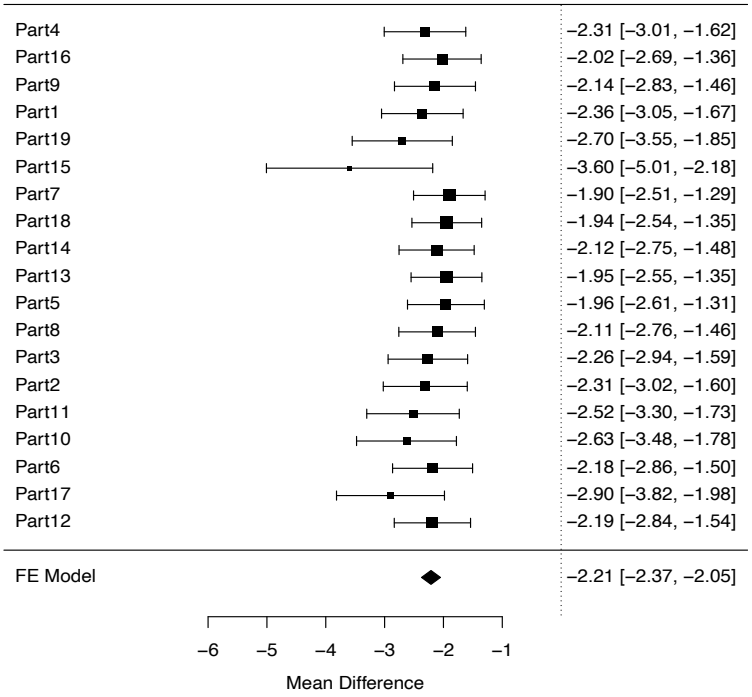


Figura D.29: Forest plot de los Interceptos de las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

Tabla D.43: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH3$zYE06,sei=resuH3$std.zYE06, method="FE", measure="MD"))

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 21.6798   24.5817  -41.3597  -40.4152  -41.1244

Test for Heterogeneity:
  Q(df = 18) = 24.5817, p-val = 0.1369

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.1915  0.0153  -12.5167  <.0001  -0.2215  -0.1615
```

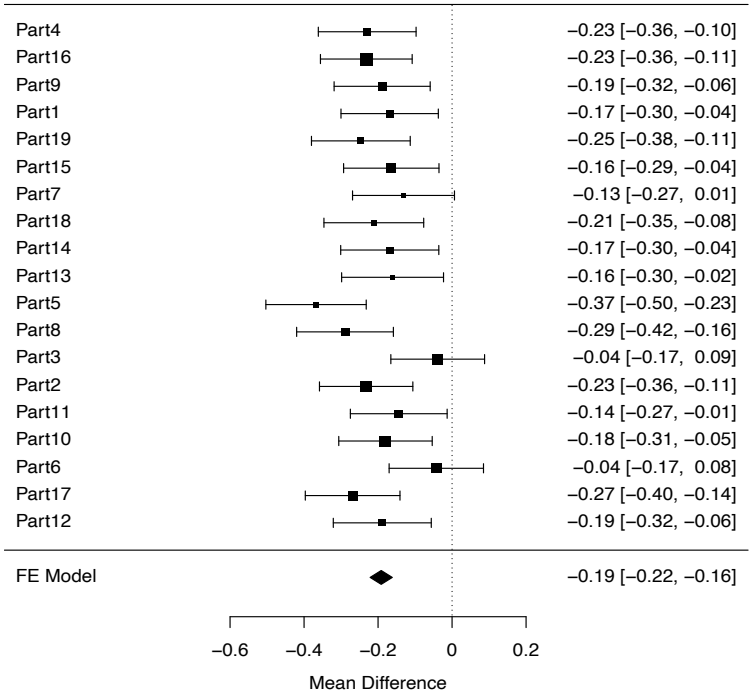


Figura D.30: Forest plot de los coeficientes de SEXO obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

Tabla D.44: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH3$zED50,sei=resuH3$std.zED50, method="FE", measure="MD"))


Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 40.5051   13.4901  -79.0101  -78.0657  -78.7748

Test for Heterogeneity:
  Q(df = 18) = 13.4901, p-val = 0.7617

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.0675  0.0076  8.8717  <.0001  0.0526  0.0824
```

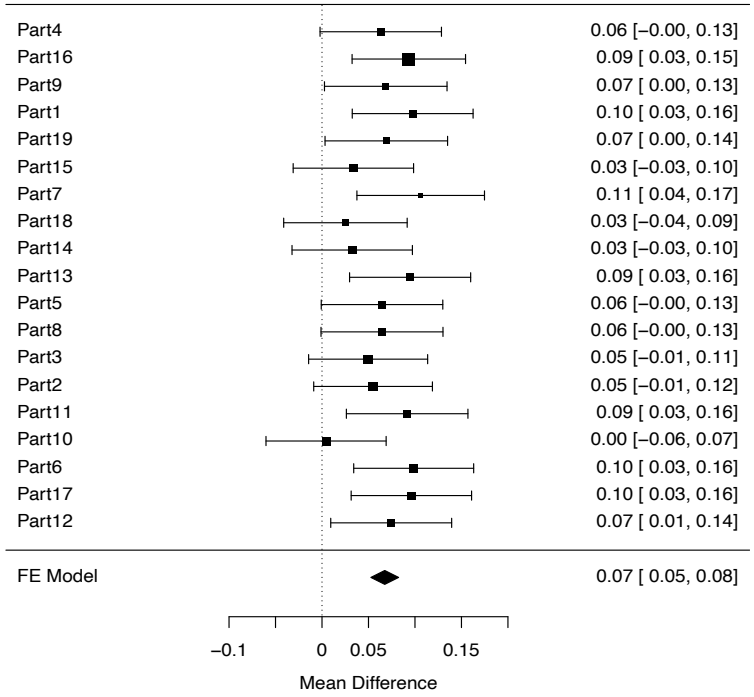


Figura D.31: Forest plot de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

Tabla D.45: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH3$zPAMED,sei=resuH3$std.zPAMED, method="FE", measure="MD"))

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 13.0764   12.4622  -24.1527  -23.2083  -23.9174

Test for Heterogeneity:
  Q(df = 18) = 12.4622, p-val = 0.8225

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.0867  0.0331  2.6188  0.0088  0.0218  0.1515
```

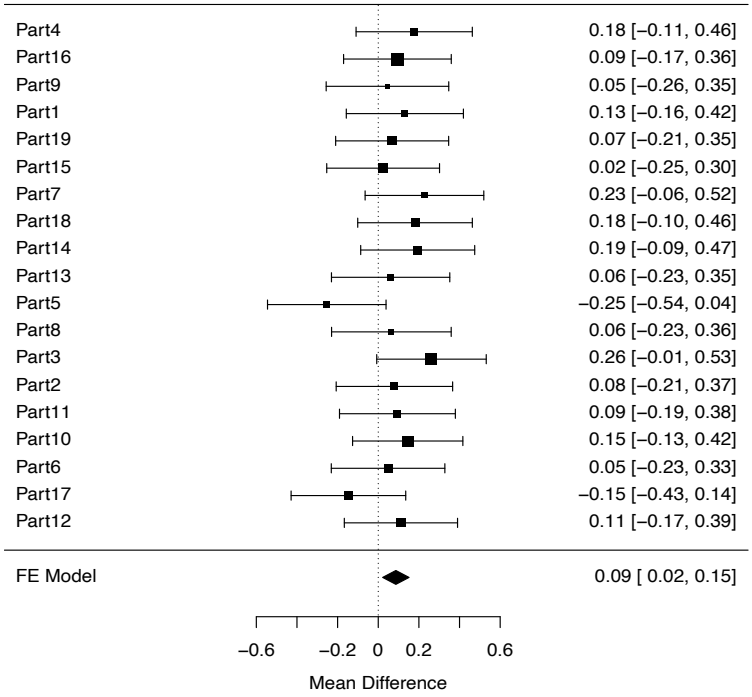


Figura D.32: Forest plot de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

Tabla D.46: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH3$zPAENF,sei=resuH3$std.zPAENF, method="FE", measure="MD"))

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
  2.0714   46.4412   -2.1427   -1.1983   -1.9074

Test for Heterogeneity:
  Q(df = 18) = 46.4412, p-val = 0.0003

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.1043  0.0241  -4.3226  <.0001  -0.1516  -0.0570
```

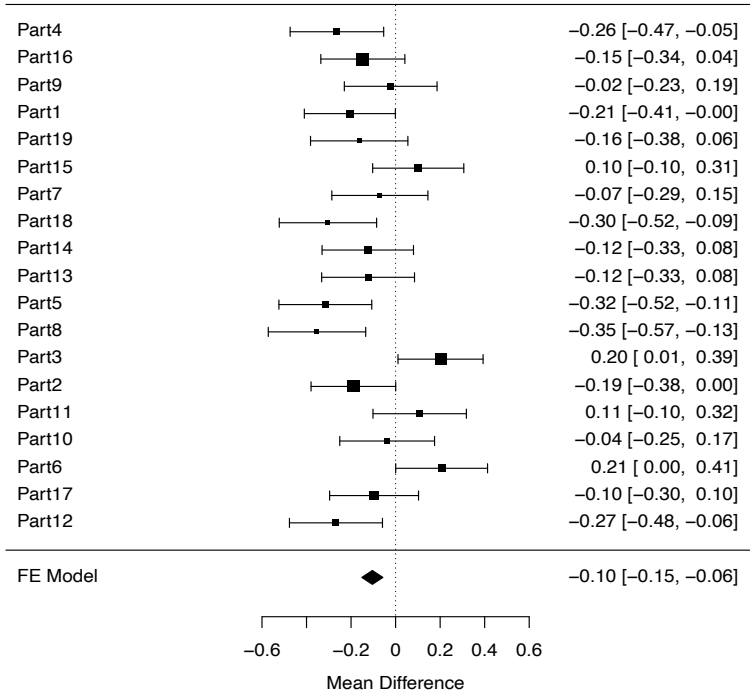


Figura D.33: Forest plot de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-partición-aleatoria para la *odds* de *missings* en CIGARRILLOS.

D.2.2.2 Modelo logístico con efectos aleatorios para los individuos.

Resultados de los análisis realizados para la los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos, sin anidar. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mCIGARRILLOS~zYE06+zED50+zPAMED+zPAENF+(1|EAP/ID),
          data=partition, family="binomial", REML=F)
```

La Tabla D.47 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción “FE”, Tabla D.48 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.47: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable CIGARRILLOS realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los individuos.

Part	Inter	YE06	ED50	PAMED	PAENF	std.err Inter	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF
4	-11.832	-3.718	0.110	0.599	-0.609	0.826	1.002	0.335	0.543	0.603
1	-11.248	-1.920	0.183	0.588	-0.775	0.649	0.344	0.297	0.421	0.499
14	-11.378	-2.543	0.174	-0.121	-0.267	0.653	0.448	0.297	0.368	0.507
10	-11.318	-1.358	0.152	0.528	-0.290	0.686	0.268	0.295	0.416	0.354
16	-12.102	-1.607	0.186	-0.339	0.504	0.739	0.332	0.311	0.399	0.517
9	-11.973	-3.187	0.131	0.644	-0.680	0.795	0.758	0.326	0.518	0.597
12	-16.019	-15.872	0.137	0.117	-0.359	1.177	1.322	0.557	0.729	0.801
13	-11.537	-1.735	0.068	0.208	-0.298	0.658	0.342	0.310	0.385	0.516
18	-11.429	-1.676	0.262	0.268	-0.523	0.619	0.332	0.291	0.383	0.428
7	-10.734	-2.214	0.105	0.651	-0.831	0.584	0.340	0.257	0.372	0.405
3	-11.764	-1.682	0.203	-0.156	0.053	0.682	0.334	0.309	0.386	0.484
8	-11.046	-1.595	0.036	0.263	-0.774	0.624	0.287	0.310	0.406	0.433
2	-11.563	-2.346	0.212	0.469	-0.336	0.730	0.430	0.311	0.447	0.492
15	-11.724	-1.535	0.030	0.273	-0.052	0.729	0.344	0.330	0.436	0.465
11	-10.910	-1.601	0.050	0.265	-0.784	0.563	0.273	0.285	0.357	0.440
6	-11.291	-2.454	0.121	0.358	-0.809	0.631	0.435	0.297	0.413	0.560
17	-11.242	-1.793	0.019	0.216	-0.566	0.599	0.308	0.287	0.381	0.419
5	-11.295	-1.183	0.054	0.756	-0.557	0.642	0.241	0.269	0.378	0.383
19	-11.401	-2.582	0.263	0.143	-0.249	0.744	0.465	0.332	0.434	0.600

Tabla D.48: Metaanálisis de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH4$Inter,sei=resuH4$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
    logLik  deviance       AIC       BIC      AICc
  -20.5210   20.1290   43.0419   43.9864   43.2772

Test for Heterogeneity:
  Q(df = 18) = 20.1290, p-val = 0.3256

Model Results:
  estimate      se      zval    pval     ci.lb     ci.ub
  -11.4551  0.1553  -73.7611  <.0001  -11.7595  -11.1507
```

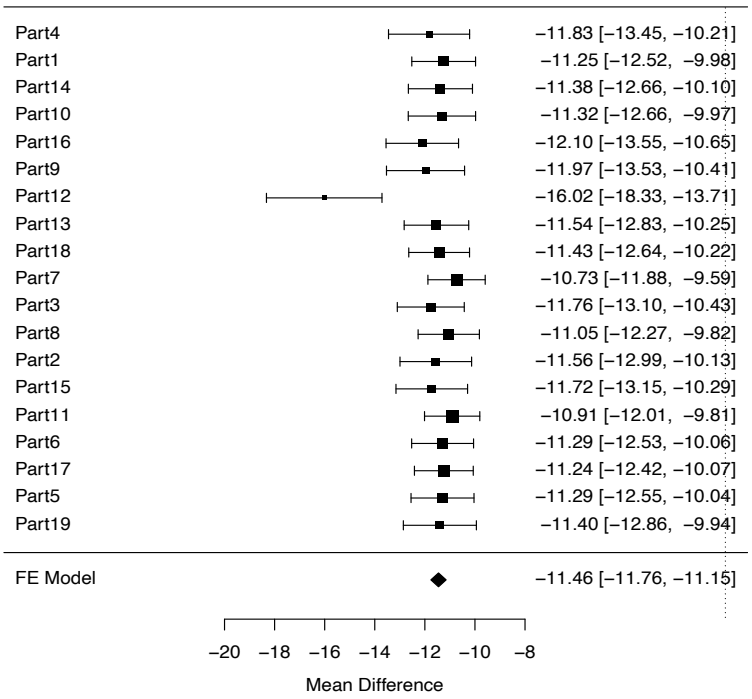


Figura D.34: Forest plot de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.49: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH4$zYE06,sei=resuH4$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
 -71.6091  142.8986  145.2182  146.1626  145.4535

Test for Heterogeneity:
  Q(df = 18) = 142.8986, p-val < .0001

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -1.8126  0.0811  -22.3639  <.0001  -1.9715  -1.6538
```

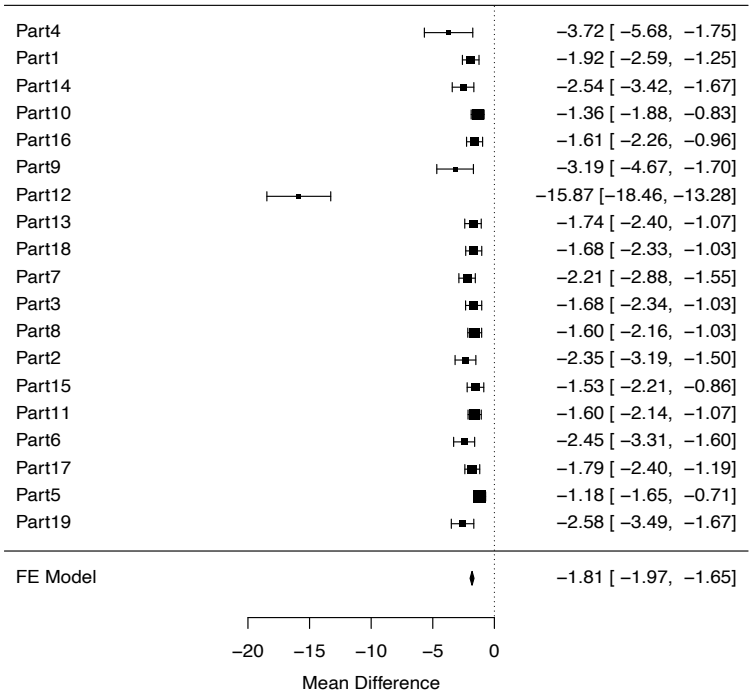


Figura D.35: Forest plot de los coeficientes de YEAR en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.50: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH4$zED50,sei=resuH4$std.zED50, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
  4.1079    1.1337   -6.2158   -5.2714   -5.9805

Test for Heterogeneity:
  Q(df = 18) = 1.1337, p-val = 1.0000

Model Results:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.1288  0.0703  1.8328  0.0668  -0.0089  0.2666
```

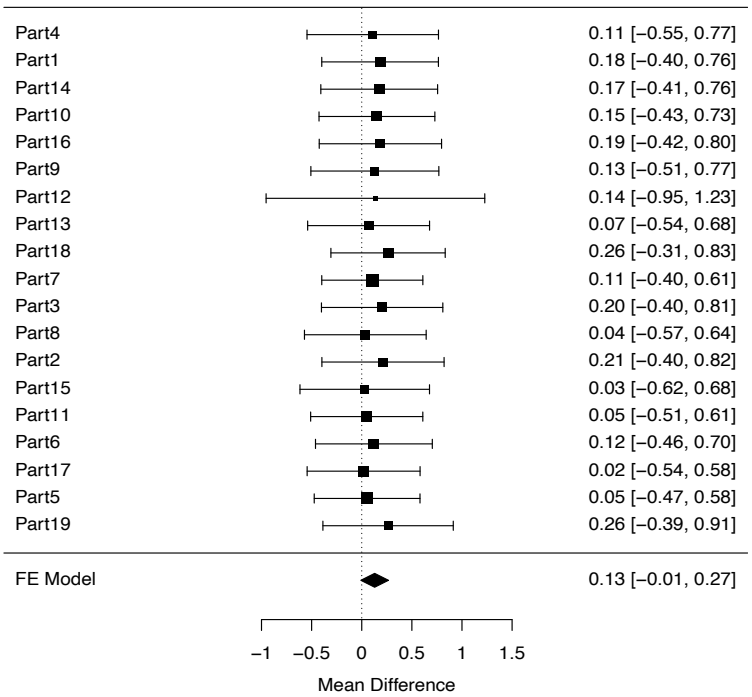


Figura D.36: Forest plot de los coeficientes de EDAD en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.51: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH4$zPAMED,sei=resuH4$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 -5.9413    9.6022   13.8826   14.8270   14.1179

Test for Heterogeneity:
  Q(df = 18) = 9.6022, p-val = 0.9441

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.2888  0.0950  3.0391  0.0024  0.1025  0.4750
```

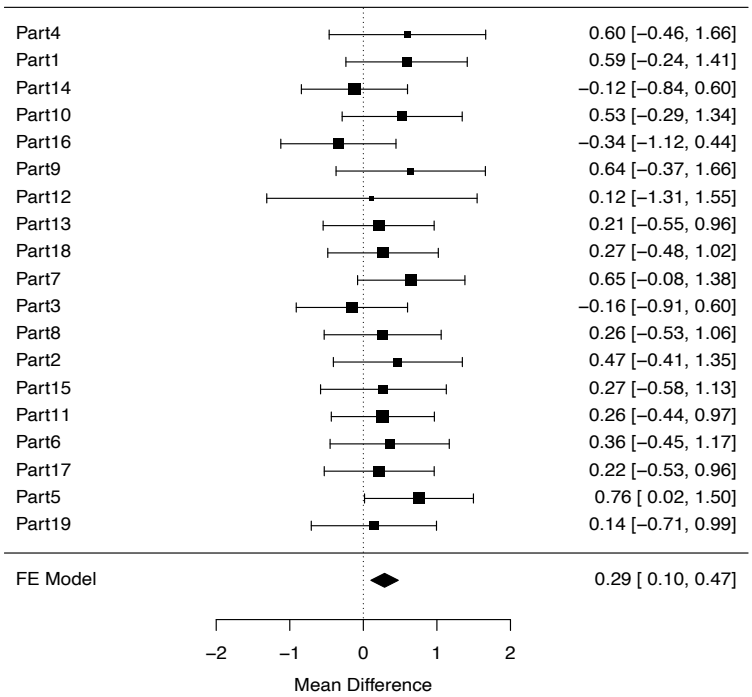


Figura D.37: Forest plot de los coeficientes de PAMED en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.52: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH4$zPAENF,sei=resuH4$std.zPAENF, method="FE", measure="MD"))

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 -8.4871    9.0656   18.9742   19.9186   19.2095

Test for Heterogeneity:
  Q(df = 18) = 9.0656, p-val = 0.9582

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.4479  0.1091  -4.1055  <.0001  -0.6617  -0.2341
```

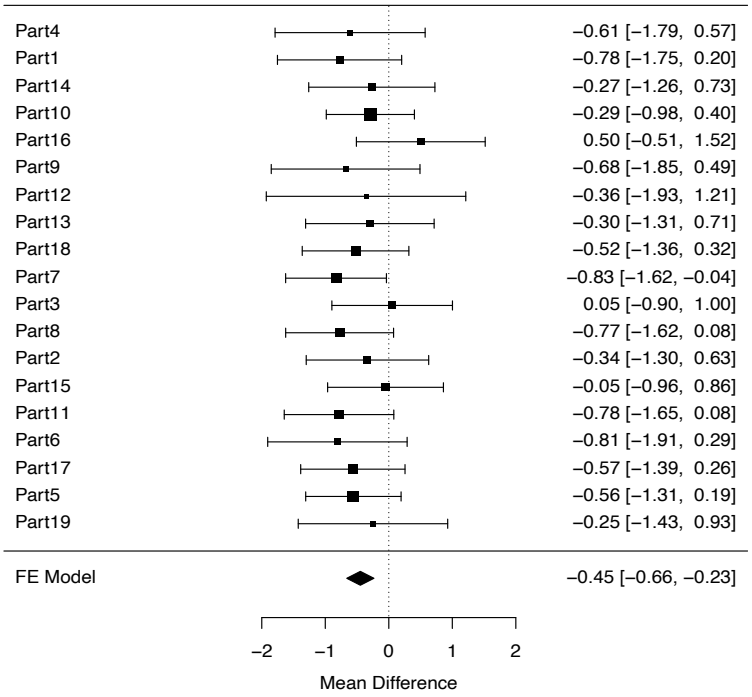


Figura D.38: Forest plot de los coeficientes de PAENF en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

D.2.2.3 Modelo logístico con efectos aleatorios para los individuos anidados a los centros.

Resultados de los análisis realizados para la los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos anidados a cada centro médico. Los datos se han particionado incluyendo todas las medidas de un individuo en la misma partición.

```
glmmTMB(mCIGARRILLOS~zYE06+zED50+zPAMED+zPAENF+(1|EAP/ID),
          data=partition, family="binomial", REML=F)
```

La Tabla D.53 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. En este ajuste R nos ha avisado de numerosos errores de convergencia, por lo que los resultados no son fiables. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.54 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.54: Metaanálisis de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH1$Inter,sei=resuH1$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
 -27.6709   32.3723   57.3418   58.2863   57.5771

Test for Heterogeneity:
  Q(df = 18) = 32.3723, p-val = 0.0199

Model Results:
  estimate      se      zval    pval     ci.lb     ci.ub
  -11.5182  0.1584  -72.7048  <.0001  -11.8287  -11.2077
```

Tabla D.53: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable CIGARRILLOS realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los individuos anidados al centro médico.

Part	EAP var	Inter	YE06	ED50	PAMED	PAENF	std.err Inter	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF
1	0.000	-11.385	-1.778	0.079	0.003	0.072	0.670	0.307	0.283	0.386	0.514
18	0.000	-11.841	-1.768	0.169	0.479	-0.684	0.681	0.390	0.322	0.435	0.547
12	0.000	-11.606	-2.274	0.122	0.241	-0.124	0.664	0.420	0.285	0.408	0.475
6	0.000	-15.322	-13.824	0.182	0.893	-1.261	1.883	1.341	0.543	0.894	1.150
9	0.000	-11.493	-1.822	0.097	0.124	-0.421	0.621	0.336	0.310	0.370	0.411
14	0.000	-11.221	-2.631	0.236	0.044	-0.597	0.657	0.434	0.311	0.399	0.538
2	0.000	-11.441	-1.602	0.073	0.320	-0.038	0.698	0.313	0.302	0.414	0.495
10	0.000	-11.204	-1.803	0.156	-0.096	-0.703	0.623	0.345	0.304	0.373	0.489
8	0.000	-15.008	-13.295	0.220	0.415	-0.682	1.128	1.256	0.551	0.757	0.907
13	0.000	-15.658	-16.253	0.229	0.340	-1.175	1.112	1.321	0.563	0.781	1.153
16	0.000	-10.885	-1.416	0.166	0.276	-0.336	0.590	0.231	0.253	0.353	0.401
19	0.002	-11.629	-2.043	0.061	-0.099	0.449	0.748	0.365	0.299	0.396	0.499
5	0.029	-11.234	-1.847	0.122	0.297	-0.433	0.612	0.315	0.284	0.363	0.416
4	0.000	-11.857	-1.460	0.103	0.671	-0.625	0.717	0.341	0.306	0.455	0.499
15	0.000	-10.842	-1.946	0.165	0.408	-0.721	0.591	0.328	0.284	0.364	0.463
17	0.000	-11.015	-2.136	-0.005	0.504	-0.781	0.605	0.341	0.293	0.384	0.442
11	0.000	-11.395	-1.693	0.191	0.579	-0.669	0.669	0.322	0.295	0.420	0.431
7	0.000	-11.679	-1.344	0.022	0.202	0.385	0.704	0.273	0.293	0.399	0.429
3	0.000	-11.121	-1.788	0.167	0.376	-0.729	0.614	0.307	0.304	0.391	0.337

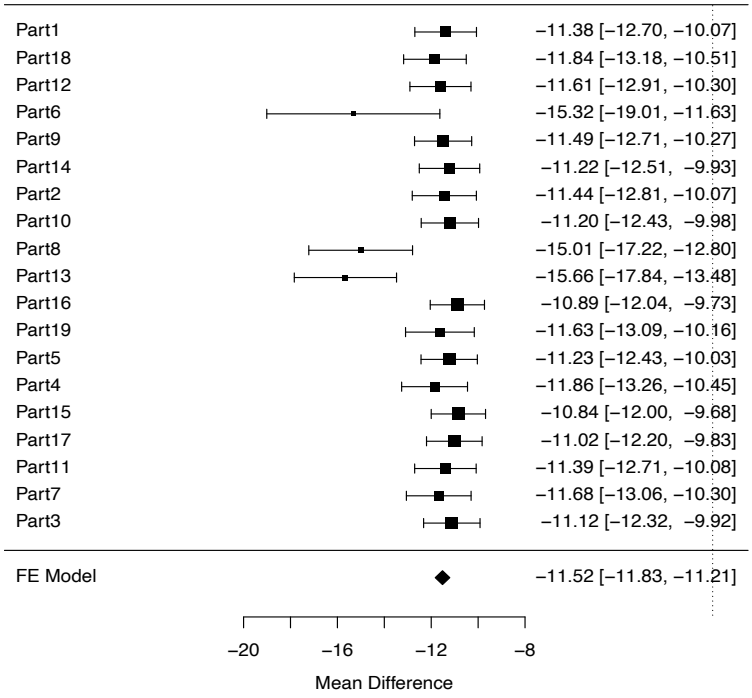


Figura D.39: Forest plot de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.55: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH1$zYE06,sei=resuH1$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
    logLik   deviance        AIC        BIC       AICc
 -148.4768   295.7353   298.9535   299.8979   299.1888

Test for Heterogeneity:
 Q(df = 18) = 295.7353, p-val < .0001

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -1.9055  0.0806  -23.6411  <.0001  -2.0635  -1.7476
```

Tabla D.56: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH1$zED50,sei=resuH1$std.zED50, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
  3.4606    0.7830   -4.9211   -3.9767   -4.6859

Test for Heterogeneity:
  Q(df = 18) = 0.7830, p-val = 1.0000

Model Results:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.1246  0.0717  1.7384  0.0821  -0.0159  0.2650
```

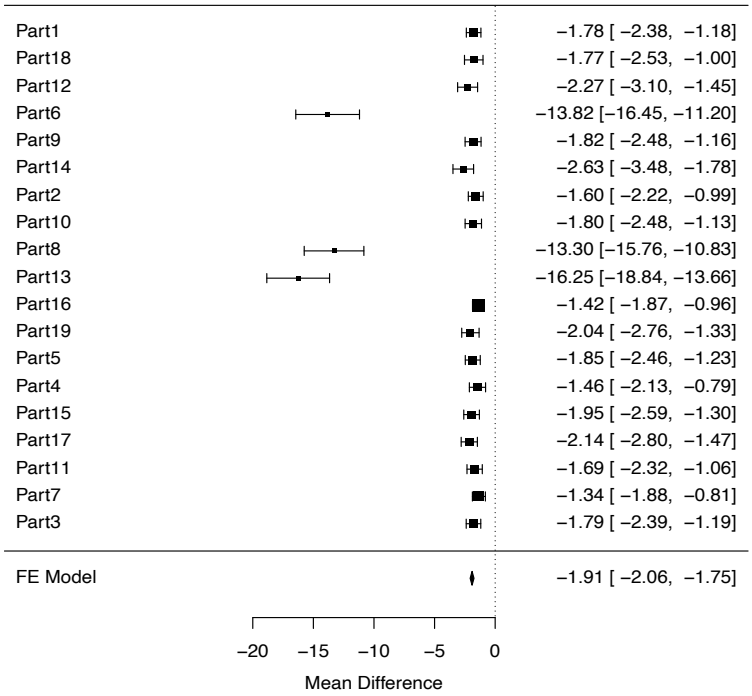


Figura D.40: Forest plot de los coeficientes de YEAR de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

.

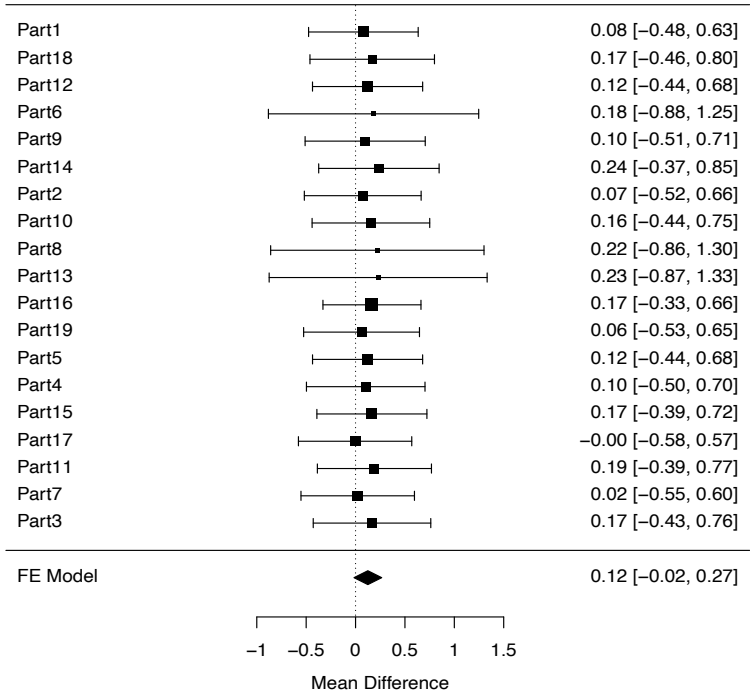


Figura D.41: Forest plot de los coeficientes de EDAD de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.57: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH1$zPAMED,sei=resuH1$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 -4.6527    5.5119   11.3055   12.2499   11.5408

Test for Heterogeneity:
  Q(df = 18) = 5.5119, p-val = 0.9978

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.2695  0.0958  2.8127  0.0049  0.0817  0.4573
```

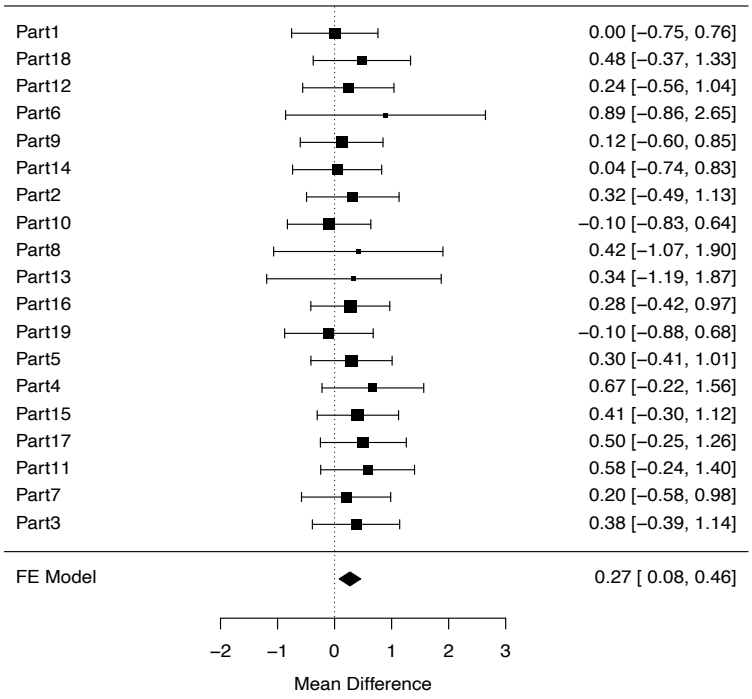


Figura D.42: Forest plot de los coeficientes de PAMED de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS..

Tabla D.58: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH1$zPAENF,sei=resuH1$std.zPAENF, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
   logLik  deviance       AIC       BIC      AICc
 -11.4912   12.6541   24.9824   25.9269   25.2177

Test for Heterogeneity:
  Q(df = 18) = 12.6541, p-val = 0.8117

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.4111  0.1109  -3.7069  0.0002  -0.6284  -0.1937
```

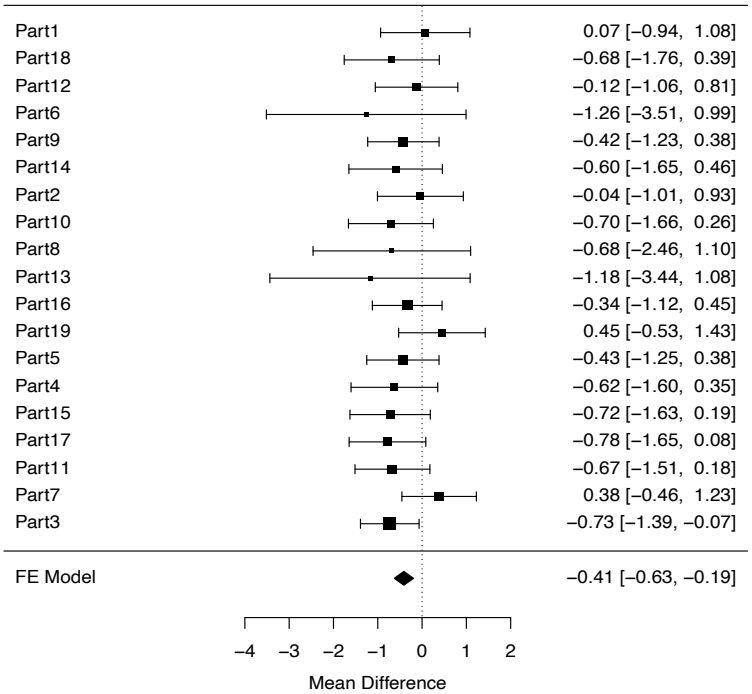


Figura D.43: Forest plot de los coeficientes de PAENF de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los individuos anidados a sus centros médicos para la *odds* de *missings* en CIGARRILLOS.

D.2.2.4 Modelo logístico con efectos aleatorios para los centros médicos.

Resultados de los análisis realizados para la los *missings* en la variable CIGARRILLOS con los datos de cada partición independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios sólo para los centros médicos, no para los individuos.

```
glmmTMB(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + (1|EAP),
         data=partition, family="binomial", REML=F)
```

La Tabla D.59 resume los coeficientes obtenidos en dichos análisis intra-partición para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todas las particiones con la opción "FE", Tabla D.60 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots.

Tabla D.60: Metaanálisis de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH2$Inter,sei=resuH2$std.Inter, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 16.7990    9.1618  -31.5980  -30.6535  -31.3627

Test for Heterogeneity:
  Q(df = 18) = 9.1618, p-val = 0.9559

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -1.9575  0.0296  -66.2257  <.0001  -2.0154  -1.8996
```

Tabla D.61: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH2$zYE06,sei=resuH2$std.zYE06, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
```

Tabla D.59: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable CIGARRILLOS realizados en cada partición independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los centros médicos.

Part	EAP var	Inter	YE06	ED50	PAMED	PAENF	std.err Inter	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF
4	0.298	-1.902	-0.218	0.066	0.132	-0.210	0.119	0.048	0.033	0.086	0.088
16	0.207	-1.773	-0.203	0.100	0.163	-0.140	0.125	0.047	0.031	0.087	0.084
9	0.264	-1.917	-0.179	0.073	0.116	-0.066	0.147	0.051	0.034	0.104	0.095
1	0.231	-2.064	-0.117	0.105	0.255	-0.182	0.118	0.047	0.033	0.085	0.085
19	0.387	-1.987	-0.186	0.076	0.220	-0.138	0.131	0.050	0.034	0.094	0.093
15	0.266	-1.949	-0.153	0.040	0.095	0.056	0.121	0.047	0.033	0.087	0.087
7	0.309	-2.086	-0.157	0.114	0.202	-0.130	0.120	0.049	0.035	0.085	0.089
18	0.245	-1.934	-0.194	0.030	0.231	-0.296	0.140	0.052	0.034	0.097	0.095
14	0.146	-1.887	-0.179	0.037	0.186	-0.153	0.128	0.049	0.033	0.090	0.088
13	0.181	-2.091	-0.112	0.100	0.179	-0.102	0.134	0.050	0.033	0.093	0.090
5	0.482	-1.896	-0.203	0.071	0.140	-0.211	0.132	0.052	0.033	0.099	0.091
8	0.327	-1.894	-0.228	0.069	0.194	-0.293	0.133	0.051	0.033	0.095	0.094
3	0.202	-1.964	-0.125	0.053	0.123	0.087	0.142	0.051	0.033	0.094	0.092
2	0.199	-1.862	-0.191	0.059	0.166	-0.164	0.139	0.050	0.032	0.098	0.088
11	0.189	-2.018	-0.150	0.096	0.122	0.052	0.133	0.050	0.033	0.095	0.092
10	0.322	-1.879	-0.162	0.009	0.186	-0.041	0.117	0.046	0.033	0.084	0.086
6	0.221	-2.109	-0.076	0.103	0.058	0.100	0.134	0.049	0.033	0.093	0.092
17	0.306	-1.923	-0.173	0.100	0.109	-0.062	0.139	0.051	0.033	0.102	0.089
12	0.200	-2.032	-0.127	0.081	0.230	-0.211	0.115	0.047	0.033	0.079	0.084

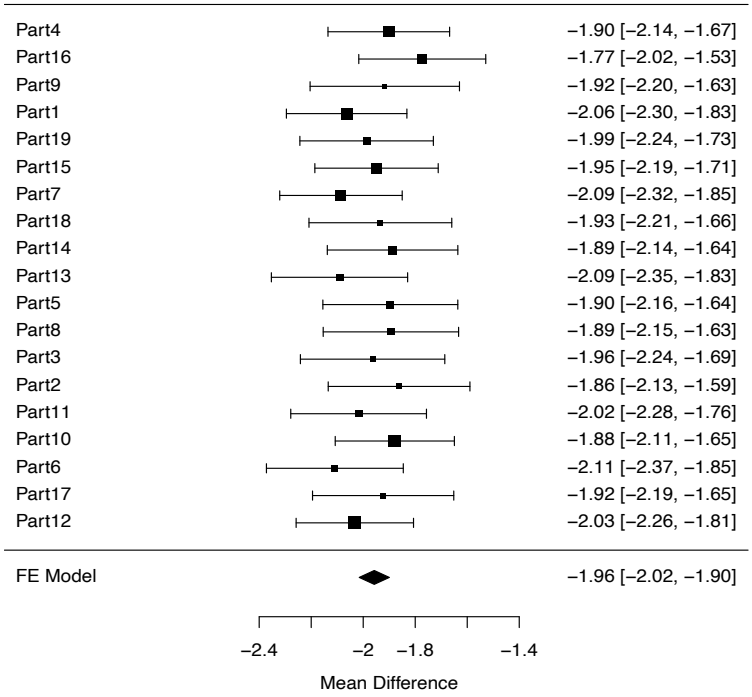


Figura D.44: Forest plot de los Interceptos de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
  logLik  deviance       AIC       BIC      AICc
 33.8438   11.7642  -65.6875  -64.7431  -65.4522

Test for Heterogeneity:
  Q(df = 18) = 11.7642, p-val = 0.8592

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.1641  0.0113  -14.5265  <.0001  -0.1862  -0.1419
```

Tabla D.62: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH2$zED50,sei=resuH2$std.zED50, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
```

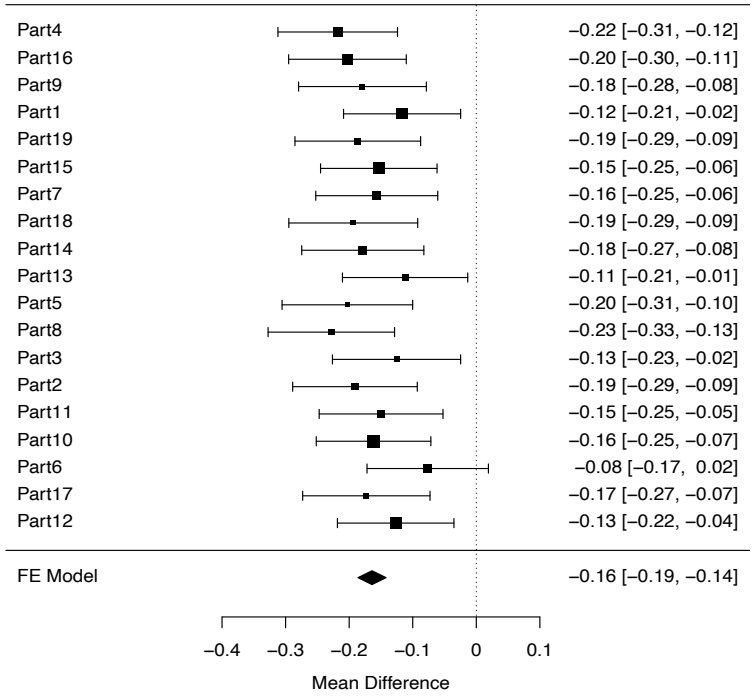


Figura D.45: Forest plot de los coeficientes de YEAR de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en CIGARRILLOS.

```
 40.2481   14.1468  -78.4961  -77.5517  -78.2608

Test for Heterogeneity:
  Q(df = 18) = 14.1468, p-val = 0.7195

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.0726  0.0076  9.5837  <.0001  0.0578  0.0875
```

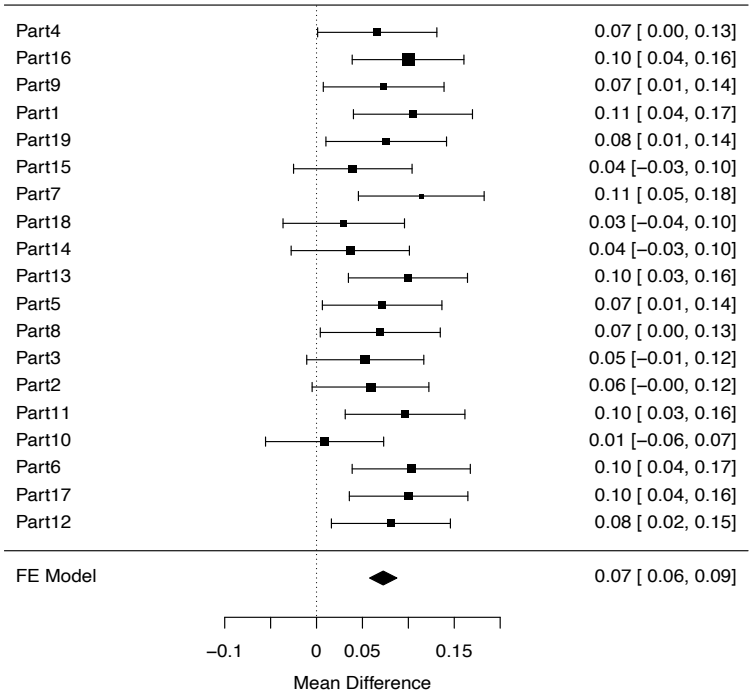


Figura D.46: Forest plot de los coeficientes de EDAD de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en CIGARRILLOS.

Tabla D.63: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH2$zPAMED,sei=resuH2$std.zPAMED, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 24.8771    6.1338  -47.7543  -46.8099  -47.5190
```

```
Test for Heterogeneity:
  Q(df = 18) = 6.1338, p-val = 0.9956

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.1664  0.0209  7.9551  <.0001  0.1254  0.2074
```

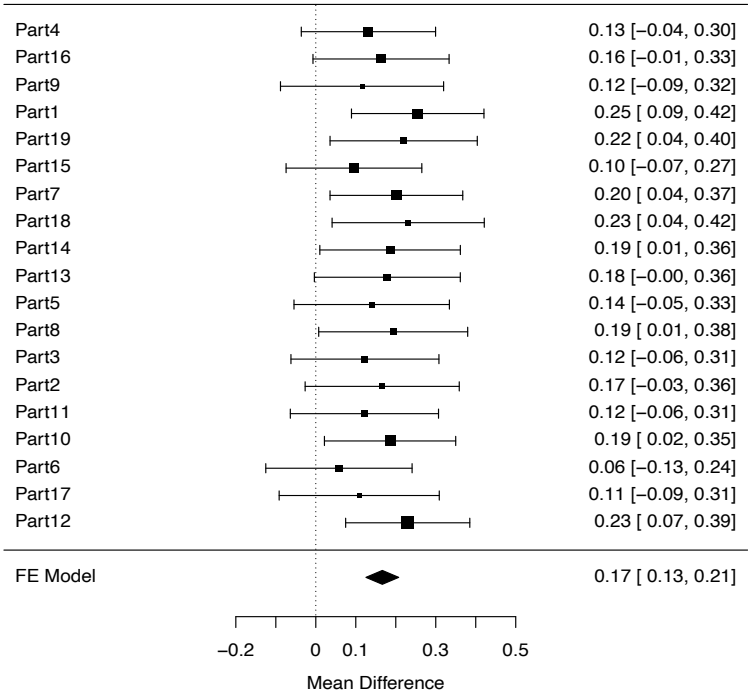


Figura D.47: Forest plot de los coeficientes de PAMED de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en CIGARRILLOS.

Tabla D.64: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros médicos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuH2$zPAENF,sei=resuH2$std.zPAENF, method="FE", measure="MD")

Fixed-Effects Model (k = 19)
  logLik  deviance       AIC       BIC      AICc
 12.9580   30.8627  -23.9160  -22.9716  -23.6807
```

```
Test for Heterogeneity:
  Q(df = 18) = 30.8627, p-val = 0.0299

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.1113  0.0205  -5.4241  <.0001  -0.1515  -0.0711
```

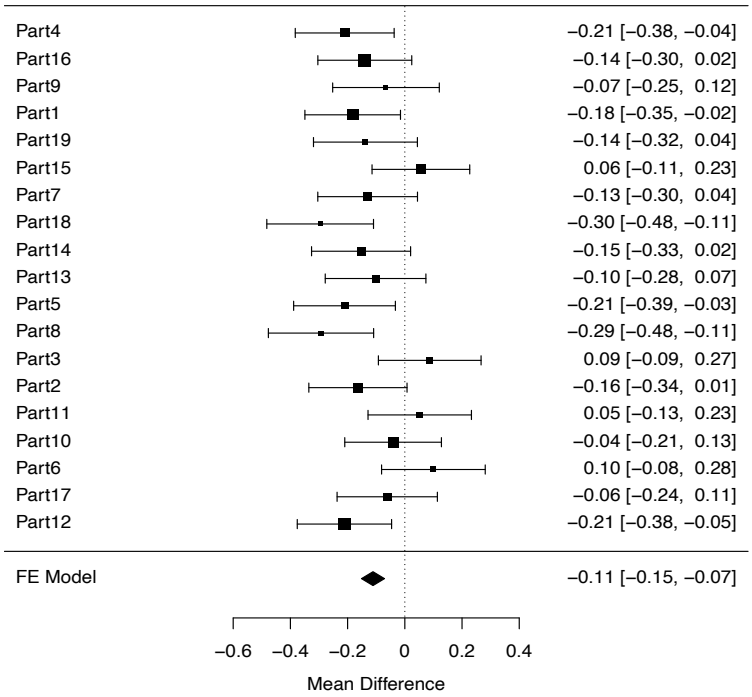


Figura D.48: Forest plot de los coeficientes de PAENF de las regresiones logísticas intra-partición-aleatoria con efectos aleatorios de los centros para la *odds* de *missings* en CIGARRILLOS.

D.3 Análisis de los datos particionados por centro

D.3.1 *Missing data* en la variable Peso

D.3.1.1 Modelo logístico simple.

Resultados de los análisis de los *missings* en la variable PESO realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios.

```
glm(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO,
    data=centro, family="binomial")
```

La Tabla D.65 resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todos los centros con la opción "ML", Tabla D.66 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, Figura D.49 y siguientes.

También realizamos un segundo set de regresiones, Tabla D.73 y siguientes, para estudiar si los coeficientes obtenidos en cada modelo intra-centro dependen de los parámetros que varían entre diferentes centros, y que no pudieron ser analizados simultáneamente en el paso previo porque la información era parcial. Sólo hemos considerado la variable inter-centro POSMAS.

Tabla D.66: Metaanálisis de los Interceptos de las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$Inter,sei=resuCH$std.Inter, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -39.8335  143.1443   83.6669   85.5558   84.4169

tau^2 (estimated amount of total heterogeneity): 3.8474 (SE = 1.2521)
I^2 (total heterogeneity / total variability):   99.91%
H^2 (total variability / sampling variability):  1104.39

Test for Heterogeneity:
  Q(df = 18) = 11100.8533, p-val < .0001

Model Results:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.8401  0.4507  1.8641  0.0623  -0.0432  1.7234
```

Tabla D.65: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable PESO realizados en cada centro médico independientemente. Coeficientes de las variables estandarizadas de la regresión logística simple.

EAP	Inter	SEXO	YE06	ED50	PAMED	PAENF	EXTRA	std.err Inter	std.err SEXO	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF	std.err EXTRA
0110	-0.807	-0.089	0.643	-0.108	1.264	0.346	0.140	0.045	0.027	0.034	0.007	0.087	0.032	0.058
0210	-1.047	0.031	0.446	-0.297	1.385	-0.072	0.371	0.171	0.027	0.039	0.007	0.126	0.016	0.056
0310	0.666	-0.163	0.263	-0.271	-0.073	0.740	0.073	0.049	0.021	0.039	0.006	0.067	0.105	0.040
0410	-1.122	0.067	0.835	-0.274	-1.105	1.279	0.064	0.046	0.017	0.033	0.005	0.042	0.054	0.029
0610	-0.506	-0.034	0.305	-0.161	-0.152	0.657	0.364	0.050	0.018	0.007	0.005	0.037	0.031	0.031
1210	-0.638	0.032	0.222	-0.238	-0.068	-0.066	0.266	0.086	0.025	0.037	0.007	0.044	0.059	0.038
1310	1.001	-0.053	0.041	-0.456	-0.396	-0.117	0.176	0.076	0.022	0.029	0.006	0.088	0.053	0.033
1410	-1.244	-0.063	0.528	-0.050	-0.059	0.563	0.398	0.025	0.020	0.015	0.006	0.024	0.022	0.031
1510	1.400	-0.203	0.193	-0.363	-0.234	0.649	0.253	0.058	0.023	0.048	0.007	0.086	0.047	0.034
1610	1.554	-0.121	-0.188	-0.316	-1.112	0.106	0.269	0.040	0.022	0.027	0.006	0.069	0.026	0.036
2110	2.996	-0.091	-0.093	-0.324	-2.157	0.002	0.180	0.115	0.019	0.013	0.005	0.073	0.061	0.035
2210	5.673	-0.135	-1.015	-0.356	-1.921	-0.864	0.228	0.248	0.020	0.064	0.006	0.078	0.078	0.042
2410	2.231	-0.009	-0.494	-0.284	-2.100	0.211	0.104	0.084	0.015	0.035	0.004	0.070	0.018	0.030
2610	2.293	-0.080	-0.132	-0.345	-1.529	0.738	0.242	0.070	0.018	0.022	0.005	0.059	0.047	0.039
2810	1.105	-0.156	0.050	-0.485	-1.170	0.531	0.554	0.062	0.022	0.017	0.006	0.074	0.034	0.041
3110	-2.652	-0.118	1.221	-0.215	-0.189	1.409	0.654	0.077	0.021	0.032	0.006	0.017	0.049	0.036
3210	1.584	-0.027	-0.042	-0.274	-0.778	-0.433	0.338	0.101	0.016	0.024	0.005	0.034	0.055	0.048
3410	3.790	-0.171	-0.594	-0.189	-1.412	-0.391	0.257	0.236	0.024	0.056	0.007	0.077	0.049	0.062
3610	-0.208	-0.023	0.149	-0.195	-0.682	0.166	0.604	0.072	0.017	0.023	0.005	0.031	0.043	0.036

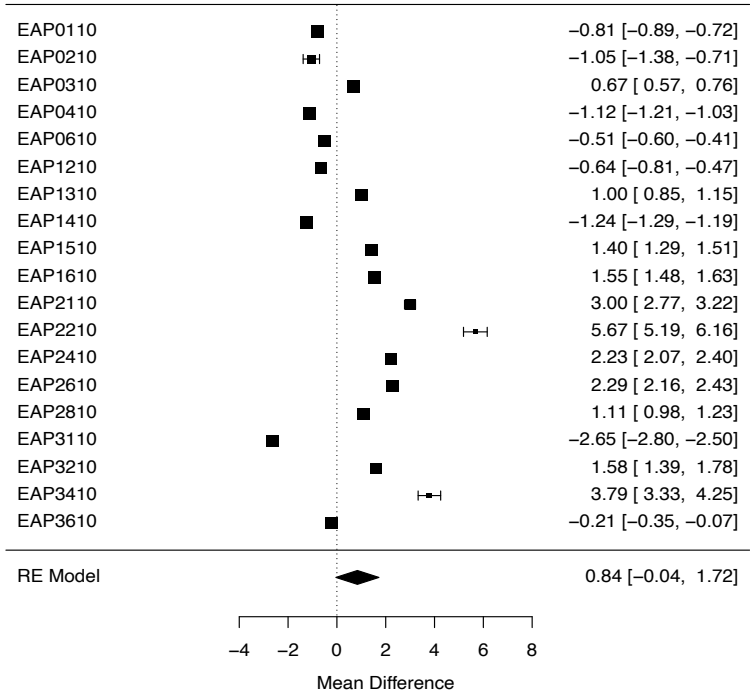


Figura D.49: Forest plot de los Interceptos de las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.67: Metaanálisis de los coeficientes de SEXO obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$SEXO,sei=resuCH$std.SEXO, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
 22.7010   67.4478  -41.4020  -39.5131  -40.6520

tau^2 (estimated amount of total heterogeneity): 0.0049 (SE = 0.0017)
I^2 (total heterogeneity / total variability):   92.45%
H^2 (total variability / sampling variability):  13.24

Test for Heterogeneity:
  Q(df = 18) = 239.7003, p-val < .0001

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.0735  0.0168  -4.3866  <.0001  -0.1064  -0.0407
```

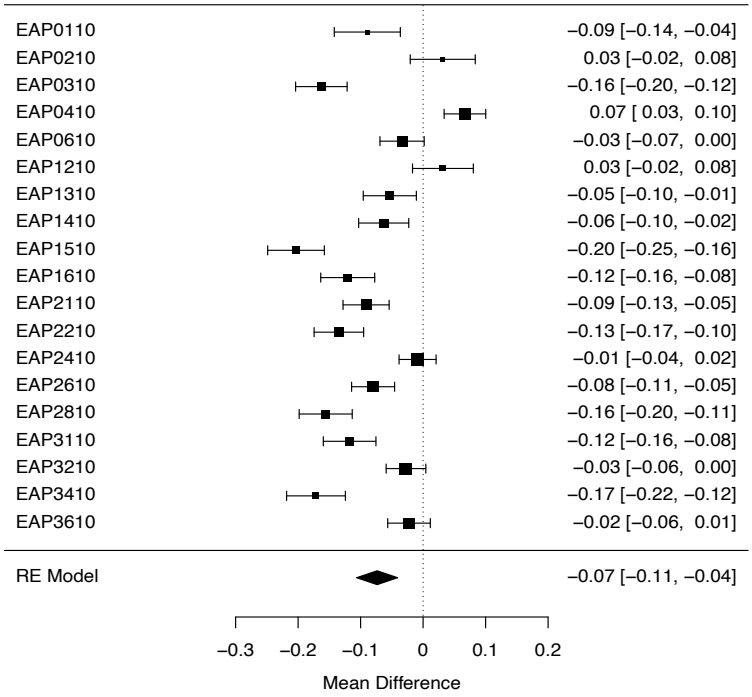


Figura D.50: Forest plot de los coeficientes de SEXO obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.68: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$zYE06,sei=resuCH$std.zYE06, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
-13.7718  128.7196   31.5437   33.4326   32.2937

tau^2 (estimated amount of total heterogeneity): 0.2474 (SE = 0.0807)
I^2 (total heterogeneity / total variability):   99.81%
H^2 (total variability / sampling variability):  526.81

Test for Heterogeneity:
  Q(df = 18) = 4358.2736, p-val < .0001

Model Results:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.1239  0.1144  1.0829  0.2788  -0.1003  0.3481
```

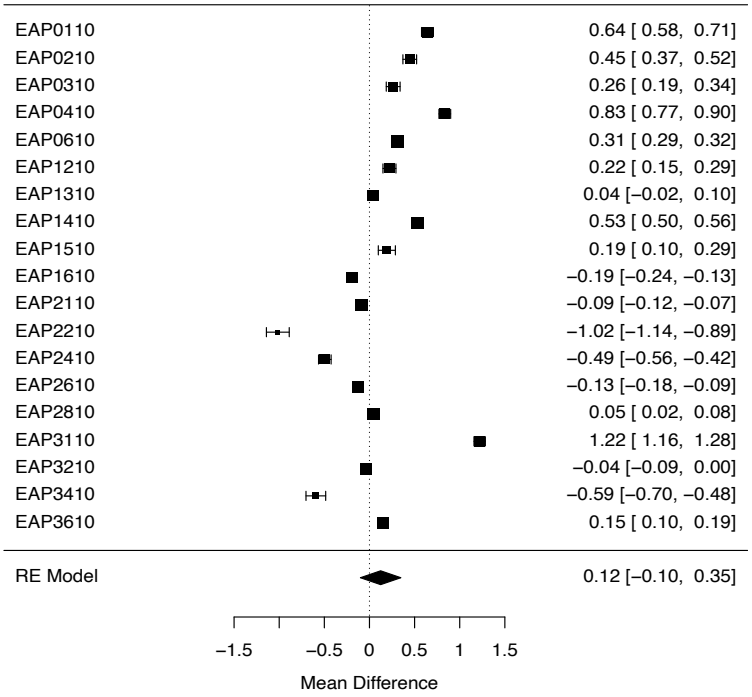


Figura D.51: Forest plot de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.69: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$zED50,sei=resuCH$std.zED50, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
 15.8142  128.7287  -27.6283  -25.7394  -26.8783

tau^2 (estimated amount of total heterogeneity): 0.0110 (SE = 0.0036)
I^2 (total heterogeneity / total variability):   99.70%
H^2 (total variability / sampling variability):  335.08

Test for Heterogeneity:
  Q(df = 18) = 5635.0369, p-val < .0001

Model Results:
  estimate      se      zval    pval    ci.lb    ci.ub
   -0.2738  0.0241  -11.3374  <.0001  -0.3211  -0.2264
```

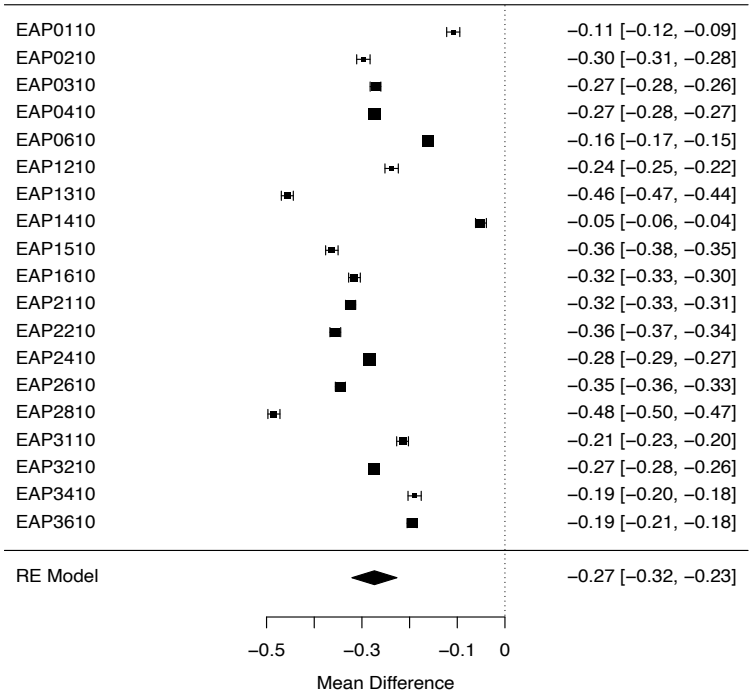


Figura D.52: Forest plot de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.70: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$zPAMED,sei=resuCH$std.zPAMED, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -26.1837  127.1067   56.3673   58.2562   57.1173

tau^2 (estimated amount of total heterogeneity): 0.9136 (SE = 0.2979)
I^2 (total heterogeneity / total variability):   99.79%
H^2 (total variability / sampling variability):  472.72

Test for Heterogeneity:
  Q(df = 18) = 3797.5875, p-val < .0001

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.6585  0.2198  -2.9954  0.0027  -1.0894  -0.2276
```

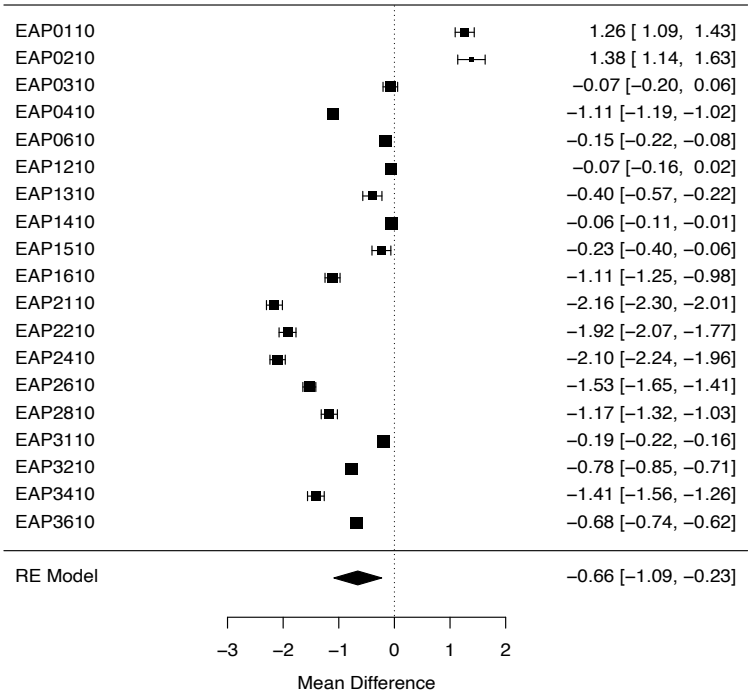


Figura D.53: Forest plot de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.71: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$zPAENF,sei=resuCH$std.zPAENF, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)

   logLik  deviance       AIC       BIC      AICc
 -15.8418  117.6197   35.6837   37.5725   36.4337

tau^2 (estimated amount of total heterogeneity): 0.3067 (SE = 0.1003)
I^2 (total heterogeneity / total variability):   99.62%
H^2 (total variability / sampling variability):  265.43

Test for Heterogeneity:
  Q(df = 18) = 2641.0283, p-val < .0001

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.2872  0.1276  2.2511  0.0244  0.0371  0.5373
```

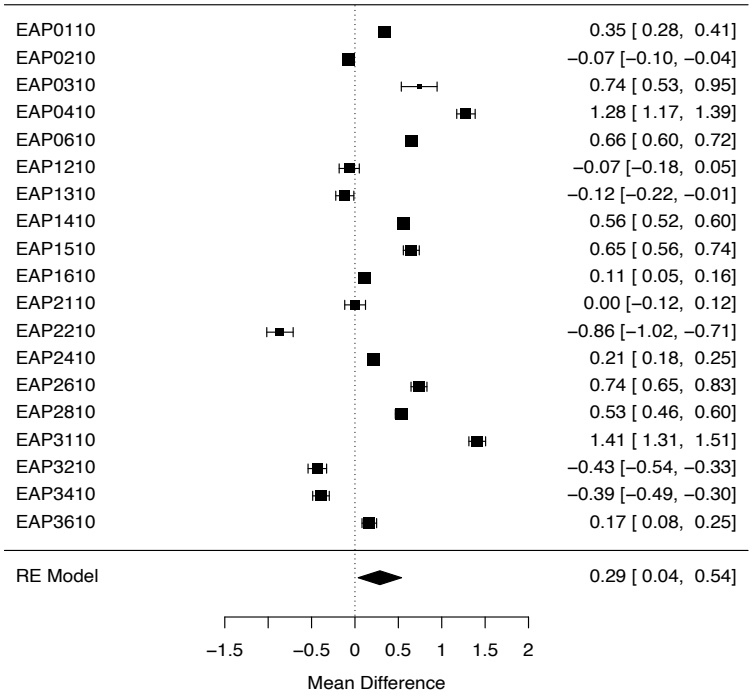


Figura D.54: Forest plot de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.72: Metaanálisis de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

```
rma(resuCH$EXTRANJERO,sei=resuCH$std.EXTRANJERO, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
  7.1944   74.1088  -10.3888   -8.4999   -9.6388

tau^2 (estimated amount of total heterogeneity): 0.0260 (SE = 0.0090)
I^2 (total heterogeneity / total variability):   94.90%
H^2 (total variability / sampling variability):  19.59

Test for Heterogeneity:
  Q(df = 18) = 398.9364, p-val < .0001

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.2913  0.0382  7.6315  <.0001  0.2165  0.3661
```

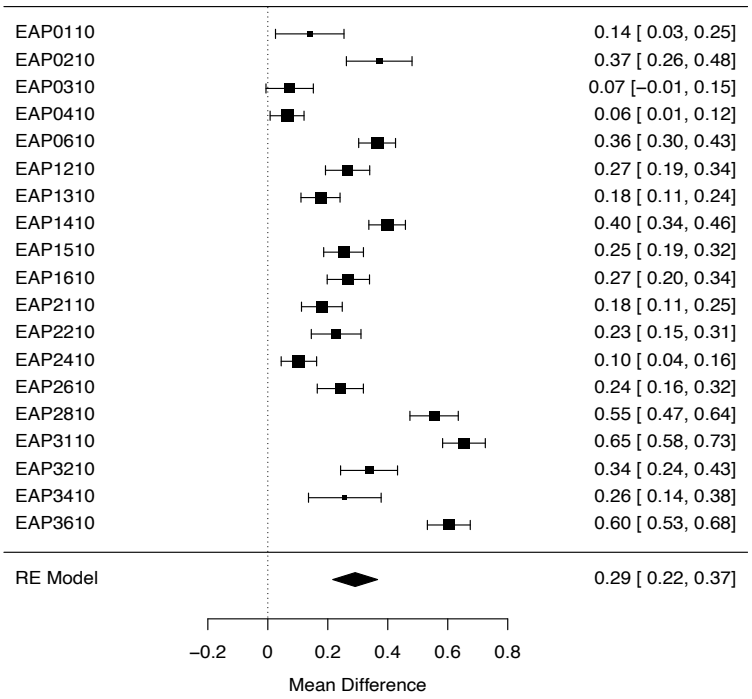


Figura D.55: Forest plot de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en PESO.

Tabla D.73: Modelo lineal simple del Intercepto (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(formula = Inter ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.Inter^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept)  0.02332    0.38616     0.060      0.953
  zPOSMAS      0.14148    0.38091     0.371      0.715

Residual standard error: 25.45 on 17 degrees of freedom
Multiple R-squared: 0.00805, Adjusted R-squared: -0.0503
F-statistic: 0.138 on 1 and 17 DF,  p-value: 0.7149
```

Tabla D.74: Modelo lineal del coeficiente SEXO (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(formula = SEXO ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.SEXO^2)

Coefficients:
             Estimate    Std.Error   z-value   Pr(>|z|)
  (Intercept) -0.0651138  0.0171930    -3.787    0.00147
  zPOSMAS     -0.0002443  0.0181643    -0.013    0.98942

Residual standard error: 3.755 on 17 degrees of freedom
Multiple R-squared: 1.064e-05, Adjusted R-squared: -0.05881
F-statistic: 0.000181 on 1 and 17 DF,  p-value: 0.9894
```

Tabla D.75: Modelo lineal del coeficiente de YEAR (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(formula = zYE06 ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.zYE06^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept)  0.20627    0.07295     2.828     0.0116
  zPOSMAS     -0.05550    0.10080    -0.551     0.5890

Residual standard error: 15.87 on 17 degrees of freedom
Multiple R-squared: 0.01752, Adjusted R-squared: -0.04027
F-statistic: 0.3032 on 1 and 17 DF,  p-value: 0.589
```

Tabla D.76: Modelo lineal del coeficiente de EDAD (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(formula = zED50 ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.zED50^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept) -0.27119    0.02378    -11.40   2.18e-09
  zPOSMAS      0.01300    0.02549      0.51      0.617

Residual standard error: 18.07 on 17 degrees of freedom
Multiple R-squared: 0.01507, Adjusted R-squared: -0.04287
```

```
F-statistic: 0.2601 on 1 and 17 DF,  p-value: 0.6166
```

Tabla D.77: Modelo lineal del coeficiente de PAMED (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(formula = zPAMED ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.zPAMED^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept)  -0.4815     0.1409    -3.419    0.00327
  zPOSMAS      -0.1677     0.1720    -0.975    0.34336

Residual standard error: 14.55 on 17 degrees of freedom
Multiple R-squared: 0.05293, Adjusted R-squared: -0.002779
F-statistic: 0.9501 on 1 and 17 DF,  p-value: 0.3434
```

Tabla D.78: Modelo lineal del coeficiente de PAENF (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(formula = zPAENF ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.zPAENF^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept)  0.24138    0.09088     2.656     0.0166
  zPOSMAS     -0.14092    0.10871    -1.296     0.2122

Residual standard error: 11.89 on 17 degrees of freedom
Multiple R-squared: 0.08995, Adjusted R-squared: 0.03642
F-statistic:  1.68 on 1 and 17 DF,  p-value: 0.2122
```

Tabla D.79: Modelo lineal del coeficiente de EXTRANJERO (de las regresiones logísticas simple intra-centro) en función de POSMAS.

```
Call: lm(EXTRANJERO ~ 1 + zPOSMAS, data = resuCH, weights = 1/std.EXTRANJERO^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept)   0.2782     0.0431     6.456   5.92e-06
  zPOSMAS      -0.0184     0.0437    -0.421      0.679
```

```
Residual standard error: 4.819 on 17 degrees of freedom
Multiple R-squared: 0.01032, Adjusted R-squared: -0.0479
F-statistic: 0.1772 on 1 and 17 DF,  p-value: 0.679
```

D.3.1.2 Modelo logístico con efectos aleatorios para los individuos.

Resultados de los análisis de los *missings* en la variable PESO realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos.

```
glmmTMB(mPESO ~ SEXO + zYE06 + zED50 + zPAMED + zPAENF + EXTRANJERO + (1|ID),
        data=centro, family="binomial", REML=F)
```

La Tabla D.80 resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. Para cada uno de los coeficientes hemos calculado un meta-análisis combinando los resultados de todos los centros con la opción "ML". La Tabla D.81 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, , Figura D.56 y siguientes.

La variable inter-centro POSMAS, porcentaje de pacientes del centro médico con al menos 65 años, no puede ser introducida directamente en los modelos intra-centro por ser un parámetro anual fijo en cada uno, ha sido analizada en un proceso posterior, modelando cada uno de los coeficientes obtenidos en cada centro en función de POSMAS mediante regresiones lineales. De este modo, podemos testear si las variables intra-centro pudieran estar confundidas con variables externas. El restulado de estos modelos se expone en la Tabla D.88 y siguientes.

A continuación mostramos el resultado del meta-análisis de los coeficientes obtenidos en los análisis intra-centro con efectos aleatorios, y los representamos gráficacamente mediante forest plots.

Tabla D.80: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable PESO realizados en cada centro médico independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los individuos.

EAP	Inter	SEXO	YE06	ED50	PAMED	PAENF	EXTRA	std.err Inter	std.err SEXO	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF	std.err EXTRA
0110	-1.900	-0.269	1.629	-0.301	3.191	0.875	0.396	0.121	0.124	0.057	0.033	0.144	0.051	0.274
0210	-1.653	0.017	0.902	-0.728	2.762	-0.146	0.766	0.254	0.095	0.056	0.028	0.180	0.023	0.181
0310	3.701	-0.703	0.950	-1.330	-0.260	2.532	0.266	0.165	0.147	0.073	0.044	0.123	0.195	0.247
0410	-2.251	0.168	2.520	-1.216	-3.254	3.779	0.168	0.119	0.103	0.062	0.036	0.077	0.098	0.167
0610	-1.052	-0.139	0.816	-0.537	-0.392	1.725	1.315	0.109	0.092	0.014	0.026	0.060	0.052	0.163
1210	-1.351	0.075	0.503	-0.554	-0.145	-0.144	0.702	0.149	0.099	0.055	0.029	0.067	0.088	0.154
1310	3.570	-0.251	0.143	-1.395	-0.938	-0.318	0.340	0.155	0.103	0.046	0.034	0.139	0.084	0.138
1410	-2.985	-0.145	1.230	-0.069	-0.137	1.314	1.080	0.079	0.085	0.026	0.024	0.037	0.035	0.134
1510	3.739	-0.522	0.418	-0.931	-0.467	1.323	0.370	0.119	0.085	0.069	0.028	0.122	0.068	0.110
1610	4.607	-0.393	-0.419	-0.984	-2.656	0.255	0.579	0.121	0.099	0.042	0.032	0.110	0.041	0.141
2110	6.946	-0.281	-0.181	-0.861	-4.693	0.004	0.468	0.188	0.074	0.020	0.024	0.114	0.089	0.132
2210	11.889	-0.358	-2.024	-0.843	-3.881	-1.729	0.453	0.370	0.072	0.092	0.023	0.115	0.110	0.136
2410	5.173	-0.088	-1.002	-0.786	-4.394	0.459	0.253	0.133	0.056	0.051	0.019	0.103	0.026	0.108
2610	5.862	-0.265	-0.261	-0.929	-3.357	1.639	0.476	0.131	0.072	0.033	0.023	0.092	0.071	0.133
2810	2.798	-0.426	0.122	-1.176	-2.499	1.139	1.251	0.115	0.084	0.024	0.029	0.110	0.051	0.142
3110	-5.711	-0.328	2.702	-0.490	-0.411	3.110	1.756	0.139	0.081	0.055	0.024	0.025	0.079	0.143
3210	3.598	-0.127	-0.067	-0.691	-1.637	-0.905	0.902	0.159	0.061	0.036	0.020	0.052	0.081	0.177
3410	8.669	-0.444	-1.320	-0.506	-3.145	-0.874	0.717	0.374	0.095	0.085	0.029	0.120	0.074	0.250
3610	-0.420	-0.070	0.350	-0.517	-1.660	0.363	1.818	0.126	0.077	0.036	0.024	0.050	0.068	0.166

Tabla D.81: Metaanálisis de los Interceptos de las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$Inter,sei=resuC1$std.Inter, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
    logLik  deviance       AIC       BIC      AICc
  -55.0469  147.0041  114.0939  115.9828  114.8439

tau^2 (estimated amount of total heterogeneity): 19.1680 (SE = 6.2297)
I^2 (total heterogeneity / total variability):   99.90%
H^2 (total variability / sampling variability):  1037.36

Test for Heterogeneity:
  Q(df = 18) = 14162.0430, p-val < .0001

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    2.2700  1.0053  2.2581  0.0239  0.2997  4.2403
```

Tabla D.82: Metaanálisis de los coeficientes de SEXO obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$SEXO,sei=resuC1$std.SEXO, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
   3.4366   50.7597   -2.8732   -0.9844   -2.1232

tau^2 (estimated amount of total heterogeneity): 0.0299 (SE = 0.0123)
I^2 (total heterogeneity / total variability):   80.94%
H^2 (total variability / sampling variability):  5.25

Test for Heterogeneity:
  Q(df = 18) = 86.8829, p-val < .0001

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.2336  0.0447  -5.2267  <.0001  -0.3212  -0.1460
```

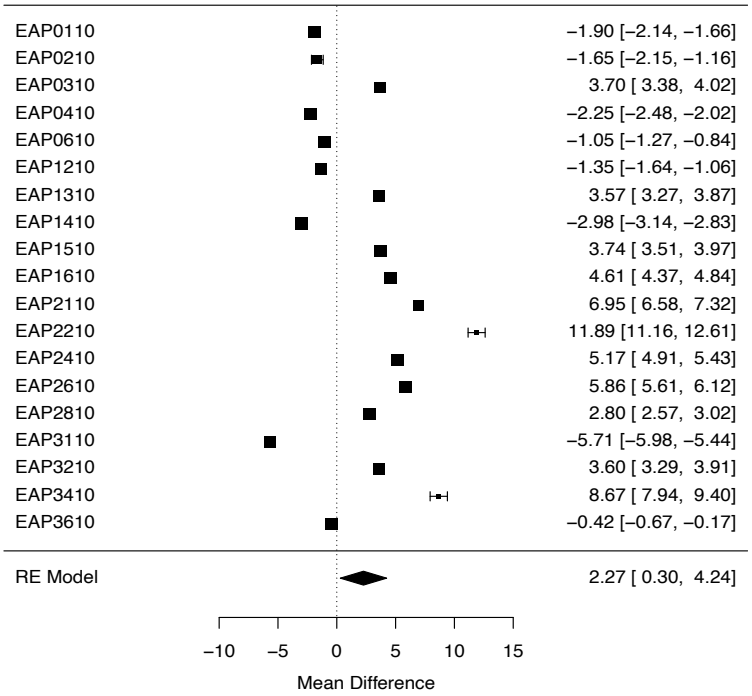


Figura D.56: Forest plot de los Interceptos de las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

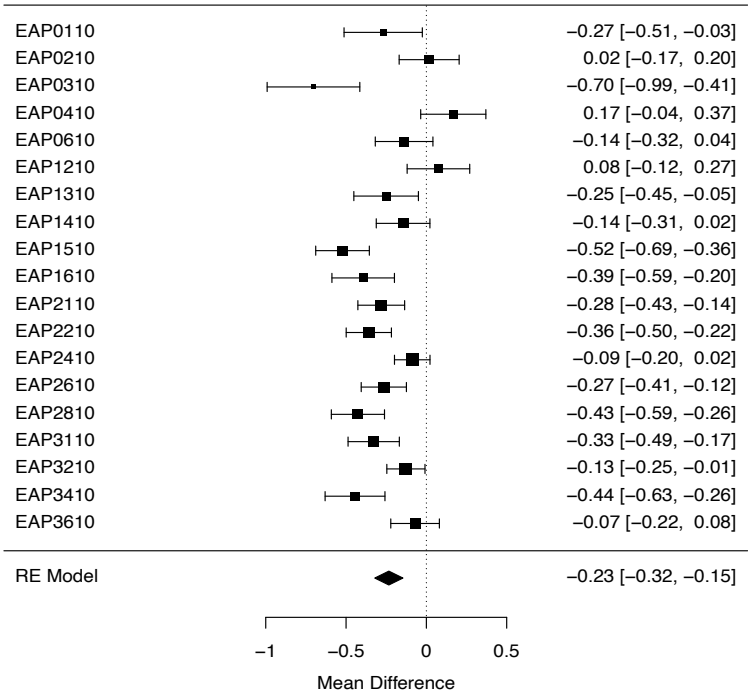


Figura D.57: Forest plot de los coeficientes de SEXO en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.83: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$zYE06,sei=resuC1$std.zYE06, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -29.6754  143.1349   63.3507   65.2396   64.1007

tau^2 (estimated amount of total heterogeneity): 1.3262 (SE = 0.4312)
I^2 (total heterogeneity / total variability):   99.91%
H^2 (total variability / sampling variability):  1074.03

Test for Heterogeneity:
  Q(df = 18) = 9293.9891, p-val < .0001

Model Results:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.3696  0.2645  1.3975  0.1623  -0.1488  0.8880
```

Tabla D.84: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$zED50,sei=resuC1$std.zED50, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
  -6.5063  115.1760   17.0127   18.9015   17.7627

tau^2 (estimated amount of total heterogeneity): 0.1151 (SE = 0.0376)
I^2 (total heterogeneity / total variability):   99.41%
H^2 (total variability / sampling variability):  169.62

Test for Heterogeneity:
  Q(df = 18) = 2533.6931, p-val < .0001

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.7807  0.0781  -9.9952  <.0001  -0.9338  -0.6276
```

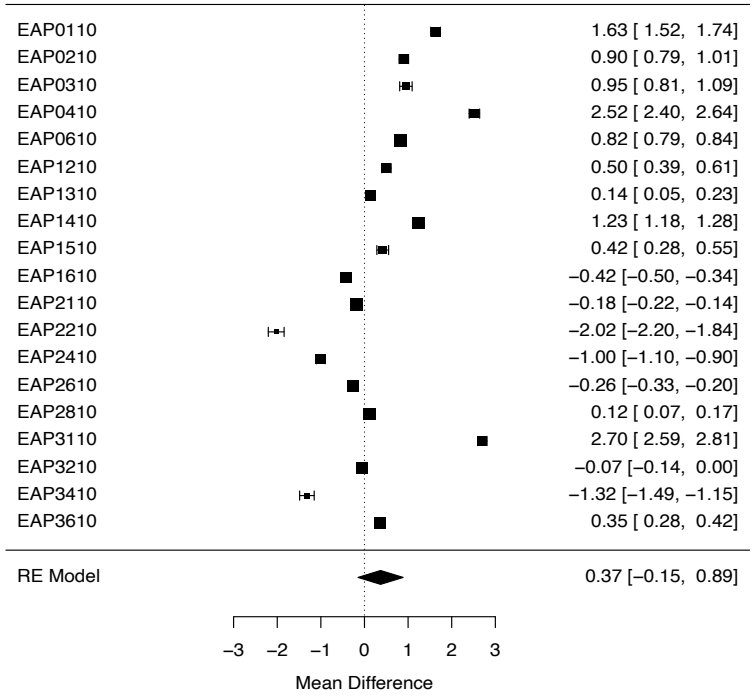


Figura D.58: Forest plot de los coeficientes de YEAR en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.85: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$zPAMED,sei=resuC1$std.zPAMED, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -41.1809  140.1162   86.3617   88.2506   87.1117

tau^2 (estimated amount of total heterogeneity): 4.4503 (SE = 1.4474)
I^2 (total heterogeneity / total variability):   99.89%
H^2 (total variability / sampling variability):  943.95

Test for Heterogeneity:
  Q(df = 18) = 8006.1498, p-val < .0001

Model Results:
  estimate      se     zval    pval    ci.lb    ci.ub
   -1.4735  0.4846  -3.0409  0.0024  -2.4232  -0.5238
```

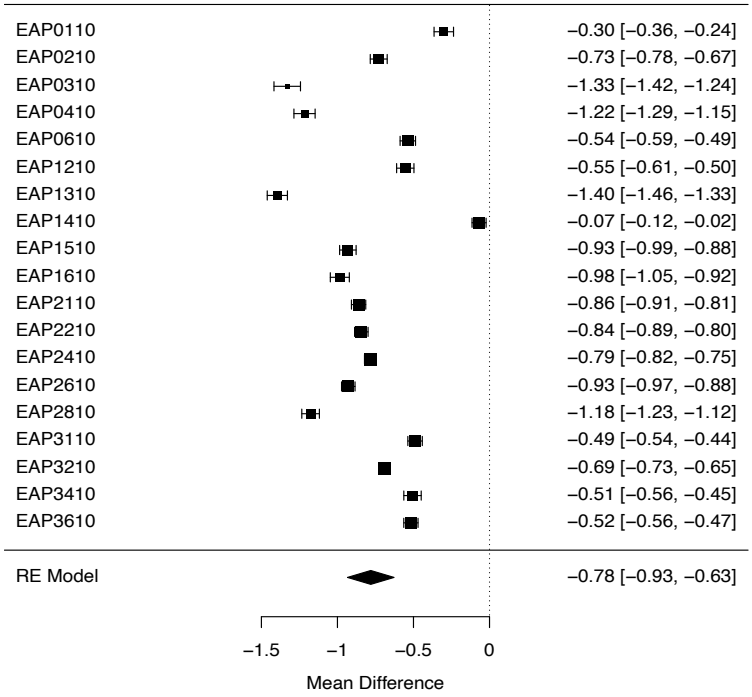


Figura D.59: Forest plot de los coeficientes de EDAD en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.86: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$zPAENF,sei=resuC1$std.zPAENF, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -32.9517  135.0098   69.9033   71.7922   70.6533

tau^2 (estimated amount of total heterogeneity): 1.8684 (SE = 0.6083)
I^2 (total heterogeneity / total variability):   99.86%
H^2 (total variability / sampling variability):  689.90

Test for Heterogeneity:
  Q(df = 18) = 6064.1487, p-val < .0001

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.7565  0.3141  2.4081  0.0160  0.1408  1.3722
```

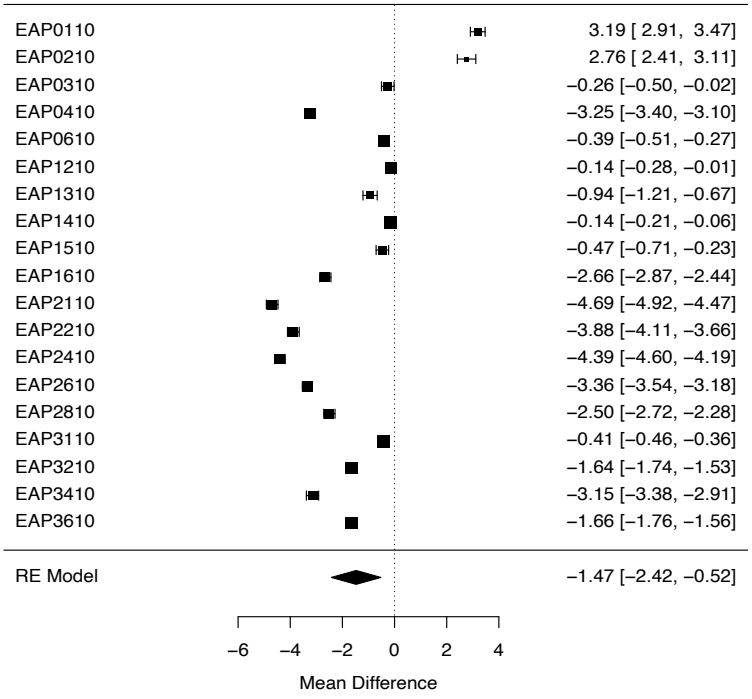


Figura D.60: Forest plot de los coeficientes de PAMED en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.87: Metaanálisis de los coeficientes de EXTRANJERO obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

```
rma(resuC1$EXTRANJERO,sei=resuC1$std.EXTRANJERO, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -13.1433   61.5865   30.2867   32.1755   31.0367

tau^2 (estimated amount of total heterogeneity): 0.2082 (SE = 0.0763)
I^2 (total heterogeneity / total variability):   90.25%
H^2 (total variability / sampling variability):  10.26

Test for Heterogeneity:
  Q(df = 18) = 195.9933, p-val < .0001

Model Results:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.7442  0.1114  6.6831  <.0001  0.5259  0.9625
```

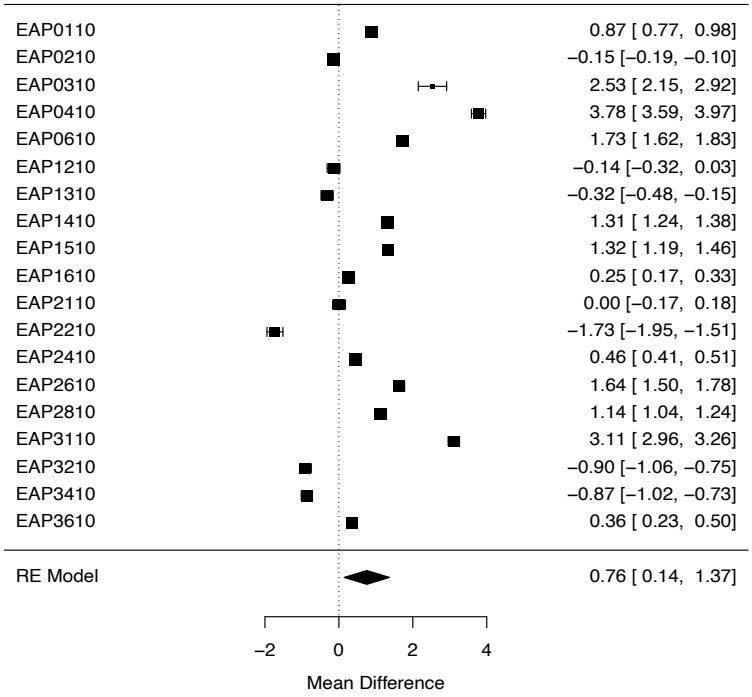


Figura D.61: Forest plot de los coeficientes de PAENF en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.88: Modelo lineal del coeficiente del Intercepto (de las regresiones logísticas intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(formula = Inter ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.Inter^2)

Coefficients:
             Estimate  Std.Error  z-value  Pr(>|z|)
  (Intercept)   1.0594     0.9615    1.102     0.286
  zPOSMAS       0.4448     0.9599    0.463     0.649

Residual standard error: 28.68 on 17 degrees of freedom
Multiple R-squared: 0.01248, Adjusted R-squared: -0.04561
F-statistic: 0.2148 on 1 and 17 DF,  p-value: 0.6489
```

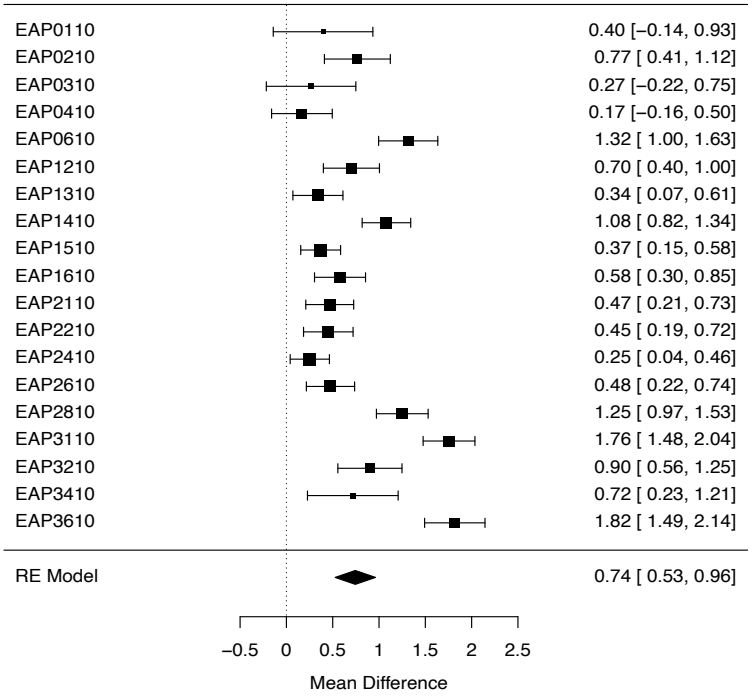


Figura D.62: Forest plot de los coeficientes de EXTRANJERO en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en PESO.

Tabla D.89: Modelo lineal del coeficiente de SEXO (de las regresiones logísticas intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(formula = SEXO ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.SEXO^2)

Coefficients:
             Estimate  Std.Error   z-value   Pr(>|z|)
  (Intercept) -0.220238   0.043224   -5.095   8.98e-05
  zPOSMAS      0.007284   0.043931    0.166       0.87

Residual standard error: 2.259 on 17 degrees of freedom
Multiple R-squared: 0.001615, Adjusted R-squared: -0.05711
F-statistic: 0.02749 on 1 and 17 DF,  p-value: 0.8703
```

Tabla D.90: Modelo lineal del coeficiente de YEAR (de las regresiones logísticas Intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(formula = zYE06 ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.zYE06^2)

Coefficients:
             Estimate  Std.Error  z-value   Pr(>|z|)
  (Intercept)   0.4323     0.1790    2.415     0.0273
  zPOSMAS      -0.1269     0.2350   -0.540     0.5962

Residual standard error: 23.18 on 17 degrees of freedom
Multiple R-squared: 0.01686, Adjusted R-squared: -0.04097
F-statistic: 0.2916 on 1 and 17 DF,  p-value: 0.5962
```

Tabla D.91: Modelo lineal del coeficiente de EDAD (de las regresiones logísticas intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(formula = zED50 ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.zED50^2)

Coefficients:
             Estimate  Std.Error  z-value   Pr(>|z|)
  (Intercept) -0.73272    0.07274  -10.073   1.39e-08
  zPOSMAS      0.01992    0.07444    0.268      0.792

Residual standard error: 12.18 on 17 degrees of freedom
Multiple R-squared: 0.004195, Adjusted R-squared: -0.05438
F-statistic: 0.07161 on 1 and 17 DF,  p-value: 0.7922
```

Tabla D.92: Modelo lineal del coeficiente de PAMED (de las regresiones logísticas intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(formula = zPAMED ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.zPAMED^2)

Coefficients:
             Estimate  Std.Error  z-value   Pr(>|z|)
  (Intercept)  -1.0624     0.3173   -3.348    0.00381
  zPOSMAS      -0.3477     0.3846   -0.904    0.37862

Residual standard error: 21.2 on 17 degrees of freedom
Multiple R-squared: 0.04587, Adjusted R-squared: -0.01026
F-statistic: 0.8173 on 1 and 17 DF,  p-value: 0.3786
```

Tabla D.93: Modelo lineal del coeficiente de PAENF (de las regresiones logísticas intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(formula = zPAENF ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.zPAENF^2)

Coefficients:
             Estimate  Std.Error  z-value  Pr(>|z|)
  (Intercept)   0.5141     0.2081    2.471    0.0244
  zPOSMAS      -0.3259     0.2527   -1.289    0.2146

Residual standard error: 18.03 on 17 degrees of freedom
Multiple R-squared: 0.08907, Adjusted R-squared: 0.03548
F-statistic: 1.662 on 1 and 17 DF,  p-value: 0.2146
```

Tabla D.94: Modelo lineal del coeficiente de EXTRANJERO (de las regresiones logísticas intra-centro con efectos aleatorios para los individuos) en función de POSMAS.

```
Call: lm(EXTRANJERO ~ 1 + zPOSMAS, data = resuC1, weights = 1/std.EXTRANJERO^2)

Coefficients:
             Estimate  Std.Error  z-value   Pr(>|z|)
  (Intercept)  0.71742    0.12064    5.947   1.59e-05
  zPOSMAS      0.01009    0.11328    0.089       0.93

Residual standard error: 3.395 on 17 degrees of freedom
Multiple R-squared: 0.0004665, Adjusted R-squared: -0.05833
F-statistic: 0.007934 on 1 and 17 DF,  p-value: 0.9301
```

D.3.2 *Missing data* en la variable Cigarrillos

D.3.2.1 Modelo logístico simple.

Resultados de los análisis de los *missings* en la variable CIGARRILLOS realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística simple, sin efectos aleatorios.

```
glm(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF,
       data=centro,family="binomial")
```

La Tabla D.95 resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todos los centros con la opción "ML", Tabla D.96 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, Figura D.63 y siguientes.

Tabla D.96: Metaanálisis de los Interceptos de las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJH$Inter,sei=resuJH$std.Inter, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -20.2784   44.6551   44.5568   46.4457   45.3068

tau^2 (estimated amount of total heterogeneity): 0.2636 (SE = 0.1248)
I^2 (total heterogeneity / total variability):   78.80%
H^2 (total variability / sampling variability):  4.72

Test for Heterogeneity:
  Q(df = 18) = 99.9960, p-val < .0001

Fixed-Effects:
  estimate      se      zval    pval    ci.lb    ci.ub
   -1.9150  0.1469  -13.0328  <.0001  -2.2030  -1.6270
```

Tabla D.95: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable CIGARRILLOS realizados en cada centro médico independientemente. Coeficientes de las variables estandarizadas de la regresión logística simple.

EAP	Inter	YE06	ED50	PAMED	PAENF	std.err Inter	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF
0110	-2.286	-0.125	0.152	0.007	-0.000	0.246	0.200	0.052	0.523	0.190
0210	0.470	-0.757	-0.022	-1.125	0.005	0.936	0.207	0.044	0.711	0.087
0310	-1.583	-0.187	0.151	0.151	-0.140	0.240	0.198	0.035	0.350	0.532
0410	-2.640	-0.028	0.014	-0.108	0.015	0.236	0.172	0.028	0.219	0.270
0610	-2.473	-0.041	0.160	0.030	-0.020	0.269	0.039	0.030	0.205	0.164
1210	-2.499	-0.043	-0.001	-0.125	0.132	0.381	0.170	0.038	0.195	0.273
1310	-1.800	-0.070	-0.003	0.437	-0.283	0.336	0.126	0.032	0.381	0.236
1410	-2.516	-0.192	0.151	0.002	-0.060	0.085	0.068	0.032	0.121	0.101
1510	-0.733	-0.656	-0.044	-0.231	-0.546	0.232	0.203	0.035	0.361	0.202
1610	-1.690	-0.229	0.175	0.156	-0.060	0.183	0.137	0.036	0.347	0.131
2110	-1.577	-0.240	-0.240	0.369	-0.104	0.513	0.062	0.031	0.334	0.279
2210	-1.928	-0.070	0.016	0.091	0.250	1.308	0.340	0.036	0.423	0.408
2410	-1.245	-0.350	0.164	-0.242	-0.244	0.376	0.160	0.025	0.317	0.094
2610	-1.444	-0.276	0.141	0.352	-0.435	0.297	0.098	0.028	0.281	0.248
2810	-2.123	-0.126	-0.074	0.258	-0.327	0.275	0.076	0.037	0.339	0.164
3110	-2.636	-0.091	-0.006	-0.007	0.012	0.402	0.170	0.041	0.091	0.264
3210	-2.569	-0.128	0.106	0.186	0.060	0.566	0.135	0.031	0.193	0.307
3410	-0.516	-0.499	0.129	-0.519	-0.242	1.194	0.283	0.038	0.390	0.244
3610	-2.001	-0.254	0.184	0.098	-0.262	0.367	0.118	0.031	0.161	0.225

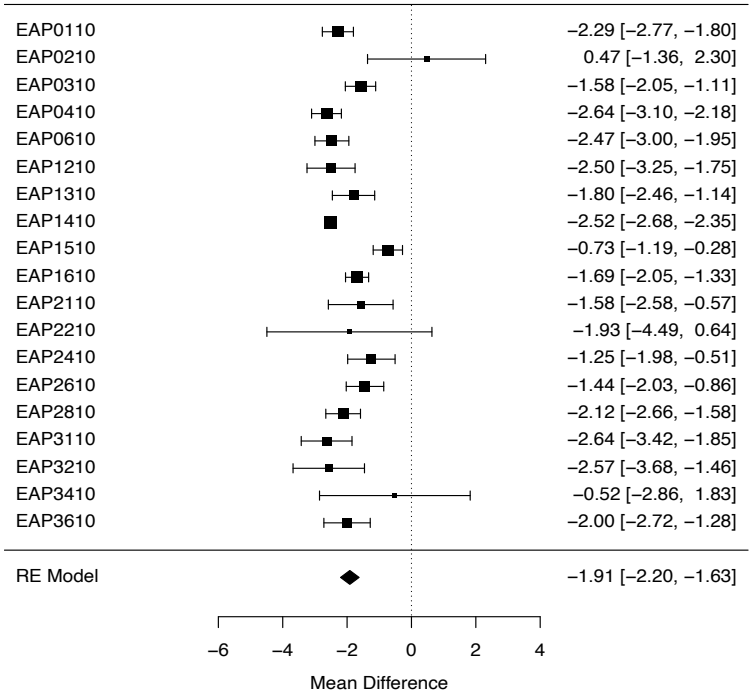


Figura D.63: Forest plot de los Interceptos de las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

Tabla D.97: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJH$zYE06,sei=resuJH$std.zYE06, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
  6.5941   27.2877   -9.1883   -7.2994   -8.4383

tau^2 (estimated amount of total heterogeneity): 0.0054 (SE = 0.0054)
I^2 (total heterogeneity / total variability):   32.64%
H^2 (total variability / sampling variability):  1.48

Test for Heterogeneity:
  Q(df = 18) = 32.5481, p-val = 0.0189

Fixed-Effects:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.1870  0.0331  -5.6536  <.0001  -0.2518  -0.1222
```

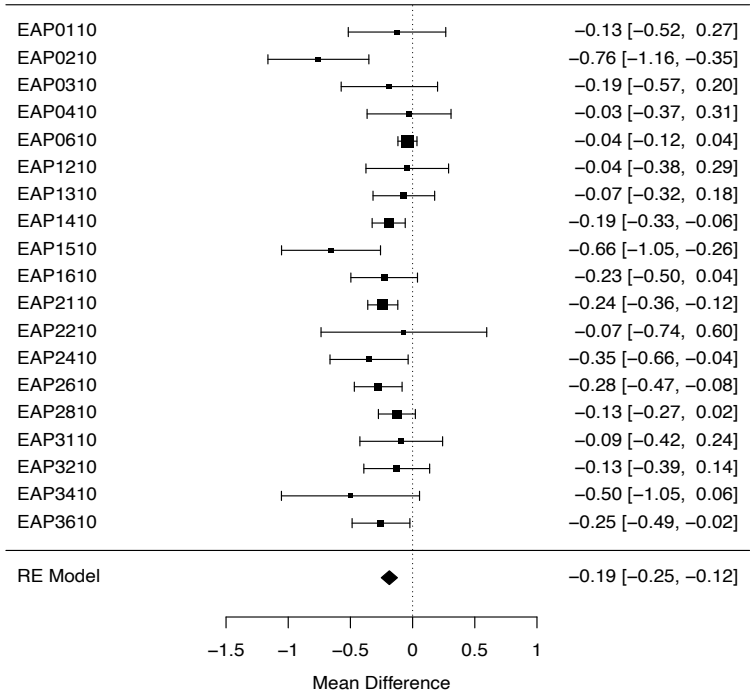


Figura D.64: Forest plot de los coeficientes de YEAR obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

Tabla D.98: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJH$zED50,sei=resuJH$std.zED50, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
 14.9390   63.4232  -25.8780  -23.9891  -25.1280

tau^2 (estimated amount of total heterogeneity): 0.0110 (SE = 0.0040)
I^2 (total heterogeneity / total variability):   90.79%
H^2 (total variability / sampling variability):  10.86

Test for Heterogeneity:
  Q(df = 18) = 214.9519, p-val < .0001

Fixed-Effects:
  estimate      se    zval    pval   ci.lb   ci.ub
    0.0610  0.0254  2.4023  0.0163  0.0112  0.1107
```

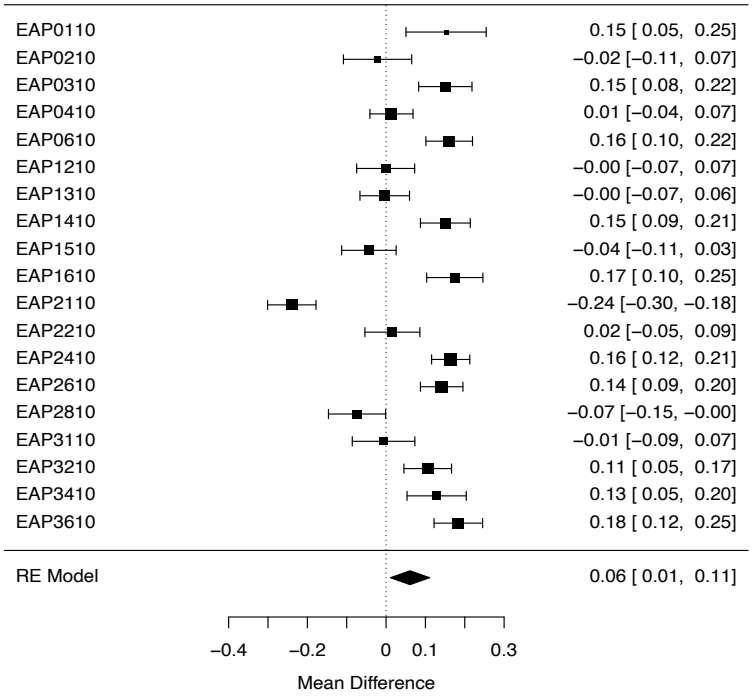


Figura D.65: Forest plot de los coeficientes de EDAD obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

Tabla D.99: Metaanálisis de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJH$zPAMED,sei=resuJH$std.zPAMED, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
  0.6691   12.1386    2.6618    4.5507    3.4118

tau^2 (estimated amount of total heterogeneity): 0.0000 (SE = 0.0092)
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00

Test for Heterogeneity:
  Q(df = 18) = 12.1383, p-val = 0.8400

Fixed-Effects:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.0232  0.0491  0.4722  0.6368  -0.0731  0.1195
```

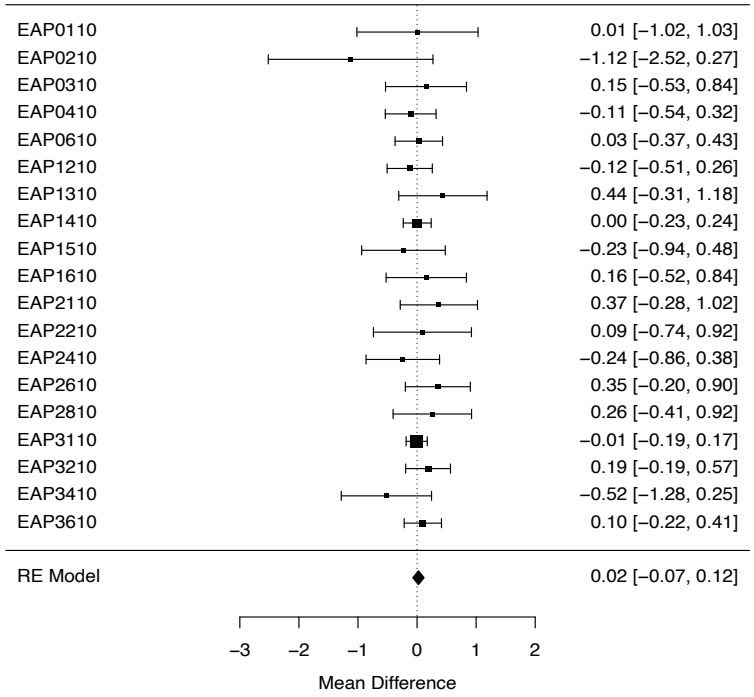


Figura D.66: Forest plot de los coeficientes de PAMED obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

Tabla D.100: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJH$zPAENF,sei=resuJH$std.zPAENF, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
  4.0386   16.3840   -4.0772   -2.1884   -3.3272

tau^2 (estimated amount of total heterogeneity): 0.0018 (SE = 0.0075)
I^2 (total heterogeneity / total variability):   5.79%
H^2 (total variability / sampling variability):  1.06

Test for Heterogeneity:
  Q(df = 18) = 16.4790, p-val = 0.5592

Fixed-Effects:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.1234  0.0409  -3.0168  0.0026  -0.2035  -0.0432
```

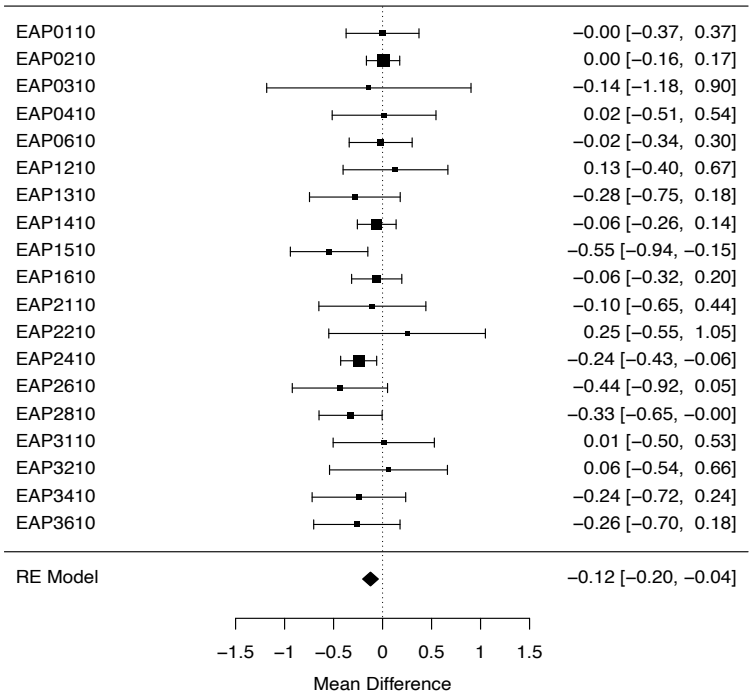


Figura D.67: Forest plot de los coeficientes de PAENF obtenidos en las regresiones logísticas simples intra-centro para la *odds* de *missings* en CIGARRILLOS.

D.3.2.2 Modelo logístico con efectos aleatorios para los individuos.

Resultados de los análisis de los *missings* en la variable CIGARRILLOS realizados con los datos de cada centro médico independientemente, dentro de cada cual se ha ajustado un modelo de regresión logística con efectos aleatorios para los individuos.

```
glmmTMB(mCIGARRILLOS ~ zYE06 + zED50 + zPAMED + zPAENF + (1|ID),
        data=centro, family="binomial", REML=F)
```

La Tabla D.101 resume los coeficientes obtenidos en dichos análisis intra-centro para las diferentes variables junto con sus errores estándar. A partir de esos datos, para cada uno de los coeficientes hemos realizado un meta-análisis combinando los resultados de todos los centros con la opción "ML", Tabla D.102 y siguientes. Los resultados, además de ser mostrados numéricamente, son mostrados gráficamente mediante forest plots, Figura D.68 y siguientes.

A continuación mostramos el resultado del meta-análisis de los coeficientes obtenidos en

Tabla D.101: Resumen de los resultados obtenidos en los análisis de la *odds* de *missing data* en la variable CIGARRILLOS realizados en cada centro médico independientemente. Coeficientes de las variables estandarizadas de la regresión logística con efectos aleatorios para los individuos.

EAP	Inter	YE06	ED50	PAMED	PAENF	std.err Inter	std.err YE06	std.err ED50	std.err PAMED	std.err PAENF
0110	-12.146	-0.731	0.171	1.961	0.224	1.260	0.829	0.445	2.157	0.791
0210	-9.317	-1.773	-0.202	-0.791	0.118	3.611	0.841	0.387	2.545	0.344
0310	-10.635	-1.508	0.266	0.644	-0.316	1.171	0.824	0.310	1.473	2.191
0410	-13.459	-1.598	0.081	-0.836	-0.765	1.283	0.867	0.273	1.052	1.400
0610	-12.127	-1.538	0.201	-0.177	-1.687	1.375	0.279	0.291	0.910	0.841
1210	-16.257	-0.791	0.093	-1.971	1.328	2.530	0.846	0.356	1.076	1.367
1310	-12.357	-0.507	0.012	2.419	-1.595	1.207	0.432	0.257	1.311	0.827
1410	-16.154	-15.874	0.359	-0.799	-0.544	0.932	1.202	0.486	1.043	0.809
1510	-15.260	-15.346	0.075	0.207	-0.634	2.166	2.051	0.580	2.774	1.732
1610	-15.882	-16.558	0.241	-1.390	-1.081	1.651	1.832	0.634	2.453	1.445
2110	-18.729	-16.340	-0.283	-0.769	5.537	3.614	1.095	0.518	2.346	3.484
2210	-10.318	-2.135	-0.136	-1.102	0.924	6.550	1.727	0.351	2.124	1.995
2410	-14.966	-13.386	0.201	0.972	-0.682	5.222	1.989	0.396	3.229	1.372
2610	-15.392	-15.123	0.139	2.078	-1.824	3.009	1.229	0.433	2.041	2.855
2810	-12.871	-1.773	-0.023	3.015	-1.792	1.872	0.534	0.326	2.038	0.786
3110	-12.764	-0.689	0.040	0.068	0.286	1.684	0.654	0.356	0.368	1.040
3210	-11.419	-1.420	0.102	0.144	-0.031	1.955	0.505	0.252	0.680	1.119
3410	2.365	-4.750	0.195	-4.727	-2.013	4.918	1.323	0.318	1.679	1.103
3610	-2.584	-6.028	0.222	-2.406	-4.975	5.074	2.304	0.346	1.601	2.702

los análisis intra-centro con efectos aleatorios, y los representamos gráficacamente mediante forest plots.

Tabla D.102: Metaanálisis de los Interceptos de las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJ1$Inter,sei=resuJ1$std.Inter, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -50.6452   35.3409  105.2904  107.1793  106.0404

tau^2 (estimated amount of total heterogeneity): 2.2143 (SE = 1.7591)
I^2 (total heterogeneity / total variability):   41.49%
H^2 (total variability / sampling variability):  1.71

Test for Heterogeneity:
  Q(df = 18) = 41.3303, p-val = 0.0014

Fixed-Effects:
  estimate      se      zval    pval     ci.lb     ci.ub
  -13.0756  0.5780  -22.6203  <.0001  -14.2086  -11.9427
```

Tabla D.103: Metaanálisis de los coeficientes de YEAR obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJ1$zYE06,sei=resuJ1$std.zYE06, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
-62.2663   90.7426  128.5325  130.4214  129.2825

tau^2 (estimated amount of total heterogeneity): 39.2838 (SE = 13.2395)
I^2 (total heterogeneity / total variability):   98.63%
H^2 (total variability / sampling variability):  73.08

Test for Heterogeneity:
  Q(df = 18) = 581.6777, p-val < .0001

Fixed-Effects:
  estimate      se     zval    pval    ci.lb    ci.ub
   -6.0742  1.4660  -4.1434  <.0001  -8.9475  -3.2009
```

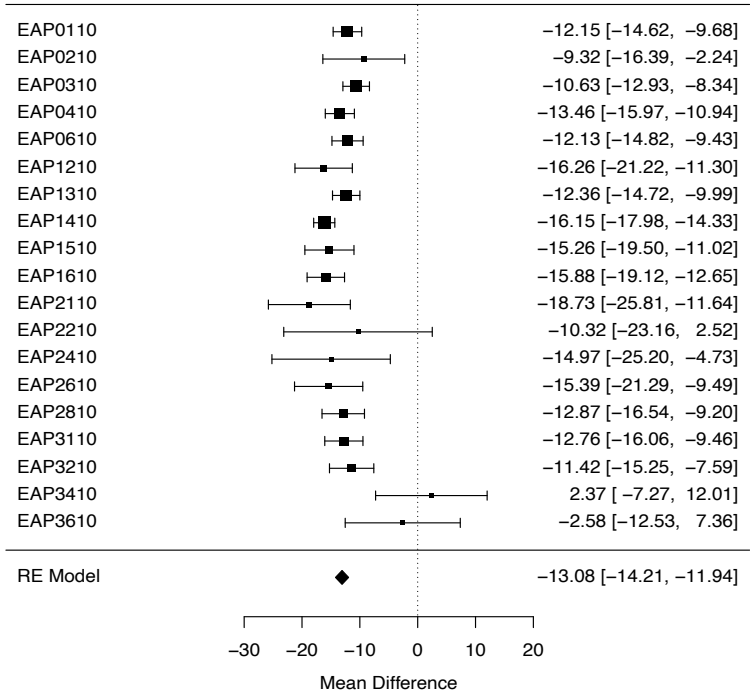


Figura D.68: Forest plot de los Interceptos de las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.104: Metaanálisis de los coeficientes de EDAD obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJ1$zED50,sei=resuJ1$std.zED50, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
 -0.1460    2.9438    4.2921    6.1810    5.0421

tau^2 (estimated amount of total heterogeneity): 0 (SE = 0.0360)
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00

Test for Heterogeneity:
  Q(df = 18) = 2.9438, p-val = 1.0000

Fixed-Effects:
  estimate      se    zval    pval    ci.lb   ci.ub
    0.0927  0.0803  1.1539  0.2485  -0.0647  0.2501
```

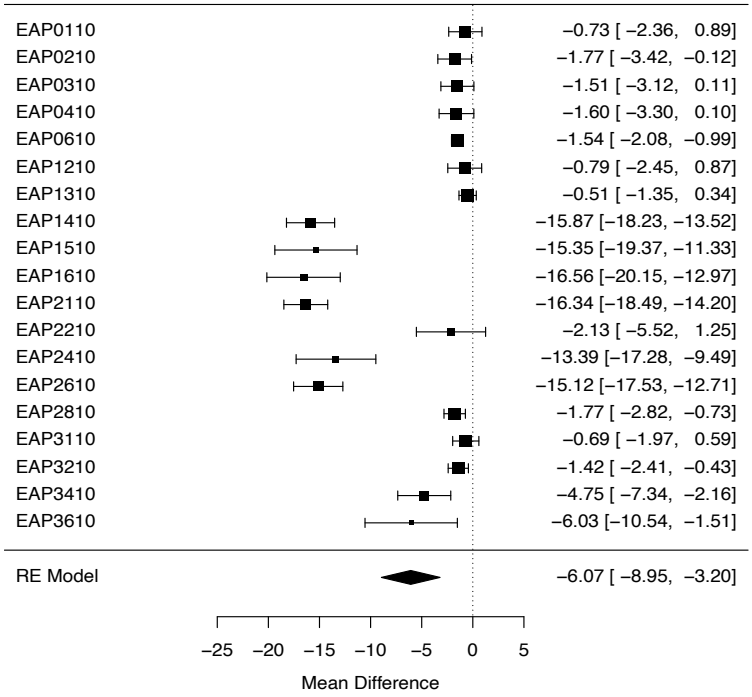


Figura D.69: Forest plot de los coeficientes de YEAR en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.105: Metaanálisis de los coeficientes de PAMED de las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJ1$zPAMED,sei=resuJ1$std.zPAMED, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
   logLik  deviance       AIC       BIC      AICc
 -37.1402   23.0437   78.2804   80.1693   79.0304

tau^2 (estimated amount of total heterogeneity): 0 (SE = 0.1764)
I^2 (total heterogeneity / total variability):   0.00%
H^2 (total variability / sampling variability):  1.00

Test for Heterogeneity:
  Q(df = 18) = 23.0437, p-val = 0.1889

Fixed-Effects:
  estimate      se     zval    pval    ci.lb   ci.ub
   -0.1465  0.2434  -0.6020  0.5472  -0.6236  0.3305
```

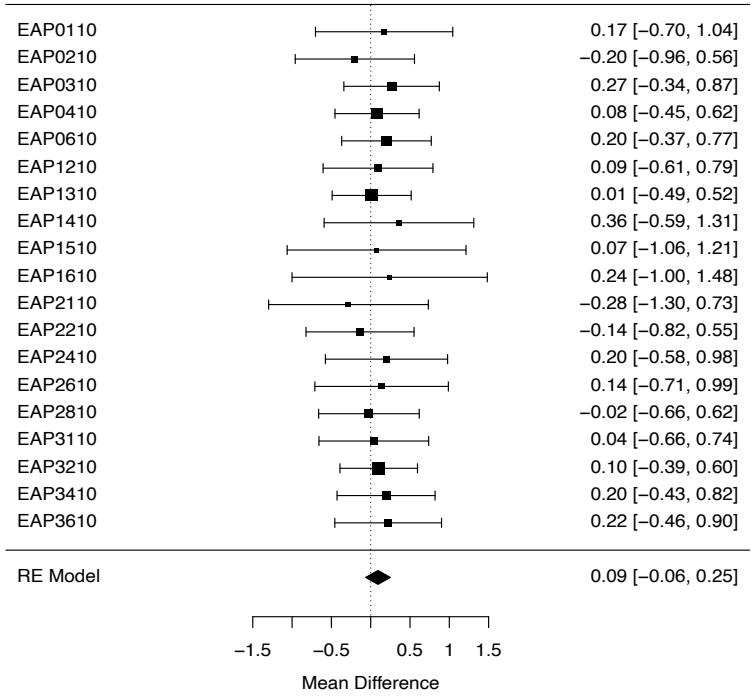


Figura D.70: Forest plot de los coeficientes de EDAD en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Tabla D.106: Metaanálisis de los coeficientes de PAENF obtenidos en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

```
rma(resuJ1$zPAENF,sei=resuJ1$std.zPAENF, method="ML", measure="MD")

Random-Effects Model (k = 19; tau^2 estimator: ML)
  logLik  deviance       AIC       BIC      AICc
-32.4960   20.5125   68.9921   70.8810   69.7421

tau^2 (estimated amount of total heterogeneity): 0.1949 (SE = 0.3177)
I^2 (total heterogeneity / total variability):   16.19%
H^2 (total variability / sampling variability):  1.19

Test for Heterogeneity:
  Q(df = 18) = 21.5529, p-val = 0.2524

Fixed-Effects:
  estimate      se     zval    pval    ci.lb    ci.ub
   -0.5696  0.2653  -2.1470  0.0318  -1.0896  -0.0496
```

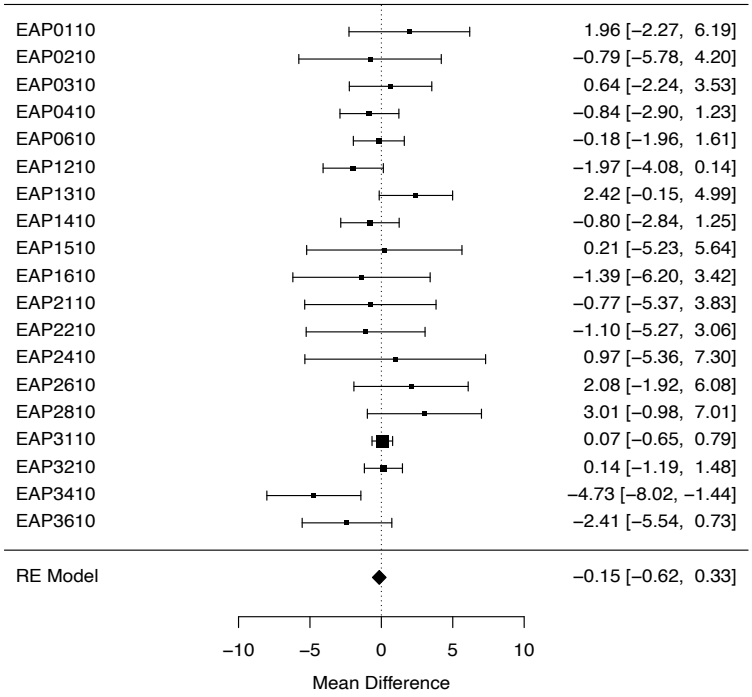


Figura D.71: Forest plot de los coeficientes de PAMED en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

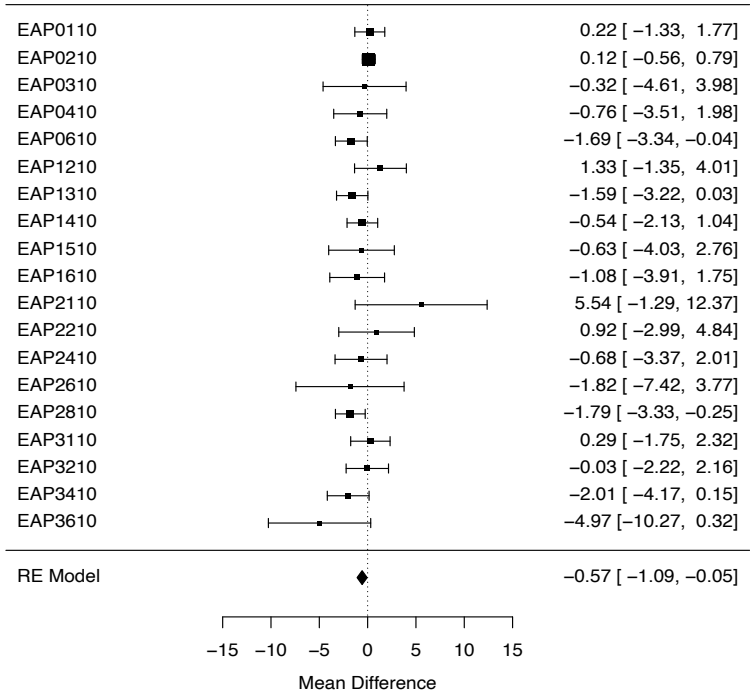


Figura D.72: Forest plot de los coeficientes de PAENF en las regresiones logísticas intra-centro con efectos aleatorios de los individuos para la *odds* de *missings* en CIGARRILLOS.

Apéndice E

DETALLES DE LOS ANÁLISIS DE LAS SECUENCIAS DE EVENTOS

A continuación adjuntamos información adicional (evento inicial, evento final, número de ocurrencias de cada par, tiempo de exposición al riesgo, IR y el $IC_{95\%}$ de IR) de los análisis de secuencias de eventos realizados en la tesis: a) utilizando datos con fechas completas, b) utilizando datos con fechas truncadas y c) simulando fechas truncadas a partir de datos con fechas completas. Los datos truncados han sido imputados de dos modos: asignando la misma probabilidad a todas las secuencias posibles y asignando probabilidades a priori proporcionales a las frecuencias calculadas con los datos interanuales.

En todos los casos sólo mostramos los 24 pares o tripletes con mayor IR y con al menos 5 ocurrencias. Los intervalos de confianza, IC, han sido calculados utilizando el método exacto de Poisson, Ecuación 3.1 del Apartado 3.2.2, tanto para los datos con fechas exactas como para los datos truncados. No se ha tenido en cuenta el error introducido por el proceso de imputación.

La Tabla E.1 muestra información detallada de los análisis del Apartado 4.2.2.1 para los pares de eventos consecutivos con mayor IR, calculados utilizando fechas completas. Vemos como el diagnóstico (al menos una vez) en una misma persona de ATERO_PER seguido en cualquier momento de un diagnóstico TR.LIPIDOS tiene un IR de 82.12 casos por cada 1000 personas-año, con un $IC_{95\%}$ entre 41.0 y 146.9.

Tabla E.1: Información de los pares de eventos consecutivos con mayor IR (ocurrencias por cada 1000 personas-año), calculados utilizando fechas completas.

From	To	Casos	Tiempo	IR	IC
ATERO_PER	TR.LIPIDOS	11	133.95	82.12	(41.0,146.9)
DIABETES	TR.LIPIDOS	5915	76860.77	76.96	(75.0,78.9)
HTA	TR.LIPIDOS	18650	246099.80	75.78	(74.7,76.9)
CARD_ISQ	HTA	2031	30799.55	65.94	(63.1,68.9)
INSF_CARD	ARRITMIAS	623	10080.89	61.80	(57,66.8)
OBESIDAD	TR.LIPIDOS	6197	113113.41	54.79	(53.4,56.2)
ARRITMIAS	HTA	2633	56267.81	46.79	(45,48.6)
CARD_ISQ	TR.LIPIDOS	1434	30799.55	46.56	(44.2,49.0)
ACV	TR.LIPIDOS	807	18868.82	42.77	(39.9,45.8)
DIABETES	HTA	3078	76861.62	40.05	(38.6,41.5)
HTA	OBESIDAD	9839	246098.60	39.98	(39.2,40.8)
VALVULOPA	ARRITMIAS	625	16270.57	38.41	(35.5,41.5)
ACV	HTA	694	18868	36.78	(34.1,39.6)
ARRITMIAS	TR.LIPIDOS	1952	56267.81	34.69	(33.2,36.3)
DIABETES	OBESIDAD	2596	76863.87	33.77	(32.5,35.1)
VALVULOPA	HTA	537	16270.57	33	(30.3,35.9)
HTA	DIABETES	7551	246101.27	30.68	(30,31.4)
OBESIDAD	HTA	3336	113113.02	29.49	(28.5,30.5)
TR.LIPIDOS	HTA	7635	268953.15	28.39	(27.8,29)
VALVULOPA	TR.LIPIDOS	461	16270.57	28.33	(25.8,31)
INSF_CARD	HTA	283	10080.89	28.07	(24.9,31.5)
ALCOHOL	TR.LIPIDOS	367	13593.87	27	(24.3,29.9)
ALCOHOL	HTA	342	13593.87	25.16	(22.6,28)
ACV	ARRITMIAS	454	18867.69	24.06	(21.9,26.4)

La Tabla E.2 muestra información detallada de los análisis del Apartado 4.2.2.2 para los pares de eventos, tanto consecutivos como no consecutivos, con mayor IR, calculados utilizando fechas completas.

Tabla E.2: Información de los pares de eventos consecutivos y no consecutivos (ocurrencias por cada 1000 personas-año), calculados utilizando fechas completas.

From	To	Casos	Tiempo	IR	IC
HTA	TR.LIPIDOS	27711	357123.18	77.60	(76.7,78.5)
DIABETES	TR.LIPIDOS	8599	112039.17	76.75	(75.1,78.4)
ATERO_PER	TR.LIPIDOS	14	200.71	69.75	(38.1,117)
CARD_ISQ	TR.LIPIDOS	3029	47458.44	63.82	(61.6,66.1)
INSF_CARD	ARRITMIAS	830	15031	55.22	(51.5,59.1)
CARD_ISQ	HTA	2716	49325.52	55.06	(53,57.2)
OBESIDAD	TR.LIPIDOS	8164	148606.14	54.94	(53.8,56.1)
ACV	TR.LIPIDOS	1359	26338.64	51.60	(48.9,54.4)
ARRITMIAS	HTA	3418	77648.65	44.02	(42.6,45.5)
ARRITMIAS	TR.LIPIDOS	3381	80674.52	41.91	(40.5,43.3)
VALVULOPA	TR.LIPIDOS	950	24607.84	38.61	(36.2,41.1)
VALVULOPA	ARRITMIAS	933	24471.87	38.13	(35.7,40.7)
DIABETES	HTA	4944	132060.33	37.44	(36.4,38.5)
HTA	OBESIDAD	15681	424873.33	36.91	(36.3,37.5)
VALVULOPA	HTA	871	24125.72	36.10	(33.7,38.6)
ACV	HTA	989	27931.21	35.41	(33.2,37.7)
ATERO_PER	ARRITMIAS	8	230.82	34.66	(15,68.3)
INSF_CARD	TR.LIPIDOS	528	16286.90	32.42	(29.7,35.3)
DIABETES	OBESIDAD	4292	135409.41	31.70	(30.8,32.7)
ALCOHOL	TR.LIPIDOS	559	17785	31.43	(28.9,34.1)
INSF_CARD	HTA	480	16018.92	29.96	(27.3,32.8)
ATERO_PER	HTA	7	235.52	29.72	(11.9,61.2)
OBESIDAD	HTA	4663	165496.33	28.18	(27.4,29)
TR.LIPIDOS	HTA	9582	340449.55	28.15	(27.6,28.7)

La Tabla E.3 muestra información detallada de los análisis del Apartado 4.2.2.3 para los tripletes de eventos, tanto consecutivos como no consecutivos, calculados utilizando fechas completas. Vemos como el diagnóstico (al menos una vez) en una misma persona de DIABETES seguido en cualquier momento de otro diagnóstico DIABETES y luego de TRASTLIPIDOS tiene un IR de 114.39 casos por cada 1000 personas-año, con un $IC_{95\%}$ entre 74.7 a 167.6.

Tabla E.3: Información de los tripletes de eventos tanto consecutivos como no consecutivos (ocurrencias por cada 1000 personas-año), calculados utilizando fechas completas.

E1	E2	E3	Casos	Tiempo	IR	IC
DIABETES	DIABETES	TR.LIPIDOS	26	227.29	114.39	(74.7,167.6)
CARD_ISQ	HTA	TR.LIPIDOS	1134	10014.61	113.23	(106.7,120)
HTA	HTA	TR.LIPIDOS	58	591.66	98.03	(74.4,126.7)
HTA	DIABETES	TR.LIPIDOS	4003	45362.78	88.24	(85.5,91)
ACV	HTA	TR.LIPIDOS	307	3625.95	84.67	(75.5,94.7)
CARD_ISQ	DIABETES	TR.LIPIDOS	466	5571.20	83.64	(76.2,91.6)
ACV	DIABETES	TR.LIPIDOS	165	1978.88	83.38	(71.1,97.1)
CARD_ISQ	CARD_ISQ	TR.LIPIDOS	12	151.07	79.43	(41,138.8)
INS_CARD	HTA	TR.LIPIDOS	130	1708.47	76.09	(63.6,90.4)
INS_CARD	VALVULO	ARRITMIA	64	860.85	74.34	(57.3,94.9)
VALVULO	DIABETES	TR.LIPIDOS	111	1508.42	73.59	(60.5,88.6)
ALCOHOL	INS_CARD	ARRITMIA	11	151.13	72.79	(36.3,130.2)
TABAQUIS	DIABETES	TR.LIPIDOS	154	2177.69	70.72	(60.0,82.8)
ARRITMIA	HTA	TR.LIPIDOS	911	13089.27	69.60	(65.2,74.3)
DIABETES	OBESIDAD	TR.LIPIDOS	1170	17401.05	67.24	(63.4,71.2)
VALVULO	HTA	TR.LIPIDOS	256	3865.76	66.22	(58.4,74.9)
DIABETES	HTA	TR.LIPIDOS	1317	20606.63	63.91	(60.5,67.5)
HTA	OBESIDAD	TR.LIPIDOS	4243	66481.03	63.82	(61.9,65.8)
ARRITMIA	DIABETES	TR.LIPIDOS	337	5345.67	63.04	(56.5,70.1)
ALCOHOL	DIABETES	TR.LIPIDOS	44	703.07	62.58	(45.5,84.0)
ACV	OBESIDAD	TR.LIPIDOS	103	1705.10	60.41	(49.3,73.3)
INS_CARD	DIABETES	TR.LIPIDOS	74	1243.50	59.51	(46.7,74.7)
TABAQUIS	HTA	TR.LIPIDOS	325	5535.44	58.71	(52.5,65.5)
ACV	INS_CARD	ARRITMIA	52	887.82	58.57	(43.7,76.8)

La Tabla E.4 muestra información detallada de los análisis del Apartado 4.2.3.1 para los pares de eventos, tanto consecutivos como no consecutivos. Los cálculos se ha realizado simulando fechas truncadas a partir de las originales y asignando la misma probabilidad de ocurrencia a todas las secuencias posibles.

Tabla E.4: Información de los pares de eventos consecutivos y no consecutivos con simulación de la fecha truncada e imputación equiprobable (ocurrencias por cada 1000 personas-año).

From	To	Casos	Tiempo	IR	IC
DIABETES	HTA	6913.83	117611.92	58.79	(57.4,60.2)
HTA	TR.LIPIDOS	22559.00	392785.92	57.43	(56.7,58.2)
DIABETES	TR.LIPIDOS	6934.83	123169.25	56.30	(55,57.6)
CARD_ISQ	TR.LIPIDOS	2502.00	51308.92	48.76	(46.9,50.7)
ACV	HTA	1265.17	26161.25	48.36	(45.7,51.1)
TR.LIPIDOS	HTA	14697.92	304057.00	48.34	(47.6,49.1)
ATERO_PER	TR.LIPIDOS	10.50	218.67	48.02	(23.5,87.1)
OBESIDAD	HTA	7052.33	148125.42	47.61	(46.5,48.7)
INSF_CARD	ARRITMIAS	736.67	15600.42	47.22	(43.9,50.8)
OBESIDAD	TR.LIPIDOS	7315.08	155208.00	47.13	(46.1,48.2)
CARD_ISQ	HTA	2353.50	51982.08	45.28	(43.5,47.1)
ACV	TR.LIPIDOS	1182.67	27412.08	43.14	(40.7,45.7)
ARRITMIAS	HTA	3055.17	80163.87	38.11	(36.8,39.5)
ARRITMIAS	TR.LIPIDOS	3068.58	82481.70	37.20	(35.9,38.5)
ATERO_PER	ARRITMIAS	8.50	230.33	36.90	(16.4,71.3)
ALCOHOL	TR.LIPIDOS	627.50	17312.17	36.25	(33.5,39.2)
ATERO_PER	HTA	8.00	229.17	34.91	(15.1,68.8)
VALVULOPA	TR.LIPIDOS	859.17	25179.50	34.12	(31.9,36.5)
VALVULOPA	HTA	808.50	24591.33	32.88	(30.6,35.2)
VALVULOPA	ARRITMIAS	816.33	25401.33	32.14	(30,34.4)
ALCOHOL	HTA	545.00	17562.33	31.03	(28.5,33.8)
HTA	OBESIDAD	13262.83	441045.00	30.07	(29.6,30.6)
INSF_CARD	TR.LIPIDOS	483.50	16472.75	29.35	(26.8,32.1)
TABAQUISMO	TR.LIPIDOS	2484.33	85324.75	29.12	(28,30.3)

La Tabla E.5 muestra información detallada de los análisis del Apartado 4.2.3.1 para los pares de eventos, tanto consecutivos como no consecutivos. Los cálculos se ha realizado simulando fechas truncadas a partir de las originales y asignando a las secuencias candidatas probabilidades de ocurrencia proporcionales a las frecuencias obtenidas previamente de los pares de eventos interanuales.

Tabla E.5: Información de los pares de eventos consecutivos y no consecutivos con simulación de la fecha truncada e imputación con probabilidad interanual (ocurrencias por cada 1000 personas-año).

From	To	Casos	Tiempo	IR	IC
DIABETES	HTA	6913.83	117611.92	58.79	(57.4,60.2)
HTA	TR.LIPIDOS	22558.99	392785.92	57.43	(56.7,58.2)
DIABETES	TR.LIPIDOS	6934.83	123169.25	56.30	(55,57.6)
CARD_ISQ	TR.LIPIDOS	2502.00	51308.92	48.76	(46.9,50.7)
ACV	HTA	1265.17	26161.25	48.36	(45.7,51.1)
TR.LIPIDOS	HTA	14697.88	304056.21	48.34	(47.6,49.1)
ATERO_PER	TR.LIPIDOS	10.50	218.67	48.02	(23.5,87.1)
OBESIDAD	HTA	7052.27	148123.94	47.61	(46.5,48.7)
INSF_CARD	ARRITMIAS	736.67	15600.42	47.22	(43.9,50.8)
OBESIDAD	TR.LIPIDOS	7314.99	155206.57	47.13	(46.1,48.2)
CARD_ISQ	HTA	2353.50	51982.08	45.28	(43.5,47.1)
ACV	TR.LIPIDOS	1182.67	27412.08	43.14	(40.7,45.7)
ARRITMIAS	HTA	3055.17	80163.55	38.11	(36.8,39.5)
ARRITMIAS	TR.LIPIDOS	3068.48	82481.83	37.20	(35.9,38.5)
ATERO_PER	ARRITMIAS	8.50	230.33	36.90	(16.4,71.3)
ALCOHOL	TR.LIPIDOS	627.50	17312.17	36.25	(33.5,39.2)
ATERO_PER	HTA	8.00	229.17	34.91	(15.1,68.8)
VALVULOPA	TR.LIPIDOS	859.17	25179.50	34.12	(31.9,36.5)
VALVULOPA	HTA	808.50	24591.33	32.88	(30.6,35.2)
VALVULOPA	ARRITMIAS	816.33	25401.33	32.14	(30,34.4)
ALCOHOL	HTA	545.00	17562.33	31.03	(28.5,33.8)
HTA	OBESIDAD	13262.83	441044.90	30.07	(29.6,30.6)
INSF_CARD	TR.LIPIDOS	483.50	16472.75	29.35	(26.8,32.1)
TABAQUISMO	TR.LIPIDOS	2484.33	85324.50	29.12	(28,30.3)

La Tabla E.6 muestra información detallada de los análisis del Apartado 4.2.3.2 para los tripletes de eventos, tanto consecutivos como no consecutivos. Los cálculos se ha realizado simulando fechas truncadas a partir de las originales y asignando la misma probabilidad de ocurrencia a todas las secuencias posibles.

Tabla E.6: Información de los tripletes de eventos simulando fechas truncadas y con imputación equiprobable (ocurrencias por cada 1000 personas-año).

E1	E2	E3	Casos	Tiempo	IR	IC
DIABETES	DIABETES	TR.LIPIDOS	17.83	224.42	79.47	(47.0,125.9)
CARD_ISQ	CARD_ISQ	TR.LIPIDOS	10.50	133.75	78.50	(38.4,142.3)
HTA	HTA	TR.LIPIDOS	42.83	601.08	71.26	(51.5,96.0)
ALCOHOL	INSF_CARD	ARRITMIA	12.00	187.33	64.06	(33.1,111.9)
ACV	INSF_CARD	ARRITMIA	68.00	1080.92	62.91	(48.9,79.8)
DIABETES	DIABETES	HTA	13.50	232.33	58.11	(31.4,98.4)
TABAQUIS	DIABETES	TR.LIPIDOS	167.75	2973.04	56.42	(48.2,65.6)
DIABETES	HTA	TR.LIPIDOS	2469.33	45132.66	54.71	(52.6,56.9)
TR.LIPIDOS	TR.LIPIDOS	HTA	16.67	316.75	52.62	(30.5,84.6)
ACV	DIABETES	TR.LIPIDOS	138.50	2648.62	52.29	(43.9,61.8)
CARD_ISQ	HTA	TR.LIPIDOS	746.88	14401.32	51.86	(48.2,55.7)
HTA	DIABETES	TR.LIPIDOS	2746.53	53363.99	51.47	(49.6,53.4)
ACV	HTA	TR.LIPIDOS	375.42	7303.04	51.41	(46.3,56.9)
VALVULOPA	DIABETES	TR.LIPIDOS	90.00	1757.92	51.20	(41.2,62.9)
ACV	DIABETES	HTA	132.54	2595.40	51.07	(42.7,60.5)
OBESIDAD	DIABETES	TR.LIPIDOS	975.33	19194.51	50.81	(47.7,54.1)
OBESIDAD	HTA	TR.LIPIDOS	2397.42	48223.65	49.71	(47.7,51.7)
ALCOHOL	HTA	TR.LIPIDOS	131.67	2655.54	49.58	(41.5,58.8)
OBESIDAD	INSF_CARD	ARRITMIA	159.83	3238.25	49.36	(42.0,57.6)
ACV	TABAQUIS	HTA	24.00	488.83	49.10	(31.5,73.1)
TABAQUIS	HTA	TR.LIPIDOS	398.33	8160.58	48.81	(44.1,53.8)
DIABETES	OBESIDAD	TR.LIPIDOS	987.58	20249.42	48.77	(45.8,51.9)
CARD_ISQ	CARD_ISQ	DIABETES	7.50	154.00	48.70	(20.3,98.0)
INSF_CARD	VALVULOPA	ARRITMIA	54.83	1131.90	48.44	(36.5,63.1)

La Tabla E.7 muestra información detallada de los análisis del Apartado 4.2.3.2 para los tripletes de eventos, tanto consecutivos como no consecutivos. Los cálculos se ha realizado simulando fechas truncadas a partir de las originales y asignando a las secuencias candidatas probabilidades de ocurrencia proporcionales a las frecuencias obtenidas previamente de los pares de eventos interanuales.

Tabla E.7: Información de los tripletes de eventos simulando fechas truncadas y con imputación con probabilidad interanual (ocurrencias por cada 1000 personas-año).

E1	E2	E3	Casos	Tiempo	IR	IC
DIABETES	DIABETES	TR.LIPIDOS	18.02	223.87	80.49	(47.7,127.2)
CARD_ISQ	CARD_ISQ	TR.LIPIDOS	10.65	133.49	79.78	(39.3,144.1)
HTA	HTA	TR.LIPIDOS	42.94	600.53	71.50	(51.7,96.3)
ALCOHOL	INSF_CARD	ARRITMIAS	12.01	187.30	64.14	(33.2,112.0)
ACV	INSF_CARD	ARRITMIAS	68.01	1080.89	62.92	(48.9,79.8)
DIABETES	DIABETES	HTA	13.78	231.80	59.45	(32.3,100.1)
TABAQUIS	DIABETES	TR.LIPIDOS	167.86	2972.52	56.47	(48.3,65.7)
DIABETES	HTA	TR.LIPIDOS	2470.42	45126.04	54.74	(52.6,56.9)
TR.LIPIDOS	TR.LIPIDOS	HTA	17.07	315.92	54.03	(31.5,86.4)
ACV	DIABETES	TR.LIPIDOS	138.50	2648.57	52.29	(43.9,61.8)
CARD_ISQ	HTA	TR.LIPIDOS	747.00	14400.77	51.87	(48.2,55.7)
HTA	DIABETES	TR.LIPIDOS	2747.19	53360.41	51.48	(49.6,53.4)
ACV	HTA	TR.LIPIDOS	375.72	7301.36	51.46	(46.4,56.9)
VALVULOPA	DIABETES	TR.LIPIDOS	90.02	1757.81	51.21	(41.2,62.9)
ACV	DIABETES	HTA	132.55	2595.37	51.07	(42.7,60.5)
OBESIDAD	DIABETES	TR.LIPIDOS	975.75	19192.16	50.84	(47.7,54.1)
OBESIDAD	HTA	TR.LIPIDOS	2398.57	48215.58	49.75	(47.8,51.8)
ALCOHOL	HTA	TR.LIPIDOS	131.70	2655.29	49.60	(41.5,58.8)
CARD_ISQ	CARD_ISQ	DIABETES	7.62	153.90	49.51	(20.8,99.1)
OBESIDAD	INSF_CARD	ARRITMIAS	159.64	3238.57	49.29	(41.9,57.6)
ACV	TABAQUIS	HTA	24.01	488.77	49.12	(31.5,73.1)
TABAQUIS	HTA	TR.LIPIDOS	398.50	8159.71	48.84	(44.2,53.9)
DIABETES	OBESIDAD	TR.LIPIDOS	987.78	20248.73	48.78	(45.8,51.9)
INSF_CARD	VALVULOPA	ARRITMIAS	54.82	1131.82	48.44	(36.5,63.1)

La Tabla E.8 muestra información detallada de los análisis del Apartado 4.2.4.1 para los pares de eventos, tanto consecutivos como no consecutivos. Los cálculos se ha realizado utilizando las variables que aparecen recogidas en la base de datos con la fecha truncada y asignando la misma probabilidad de ocurrencia a todas las secuencias posibles. Sólo se muestran los resultados para los pares con al menos 5 ocurrencias y en los que el evento inicial y final son de distinto tipo.

Tabla E.8: Información de los pares de eventos consecutivos y no consecutivos con fechas truncadas e imputación equiprobable (ocurrencias por cada 1000 personas-año). Se han excluido los pares en los que el evento inicial y final son del mismo tipo.

From	To	Casos	Tiempo	IR	IC
COLEST	CARD_ISQ	940.50	1219.50	771.22	(722.7,822.1)
INSF_CARD	GRIPE	1690.00	4584.25	368.65	(351.3,386.7)
INSU_RENA	HTA	1904.00	5501.25	346.10	(330.7,362.0)
COLEST	GRIPE	839.00	2624.50	319.68	(298.4,342.1)
INSU_RENA	GRIPE	2045.50	6426.42	318.30	(304.6,332.4)
ATERO_PER	TR.LIPIDOS	27.00	86.58	311.84	(205.5,453.7)
CARD_ISQ	GRIPE	4183.50	14463.92	289.24	(280.5,298.1)
COLEST	TR.LIPIDOS	727.50	2531.83	287.34	(266.8,309.0)
INSF_CARD	HTA	1375.00	4839.42	284.13	(269.3,299.5)
ATERO_PER	HTA	25.50	89.75	284.12	(184.7,417.8)
ACV	GRIPE	2397.00	8552.67	280.26	(269.2,291.7)
VALVULOPA	GRIPE	1736.50	6479.83	267.99	(255.5,280.9)
RETI_DIAB	GRIPE	363.50	1361.83	266.92	(240.2,295.8)
RETI_DIAB	HTA	316.50	1194.33	265.00	(236.6,295.9)
COLEST	HTA	680.00	2675.92	254.12	(235.4,274)
ACV	HTA	1970.50	7979.00	246.96	(236.2,258.1)
DIABETES	GRIPE	10381.00	42246.00	245.73	(241,250.5)
ATERO_PER	GRIPE	24.50	99.83	245.41	(158.0,363.7)
ARRITMIAS	GRIPE	7475.50	33720.00	221.69	(216.7,226.8)
DIABETES	HTA	8412.00	38811.17	216.74	(212.1,221.4)
HTA	GRIPE	29548.00	142307.75	207.63	(205.3,210.0)
CARD_ISQ	HTA	3030.50	15861.58	191.06	(184.3,198)
INSU_RENA	TR.LIPIDOS	1446.00	7849.08	184.23	(174.9,194.0)
RETI_DIAB	DIABETES	263.00	1451.92	181.14	(159.9,204.4)

La Tabla E.9 muestra información detallada de los análisis del Apartado 4.2.4.2 para los tripletes de eventos, tanto consecutivos como no consecutivos. Los cálculos se ha realizado utilizando las variables que aparecen recogidas en la base de datos con la fecha truncada y asignando la misma probabilidad de ocurrencia a todas las secuencias posibles. En este caso no se han excluido tripletes con repetición de eventos.

Tabla E.9: Información de los tripletes de eventos con fechas truncadas e imputación equiprobable (ocurrencias por cada 1000 personas-año).

E1	E2	E3	Casos	Tiempo	IR	IC
COLESTEROL	COLESTEROL	CARDIO_ISQ	461.50	280.08	1647.7	(1501,1805)
ALCOHOL	COLESTEROL	CARDIO_ISQ	26.67	16.83	1584.2	(1041,2310)
INSUF_RENAL	COLESTEROL	CARDIO_ISQ	77.25	51.87	1489.2	(1176,1861)
ANEMIA	COLESTEROL	CARDIO_ISQ	144.50	99.33	1454.7	(1227,1712)
HTA	COLESTEROL	HTA	972.42	668.83	1453.9	(1364,1548)
INSF_CARD	COLESTEROL	CARDIO_ISQ	82.08	56.58	1450.7	(1154,1801)
GRIPE	COLESTEROL	CARDIO_ISQ	960.42	666.12	1441.8	(1352,1536)
VALVULOPA	COLESTEROL	CARDIO_ISQ	68.83	47.79	1440.3	(1120,1823)
DIABETES	COLESTEROL	DIABETES	372.50	260.54	1429.7	(1288,1583)
TR.LIPIDOS	COLESTEROL	TR.LIPIDOS	976.42	683.92	1427.7	(1340,1520)
HTA	COLESTEROL	CARDIO_ISQ	967.58	681.58	1419.6	(1332,1512)
DIABETES	COLESTEROL	CARDIO_ISQ	372.17	263.29	1413.5	(1274,1565)
TR.LIPIDOS	COLESTEROL	CARDIO_ISQ	973.92	689.33	1412.8	(1326,1504)
CARDIO_ISQ	COLESTEROL	CARDIO_ISQ	1414.33	1002.58	1410.7	(1338,1486)
INSF_CARD	COLESTEROL	INSF_CARD	80.42	57.25	1404.7	(1115,1747)
VALVULOPA	COLESTEROL	VALVULOPA	68.17	48.54	1404.3	(1091,1780)
RETINO_DIAB	COLESTEROL	CARDIO_ISQ	11.25	8.12	1384.6	(698,2462)
RETINO_DIAB	COLESTEROL	RETINO_DIAB	11.25	8.12	1384.6	(698,2462)
OBESIDAD	COLESTEROL	CARDIO_ISQ	329.58	238.08	1384.32	(1239,1542)
ACV	COLESTEROL	CARDIO_ISQ	82.83	59.87	1383.44	(1102,1715)
TABAQUISMO	COLESTEROL	CARDIO_ISQ	86.00	62.29	1380.60	(1104,1705)
ALCOHOL	COLESTEROL	ALCOHOL	25.67	18.75	1368.89	(892,2011)
ARRITMIAS	COLESTEROL	CARDIO_ISQ	273.25	200.00	1366.25	(1209,1538)
OBESIDAD	COLESTEROL	OBESIDAD	327.17	241.00	1357.54	(1214,1513)

Apéndice F

RESUMEN DE LOS CÓDIGOS UTILIZADOS PARA REALIZAR LOS ANÁLISIS

A continuación incluimos un resumen de los códigos más importantes utilizados en esta tesis para transformar los datos y calcular las secuencias de modo eficiente. Este código es sólo una pequeña parte de las más de 3800 líneas que han sido necesarias para cargar, limpiar, transformar, calcular, guardar y representar gráficamente los datos. Especialmente laborioso ha sido el proceso de exploración y reparación de errores.

Para realizar la tesis hemos utilizado las librerías: data.table, openxlsx, lubridate, microbenchmark, lme4, lmerTest, ggplot2, glmmTMB, metafor, xtable, R.utils, stringr, arrangements, bench, visNetwork, htmlwidgets, DiagrammeR, DiagrammeRsvg, svglite, rsvg, RColorBrewer, bit64, lubridate, microbenchmark, cowplot, scales, fractional, doParallel y fst.

Código F.1: Transformación de formato Wide a Long de las columnas cuyo nombre contiene el año.

```
# Input:  (Una línea para cada ID)
#    ID, VarA1, VarA2, VarA3,... ,VarB_year1, VarB_year2, VarB_year3,...
# Output: (Una para cada ID y Año)
#   ID, YEAR, VarA1, VarA2, VarA3,... , VarB1, VarB2, VarB3
# Para minimizar el consumo de memoria leemos los datos en bloques de tamaño
↪  reducido y se van almacenando los resultados intermedios en el disco duro.

# Lectura de los datos comprimidos en formato CSV.
my <- fread("7z e -y -bso0 -so redux.7z",stringsAsFactors=F,na.strings=c("","NA"))
# Guarda datos en formato fst para acceso rápido a disco
write.fst(my,"todowide.fst",80)
lineas <- 224320 # Nº total de líneas
cacho <- 5608 # Tamaño de cada bloque
veces <- ceiling(lineas/cacho) # Nº veces que repito el bucle. (40)
nombres <- names(fread("7z e -y -bso0 -so unodoswide.7z", encoding="UTF-8",
↪  stringsAsFactors=F,na.strings=c("", "NA")))
# Elimina texto indeseado de los nombres.
nombres <- gsub("(_BIN)?_OK","",nombres)
# Sustituyo en los nombres el caracter "_" por "ñ" delante de la fecha.
newnames <- sub("_(20[0-9][0-9])$","ñ\\1" ,nombres)
# Columnas que vamos a transformar
measure_cols <- grep("ñ20[0-9][0-9]$",newnames, value = T )
# Barra de progreso
pb <- txtProgressBar(min = 0, max = veces, style = 3)

for(iter in 0:(veces-1)) {
   setTxtProgressBar(pb, iter)  # Actualiza barra progreso
   my<- read.fst("todowide.fst", from=1+(iter*cacho), to=(iter*cacho)+cacho,
   ↪  as.data.table = T)
   setnames(my, newnames ) # Sustituyo "_" por ñ ante fecha.
   my[, rn := seq_len(.N)] # Añado indice único para luego unir
   # Transforma a long las variables con fecha. Ignora el resto.
   largo <- melt(my, id.vars = "rn", measure.vars = measure_cols)
   # Divido los nombres_fecha en dos columnas: nombre y fecha
   largo[, c("variable", "YEAR") := tstrsplit(variable, "ñ")]
   # Convierto los nombres a factores en el mismo orden que original
   largo[, variable:=factor(variable, levels=unique(variable))]
   # borra antiguas coulmnas ya no necesarias.
```

```
    my[, (measure_cols) := NULL]
    # Reagrupa por años y lo une a las fijas.
    result <- my[dcast(largo, ... ~ variable), on = "rn"]
    # elimina índice
    result[, rn := NULL]
    # Voy grabando los resultados.
    write.fst(result, "output.fst", append = T)
    # borro variables y limpio memoria.
    rm(result); rm(largo); rm(my); gc()
}
close(pb)
```

Código F.2: Generación de pares consecutivos y no consecutivos de eventos.

```
# Input:  ID, evetype, evedate
# Output: from, to, casos, tiempo, IC, ICLow, ICHigh, RatioIC
generalejos <- function(todo, extra=NULL) {
    # Fecha fin de estudio o Fallecimiento de esa persona.
    todo[,deathORend:=ifelse(any(evetype=="Death"),.SD[evetype=="Death",evedate],
    ↪  "2010-12-31"), by=ID]
    todo[,time1toend:=ymd(deathORend)-ymd(evedate)]
    ##  Calculo las combinaciones buenas de mis datos.
    lejos <- todo[,c(.SD,.N), by=ID][N>1,-"N"][,{ ii=combinations(v=1:.N,k=2);
    ↪  .(evetype[ii[,1]], evetype[ii[,2]], evedate[ii[,1]], evedate[ii[,2]],
    ↪  deathORend[ii[,1]], deathORend[ii[,2]]) },by=ID]
    setnames(lejos,  c("ID", "from", "to", "datefrom", "dateto","deathORend",
    ↪  "deathORend2")) # Asigno nombres correctos
    lejos[,datediff:= ymd(dateto)-ymd(datefrom)]  # Tiempo entre 1er y 2º eventos
     # Tiempo entre 2º evento y fin estudio
    lejos[,time1toend:=ymd(deathORend)-ymd(datefrom)]
    lejos[,c("datefrom","dateto","deathORend", "deathORend2") := NULL] # Borro
    lejos[,time1toend:=NULL]  # Borro, variables ya no necesarias.
    # Intervalo entre evento inicial y final
    lejos[,datediff:=as.difftime(datediff,units="days")]
    lejos[,occur:=1] # Añado columna de 1 para indicar ocurrencia real
    # Combinaciones posibles que no salieron, para computar tiempo hasta fin.
    TT <- todo[,unique(evetype)]   # Tipos de  eventos en la base de datos
    # Calculo combinaciones posibles que no salieron pero hay exposición.
    ug <- todo[, unique(evetype), ID]  # Tipos de  eventos en cada ID
    # Indices para crear combinaciones
```

```
    idx <- CJ(ug[,seq_len(.N)], seq_along(TT))
    combi <- ug[idx$V1, c(.SD, .(TT=TT[idx$V2]))]  # Combinaciones
    setnames(combi, c("ID", "from", "to"))
    # Coge la 1ª aparición de cada evento
    temp <-todo[,time1toend[1], by=c("ID","evetype")]
    setnames(temp, "evetype", "from")
    # Asigno a cada evento from su datimetoend de la 1ª aparición
    combi <- merge(combi,temp ,  by=c("ID","from"), all.x=T,all.y=F)
    rm(temp, idx, ug)
    setnames(combi, "V1", "datediff")
    # Indica pares posibles que no han sucedido pero hay exposición
    combi[,occur:=0]
    junto <- rbind(lejos, combi)    # Unimos pares reales con posibles
    # Extrae la 1ª ocurrencia del par y se queda con la real.
    junto <- junto[,.SD[1],by=c("ID","from","to")]
    # Calculamos las sumas, divisiones y tiempos.
    oneyear <- as.duration(years(1))
    # Calculo suma de ocurrencias y suma de tiempos para cada par.
    resumen <- junto[,.(casos=sum(occur),tiempo=sum(datediff)/oneyear),
    ↪  by=.(from,to)]
    resumen[,propo:=1000*casos/tiempo]  # IR*1000
    # Ofrece datos detallados si se piden.
    if(!is.null(extra)) assign(extra,junto, envir=parent.frame())
    alf <- 0.05 # Intervalo de confianza.
    resumen[, PropLow := 1000*qchisq(alf/2, 2*casos)/2/tiempo ]
    resumen[, PropHigh := 1000*qchisq(1-alf/2, 2*(casos+1))/2/tiempo ]
    resumen[,RatioIC := paste0("(",round(PropLow,1), ",",round(PropHigh,1),")")]
    return(resumen)
}
```

Código F.3: Generación de pares consecutivos de eventos con fechas exactas.

```
generacerca <- function(todo, extra=NULL) {
    TT <- todo[,unique(evetype)] # Tipos de  eventos en la base de datos
    cerca <- todo[,.(ID,evetype, evedate)] # Selecciona columnas útiles.
    cerca[,nn:=1:.N]  # Asigna índice único a cada fila.
    # Convierte texto a fecha.
    cerca[, evedate:=parse_date_time2(evedate, "%Y-%m-%d")]
    # Desplaza una posición cada ocurrencia para crear eventos finales.
    cerca[,c("to","dateto","ID2"):=shift(.(evetype,evedate,ID),1,type="lead")]
```

```
    # Evita mezclar datos de diferentes ID.
    cerca[ID!=ID2, c("to",  "dateto", "ID2") := NA][,ID2:=NULL]
    cerca[,datediff := difftime(dateto, evedate, units="days")]
    cerca[is.na(to),datediff:=difftime(parse_date_time2("2010-12-31","%Y-%m-%d"),
    ↪  evedate, units="days")]  # Calcula diferencias de fechas.
    setnames(cerca, c("evetype", "evedate"), c("from","datefrom"))
    # generamos las combinaciones faltantes y quitamos las imposibles.
    combi <- rbindlist(lapply(TT, function(x) cerca[ (is.na(to)) | (x != to)
    ↪  ,.(nn,ID, from, x, to)] ))[,to:=NULL]
    setnames(combi, "x", "to")
    # Añade a las nuevas combinaciones información temporal
    combi <- merge(combi,cerca[,.(nn, datefrom, dateto, datediff)] ,  by="nn",
    ↪  all.x=T,all.y=F)
    combi[,occur:=0]  # Marca no ocurrencia de ese par.
    combi[,nn:=NULL]
    cerca[,occur:=1]  # Marca ocurrencias que sí sucedieron.
    cerca[,nn:=NULL]
    cerca <- cerca[!is.na(to),]  # Elimino pares sin evento final.
    setcolorder(cerca,c("ID", "from", "to")) # Reordeno columnas
    junto <- rbind(cerca, combi)
    setorder(junto, ID, from, to, datefrom) # Reordeno filas.
    # Cojo 1ª aparición, que será la real si existe.
    junto <- junto[,.SD[1],by=c("id","from","to")]
    # Suma acumulada para filtrar las primeras ocurrencias.
    junto[,V2:=cumsum(occur), by=c("ID","from","to")]
    # Constante para tener unidades temporales correctas.
    oneyear <- as.duration(years(1))
    resumen <- junto[V2<=1,.(casos=sum(occur),tiempo =
    ↪  sum(datediff)/oneyear),by=.(from, to)]
    resumen[,propo:=1000*casos/tiempo]
    if(!is.null(extra)) assign(extra,junto, envir=parent.frame())
    alf <- 0.05   # Intervalo de confianza.
    resumen[, PropLow := 1000*qchisq(alf/2, 2*casos)/2/tiempo ]
    resumen[, PropHigh := 1000*qchisq(1-alf/2, 2*(casos+1))/2/tiempo ]
    resumen[,RatioIC := paste0("(",round(PropLow,1), ",",round(PropHigh,1),")")]
    return(resumen)
}
```

Código F.4: Generación de pares consecutivos y no consecutivos con imputación aleatoria.

```
gN <- function(todo, longi=10L) {
   todo2 <- todo[,c("ID","evetype","ano")]
   setnames(todo2,"ano", "evedate")   # Uso fecha truncada
   TT <- todo2[,unique(evetype)] # Tipos de  eventos en la base de datos
   setorder(todo2,ID,evedate)   # Reordeno filas por ID y tipos de evento.
   # Defino las funciones que usaré dentro de cada ID
   # Genero permutaciones intraanuales y las combino.
   CJ2 <- function(x,y) {cbind(x[rep(1:nrow(x), times=nrow(y)),],y[rep(1:nrow(y),
   ↪  each=nrow(x)),])}
   rere <- function(z) { Reduce(CJ2 , tapply(seq_along(z$evedate),z$evedate,
   ↪  function(x) permutations(v=x), simplify = F)) }
   # Crea subgrupos utilizando la primera parte no fija del ID.
   todo2[,jjj := str_sub(ID,7L,longi)]

   divide <- function(kk) {    # Función principal
      multi <- kk[,rbindlist(apply(rere(.SD), 1, function(x)
      ↪  .(evetype[x],evedate[x])), idcol = T) ,by=ID]
      setnames(multi, c("ID", "rr", "evetype", "evedate"))
      multi[,evedate:=evedate+(1:.N)/(.N+1),by=.(ID, rr,evedate)] # Fecha a decimal
      # creamos las combinaciones y calculamos nº de ocurrencias y tiempo
      lejos <- multi[, if(.N>1) { ii=combinations(v=1:.N,k=2); .(evetype[ii[,1]],
         evetype[ii[,2]], evedate[ii[,1]], evedate[ii[,2]]) },by=.(ID,rr)]
      setnames(lejos,  c("ID","rr", "from", "to", "datefrom", "dateto"))
      # Intervalo entre evento inicial y final
      lejos[,datediff:= as.numeric(dateto)-as.numeric(datefrom)]
      lejos[,occur:=1] # Añado columna de 1 para indicar ocurrencia real
      # Calculo combinaciones posibles que no salieron pero hay exposición.
      ug <- multi[, unique(evetype), by=.(ID,rr)] # Tipos de  eventos en cada ID
      idx <- CJ(ug[,seq_len(.N)], seq_along(TT))  # Indices para crear
      ↪  combinaciones
      combi <- ug[idx$V1, c(.SD, .(TT=TT[idx$V2]))] # Combinaciones
      setnames(combi, c("ID","rr", "from", "to"))
      # Coge la 1ª aparición de cada evento
      temp <-multi[,2011-evedate[1], by=.(ID,rr, evetype)]
      setnames(temp, "evetype", "from")
      # Añade a las nuevas combinaciones información temporal
      combi <- merge(combi, temp, by=c("ID","rr","from"), all.x=T,all.y=F)
      rm(temp, idx, ug)
      setnames(combi, "V1", "datediff")
```

```
      combi[,occur:=0]    # Marca pares que no ocurrieron
      lejos[,c("datefrom","dateto"):=NULL]
      junto <- rbind(lejos, combi) # Unimos
      # Cojo 1ª aparición, que será la real si existe.
      junto <- junto[,.SD[1],by=.(ID,rr,from,to)]
      # Ponderamos haciendo la media.
      return(junto[,.(datediff=mean(datediff), occur=mean(occur)),by=.(ID, from,
      ↪  to)])
   }
   # Aplicamos función "divide" a los datos por bloques.
   junto <-  rbindlist(lapply(split(todo2,by="jjj" ), function(x) divide(x)))
   return(junto)
}
```

Código F.5: Generación de pares consecutivos y no consecutivos con imputación utilizando probabilidad a priori interanual.

```
# Reutilizar el Código F.4 pero sustituyendo la linea

return(junto[,.(datediff=mean(datediff), occur=mean(occur)), by=.(ID,from,to)])

# por el siguiente código

# Relaciona cada par con el IR interanual de la tabla.
junto <- merge(junto, tabla, by=c("from", "to"), all.x=F,all.y=F)
junto[is.na(IRtab), IRtab := 0,]   # Sustituimos missings por 0.
junto[,IR:=occur/datediff]   # IR del par
junto[,peso := 1/sqrt(sum( (IR-IRtab)^2  ,na.rm=T)), by=.(rr,ID)]
# Calculamos la media ponderada utilizando esos pesos para cada secuencia posible.
return(junto[,.(datediff=weighted.mean(datediff,peso), occur=weighted.mean(occur,
↪  peso)),by=.(ID, from, to)])

# Donde tabla ha sido calculada previamente y contiene las IR calculadas utilizando
↪  sólo on los datos interanuales.

tabla <- gN(todo)
tabla <- tabla[datediff>0,] # eliminamos repes intranuales.
tabla <- tabla[,.(casostab=sum(occur),tiempotab = sum(datediff)),by=.( from, to)]
tabla[,IRtab:=casostab/tiempotab]
```

Código F.6: Generación de tripletes de eventos con fechas exactas.

```
generatrip <- function(datos){
   TT <- datos[,unique(evetype)]
   todo2 <- datos[,if(.N>1) .SD, by=ID]  # Personas con al menos 2 eventos
   todo2[,deathORend:="2010-12-31"] # Fecha fin o añadir fallecimiento
   todo2[,time1toend:=ymd(deathORend)-ymd(evedate)]
   ##  Calculo las combinaciones buenas de mis datos, LEJOS.
   lejos <- todo2[,if(.N>2) { ii=combinations(v=1:.N,k=3); .(evetype[ii[,1]],
   ↪  evetype[ii[,2]], evetype[ii[,3]],
   ↪  evedate[ii[,1]],evedate[ii[,2]],evedate[ii[,3]]) },by=ID]
   setnames(lejos,  c("ID", "EV1", "EV2","EV3", "EVD1", "EVD2","EVD3"))
   lejos[,datediff:= ymd(EVD3)-ymd(EVD2)]  # Desde  el 2º evento 3º
   lejos[,c("EVD1","EVD2","EVD3") :=NULL]
   lejos[,datediff:=as.difftime(datediff,units="days")]
   lejos[,occur:=1] # para contar los que sí sucedieron
   # Calculo las combinaciones posibles que no salieron pero sí hubo exposición
   # Parto de los pares que sí existieron.
   paresx <- todo2[,{ ii=combinations(v=1:.N,k=2); .(evetype[ii[,1]],
   ↪  evetype[ii[,2]],evedate[ii[,2]], deathORend[ii[,2]]) },by=ID]
   paresx[,datediff:=as.difftime((ymd(V4)-ymd(V3)),units="days")]
   pares <- paresx[,.(datediff=datediff[1]), by=c("ID","V1","V2")] # 1ª aparición
   # Y con esos pares genero posibles tripletes.
   combix <- pares[,TT, by=c("ID","V1","V2", "datediff")] # crea combis
   setnames(combix,c("V1", "V2","TT"),c("EV1","EV2","EV3"))
   combi <- combix  # para más restricciones
   setcolorder(combi, c("ID", "EV1", "EV2", "EV3"))
   combi[,occur:=0]   # para no contabizarlo en el sumatorio.
   juntox <- rbind(lejos, combi) # Unimos tripletes reales y virtuales.
   # Si ha sucedido ese triplete cojo el bueno, si no el virtual.
   junto <- juntox[,.SD[1],by=c("ID","EV1","EV2","EV3")]
   oneyear <- as.duration(years(1))  # para normalizar intervalos
   resumen <- junto[,.(casos=sum(occur), tiempo=sum(datediff)/oneyear),
   ↪  by=.(EV1,EV2,EV3)]
   resumen <- resumen[casos>0,]
   resumen[,propo:=1000*casos/tiempo]    # IR*1000
}
```

Código F.7: Dibujo de la red de conexiones con los IR de los tripletes de eventos con colores.

```
# Asignación de colores no solapados a cada transición.
# Desglosa el resultado en pares e indica la dirección de cada uno.
# Input: EV1, EV2, EV3, IR, colo.
colores <- function(inp, colo) {
   tri <- inp[,.(EV1,EV2,EV3,propo,idd=1:.N, colo=NA_character_)]
   for(ii in tri$idd) {    # Recorre cada transición
      bus <-  tri[idd==ii,.(EV1,EV2,EV3)]
      prohib <- unique(tri[EV1 %in% bus | EV2 %in% bus | EV3 %in% bus, colo])
      nuevo <- setdiff(colo, prohib)[1] # Elimina colores ya usados.
      tri[idd==ii,colo:=nuevo]
   }
   # Une primeros tramos y segundos tramos, que se dibujan diferente.
   dibu <- rbind(tri[,.(from=EV1,to=EV2,propo,idd,colo,parte="ini")],tri[,.(from=
       EV2, to=EV3, propo, idd, colo,parte="fin")], use.names=F)[order(idd),]
   return(dibu)
}
# Dibujo del grafo a partir de los tripletes con los colores incluidos.
# Utiliza el resultado de colores, la lista de nodos, dirección, tipo de dibujo...
Dia3 <- function(salt, nodesd, sentido, lay, dirg, nombre=NA, labe=nodesd) {
   if( !("colo" %in% names(salt) )) {salt[,colo := "blue"]}
   # Crea los nodos con sus etiquetas
   nodes <-   create_node_df(  n=length(nodesd), label=labe, width=0.45,
   ↪  shape="circle", penwidth=1, style="filled", fillcolor="azure",
   ↪  fontcolor="black", fontsize=7)
   # Crea las linas de las transiciones. Sólo etiqueta las finales.
   edges <- create_edge_df(from = factor(salt$from, levels=nodesd), to =
   ↪  factor(salt$to, levels=nodesd),
   rel = "leading_to", color=salt$colo, label=salt[,ifelse( salt$parte==sentido,"",
   ↪  round(propo,0))], penwidth=salt[,1+sqrt(propo)/10])
   # Añade los atributos necesarios para el gráfico.
   # Opciones: neato, dot, circo, twopi, curved, spline, ortho, TB, LR
   graph <- create_graph(  nodes_df = nodes, edges_df = edges)
   graph <- add_global_graph_attrs(graph, "layout", lay, "graph")
   graph <- add_global_graph_attrs(graph, "splines","spline", "graph")
   graph <- add_global_graph_attrs(graph, "rankdir", dirg, "graph")
   if (!is.na(nombre)) export_graph(graph, nombre) # Para exportar archivo.
   render_graph(graph)
}
```

Código F.8: Generación de pares de eventos a partir de fechas truncadas. Versión recursiva que utiliza poca memoria y asume secuencias equiprobables.

```
# Promedio ocurrencias y tiempos exposición desde A hasta 1er B tras A
info <-  function(x,eveini, evefin) {     # x contiene ID, evetype, evedate
   # Tiempo 1ª ocurrencia de A
   Tini <- x[evetype==eveini,ano][1]
   if(!is.na(Tini)) { # Existe A
      if(eveini!=evefin){ # A y B diferente tipo.
         # Tiempo 1ª ocurrencia B tras A. Puede coincidir año.
         Tfin <- x[ano>=Tini & evetype==evefin,ano][1]
      } else { # A y B mismo tipo. No pueden coincidir año.
          Tfin <- x[ano>Tini & evetype==evefin,ano][1]
       }
      if(!is.na(Tfin)) { # Existe B
         if(Tini==Tfin) { # A y B mismo año.
            # Resultado combinación:
            # Cuando dentro de mismo año A es anterior a B.
            tem <- x[ano>Tfin & evetype==evefin,ano][1]
            # Cuando A es posterior. Tiempo del siguente B o fin estudio.
            Ttres <- ifelse(is.na(tem),2011,tem+0.5)
            c(0.5, (0.16666666666+(Ttres-(Tini+0.66666666666))/2)) # caso2
         } else { c(1, (Tfin-Tini))} # A y B distinto año.  # caso3
      } else { # No existe B
         c(0, (2011-(0.5+Tini)))}  # caso4
   } else {c(0,0)} # No existe A  # caso1
}

# Aplicamos la función info a todos los pares de evento...
todosfin <- todo[,unique(evetype)] # Eventos finales para el bucle
todo[, { todoini=unique(evetype); transpose(sapply(todoini, function(ini)
↪  (lapply(todosfin, function(fin) c(ini, fin, info(.SD,ini,fin))))))}, by=ID]
```

Código F.9: Generación de tripletes de eventos a partir de fechas truncadas. Versión recursiva que utiliza poca memoria y asume secuencias equiprobables.

```
# Localiza el 3er evento, C y da su tiempo. Venimos de AB.
Hc <- function(x,evec,tb) { # Datos, evento buscado, tiempo inicial.
   # Tiempo de 1ª ocurrencia de C tras B
   tc <- x[ano>tb & evetype==evec, ano][1]
   if(is.na(tc)) {
```

```
      c(0,2011-tb)   # No existe C
   } else {         # Existe C
      c(1,tc+0.5-tb)   # 1 ocurrencia. Diferencia de tiempos.
   }
}

# Localiza 2º y 3er eventos, BC, da tiempo entre ellos, venimos de A.
Gbc <- function(x, eveb, evec, ta) {
   tb <- x[ano>ta & evetype==eveb,ano][1] # Guarda tiempo del 1er B tras A
   if(!is.na(tb)) {     # Si existe B
      if(eveb != evec){ B y C son de distinto tipo
         tc <- x[ano>=tb & evetype==evec, ano][1] # Tiempo de C
      } else { B y C son del mismo tipo
         tc <- x[ano>tb & evetype==evec, ano][1]  # Tiempo de C
      }  # no permite 2 eventos iguales el mismo año
      if(!is.na(tc)) { # Existe C
         if(tb==tc) { # B y C ocurren el mismo año
         # Promedio 2 opciones: B antes y C antes (calcula siguiente C tras B)
            c(0.5,0.166666)+0.5*Hc(x,evec,tb+0.666666)
         } else { # B y C ocurren en diferentes años
            c(1, (tc-tb))}   # Una ocurrencia, intervalo entre B y C.
      } else {  # No existe c
         c(0, (2011-(0.5+tb)))  # 0 Ocurrencias. Tiempo hasta fin
      }
   } else {   # No existe b.
      c(0,0)
   }
}


# Función principal. Localiza ABC y da intervalo de B a C.
# No acepta dos eventos iguales el mismo año.
Fabc <-  function(x, evea, eveb, evec) {
   ta <- x[evetype==evea,ano][1]  # Tiempo 1ª ocurrencia de A
   if(!is.na(ta)) {   # Existe A
      if(evea != eveb){
         tb <- x[ano>=ta & evetype==eveb, ano][1]
      } else {
       tb <- x[ano>ta & evetype==eveb, ano][1]
    }
      if(!is.na(tb)) { # No existe B
```

```
        if(eveb != evec) {   # B y C distinto tipo
           tc <- x[ano>=tb & evetype==evec, ano][1]
        } else {   # B y C igual tipo
           tc <- x[ano>tb & evetype==evec, ano][1]
        }
        if(!is.na(tc)) {   # No existe C
           # Aquí ponemos cuando existen a, b y c
           # Diferentes casos según coincidan los años o no.
           if(ta!=tb & tb!=tc) {   # Todos diferente años.
              c(1,tc-tb)
           } else if(ta==tb & tb!=tc) { # A mismo año que B
     0.5*c(1,tc+0.5-(tb+0.66666666)) + 0.5*Gbc(x,eveb,evec,ta+0.3333333)
           } else if(ta!=tb & tb==tc) { # B mismo año que C
     0.5*c(1,0.33333333)+0.5*Hc(x,evec,tb+0.666666666)
           } else {   # A, B y C mismo año
     0.16666666*(c(1,0.25) + Hc(x,evec,tb+0.75) + Gbc(x,eveb,evec,ta+0.5) +
  ↪ Gbc(x,eveb,evec,ta+0.75) + Hc(x,evec,tb+0.75) + Gbc(x,eveb,evec,ta+0.75))
           }
        } else {
           if(ta==tb) {
              c(0, 0.5*(2011-(tb+0.6666666))) + 0.5*Gbc(x, eveb, evec,
              ↪  ta+0.6666666) # No existe C.  A y B mismo año
           } else {
              c(0,2011-(tb+0.5))   # No existe C.  A y B diferente año.
           }
        }
     } else {   # No existe B
        c(0,0)
     }
  } else {   # No existe A
     {c(0,0)}
  }
}


todotres <- todo[,unique(evetype)]  # Extrae lista de eventos finales

# Calcula las ocurrencias y tiempos de todos los tripletes.
todo[,{ todouno=unique(evetype); transpose(sapply(todouno, function(uno)
↪  sapply(todouno, function(dos) lapply(todotres, function(tres) c(uno, dos, tres,
↪  Fabc(.SD,uno,dos,tres))))))},by=ID]
```

```
#  Versión paralelizada con la librería ``parallel''.
cl <- makeCluster(8)  # Inicializo clusters y comparto datos y funciones.
clusterExport(cl=cl, varlist=c("todo", "Fabc", "Gbc", "Hc", "todotres"),
↪  envir=environment())

todo[,{ todouno=unique(evetype); transpose(parSapply(cl, todouno, function(uno)
↪  sapply(todouno,    function(dos) lapply(todotres, function(tres) c(uno, dos,
↪  tres, Fabc(.SD,uno,dos,tres))))))},by=ID]
stopCluster(cl)
```

Apéndice G

PUBLICACIONES Y CONGRESOS

PUBLICACIONES

- Silvia Montoro-García, María Pilar Zafrilla-Rentero, Francisco Miguel Celdrán-de Haro, Juan José Piñero-de Armas, Fidel Toldrá, Luis Tejada-Portero y José Abellán-Alemán. «Effects of dry-cured ham rich in bioactive peptides on cardiovascular health: A randomized controlled trial». *Journal of Functional Foods* 38 (2017), págs. 160-167
- Andrés Parrilla-Almansa, Carlos Alberto González-Bermúdez, Silvia Sánchez-Sánchez, Luis Meseguer-Olmo, Carlos Manuel Martínez-Cáceres, Francisco Martínez-Martínez, José Luis Calvo-Guirado, Juan José Piñero de Armas, Juan Manuel Aragoneses, Nuria García-Carrillo y col. «Intraosteal Behavior of Porous Scaffolds: The mCT Raw-Data Analysis as a Tool for Better Understanding». *Symmetry* 11.4 (2019), pág. 532

CONGRESOS

- Juan Jose Piñero de Armas. «Analysis of Sequences of Cardiovascular Events». Congreso Internacional "XVII Conferencia Española y VII Encuentro Iberoamericano de Biometría CEB-EIB 2019". Valencia, jun. de 2019