Tesis doctoral

Programa de Doctorado en Investigación en Salud

Universidad Internacional de Catalunya

Estrategias de emparejamiento de muestras (matching) para eliminar la confusión en los estudios observacionales: Aplicación en fármaco-epidemiología con grandes bases de datos de registros clínicos

Doctorando: Jordi Real Gatius

Director: Dr. Jose M Martínez-Sánchez

Àrea de Bioestadística
Departament de Ciències Bàsiques
Universitat Internacional de Catalunya

Unitat de Control del Tabaquisme
Institut Català d'Oncologia

Línea de investigación: Epidemiología; Salud Pública; Metodología

Tenemos que comprender el cosmos tal como es
y no confundirlo con lo que queremos que sea

Carl Sagan

Índice

Agradecimientos

Aquesta tesis no s'hauria realitzat sense l'ajuda dels membres del reconegut grup QuALiStat (Carles Forné, Albert Roso). Gràcies, Litus y Albert, per unir-vos a aquesta idea boja, i per la gran currada que vau fer de revisar tants i tants articles, i també per les valuoses aportacions en un dels articles centrals d'aquesta tesis. Les intenses sessions gastro-metodològiques d'estadística van ser molt útils i fructíferes a tots els nivells (Altrament dit reunions QualiStateres), encara que alguna vegada vaig acabar una mica perjudicat. Gracies amics.

Inevitablement després d'aquestes reunions QualiStateres, qui va haver de suportar o patir algunes de les conclusions qualistatils i efectes secundaris de la nit anterior, va ser la meva estimada Flor. Gracies, Laura, i perdona pels rollassos varis d'estadística que has hagut d'aguantar. Espero que alguns d'ells t'hagin servit pel teu present professional, i no m'ho tinguis en compte en un futur. Si es així, ho sento, t'estimaré igualment.

Agrair explícitament al meu director de tesis i amic, Dr. Jose María Martínez, sobretot pel seu grau interès i implicació en tot moment. Des del principi va creure en mi i en la idea boja d'aquesta tesis i durant tot el procés ha sabut donar-li forma aportant brillants idees i també, dures però necessàries correccions. Jose, gracies per compartir la teva experiència, encomanar-me el teu optimisme, i donar-me la energia i empenta que just necessitava. Sense tu no hauria estat possible.

Als companys de l'institut de Salut Pública de la Universitat Internacional de Catalunya, Maria Dolors, Alicia, Pau, Homar, Lluïsa , i en especial al Lluis Gonzalez de Paz, investigador d'un dels estudis inclosos en aquesta tesis, qui sorprenentment va deixar a les meves mans el disseny de la seva interesant pregunta d'investigació. No abunden investigadors que confiïn des d'un principi en el criteri estadístic.

A tots els companys i investigadors de l'IDIAP Jordi Gol i de les USRs de Lleida i de Barcelona que m'han donat la oportunitat d'ajudar a respondre interesants preguntes clíniques mitjançant eines estadístiques. Concretament a la Inés Cruz la meva "ex-jefa" per la confiança que m'ha regalat, sense res a canvi durant els últims 8 anys a la USR-Lleida, i especialment també a la Gisela Galindo, amiga, brillant i crítica investigadora que sempre ha confiat en el meu criteri metodològic, més enllà de l'estadística, i també em compartit complicitat existencial i sentit de l'humor. I junts, vam lluitar contra els múltiples biaixos a que ens enfrontàvem per saber el que amagaven les dades i esbrinar la veritat.

Gracies a tots.

1. Resumen

Los estudios observacionales tienen un papel importante en la investigación médica. Sin embargo, una de las limitaciones comunes de los estudios observacionales analíticos es la que afecta a la validez interna, principalmente debido al potencial sesgo de confusión causada por la asignación no controlada de los individuos a los grupos de exposición. Las técnicas más habituales en la investigación médica en general, y la epidemiología en particular, para corregir el sesgo de confusión son los modelos de regresión multivariables (MRMs) (tales como Regresión logística, lineal, Cox etc...). Estas técnicas de ajuste requieren de una adecuada especificación del modelo para que la las medidas de asociación de interés (efecto, riesgo relativo, Odds ratio, razón de prevalencias, etc.) estén correctamente estimadas. En la actualidad, existe un creciente interés en otras alternativas a los modelos multivariables como son enfoques no paramétricos utilizando algoritmos de emparejamiento (matching). Estas técnicas prometen inferencias más robustas al no depender de la correcta especificación del modelo.

Los objetivos generales de la presente tesis doctoral son: 1) Evaluar el reporte, en cuanto a presentación de medidas de bondad de ajuste o diagnóstico de los MRMs utilizados en estudios observacionales analíticos publicados e indexados en PubMed; 2) Comparar la robustez de los métodos matching enfrente los MRMs como métodos de ajuste mediante un estudio de simulación; y 3) Aplicar la metodología matching en estudios observacionales analíticos con una hipótesis clínica o de salud pública.

Los resultados de la presente tesis doctoral han mostrado que existe un extensivo y creciente uso de los MRMs en el ámbito de la investigación biomédica con diseño observacional. A la vez, también se ha observado un bajo reporte en la verificación de las hipótesis de tres técnicas de regresión muy comunes (Regresión logística, lineal y Cox) en artículos publicados e indexados en PubMed. En este sentido, tan solo uno de cada 4 artículos revisados mostró o declaró realizar un análisis de validación de las hipótesis de los modelos o aportó estadísticos de bondad de ajuste. Por otro lado, mediante un estudio de simulación se ha mostrado, como una técnica de regresión, ampliamente utilizada, como es la regresión logística multivariable, puede generar estimaciones sesgadas, y consecuentemente un elevado error de tipo I si la comprobación de las asunciones del modelo es ignorada. En este sentido, los algoritmos matching, presentaron una mayor robustez en comparación con los MRMs.

Por último, se presenta la utilidad práctica con la aplicación de los métodos matching en investigación clínica y de salud pública en estudios basados en registros clínicos.

En conclusión, el reporte de la comprobación de las asunciones formales de los MRMs es bajo en los artículos científicos publicados e indexados en PubMed. Dada la importancia de estas, y la sensibilidad de las estimaciones de los MRMs paramétricos, sería deseable una mayor transparencia en la declaración de las asunciones de los MRMs, especialmente en estudios observacionales analíticos. Por otro lado, los métodos matching controlan mucho mejor la reducción del sesgo de confusión por lo que se deberían considerar más a menudo en la investigación clínica y de salud pública como alternativa a los MRMs, especialmente en estudios basados en registros clínicos o grandes muestras disponibles.

2. Abreviaturas

Acrónimo	Descripción
CI	Cuidador informal
CV	Cardiovascular
ECM	Error cuadrático medio
FRCV	Factor de riesgo cardiovascular
GAM	Modelos aditivos generalizados
GLM	Modelos lineales generalizados
H&L	Prueba de Hosmer & Lemeshow
HR	Hazard ratio
IC95%	Intervalo de confianza al 95%
IPTW-PS	Inverse probablity of treatment weighting using PS
K-S	Prueba de Kolmogorov-Smirnov
MRM	Modelo de Regresión Multivariable
N-N	Nearest Neighbour
OR	Odds ratio
PS	Propensity score
PS-M	Propensity Score Matching
RCT	Ensayo clínico aleatorio
ST-PS	Estratificación por propensity score

3. Introducción

Muchas de las preguntas en la investigación médica se contrastan mediante estudios observacionales. Los estudios observacionales tienen un papel importante en la investigación sobre los beneficios y los daños de las intervenciones médicas o fármacos, ya que los ensayos clínicos no pueden responder todas las cuestiones importantes sobre intervenciones y tratamientos. También es más probable encontrar una indicación de un tratamiento en la práctica clínica habitual(1), ya que a diferencia de los ensayos clínicos, la evaluación de la potencial eficacia y seguridad de los tratamientos se realiza en las condiciones de la práctica habitual, no excluyéndose aquellos sujetos o condiciones que habitualmente no se incluyen en los ensayos clínicos (edades extremas, pacientes con elevada comorbilidad, polimedicación, etc...). Gran parte de la investigación epidemiológica sobre las causas de las enfermedades depende de estudios de cohortes, de casos y controles, o estudios transversales. Además los estudios observacionales también son adecuados para detectar efectos infrecuentes o tardíos de tratamientos(1) .

La validez interna de todo trabajo observacional, como señala la guía STROBE (StrengtheningTheReporting of OBservationalstudies in Epidemiology)(1), puede verse afectada por dos aspectos muy importantes a menudo presentes en toda investigación observacional, como es el sesgo de información y confusión(2). El sesgo de información resulta de una determinación incorrecta de la exposición, del resultado, o de ambos. La confusión es una “mezcla” o “difuminación” de efectos: un investigador trata de relacionar una exposición (X) a un resultado Y, pero en realidad mide el efecto de un tercer factor Z (la variable de confusión), en ocasiones no observada (Figura 1).

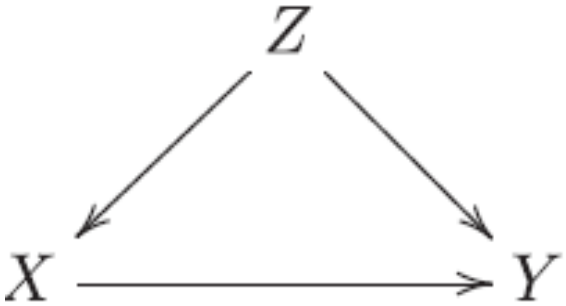


Figura 1. Diagrama de relación causal entre variables X-Y, en presencia de una variable de confusión Z

El sesgo de confusión puede ser corregido, siempre y cuando el factor de confusión sea anticipado. La guía STROBE incluye el control de la confusión como uno de los aspectos cruciales de este tipo de diseño(3). Además, la confusión puede ser controlada de diversas formas: restricción, matching, estratificación, estandarización y técnicas multivariables. El propósito de todos estos enfoques es conseguir homogeneidad entre los grupos de estudio y por lo tanto minimizar la posibilidad del sesgo de confusión(2).

3.1 Métodos de regresión multivariable

Los modelos de regresión multivariable (MRMs) se utilizan ampliamente en la investigación de ciencias de la salud(4-6). Este elevado uso puede deberse a los avances importantes en la potencia de cálculo, al aumento en la disponibilidad de datos y sobre todo a que el software estadístico estándar facilita el acceso a la realización de estas técnicas (7). Con frecuencia, el objetivo en la recolección de datos obedece al afán de explicar las interrelaciones que existen entre ciertas variables o a determinarlos factores que afectan a la presencia o ausencia de un episodio adverso determinado. Es ahí donde los modelos de regresión multivariables también pasan a ser un instrumento útil, al suministrar una explicación matemática simplificada de dicha relación. El objetivo final será obtener un modelo simplificado que tenga sentido desde una perspectiva clínica, cercano a los datos disponibles y que aporte asociaciones válidas al aplicarlo a datos independientes(8).

3.2 Tipos de modelos de regresión multivariable

En esencia un modelo de regresión es una ecuación matemática que describe la relación de una o más variables independientes sobre una variable llamada dependiente. La elección del tipo de modelo (Lineal, Logístico, o de Cox) depende de la naturaleza de la variable respuesta (continua, dicotómica, o tiempo hasta un evento). La elaboración de un modelo de regresión permite determinar los coeficientes, y estimar las medidas de asociación ajustadas de cada variable incluida.

En la tabla 1 se describen ejemplos de tres formas funcionales de los tipos de modelos de regresión multivariable más conocidos en investigación médica y la forma de interpretación de sus coeficientes.

Estos modelos matemáticos de regresión asumen: 1) linealidad: efectos (β_i) proporcionales por unidad de x_i y 2) inexistencia de interacción: Efecto constante β_i en todo el rango de valores de x_j. La violación de estas asunciones invalidaría la certeza

del modelo. En el Anexo I se puede ver un ejemplo con datos reales de un ajuste de un modelo lineal con su interpretación.

Tabla 1. Tipos de modelos de regresión multivariable con dos variables explicativas con un efecto aditivo, según la variable respuesta (Outcome)

Outcome	Modelo	Forma estructural	Coeficiente (β_1)	Interpretación
Continua	Lineal	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	Pendiente: β_1	Incremento de y por unidad de x_1
Dicotómica	Logístico	$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	Odds Ratio: $\text{Exp}(\beta_1)$	Incremento del Odds por unidad de x_1
Tiempo hasta evento	Cox	$h(t) = h_0(t) e^{\beta 1 * x1 + \beta 2 * x2}$	Hazard ratio: $\text{Exp}(\beta_1)$	Incremento del riesgo por unidad de x_1

t: tiempo;y=variable dependiente; x_i=variables independientes; p: probabilidad del evento, β_i: Coeficientes del modelo; h(t)=función de riesgo.

Esencialmente los tres modelos presentados en la Tabla 1 tienen las mismas asunciones en la parte derecha de la ecuación, a menos que se incluían términos no lineales además de las interacciones. Una de las ventajas de los MRMs es que permiten controlar más factores de confusión que la estratificación y simultáneamente evaluar la relación entre diversos factores de exposición y variables respuesta de distinta naturaleza (continuos, dicotómicos, contajes o eventos tiempo-dependientes, entre otras)(9,10). La estimación del efecto de cada variable refleja su asociación individual con el resultado, teniendo en cuenta la contribución del resto de variables introducidas en el modelo. Pero la modificación del efecto no es identificable por la simple inclusión de la variable en el modelo de regresión, sino que requiere la inclusión de términos de interacción entre la exposición y la variable modificadora de efecto o la variable confusora en el modelo.

3.3 Asunciones de los modelos de regresión multivariable

Los MRMs asumen distribuciones de probabilidades que conllevan ciertas asunciones de base (ej.: linealidad, asunciones de normalidad, homocedasticidad e independencia de los errores, riesgos proporcionales en los modelos de COX). Por lo tanto, antes de la selección del modelo final se han de realizar ciertas comprobaciones formales. Un modelo se dice que presenta un buen ajuste a los datos si los valores predichos por el

modelo se dice que presenta un buen ajuste a los datos si los valores predichos por el reflejan de forma adecuada a los valores observados. Si el modelo presenta un mal ajuste, este no puede ser utilizado para extraer conclusiones ni efectuar predicciones. En la tabla 2 se puede ver un esquema resumen de algunos aspectos a tener en consideración para la construcción de un buen modelo de regresión, el riesgo potencial en caso de incumplimiento, un método habitual de evaluación y una posible solución.

Tabla 2. Aspectos a tener en cuenta para la construcción de un MRM

Descripción	Riesgo potencial*	Posible evaluación	Posible solución
Linealidad	S, α, β	Gráficos de residuos	Añadir términos cuadráticos
Interacciones	S, α, β	Gráficos de residuos	Añadir interacción
Homocedasticidad	V, α, β	Gráficos de residuos	Añadir predictor
Multicolinealidad	V, β	Cambios en los coeficientes	Eliminar predictor
Observaciones anómalas	S, V, α, β	Gráficos de residuos	Eliminar observaciones
Independencia de los errores	V, α, β	Gráficos de residuos	Remodelar
Hipótesis del modelo	S, V, α, β	Bondad de ajuste **	Remodelar
Lineal: Normalidad	S, V, α, β	Test K-S	Remodelar
Cox: Riesgos proporcionales	S, V, α, β	Residuos de Shoenfield	Remodelar
Logístico: Bondad de ajuste	S, V, α, β	Test H&S	Remodelar

* Riesgo potencial en caso de incumplimiento de la hipótesis: S=Sesgo en la estimación del parámetro; V=Sesgo en la estimación de la varianza; α = Error de tipo I; β =Error de tipo II.

** Modelo lineal: Test de normalidad de Kolmogorov-Smirnov sobre los residuos (K-S), Q-Q plots, R^2; Gráficos de residuos; Modelo logístico: Test de Hosmer & Lemeshow (H&L), Estadístico C, área debajo la curva ROC, calibración-discriminación del modelo. Modelo de Cox: Test de los residuos de Shoenfield, gráficos de residuos.

Una forma de medir la adecuación de un modelo es proporcionando medidas globales de bondad de ajuste a través de test estadísticos construidos para tal fin. En este sentido cada tipo de modelo tiene sus herramientas estadísticas analíticas y gráficas específicas para evaluar el cumplimiento de las condiciones necesarias de

aplicabilidad y la idoneidad del modelo seleccionado(10-12). Algunas de las características comunes que se han de considerar durante la elaboración de un modelo de regresión son las siguientes:

- **Linealidad:**

Se supone que la variable respuesta depende linealmente de las variables explicativas. Si la respuesta no aparenta ser lineal, debemos introducir en el modelo componentes no lineales (como incluir transformaciones no lineales de las variables independientes en el modelo), o bien categorizar las variables continuas, estrategia recomendada por algunos autores (13,14). En el anexo I se puede ver un ejemplo de un ajuste de un modelo de regresión lineal ajustado, asumiendo linealidad de los predictores con datos reales de un estudio.

- **Interacción:**

Otro tipo de respuesta no lineal es la interacción. Decimos que existe interacción en la relación entre dos variables cuando los valores de una tercera afectan a esa relación, magnificándola o disminuyéndola dependiendo del nivel de la tercera variable. Es decir que la magnitud de la relación, (coeficiente del modelo), es diferente según los niveles de una tercera variable. Para tenerla en cuenta en un modelo se ha de incluir los términos de interacción, que equivaldría a introducir nuevas variables explicativas que en realidad son el producto de dos o más de las independientes.

- **Normalidad y homocedasticidad de los residuos:**

Se llaman residuos a las diferencias entre los valores calculados por el modelo y los realmente observados en la variable dependiente. Para obtener un buen modelo de regresión no es suficiente con que los residuos sean pequeños. La validez del modelo requiere que los mismos se distribuyan de modo normal en regresión lineal, y con la misma dispersión de los errores para cada combinación de valores de las variables independientes. Las distribuciones binomiales se utilizan para el manejo de los errores asociados a los modelos de regresión logística para respuestas binarias de la misma manera que la distribución normal estándar se utiliza en la regresión lineal general. La existencia de sobredispersión o infradispersión provoca sesgo en el cálculo de la estimación de la varianza del estimador. Esta condición en la práctica suele ser inverificable, sobre todo cuando existen muchas variables explicativas, ya que para todas las combinaciones de variables independientes no siempre tendremos respuestas. Lo que se suele hacer es examinar una serie de gráficos de residuos que

nos hagan sospechar de un mal ajuste (9,11). Por ejemplo si los residuos aumentan al aumentar la respuesta, o vemos que aparecen tendencias. Es decir, hay una serie de reglas heurísticas que nos ayudan a decidir si aceptar o no si el modelo de regresión es válido.

- **Número de variables independientes:**

Hay reglas generales para incluir las variables en los modelos de regresión (12). Una estrategia en elaboración de los modelos confirmatorios es incluir las variables confusoras, pero hay que tener cuidado de no ser redundante para no incurrir en multicolinealidad.

- **Multicolinealidad:**

Si dos o más variables independientes están estrechamente relacionadas (por ejemplo: consumo de refrescos y temperatura ambiente) y ambas son incluidas en un modelo, muy posiblemente ninguna de las dos sea considerada significativa, ya que los errores estándar pueden verse incrementados considerablemente, generado por una incorrecta estimación de la varianza del parámetro.

- **Observaciones anómalas:**

Debemos poner especial cuidado en identificarlas (y destacarlas si procede), pues pueden tener gran influencia en el resultado final. A veces solo son errores en la entrada de datos, pero de gran consecuencia en el análisis y el resultado final.

- **Independencia de los errores:**

Otra de las hipótesis en las que se basan los modelos de regresión es que las observaciones muestrales sean completamente independientes. Con ello se entiende que los errores son variables aleatorias independientes. La falta de independencia, se produce fundamentalmente cuando se trabaja con variables aleatorias que se observan a lo largo del tiempo, como por ejemplo cuando se trabaja con series temporales. La ausencia de aleatoriedad entre las observaciones es difícil de corregir y es especialmente grave ya que puede invalidar por completo las conclusiones del análisis estadístico. Todas las expresiones utilizadas para las varianzas son incorrectas y, por tanto, los intervalos de confianza y las pruebas de hipótesis deducidos a partir de ellas, tendrán una confianza o una potencia estadística distinta a la supuesta. La forma de evaluar la independencia de los errores seria mediante gráficos de residuos.

❖ **Riesgos proporcionales en los modelos de COX:**

El análisis de supervivencia nos permite construir modelos para analizar el tiempo que un suceso tarda en ocurrir, y en los que diferentes variables pronóstico permiten estimar el tiempo y riesgo de aparición del suceso. Entre los diferentes tipos de modelos que se pueden emplear, uno de los más extendidos en medicina es el modelo de riesgos proporcionales, también conocido como modelo de Cox, y del que podemos encontrar publicados gran número de estudios, sobre todo en áreas relativas a enfermedades crónicas (en cardiología es bien conocido el modelo de Framingham).

Al igual que en los modelos de regresión lineal, también en los modelos de COX la mayor parte de los procedimientos de verificación del modelo se basan en los residuos. Algunos de los procedimientos para determinar si el modelo está bien construido se basan en representar gráficamente estos residuos y evaluar si presentan patrones anómalos frente a la forma que teóricamente deberían presentar.

Recordemos que en el modelo de regresión de Cox la función de riesgo (hazard) con dos variables explicativas tendría la siguiente forma:

$$h(t) = h_0(t)e^{\beta 1\, x1 + \beta 2\, x2}$$

Este modelo también se denomina semiparamétrico, ya que la función de riesgo base o de referencia $h_0(t)$ no queda especificada y puede tomar cualquier forma sin embargo el término denominado perfil riesgo ($\beta 1\, x1 + \beta 2\, x2$) si viene representado como una función totalmente parametrizada. La medida de asociación de la variable introducida en el modelo de Cox se suele expresar en forma de hazard ratio (HR), que es el cociente entre dos perfiles de riesgo (categorías de variables) y no depende del tiempo, y por ello a los modelos de Cox se les denomina modelos de riesgo proporcionales.

La verificación de la hipótesis de riesgos proporcionales se realiza mediante gráficos de residuos. Normalmente se suele superponer una curva de ajuste, utilizando alguna función de ajuste local, de alisado, que suelen estar disponibles en la mayor parte de los programas estadísticos. Una de las pruebas de verificación más conocidas son los residuos Schoenfeld para cada uno de los factores pronóstico del modelo. Si se cumple la hipótesis de riesgos proporcionales los residuos debieran agruparse de

forma aleatoria a ambos lados del valor 0 del eje Y, y la curva ajustada debería ser próxima a una línea recta.

❖ **Bondad de ajuste en regresión logística:**

Existen distintas formas analíticas para evaluar si el modelo logístico ajusta a los datos (Pearson, Deviance, Hosmer-Lemeshow). En la mayoría de estos test, un p-valor bajo indica que las probabilidades predichas se desvían de las probabilidades observadas. Por ejemplo la prueba de Hosmer and Lemeshow (H&L) evalúa el ajuste del modelo comparando las frecuencias observadas con las esperadas. La estimación de las probabilidades se agrupan de menor a la mayor, y entonces se realiza un test de la Chi cuadrado para determinar si las frecuencias observadas y esperadas son significativamente diferentes. Cuando el p-valor del test de H&L es significativo el modelo no describe bien los datos. (Ver anexo II: Ajuste de un modelo logístico)

3.4 Errores estadísticos en artículos científicos publicados

Todas estas consideraciones y comprobaciones formales de los modelos de regresión multivariable ayudan al investigador a seleccionar un buen modelo. En caso de ignorarlas, el análisis puede ocasionar estimaciones incorrectas causadas por un uso inapropiado de la metodología estadística. En este sentido, algunos estudios han descrito que el inapropiado uso de herramientas estadísticas es uno de los errores más comunes en publicaciones biomédicas(5,15,16). En el año 2008, Groenwold et al.(5) realizaron una revisión sistemática de estudios observacionales publicados en revistas de medicina general y epidemiología de alto factor de impacto, indicando que la calidad del reporte de los métodos estadísticos de ajuste fue pobre. Posteriormente en el año 2014, Casals et al.(15) mostró, en otra revisión del uso de modelos lineales generalizados mixtos, que la calidad del reporte tenía margen de mejora. Sin embargo, no se encontraron trabajos de revisión publicados sobre la calidad y el reporte en relación a la validación en la aplicación de MRM's donde su uso como técnica de ajuste se ha convertido en muy habitual en la investigación clínica (4,17).

3.5 Metodología matching

Otras soluciones para corregir el sesgo de confusión con interés creciente son los métodos matching, muy utilizados en el ámbito de la farmacoepidemiología(18-20) y en el contexto de la inferencia causal. Estos métodos consisten en, mediante un pre-proceso de datos, seleccionar una submuestra de observaciones, de forma que minimicen los desequilibrios iniciales entre los grupos de comparación según las

potenciales variables de confusión y a posteriori estimar la medida de asociación de interés con la nueva submuestra seleccionada. Los métodos de emparejamiento según factores como la edad y el sexo ya eran particularmente habituales en los estudios de casos y controles (21).

Existen distintos algoritmos en el proceso de datos para la selección de las observaciones y formación de grupos (exact, subclassification, nearest-neighbour, optimal, genetic matching, Coarsened Exact matching etc…). La mayoría de estos algoritmos están implementados en paquetes estadísticos estándar como STATA, SAS, o R. Estas metodologías ofrecen la promesa de realizar inferencias más robustas en comparación a los modelos paramétricos tradicionales(22,23). Por lo tanto siempre que se disponga de una cantidad de datos suficiente puede ser una buena alternativa para controlar el sesgo de confusión. Formalmente se llaman unidades tratadas a los individuos del grupo exposición y unidades control a los individuos del otro grupo, aunque no siempre sea un tratamiento lo que estemos evaluando.

3.5.1 Origen de los métodos matching

Los métodos matching han sido utilizados desde la primera mitad del siglo 20th (24,25), sin embargo, la base teórica para estos métodos no fue desarrollada hasta los años 70. Éste desarrollo se inició con los trabajos de Cochran y Rubin (26) para situaciones con una única covariable y enfocado a estimar el efecto promedio de un tratamiento. Althauser y Rubin (27) han discutido sobre algunos problemas prácticos relacionados con el matching: Que tamaño debería tener el grupo control para conseguir buenas parejas, como determinar la calidad del matching o como definir la cercanía entre observaciones. Muchos de estos temas aún siguen en debate. Los últimos trabajos mostraron que los mejores escenarios de matchings se conseguían cuando se disponía de muchos más controles que tratados, poco sesgo inicial entre grupos, y una varianza menor en el grupo tratado en comparación con el grupo control.

El tratamiento con múltiples covariables fue todo un reto para la metodología matching por los problemas computacionales que representa y por la cantidad de datos a manejar. Con solo unas pocas covariables era muy difícil encontrar coincidencias cercanas o exactas. Por ejemplo, Chapin (25), con una muestra inicial de 671 unidades tratadas y 523 controles, solamente encontró 23 parejas que emparejasen exactamente según seis covariables categóricas. En 1983 se hizo un avance importante con la introducción de la puntuación de propensión (Propensity score), definido como la probabilidad de recibir el tratamiento, dadas unas covariables

observadas(28). El propensity score facilitó la construcción de grupos equilibrados con similares distribuciones de las covariables, sin requerir pares cercanos o exactos en todas las variables individuales.

3.5.2 Etapas en la aplicación del matching

Los métodos matching tienen cuatro etapas clave, donde las tres primeras representan la fase de "diseño" y la última la etapa es la de "análisis". Las etapas de los métodos matching son:

1. Definición de la medida de proximidad o cercanía utilizada para determinar si dos observaciones son buenos pares.

2. Aplicación del algoritmo de matching para la elección de observaciones y conformación de nuevos grupos, en base a la medida de proximidad elegida.

3. Evaluación del equilibrio (calidad del matching) de los grupos emparejados. Si la calidad del matching no se satisface se repiten las etapas 1 y 2 hasta que los grupos puedan considerarse coincidentes.

4. Análisis de los resultados y estimación del efecto o medida de asociación de interés con la muestra finalmente emparejada.

En el anexo III se pueden ver 2 ejemplos ilustrativos de los cambios después del proceso del matching en las variables involucradas.

3.5.3 Definición de proximidad

Hay dos aspectos principales que determinarán la medida de proximidad (o cercanía). El primero involucra a las covariables que incluimos, y el segundo, la combinación de éstas resumidas en una medida de la distancia. La selección de las variables a incluir implica la mayoría de principios comunes que requieren las técnicas multivariables de ajuste. Aunque depende de la medida de distancia que usemos, ya que si hay una gran cantidad de covariables pueden quedar pocas observaciones y perder eficiencia en términos de varianza a la hora de estimar la medida de asociación de interés.

Existen varias medidas de proximidad para determinar la distancia D_{ij} entre dos individuos i y j: Exacta, Mahalanobis, y basadas en el Propensity score. La distancia exacta, asigna el valor 0 a las observaciones idénticas según sus covariables, y valor=∞ a aquellas que no son exactamente iguales. La distancia de Mahalanobis es una forma de determinar la similitud entre dos variables aleatorias multidimensionales parecida a la distancia euclídea pero que tiene en cuenta la correlación entre las

variables. Y la distancia según el propensity score (PS) que es un resumen de todas las covariables mediante un escalar que representa la probabilidad de estar en el grupo tratado, estimado por ejemplo, mediante un modelo multivariable de regresión logistica.

3.5.4 Algoritmos

Una vez determinada la distancia el siguiente paso es usarla para la elección de las observaciones y formación de grupos. Para ello se puede utilizar cualquiera de los algoritmos existentes (exact, subclassification, nearest-neighbour, optimal, geneticmatching, full matching, Coarsened Exact Matching etc.). Una buena revisión de estos métodos la realiza Stuart E. et al.(29) en el año 2010. La mayoría de estos algoritmos se incluyen en la librería MatchIt del software R (30). Brevemente detallamos algunos de ellos:

- *Exact Matching* (*Exact*): Algoritmo que empareja cada unidad tratada con todas las posibles unidades del grupo control de manera que ambos grupos contengan exactamente los mismos valores según las covariables especificadas. Cuando hay muchas covariables y/o las covariables pueden tomar un amplio rango de valores, exact matching puede no ser posible.

- *Subclassification* (*Subclas*): Algoritmo que forma estratos, en función de la distribución de las distancias estimadas de tal manera que asegura la igualdad de distribuciones dentro de cada estrato según las covariables seleccionadas. Se pueden descartar observaciones para mejorar la igualdad de distribuciones dentro de cada estrato.

- *Nearest Neighbour* (N-N): Este algoritmo selecciona los mejores controles emparejados para cada individuo tratado. En cada etapa del matching se elige la observación del grupo control que esté lo más cercana a la observación tratada según la distancia especificada. Si múltiples individuos controles se encuentran a la misma distancia del sujeto tratado, se selecciona aleatoriamente a uno de estos. La opción caliper (número de desviaciones estándar de la medida de la distancia) establece una distancia máxima entre tratados y controles para ser seleccionados asegurando una igualdad mínima entre observaciones.

- *Optimal Matching*: Este algoritmo se centra en minimizar la distancia absoluta en promedio a través de todos los pares. Este método de matching también requiere la librería optmatch del sofware R.

- *Genetic Matching*: Esta técnica utiliza un algoritmo de búsqueda computacionalmente intensivo para hallar unidades apareadas del grupo control. Este algoritmo también requiere la librería Matching del sofware R.

3.5.5 Tamaño de la muestra y ratio control-tratamiento

Además de la distancia especificada, los algoritmos matching tienen varias opciones y configuraciones que implicaran resultados matching distintos. Una de ellas es el ratio tratado/control que por defecto es 1, pero a veces la relación del número de unidades control por cada tratado no necesariamente tiene que ser 1 a 1. Cuando hay un número mucho mayor de individuos en el grupo control que en el grupo tratado puede ser mejor seleccionar múltiples controles por cada tratado. Aunque esta opción generalmente aumenta el sesgo ya que los siguientes pares (2º, 3º, 4 etc…) más cercanos, generalmente, si no hay empates estarán más lejos del primer individuo tratado seleccionado, que es el que tendrá la mayor coincidencia. Pero, por otra parte, la utilización de múltiples pares puede reducir la varianza debido al incremento del tamaño de la muestra. Aproximaciones de Rubin y Thomas en 1996 (31) pueden ayudar a determinar el mejor ratio. Más recientemente King G. (32) propuso en el año 2010 comparar distintas soluciones matching mediante gráficos espaciales que ayudan a examinar el número de pares, y tamaño de muestra en relación a la reducción del sesgo. En este sentido King G. (32) discuten sobre la falta de optimización de la metodología matching en ambos sentidos; mejora de equilibrio y eliminación de observaciones. Recientemente (2015) el mismo King G., publica un paquete de R 'MatchingFrontier' (http://projects.iq.harvard.edu/frontier) que ayuda a encontrar una solución matching óptima, devolviendo subconjuntos de datos con el mínimo desequilibrio para cada posible tamaño de muestra (33). Pero aún hacen falta más trabajos metodológicos para cuantificar de manera más formal las compensaciones involucradas.

3.5.6 Evaluación y diagnóstico del matching

En cualquier caso, y al igual que en la construcción de un buen MRM, un paso requerido durante la aplicación de los métodos matching, es realizar la validación del resultado del algoritmo y por lo tanto hacer un buen diagnóstico de la calidad del matching construido. Si el método nos ha proporcionado una muestra con grupos desequilibrados, esta debe ser rechazada, y se tiene que repetir el proceso hasta conseguir un mejor equilibrio. En algunas ocasiones un mal diagnóstico puede indicar que ambos grupos (tratado y control) están demasiado separados para proporcionar estimaciones fiables (34).

Una de las medidas numéricas diagnósticas comunes para valorar el equilibrado después del matching es calcular la diferencia de medias estandarizada que es la

diferencia de medias de cada covariable entre grupos, dividido por la desviación estándar del grupo tratado. Esta medida es similar al tamaño del efecto y se compara antes y después del matching (28). Para las covariables binarias, se puede utilizar la misma fórmula, o bien se puede calcular la diferencia de proporciones(35). Este análisis se puede representar gráficamente comparando las diferencias antes y después del matching. Se puede un ejemplo en anexo IV correspondiente a las diferencias de cada covariable antes y después del matching de un estudio publicado(36).

Otras métricas de equilibrio global que se pueden aportar son: La diferencia de medias estandarizada de la distancia calculada (D_{ij}), o el ratio de las varianzas de la distancia entre el grupo tratado y control. Además también puede ser de utilidad realizar cualquier análisis cuantitativo, a nivel descriptivo de cada variable introducida en el cálculo de la distancia. Los test estadísticos de significación, tales como el t-test, no son recomendables como regla general para valorar el equilibrio ya que estos son sensibles al tamaño de la muestra.

3.5.7 Análisis del outcome

Finalmente, después de la aplicación del matching con la creación de la nueva muestra, y del control adecuado del equilibrio (y por lo tanto el estudio observacional “diseñado”), ya se puede pasar a la fase de análisis y calcular la estimación de las medida de asociación de interés tales como el efecto, odds ratio, riesgo atribuible, hazard ratio, etc. Ésta etapa generalmente puede implicar ciertos ajustes de regresión, con la particularidad de que ahora la muestra tiene una estructura de datos emparejada.

Una característica de los métodos de matching es que, aunque se haya reducido en gran medida la dependencia al modelo, ambos métodos pueden ser complementarios si existe cierto desequilibrio después del matching (22). De hecho se ha demostrado que los dos métodos funcionan mejor en combinación (37,38). Esto es similar a la idea de "doble robustez", y en cierto sentido sería lo mismo que el ajuste con regresión después de realizar un ensayo clínico aleatorio, donde se puede utilizar un ajuste por regresión para "limpiar” el pequeño desequilibrio de covarianza residual entre los grupos (29).

El método de análisis posterior al matching con la nueva muestra se puede realizar simplemente, como si estas muestras se hayan generado a través de la asignación al azar a los grupos. Sin embargo, existe cierto debate y controversia sobre si en el

análisis requiere o no tener en cuenta la naturaleza emparejada de los datos (23,39-43). Algunos autores afirman que existen suficientes razones por las que no se necesita tener en cuenta el emparejamiento para el análisis posterior (41-43). Una de ellas es que el procedimiento matching, no garantiza que las parejas de individuos estén bien emparejadas según el conjunto completo de las covariables. En cualquier caso sería suficiente condicionar por las mismas variables que se utilizaron en el proceso del matching (Similar a un modelo de regresión). Según Ho et al. (40), en esencia, los investigadores pueden hacer exactamente el mismo análisis que hubieran hecho con los datos originales, pero utilizando la muestra emparejada en su lugar.

Del mismo modo, los investigadores tampoco se ponen de acuerdo en si hay que tener en cuenta la incertidumbre de la estimación de la medida de la distancia (PS) en la estimación de la varianza del estimador, y de qué manera. Ho et al. (40) también adopta el enfoque similar a los experimentos aleatorios, condicionando solo por las covariables, las cuales son tratadas como fijas y exógenas. Otros investigadores sostienen que la incertidumbre de la estimación del PS se debe tener en cuenta de alguna forma (31,42). Métodos de Bootstrap más robustos (39) han mostrado mejores resultados en rendimiento, y además hay fórmulas empíricas (37) para la estimación de los efectos teniendo en cuenta el matching. Recientemente, Pearce N (43) en 2016, discute, algunas malas concepciones sobre los diseños matching, particularmente en estudios caso-control y sus métodos de análisis. Pearcen N, afirma que particularmente en el análisis de los diseños de casos y controles matching, requieren el control de los factores utilizados en el proceso del matching, pero el tipo de análisis estándar (no condicional), no es menos válido, y es más preciso que los que tienen en cuenta la naturaleza emparejada de la muestra (como regresión logística condicional). Por lo tanto, ésta es aún, un área para futuras investigaciones.

3.6 Comparativa de los métodos matching con ajuste por regresión

Sobre los métodos basados en matching utilizando el PS (PS-M), Austin P. (23) señala varias razones prácticas para preferir el uso de estos métodos en comparación con el ajuste de regresión habitual. Primero, es más simple determinar si el modelo ha sido adecuadamente especificado (solo con verificar la homogeneidad de los grupos después del matching o creación del PS). En cambio, es mucho más difícil determinar si un modelo de regresión relativo a la variable a evaluar, más los confusores basales sobre el outcome han sido correctamente especificados(44).

Segundo, estos métodos permiten separar el diseño del estudio del análisis, de la misma forma que lo hace un ensayo clínico aleatorizado, donde solamente se puede estimar el efecto del tratamiento estudiado una vez el estudio ha finalizado. Cuando se usan las técnicas de matching, la construcción de los grupos se puede realizar sin ninguna referencia al outcome. Solamente una vez se acepta el equilibrio de los grupos se realiza la estimación del efecto del tratamiento sobre el outcome. Sin embargo, cuando utilizamos técnicas de regresión, el outcome siempre está a la vista, y el investigador se enfrenta continuamente a la sutil tentación de modificar el modelo de regresión hasta que se alcanza el resultado esperado.

Tercero, se puede examinar explícitamente el grado de solapamiento o superposición de la distribución de las covariables entre los grupos. Si hay substanciales diferencias de las covariables entre grupos, estas serán evidentes, dado el pequeño número de sujetos emparejados. Este hecho es importante ya que, por ejemplo, si casi la totalidad de fumadores están en uno de los grupos y la mayoría de los no fumadores están en el otro, ajustar por tabaquismo no eliminará la confusión causada por esta variable(45). En el caso de que haya problemas de solapamiento, se generarán insuficientes sujetos en los grupos, el analista o investigador puede concluir que los grupos son tan diferentes que una comparativa de la respuesta no tiene sentido y por lo tanto no será plausible el análisis. En cambio, al utilizar enfoques basados en la regresión, puede ser difícil evaluar el grado de solapamiento entre grupos. Y en un contexto en el que exista una fuerte separación entre los grupos, se puede cometer el error de proceder a un análisis basado en la regresión sin ser consciente que el modelo de regresión ajustado está interpolado en dos poblaciones distintas. Este problema se dará en menor frecuencia en estudios con muestras grandes sin escasez de datos.

3.7 Grandes bases de datos en epidemiologia

Actualmente, existe una cuestión de moda en el ámbito de la epidemiologia clínica y la fármacoepidemiología que es la gestión y utilización de los *real world data* (datos de práctica clínica). Un término que se corresponde, como explicó Jessamy Baird, directora de Real World Evidence de Eli Lilly en Europa, Australia y Canadá a "todo aquello que excede el ensayo clínico aleatorizado (ECA)". En este grupo habría que meter a los estudios suplementarios de los ECA, los ensayos clínicos pragmáticos, los registros de pacientes, los datos administrativos, las encuestas de salud o las historias clínicas electrónicas, entre otros. En este sentido, la informatización de la historia clínica en la era del *Big Data* ha proporcionado un gran potencial para la investigación clínica ya que a partir de grandes bases de datos existentes se pueden realizar

estudios epidemiológicos con grandes muestras a un coste muy inferior a los estudios convencionales(46). Este hecho se ha podido demostrar en la organización de redes informatizadas o de gran bases de datos en diversos países como en Inglaterra, Holanda y España (General Practice Research Database: http://www.gprd.com; QRESEARCH: http://www.qresearch.org; THIN Database: http://www.thin-uk.com; Pharmo: http://www.pharmo.nl; y SIDIAP). A partir de estos datos provenientes de grandes bases de datos se están llevando a cabo múltiples estudios epidemiológicos, con diseños observacionales muy diversos.

Las bases de datos de registros clínicos presentan una serie de ventajas respecto de otras fuentes de información. Entre las ventajas de estas base de datos podemos destacar que aportan grandes muestras de personas y seguimientos de larga duración a un coste muy inferior al de los estudios de cohortes o caso-control convencionales; los datos se relacionan con múltiples fuentes de información (altas hospitalarias, facturación de farmacia, registros censales, etc.) que permiten una información más rica y completa; no supone una participación activa del paciente cuando se recoge la información y los datos son muy representativos de la práctica clínica real puesto que son recogidos durante la puesta en escena de la misma.

La principal limitación de estas fuentes de datos es la falta de validación individual y el posible infra registro en determinadas situaciones. En este sentido, la aplicación de métodos de emparejamiento puede aumentar la eficiencia del estudio al eliminar la heterogeneidad o eliminar observaciones fuera del área donde un modelo suele utilizar para extrapolar (22).

3.8 Retos y oportunidades del matching

En el contexto de los *data real world*, la metodología matching puede ser una buena alternativa válida a los MRMs, ya que una reducción de la muestra puede no traducirse en una pérdida de potencia estadística determinante para evaluar la hipótesis de investigación, y en cambio puede mejorar la calidad de la inferencia realizada sobre las asociaciones examinadas.

En cuanto al rendimiento que ofrecen este tipo de metodología de ajuste respecto a las metodología más tradicional, existen algunos estudios analíticos y de simulación que han comparado alguno de estos métodos (47-52), sobretodo centrándolos en las metodología relacionada con el PS, presentando resultados dispares. Entre estos trabajos no encontramos ninguno que compare distintos algoritmos matching en condiciones de mala especificación de los modelos paramétricos tradicionales,

situación donde el beneficio que aporta la metodología matching puede ser más evidente. Por lo tanto, también puede ser de interés evaluar el rendimiento de distintos algoritmos de matching con su análisis posterior como herramienta de ajuste, en comparación con métodos de regresión más tradicionales.

4. Hipótesis y objetivos

4.1 Hipótesis

1) La mayoría de las publicaciones científicas indexadas en PubMed que usan las técnicas de regresión multivariables más conocidas (Modelo logísticos, regresión lineal o de COX), no reportan estadísticos de validación diagnostica de estos métodos, tales como medidas de bondad de ajuste.

    a. Existe una alta y creciente frecuencia de publicaciones científicas que usan técnicas de regresión multivariables como método de ajuste en estudios observacionales.

    b. Existe un bajo reporte de la validación o utilización de estadísticos de bondad de ajuste de los MRMs más comunes utilizados en trabajos observacionales analíticos publicados en PubMed.

    c. Existe un reporte más completo de los MRMs en aquellos artículos publicados en revistas con un mayor factor de impacto, y también puede variar en función de características del estudio, como el diseño, tamaño de muestra del estudio, y tipo de modelo utilizado.

2) Los métodos matching, como herramienta de ajuste ofrecen inferencias más robustas en comparación con una técnica de ajuste más tradicional y conocida como es la regresión logística multivariable en condiciones de efecto nulo y presencia de confusión.

    a. Los métodos de emparejamiento “matching” presentan estimaciones más cercanas al efecto real en comparación con distintas especificadores del modelo de regresión logística, en escenarios donde la relación del confusor con la respuesta no es lineal.

    b. Diferentes especificaciones de un modelo de regresión logístico como método de ajuste presentan estimaciones muy distintas de una misma medida de asociación.

3) Las técnicas de emparejamiento de datos se pueden aplicar especialmente en los estudios observacionales de fármaco-epidemiología y salud pública con grandes bases de datos existentes. La utilización de técnicas de emparejamiento con reducción de la muestra no implica una pérdida de potencia estadística determinante que dificulte la evaluación de las hipótesis.

4.2 Objetivos

1) Describir la frecuencia de uso de las técnicas de regresión más habituales (Lineal, logística, Cox) en estudios observacionales analíticos publicados en PubMed, y evaluar la información que presentan cuando se aplican estas técnicas como método de ajuste:

    a. Describir la frecuencia de uso de los MRM y su crecimiento mediante el motor de búsqueda de PubMed según términos referidos a modelos de regresión.
    b. Verificar el reporte de pruebas estadísticas, principalmente bondad de ajuste, además de la evaluación de interacciones, informar de medidas crudas y ajustadas, análisis de sensibilidad, y/o ajuste de más de un modelo.
    c. Evaluar si existe relación entre el reporte de información del uso de MRMs con el factor de impacto de la revista donde se publica, y datos relativos al estudio como el tamaño de muestra, diseño, y el tipo de modelo utilizado.

2) Comparar métodos matching con otras técnicas de ajuste conocidas (Regresión logística) en escenarios donde la relación confusor-respuesta no sea lineal mediante un estudio de simulación. Particularmente, examinar el sesgo residual y el error de tipo I empírico en muestras grandes (n=10.000) de 7 técnicas de ajuste:

    a. 3 algoritmos matching: *Exact*, *Nearest-Neighbour*, *Subclassification* con su posterior análisis univariable convencional.
    b. 4 especificaciones de un modelo lineal generalizado (regresión logística): Introduciendo el confusor de forma aditiva y continua, polinomial, categorizado en quintiles, y mediante una función *smothing spline*.

3) Aplicar técnicas de matching en estudios observacionales analíticos de fármaco-epidemiología y salud pública con grandes bases de datos existentes, en tres diseños distintos (Cohorte retrospectiva, transversal, casos y controles). A continuación se exponen los objetivos de los tres estudios observacionales análiticos donde se han aplicado las técnicas matching:

    a. Estudio OSTEOFRAC: Estimar la incidencia acumulada a los 5 años de primera fractura osteoporótica en dos cohortes de mujeres postmenopáusicas, unas tratadas con bifosfonatos y las otras tratadas solo con calcio y vitamina D.

b. Estudio CUIDADORES: Evaluar la posible asociación de estar en una situación de cuidador informal y tener peores resultados en salud.

c. Estudio DAMOCLES: Evaluar si un mejor control de los factores de riesgo cardiovasculares (tales como: IMC, presión arterial, colesterol, frecuencia cardiaca) se asocia a un menor riesgo de la primera hospitalización por insuficiencia cardiaca a corto-medio plazo (antes de 1 o 2 años).

5. Objetivos y resumen de los resultados de los artículos

La presente tesis doctoral lo forma un compendio de seis trabajos científicos de los cuales, cinco son artículos aceptados y publicados en revistas indexadas en PubMed y Web of Science, y uno se presenta en forma de manuscrito. Los 6 trabajos de la tesis se pueden separar en dos grupos: A) Tres trabajos metodológicos que corresponden a los 2 primeros objetivos de la presente tesis doctoral; y B) tres artículos aplicados que responden al último objetivo de la presente tesis doctoral.

5.1 A) Trabajos metodológicos

A continuación se exponen los tres trabajos metodológicos de la tesis doctoral, así como el factor de impacto, si procede, de la revista donde se ha publicado.

1) Estudio BIBLIOMÉTRICO: Real J, Cleries R, Forné C, Roso-Llorach A, Martínez-Sánchez JM. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. SEMERGEN. 2015 (en prensa). (Artículo científico).

SEMERGEN está indexada en el repositorio PubMed y no tiene factor de impacto.

2) Estudio de REVISIÓN: Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine (Baltimore). 2016 May;95(20):e3653. (Artículo científico).

MEDICINE está incluida en los Journal Citation Report de ISI-Web of Science con un factor de impacto en 2014 de 5,723, Q1 (posición 40/155 en la categoría Medicine, General & Internal).

3) Estudio de SIMULACIÓN: Real J, Forné C, Martínez-Sánchez JM. "Comparación del error tipo I entre los métodos matching y los modelos paramétricos: estudio de simulación". XXXIV Reunión científica de la SEE. 15 de septiembre 2016, Sevilla. Sociedad española de epidemiología. (Manuscrito).

Se adjunta una primera versión del manuscrito que se enviará a publicar a una revista indexada. Cabe destacar que parte de este trabajo se presentó en formato de

comunicación oral en la XXXIV reunión científica de la Sociedad Española de Epidemiologia (anexo VI).

5.2 B) Trabajos aplicados

A continuación se exponen los tres artículos de la tesis doctoral donde se aplicaron las técnicas matching, así como el factor de impacto, si procede, de la revista donde se ha publicado.

1) Estudio OSTEOFRAC: Real J, Galindo G, Galván L, Lafarga MA, Rodrigo MD, Ortega M. Use of oral bisphosphonates in primary prevention of fractures in postmenopausal women: a population-based cohort study. PLoS One. 2015;10(4):e0118178. (Artículo científico).

PLoS ONE está incluida en los Journal Citation Report de ISI-Web of Science con un factor de impacto en 2015 de 3,057, Q1 (posición 11/63 en la categoría Mutidisciplinary Science).

2) Estudio CUIDADORES: Gonzalez-de Paz L, Real J, Borras-Santos A, Martinez-Sanchez JM, Rodrigo-Banos V, Dolores Navarro-Rubio M. Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study". J Public Health Policy.2016 May;37(2):173-89. (Artículo científico).

Journal of Public Health Policy está incluida en los Journal Citation Report de ISI-Web of Science con un factor de impacto en 2014 de 1,775, Q2 (posición 44/88 en la categoría Health Care Science & Services).

3) Estudio DAMOCLES: Muñoz MA, Real J, Del Val JL, Vinyoles E, Mundet X, Domingo M, Enjuanes C, Verdú-Rotellar JM. Impact of the sustained control of cardiovascular risk factors on first episode heart failure: The relevant role of primary care. E J Gen Pract. 2015;21(4):224-30. (Artículo científico).

The European Journal of General Practice está incluida en los Journal Citation Report de ISI-Web of Science con un factor de impacto en 2015 de 1,364, Q1 (posición 72/155 en la categoría Medical, General & Internal).

5.3 Resumen de trabajos de la tesis doctoral

A continuación se expone un breve resumen y principales resultados de los seis trabajos que conforman la tesis doctoral.

1) Estudio BIBLIOMÉTRICO: Real, J., Cleries, R., Forné, C., Roso-Llorach, A., & Martínez-Sánchez, J. M. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. SEMERGEN. 2015 (en prensa). (Artículo científico).

El objetivo principal de este trabajo fue describir la evolución de las técnicas de regresión en los artículos observacionales indexados en PubMed (1970-2013). El estudio se realizó mediante estrategias de búsqueda en PubMed para identificar artículos que utilizasen o referenciasen modelos de regresión logística, lineal, Cox y Poisson en sus estudios. Además, se revisaron las normas de autor de las revistas editadas en España indexadas en PubMed e incluidas en Web Of Science para conocer si recomendaban la guía STROBE para la preparación de los trabajos. Los resultados principales mostraron que el 6,1% de los artículos de estudios observacionales contenían algún término relativo a los modelos seleccionados, pasando del 0,14% en 1980 hasta un 12,3% en 2013. Este último año, un 6,7% de los artículos contenían algún término referido a regresión logística, un 2,5% lineal, un 3,5% de Cox y un 0,31% Poisson. Además, el 12,8% de las revistas editadas en España solicitaban la guía STROBE.

2) Estudio de REVISIÓN: Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine" (Baltimore). 2016 May;95(20):e3653. (Artículo científico).

El objetivo de este estudio fue describir la información estadística reportada sobre la aplicación de los MRMs más comúnmente utilizados (logística, lineal y regresión de Cox) en artículos observacionales analíticos publicados durante los últimos 10 años en revistas indexadas en MEDLINE. Particularmente se evaluó si los artículos informaban de cada uno de los siguientes aspectos: Validación de las hipótesis del modelo utilizado y/o comprobación de la bondad de ajuste, evaluación de las interacciones, análisis de sensibilidad, aportar medidas de asociación crudas y ajustadas, y mostrar más de 1 modelo de ajuste (En el anexo V de la presente tesis se adjunta el cuestionario completo de recogida de datos). Para ello se revisó un muestra aleatoria de 428 artículos con diseño observacional que utilizaran cualquiera de los MRMs. El

reporte de la validación de los MRM en los artículos observacionales indexados en PubMed fue bajo, ya que solamente 1 de cada 4 artículos revisados aportó información sobre la validación de las asunciones o la bondad de ajuste del MRM utilizado. Además sólo 1 de cada 3 trabajos revisados mostró los efectos crudos junto con los ajustados. La evaluación de las interacciones se detectó en 1 de cada 5 trabajos, aunque aumentó en trabajos con muestras más grandes y publicados en revistas con mayor Factor de impacto.

3) Estudio de SIMULACIÓN: Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM. "Comparación del error tipo I y sesgo residual entre métodos matching y los modelos de regresión logística como método de ajuste en presencia de confusión: estudio de simulación". (manuscrito).

El objetivo de este estudio fue comparar el sesgo de confusión residual de distintos métodos de ajuste (3 algoritmos matching: Exact, Nearest-Neighbour(N-N) y Subclassification versus; y 4 especificaciones de modelos logísticos) simulando escenarios donde la relación confusor-respuesta no fue lineal. Para ello se realizó un estudio de simulación con 7800 réplicas de muestras de tamaño, n=10.000, asumiendo distintos escenarios, para la comparación de la confusión residual en la estimación de un efecto nulo, con cada uno de los métodos utilizados. Los 3 métodos matching presentaron un menor sesgo residual y consecuentemente errores tipo I empíricos ($\hat{\alpha}$) más cercanos al nominal (0,05), en las 8 formas funcionales de la relación confusor–respuesta consideradas. El método matching Exact presentó la menor tasa de error tipo I, incluso inferior al nivel teórico en los dos niveles de correlación X-Z considerados ($\hat{\alpha}$=0,041 y 0,039, para r_{XZ}=0,5 y 0,3 respectivamente), mientras que el método de Subclassification fue la estrategia matching que presentó una mayor tasa de error en ambos escenarios de asociación ($\hat{\alpha}$=0,084 y 0,061). Con el modelo logístico GAM se observaron errores empíricos del $\hat{\alpha}$ = 0,053 y 0,056, y con el modelo logístico categorizando el confusor $\hat{\alpha}$=0,085 y 0,093, mientras que incluyendo el confusor de manera directamente lineal la tasa de error fue la que obtuvo peores resultados medios ($\hat{\alpha}$=0,176 y 0,474).

4) Estudio OSTEOFRAC: Real J, Galindo G, Galván L, Lafarga MA, Rodrigo MD, Ortega M. Use of oral bisphosphonates in primary prevention of fractures in postmenopausal women: a population-based cohort study. PLoS One. 2015;10(4):e0118178. (Artículo científico).

El objetivo de este estudio fue evaluar y comparar las diferencias en la incidencia de primera fractura osteoporótica en dos cohortes de mujeres postmenopáusicas, unas

tratadas con bifosfonatos y otras tratadas solo con calcio y vitamina D. Las cohortes se formaron utilizando técnicas matching (Algoritmo Nearest-Neighbour) para asegurar equilibrio en términos de comorbilidad basal, edad y uso de otros fármacos modificadores oseos. Se usó la libreria “MatchIt" del paquete estadístico R (v3.0.1). El endpoint principal fue la primera fractura osteoporótica registrada durante el seguimiento. Los resultados principales fueron que el riesgo estimado de fractura osteoporótica en ambas cohortes durante el seguimiento fue del 11,4% (IC95%: 9,6%-13,2%). En el grupo tratado con bifosfonatos obtuvo una incidencia acumulada del 11,8% (9,2%-14,3%), y en el grupo de mujeres tratadas únicamente con Calcio y vitamina D del 11,1% (8,6%-13,6%). No se encontraron diferencias estadísticamente significativas entre las dos cohortes en las fracturas globales (HR = 0,934; IC95%:0,67-1,31) ni por localización (vertebral, femoral, radial o humeral).

5) Estudio CUIDADORES: Gonzalez-de Paz L, Real J, Borras-Santos A, Martinez-Sanchez JM, Rodrigo-Banos V, Dolores Navarro-Rubio M. Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study. J Public Health Policy 2016 May;37(2):173-189. (Artículo científico).

El objetivo de este trabajo fue medir el impacto sobre la salud y factores de riesgo asociados a los cuidados informales. Se utilizaron técnicas matching (Algoritmo N-N) para seleccionar controles parecidos a los cuidadores informales en cuanto a variables sociodemográficas y económicas. Los principales resultados encontrados fueron que los cuidadores informales presentaron peores resultados en salud y factores relacionados. Concretamente se encontró que los cuidadores informales tenían un 3,4% más de depresión (OR: 1,33; IC95%: 1,06-1,68), un menor apoyo social (Efecto: 2,46; IC95%: 1,64-3.28), y un mayor estrés (Efecto: 0,48: IC95%: 0,13 a 0,83), independientemente de su edad, sexo, composición del hogar, nivel de estudios, nivel ingresos familiares y tamaño de la población.

6) Estudio DAMOCLES: Muñoz MA, Real J, Del Val JL, Vinyoles E, Mundet X, Domingo M, Enjuanes C, Verdú-Rotellar JM. Impact of the sustained control of cardiovascular risk factors on first episode heart failure: The relevant role of primary care. E J Gen Pract. 2015;21(4):224-30. (Artículo científico).

El objetivo de este estudio fue examinar si un mejor control de los factores de riesgo cardiovasculares (tales como, índice de masa corporal, presión arterial, colesterol, frecuencia cardiaca, etc.) se asociaba a un menor riesgo de la primera hospitalización por insuficiencia cardiaca a corto-medio plazo (antes de 1 o 2 años). Para ello se diseñó un estudio de casos y controles retrospectivo con datos provenientes de

registros clínicos de la misma región sanitaria y con un único hospital de referencia. Se seleccionaron, casos (personas que sufrieron un ingreso hospitalario por insuficiencia cardiaca), y potenciales controles (personas sin haber sufrido el evento de interés durante el mismo periodo de tiempo de la misma región sanitaria). Después los controles y los casos fueron seleccionados y emparejados, utilizando técnicas de matching (Algoritmo Nearest-Neighbour), por edad, sexo, factores de riesgo cardiovascular, tratamiento para la prevención cardiovascular, y presencia de enfermedad coronaria previa. Los resultados principales fueron que aquellos pacientes que sufrieron un primer episodio de insuficiencia cardiaca tenían superiores cifras previas, aunque moderadas, del índice de masa corporal y en los niveles de la presión arterial en los dos años previos a la hospitalización. Los odds ratio ajustados de sufrir un primer evento según las cifras de presión arterial sistólica y valores de índice de masa corporal previos fueron de 1,031 (IC95%:1,001–1,04), y 1,09 (IC95%:1,03-1,15) respectivamente.

6. Artículos científicos de la tesis doctoral

6.1 Estudio BIBLIOMÉTRICO: Trabajo publicado en SEMERGEN

Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. SEMERGEN. 2015

Semergen. 2015;xxx(xx):xxx–xxx

Medicina de Familia
SEMERGEN
www.elsevier.es/semergen

ORIGINAL

Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas

J. Real[a,b], R. Cleries[c,d], C. Forné[e,f], A. Roso-Llorach[a,g] y J.M. Martínez-Sánchez[b,h,i,*]

[a] *Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona, España*
[b] *Facultat de Medicina i Ciències de la Salut, Universitat Intenacional de Catalunya, Sant Cugat, Barcelona, España*
[c] *Pla Director d'Oncologia de Catalunya, Institut Català d'Oncologia, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, España*
[d] *Departament de Ciències Clíniques, Universitat de Barcelona, Campus de Bellvitge, L'Hospitalet de Llobregat, Barcelona, España*
[e] *Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, Lleida, España*
[f] *Oblikue Consulting, Barcelona, España*
[g] *Universitat Autònoma de Barcelona, Bellaterra, Barcelona, España*
[h] *Unitat de Control del Tabaquisme, Programa de Prevenció i Control del Càncer, Institut Català d'Oncologia, L'Hospitalet de Llobregat, Barcelona, España*
[i] *Grup de Prevenció i Control del Càncer, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, España*

Recibido el 10 de diciembre de 2014; aceptado el 5 de junio de 2015

PALABRAS CLAVE
Análisis multivariante;
Análisis de regresión;
Modelos logísticos;
Modelos lineales;
Modelos de riesgos proporcionales;
Modelos de Poisson;
Estudios observacionales;
Epidemiología

Resumen

Fundamentos: En el ámbito de la investigación médica los modelos de regresión logística, lineal, Cox y Poisson son técnicas estadísticas ampliamente conocidas. El objetivo de este trabajo es describir la evolución de estas técnicas de regresión en los artículos observacionales indexados en PubMed (1970-2013) y revisar los requerimientos de las normas de autor de revistas españolas para conocer si requieren el cumplimiento de la guía STROBE.

Métodos: Se realizó una búsqueda dirigida en PubMed para identificar los artículos que utilizaron modelos de regresión logística, lineal, Cox y Poisson. Además, se revisaron las normas de autor de las revistas editadas en España indexadas en PubMed e incluidas en Web Of Science.

Resultados: El 6,1% de los artículos de estudios observacionales contenían algún término relativo a los modelos seleccionados, pasando del 0,14% en 1980 hasta un 12,3% en 2013. Este último año, un 6,7% de los artículos contenían algún término referido a regresión logística, un 2,5% a lineal, un 3,5% a Cox y un 0,31% a Poisson. Por otro lado, el 12,8% de las normas de autor de las revistas revisadas recomendaban explícitamente seguir la guía STROBE, y el 35,9%, la guía CONSORT.

* Autor para correspondencia.
Correo electrónico: jmmartinez@iconcologia.net (J.M. Martínez-Sánchez).

http://dx.doi.org/10.1016/j.semerg.2015.06.020

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

Conclusiones: Los modelos de regresión multivariantes en estudios observacionales publicados, tales como la regresión logística, lineal, Cox y Poisson, son cada vez más utilizados tanto a nivel global como en revistas publicadas en lengua española. Además, un porcentaje bajo de las revistas científicas españolas indexadas en PubMed incluyen en las normas de autoría el requerimiento de la guía STROBE.

KEYWORDS
Multivariate analysis;
Regression analysis;
Logistic models;
Linear models;
Proportional hazards models;
Poisson models;
Observational studies;
Epidemiology

Use of multiple regression models in observational studies (1970-2013) and requirements of the STROBE guidelines in Spanish scientific journals

Abstract
Background: In medicine and biomedical research, statistical techniques like logistic, linear, Cox and Poisson regression are widely known. The main objective is to describe the evolution of multivariate techniques used in observational studies indexed in PubMed (1970-2013), and to check the requirements of the STROBE guidelines in the author guidelines in Spanish journals indexed in PubMed.
Methods: A targeted PubMed search was performed to identify papers that used logistic linear Cox and Poisson models. Furthermore, a review was also made of the author guidelines of journals published in Spain and indexed in PubMed and Web of Science.
Results: Only 6.1% of the indexed manuscripts included a term related to multivariate analysis, increasing from 0.14% in 1980 to 12.3% in 2013. In 2013, 6.7, 2.5, 3.5, and 0.31% of the manuscripts contained terms related to logistic, linear, Cox and Poisson regression, respectively. On the other hand, 12.8% of journals author guidelines explicitly recommend to follow the STROBE guidelines, and 35.9% recommend the CONSORT guideline.
Conclusions: A low percentage of Spanish scientific journals indexed in PubMed include the STROBE statement requirement in the author guidelines. Multivariate regression models in published observational studies such as logistic regression, linear, Cox and Poisson are increasingly used both at international level, as well as in journals published in Spanish.

Introducción

En el ámbito de la investigación médica los modelos de regresión logística, lineal, de Cox y de Poisson son técnicas estadísticas ampliamente conocidas y utilizadas, ya que permiten evaluar las relaciones entre distintos factores de exposición e indicadores de salud de diversa naturaleza (dicotómicos, continuos, eventos dependientes del tiempo o recuentos)[1,2]. Además, en los estudios observacionales es habitual utilizarlos como herramienta de ajuste debido al potencial sesgo de confusión existente en este tipo de diseño[2,3]. Sin embargo, la mayoría de estos modelos demandan asunciones muy estrictas sobre el ajuste de los datos (linealidad de los predictores, normalidad, homocedasticidad, incolinealidad, etc.), cuyo incumplimiento puede invalidar el modelo y la inferencia realizada[1,4].

En los últimos años, a fin de mejorar la comunicación y transparencia de los trabajos científicos, han surgido distintas guías de recomendaciones sobre cómo reportar los resultados científicos (CONsolidated Standards Of Reporting Trials [CONSORT][5], Statistical Analyses and Methods in the Published Literature [SAMPL][6], Strengthening The Reporting of OBservational studies in Epidemiology [STROBE][7]). La guía STROBE referencia los aspectos metodológicos esenciales que deben reportar los estudios epidemiológicos observacionales[7]. Actualmente, revistas de reconocido prestigio internacional, como *The Lancet* o *The British Medical Journal*, en sus guías de autor recomiendan seguir la guía STROBE e incluso exigen proporcionar una lista de verificación del cumplimiento de sus 22 puntos para poder enviar los trabajos.

El objetivo del presente trabajo es describir la evolución del uso de modelos de regresión estándares (logística, lineal, Cox y Poisson) en estudios observacionales indexados en PubMed y revisar las normas de autor de revistas españolas con el fin de conocer si requieren el cumplimiento de la guía STROBE.

Material y métodos

Se realizó una búsqueda en PubMed dirigida a identificar todos los trabajos originales indexados en el repositorio bibliográfico con un diseño observacional que incluyeran entre sus métodos de análisis los modelos multivariantes diferenciando por tipo de modelo de regresión (logístico, lineal, Cox y Poisson). La búsqueda fue limitada a todos los

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

estudios realizados en la especie humana incluidos desde el 1 de enero de 1970 hasta el 31 de diciembre de 2013, excluyendo ensayos clínicos, editoriales, comentarios o series de casos. Para discriminar entre los tipos de regresión utilizados se añadieron a la búsqueda los términos que hacían referencia a los tipos de análisis multivariante principales: «logistic», «linear», «Cox», «Poisson» y/o «multivariate», «regression», «statistical regression» (tabla 1). Finalmente, y con el fin de evaluar la sensibilidad de la estrategia de búsqueda, se incorporaron como texto libre los siguientes sinónimos de las técnicas multivariantes: «adjusted odds ratio», «adjusted OR», «adjusted relative risk», «adjusted RR», «adjusted hazard ratio» y «adjusted HR».

Se describió la tendencia temporal de la frecuencia del uso de estos modelos de regresión y se calculó el porcentaje de cambio anual (PCA). También se realizó una revisión de las normas de autor de las 39 revistas editadas en España indexadas en PubMed e incluidas en Web Of Science para cuantificar las revistas que recomiendan el uso de la guías STROBE, CONSORT o las recomendaciones del Comité Internacional de Directores de Revistas Médicas, que incluye la guía STROBE, para la elaboración y presentación de los manuscritos[8].

La gestión de datos se realizó mediante el paquete estadístico IBM® SPSS® v22, y para el análisis gráfico se utilizó la librería ggplot2 del paquete estadístico R3.1.2[9].

Resultados

De un total de 2.559.903 artículos observacionales indexados en PubMed, un 9,3% contenía un término relativo al análisis multivariante, pasando del 0,17% en 1970 al 16% en 2013 (PCA 11,15%) (tabla 2). Esta tendencia también se observó en los trabajos en lengua española (PCA 8,56%). El 6,1% de los trabajos reportaron la utilización de modelos de regresión lineal, logística, Cox o Poisson (del 0,14% en 1980 hasta el 12,3% en 2013; PCA 14,5%). Este porcentaje fue inferior para los trabajos en lengua española en todos los años de estudio (fig. 1 y tabla 2). Por otro lado, entre los 238.093 trabajos que contenían el término multivariante, el 48,5% contenía alguno de los modelos estándares estudiados (logístico, lineal, Cox o Poisson), presentando también una tendencia ascendente (del 9,5% en 1980 al 75% en 2013; PCA 6,5%). Se observó una tendencia similar para los artículos en lengua española (datos no mostrados).

El uso del término multivariante aumentó en los artículos indexados en PubMed en los años de estudio de 0,96 a 14,55% (PCA 11,15%). El término relativo al modelo logístico fue el modelo más utilizado durante todo el periodo de estudio (3,39%; PCA 18,47%). Le siguieron los términos referentes al modelo de Cox (1,51%; PCA 20,09%), al modelo lineal (1,31%; PCA 15,37%) y al de Poisson (0,13%; PCA 17,73%) (tabla 2). El último año analizado (2013), un 6,7% contenía términos referidos a regresión logística, un 2,5% a regresión lineal, un 3,49% a regresión de Cox y un 0,31% a regresión de Poisson. Por otro lado, al incluir los sinónimos de las técnicas multivariantes en el texto libre el número de artículos indexados aumentaron en un 0,99%, siendo el aumento del 0,25% en el modelo logístico y del 0,09% en el modelo de Cox.

El 12,8% de todas las revistas indexadas en Web Of Science y PubMed y editadas en España (n = 39) recomendaban explícitamente seguir la guía STROBE antes de enviar el manuscrito, y el 35,9%, seguir la guía CONSORT. El 30,8% recomendaban implícitamente la guía STROBE porque aconsejaban cumplir los requisitos de uniformidad de presentación de manuscritos del Comité Internacional de Directores de Revistas Médicas.

Discusión

Nuestro trabajo refleja un aumento del uso de las técnicas estadísticas multivariantes en los estudios observacionales indexados en PubMed, especialmente los modelos de regresión logística. Por otro lado, las recomendaciones STROBE para mejorar la comunicación de los resultados científicos de los estudios observacionales solo se indican en el 12,8% de las normas de autor de las revistas editadas en España. En contraste, la guía de recomendaciones para los ensayos clínicos (CONSORT) se propone en un 35,9% de las mismas normas de autor.

El aumento observado del uso de las técnicas de regresión en los estudios observacionales puede deberse a la capacidad computacional de los ordenadores actuales y de los paquetes estadísticos para realizar estos análisis. Una revisión de la metodología estadística empleada en artículos publicados en 2 revistas con alto factor de impacto mostró que en un 16% de los artículos revisados se había utilizado metodología multivariante, siendo la regresión logística la más utilizada (10%) entre los años 2000 y 2007[10], resultados consistentes con los de nuestro estudio. Por otro lado, una revisión más reciente de estudios donde la fuente de datos primaria fue la Encuesta Nacional de Salud de Canadá también indicó un predominio de la utilización del modelo logístico y un incremento del uso de las técnicas de regresión a lo largo de los años, como nuestro estudio[11]. Sin embargo, el porcentaje global que observaron de la utilización de técnicas de regresión fue notablemente superior al de nuestro estudio (80 vs. 14,5%). Existen motivos que pueden explicar tal discrepancia, como la diferencia en la estrategia de búsqueda o el tipo de referencias incluidas entre ambos estudios. Además, nuestro estudio englobó un universo de artículos de mayor variedad de estudios, especialidades, diseños, muestras y tipos de revistas.

El aumento del uso de las técnicas multivariantes en los artículos científicos observacionales en nuestro estudio también coincide con la cada vez mayor disponibilidad del software para llevarlas a cabo en los últimos años. Sin embargo, el uso de estas técnicas no está libre de potenciales errores y no siempre es apropiado, pues sobre ellas descasan fuertes asunciones que no siempre se cumplen. El incumplimiento de las asunciones formales de los modelos puede invalidar los resultados que se derivan del estudio, produciendo los conocidos errores de tipo I y/o tipo II[1], o importantes sesgos de las estimaciones[12]. En este sentido, existen herramientas estadísticas que permiten evaluar si los modelos cumplen las condiciones de aplicación[1,4].

La necesidad de evaluar críticamente la calidad metodológica de los estudios ha puesto de manifiesto graves deficiencias en los artículos de investigación. Estas deficiencias dificultan el desarrollo de revisiones sistemáticas, que posteriormente influyen en el desarrollo de guías de práctica clínica y, en última instancia, sobre el cuidado de los

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

Tabla 1 Lista de descriptores utilizados en las distintas estrategias de búsqueda en PubMed

Criterio de búsqueda	Campos PubMed			
Términos o palabras clave	MeSH	ptyp	tiab	text
Estudios observacionales (+)				
Evaluation studies		+		
Evaluation studies as topic	+			
Evaluation study			+	
Evaluation studies			+	
Intervention studies	+			
Intervention study			+	
Intervention studies			+	
Case-control studies	+			
Case-control		+		
Cohort studies	+			
Cohort			+	
Longitudinal studies	+			
Longitudinal			+	
Longitudinally			+	
Prospective			+	
Prospectively			+	
Retrospective studies	+			
Retrospective			+	
Follow up			+	
Comparative study		+		
Comparative study			+	
Observational			+	
Humanos (+)				
Humans	+			
Exclusión de ensayos clínicos, editoriales, comentarios o series de casos (−)				
Editorial/Letter/Comment/Case report		−		
Case report			−	
Case series			−	
Clinical trial		−		
Modelo multivariante				
Multivariate analyses				+
Multivariate/statistical regression				+
Regression/regressions				+
Modelo logístico				
Logistic model/s				+
Logistic regression				+
Logistic regressions				+
Regresión lineal				
Linear model/s				+
Linear regression				+
Linear regressions				+
Modelo de regresión de Cox				
Cox regression				+
Cox regressions				+
Cox models				+
Regresión de Poisson				
Poisson model/s				+
Poisson regression				+
Poisson regressions				+

MeSH: términos MeSH; ptyp: tipo de publicación; text: texto completo; tiab: título o resumen; +: término incluido; −: término excluido; (+): incluido en todas las búsquedas; (−): excluido en todas las búsquedas.
Fecha de publicación entre el 1 de enero de 1970 hasta el 31 de diciembre de 2013. Estrategia lanzada en PubMed el 16 de abril de 2014.

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

Tabla 2 Evolución porcentual de la utilización de los modelos de regresión logística, lineal, Cox y Poisson en los artículos observacionales indexados en PubMed a nivel global y en lengua española (1970-2013)

Periodo	Número de citas	Logístico, %	Lineal, %	Cox, %	Poisson, %	Término «multivariante», %
Global						
1970-1979	134.839	0,01	0,05	0,00	0,00	0,96
1980-1989	300.325	0,32	0,23	0,09	0,01	2,75
1990-1999	574.794	2,11	1,01	0,84	0,07	7,16
2000-2009	1.013.344	4,07	1,44	1,78	0,15	10,78
2010-2013	536.601	6,02	2,30	2,92	0,27	14,55
1970-2013	2.559.903	3,39	1,31	1,51	0,13	9,30
PCA		*18,47*	*15,37*	*20,09*	*17,73*	*11,15*
Lengua española						
1970-1979	3.132	0,00	0,03	0,00	0,00	0,13
1980-1989	4.950	0,08	0,02	0,00	0,00	0,83
1990-1999	10.809	2,13	0,84	0,33	0,13	5,55
2000-2009	14.086	3,59	0,69	0,85	0,19	9,41
2010-2013	4.937	4,15	0,85	1,28	0,14	10,51
1970-2013	37.914	2,49	0,61	0,58	0,13	6,57
PCA		*13,27*	*7,42*	*4,70*	*2,72*	*8,56*

PCA: porcentaje de cambio anual de todo el periodo respecto al primer año con porcentaje superior a 0.
Número de citas de estudios observacionales según la estrategia establecida en PubMed el 16/04/2014.

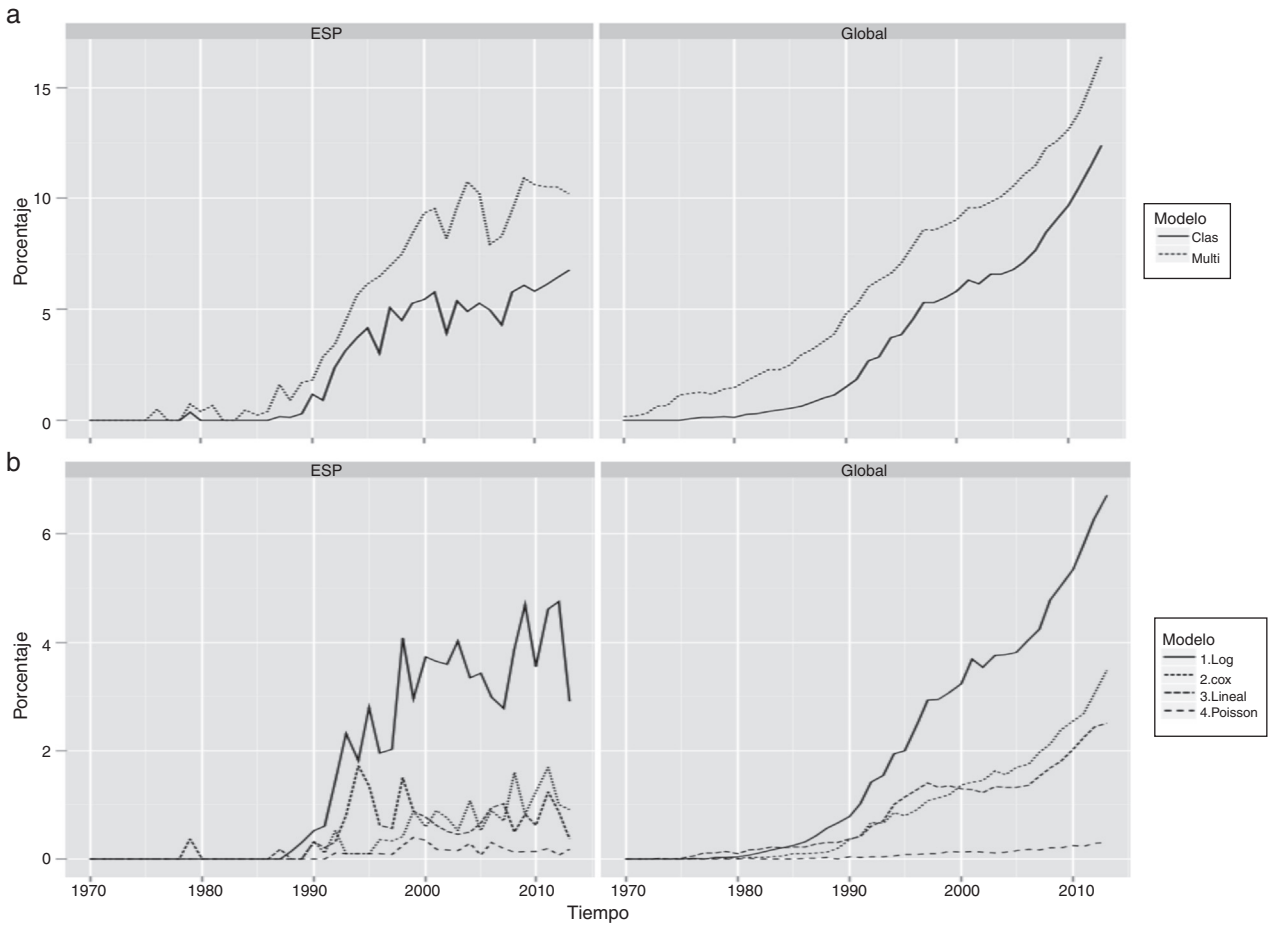


Figura 1 Evolución del porcentaje de los términos referentes a modelos estadísticos en artículos observacionales indexados en PubMed en lengua española (ESP) y a nivel internacional (Global): (a) término multivariante (Multi) respecto a término modelos clásicos (Clas: Logístico, Cox, Lineal o Poisson) y (b) términos específicos relativos a cada modelo.

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

pacientes. En ese sentido, las guías de recomendaciones son herramientas desarrolladas para facilitar información más exacta y completa de los aspectos clave de los estudios de investigación[13]. De hecho, un estudio realizado en España demostró que su requerimiento mejoraba la calidad de los manuscritos publicados[14]. En este sentido, que los editores y revisores dispongan y requieran guías o herramientas estandarizadas para evaluar la calidad metodológica es clave para mejorar la presentación de los trabajos científicos.

El control de la confusión es uno de los aspectos esenciales incluidos en la guía STROBE[7]. Sin embargo, no incluye explícitamente la validación del método estadístico. Otras guías, como la SAMPL[6], de carácter más metodológico, sí que incluyen como recomendación la validación de la metodología estadística empleada, aunque también es incompleta en lo que se refiere a metodología multivariante.

Una de las limitaciones de esta revisión deriva en que la fuente analítica principal se basa en el motor de búsqueda de PubMed. Otra limitación del estudio es la estrategia de búsqueda utilizada, ya que los resultados de la búsqueda dependen de que los autores hayan mencionado las técnicas de regresión en el resumen. En este sentido, nuestros resultados globales podrían estar infraestimados. Por otro lado, también puede haber varios artículos derivados del mismo proyecto, que puede ocasionar cierta sobreestimación del resultado. Este tipo de limitación es común en los estudios bibliométricos basados en motores de búsqueda por palabras como PubMed[10]. Sin embargo, teniendo en cuenta que a partir del año 1990 todos los descriptores ya estaban incorporados en PubMed y que nos hemos basado en todo su repositorio, en términos de evolución, el efecto de esta limitación probablemente sea reducido.

Por otro lado, en relación con el requerimiento sobre el cumplimiento de STROBE, se han revisado todas las revistas españolas que en la última actualización del año 2012 estaban indexadas en la Web Of Science y PubMed. De todos modos, este análisis no está directamente vinculado con el estudio bibliométrico, y no tiene una referencia temporal, ya que la guía STROBE se publica por primera vez en 2007, más tarde que la CONSORT, por lo que las revistas han tenido más tiempo en adaptar las normas de publicación a esta última.

En conclusión, los modelos de regresión multivariantes (logística, lineal, Cox y Poisson) en estudios observacionales publicados e indexados en PubMed son cada vez más utilizados tanto a nivel global como en revistas publicadas en lengua española. Debido al aumento de la utilización de los métodos multivariantes parece necesario establecer filtros que garanticen el correcto uso de estos métodos. Además, un porcentaje bajo de las revistas científicas españolas indexadas en PubMed incluyen en las normas de autoría el requerimiento de la guía STROBE.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que han seguido los protocolos de su centro de trabajo sobre la publicación de datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación

Los autores no han recibido financiación específica para realizar este estudio.

Autoría

JMMS concibió el trabajo. JR realizó todos los análisis. JR y JMMS escribieron el primer borrador del manuscrito y todos los autores contribuyeron significativamente en sus versiones posteriores. Todos los autores han aprobado la versión final del manuscrito.

Conflicto de intereses

Los autores declaran no tener conflictos de intereses.

Bibliografía

1. Dobson AJ. An introduction to generalized linear models. 2nd ed. United States of America: Chapman and Hall; 2001.
2. Bender R. Introduction to the use of regression models in epidemiology. En: Mukesh V, editor. Methods in Molecular Biology, Cancer Epidemiology. Totowa, NJ: Springer Science; 2009. p. 179–95.
3. Szklo M, Nieto FJ. Epidemiología intermedia. Conceptos y aplicaciones. Madrid: Díaz de Santos; 2003.
4. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med. 1995;14:1707–23.
5. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. Ann Intern Med. 2010;152:726–32.
6. Lang TA, Altman DG. Statistical analyses and methods in the published literature: The SAMPL Guidelines. John Wiley & Sons; 2014.
7. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. Gac Sanit. 2009;23:158, e1–28.
8. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication. ICMJE; 2011 [consultado 15 Jun 2014]. Disponible en: www.icmje.org
9. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer; 2009.
10. Scotch M, Duggal M, Brandt C, Lin Z, Shiffman R. Use of statistical analysis in the biomedical informatics literature. J Am Med Inform Assoc. 2010;17:3–5.
11. Yergens DW, Dutton DJ, Patten SB. An overview of the statistical methods reported by studies using the Canadian community health survey. BMC Med Res Methodol. 2014;14:15.
12. Liang W, Zhao Y, Lee AH. An investigation of the significance of residual confounding effect. Biomed Res Int. 2014;2014:658056.

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

13. Altman D, Hoey J, Marušić A, Moher D, Schulz KF. EQUATOR Network. Enhancing the QUAlity and Transparency Of health Research. 2014 [consultado 5 Nov 2014]. Disponible en: http://www.equator-network.org

14. Cobo E, Cortés J, Ribera JM, Cardellach F, Selva-O'Callaghan A, Kostov B, et al. Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: Masked randomised trial. BMJ. 2011;343:d6783.

Cómo citar este artículo: Real J, et al. Utilización de los modelos de regresión múltiple en estudios observacionales (1970-2013) y requerimiento de la guía STROBE en revistas científicas españolas. Semergen. 2015. http://dx.doi.org/10.1016/j.semerg.2015.06.020

6.2 Estudio REVISIÓN: Trabajo publicado en Medicine (Baltimore)

Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine (Baltimore). 2016 May;95(20):e3653

Medicine®

OBSERVATIONAL STUDY

OPEN

Quality Reporting of Multivariable Regression Models in Observational Studies

Review of a Representative Sample of Articles Published in Biomedical Journals

Jordi Real, BSc, Carles Forné, MSc, Albert Roso-Llorach, MSc, and Jose M. Martínez-Sánchez, PhD

Abstract: Controlling for confounders is a crucial step in analytical observational studies, and multivariable models are widely used as statistical adjustment techniques. However, the validation of the assumptions of the multivariable regression models (MRMs) should be made clear in scientific reporting. The objective of this study is to review the quality of statistical reporting of the most commonly used MRMs (logistic, linear, and Cox regression) that were applied in analytical observational studies published between 2003 and 2014 by journals indexed in MEDLINE.

Review of a representative sample of articles indexed in MEDLINE (n = 428) with observational design and use of MRMs (logistic, linear, and Cox regression). We assessed the quality of reporting about: model assumptions and goodness-of-fit, interactions, sensitivity analysis, crude and adjusted effect estimate, and specification of more than 1 adjusted model.

The tests of underlying assumptions or goodness-of-fit of the MRMs used were described in 26.2% (95% CI: 22.0–30.3) of the articles and 18.5% (95% CI: 14.8–22.1) reported the interaction analysis. Reporting of all items assessed was higher in articles published in journals with a higher impact factor.

A low percentage of articles indexed in MEDLINE that used multivariable techniques provided information demonstrating rigorous application of the model selected as an adjustment method. Given the importance of these methods to the final results and conclusions of observational studies, greater rigor is required in reporting the use of MRMs in the scientific literature.

(*Medicine* 95(20):e3653)

Abbreviations: CI = confidence interval, MRM = multivariable regression model.

Editor: Zhiyong Liu.
Received: October 8, 2015; revised: April 7, 2016; accepted: April 18, 2016.
From the Unitat de Suport a la Recerca-Lleida, Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona (JR); Universitat International de Catalunya, Facultat de Medicina i Ciències de la Salut, Sant Cugat (JR, JMM-S); Department of Basic Medical Sciences, Universitat de Lleida, Lleida (CF); Oblikue Consulting (CF); Institut Universitari d'Investigació en Atenció Primària Jordi Gol, Barcelona (AR-L); and Tobacco Control Unit, Catalan Istitute of Oncology, Hospitalet de Llobregat (JMM-S), Spain.
Correspondence: Dr. Jose M. Martínez-Sánchez, Departament de Ciències Bàsiques, Universitat Internacional de Catalunya, Carrer de Josep Trueta s/n, 08195 Sant Cugat del Vallès, Barcelona, Spain (e-mail: jmmartinez@uic.es).
The authors have no conflicts of interest to disclose.
Supplemental Digital Content is available for this article.

ISSN: 0025-7974
DOI: 10.1097/MD.0000000000003653

INTRODUCTION

Two important aspects of biomedical research are the internal and external validity of the study design.[1] Information bias and confounding variables affect internal validity and are present to some extent in all observational research. Information bias results from incorrect determination of the exposure, outcome, or both. Confounding is a ''mixture'' or ''diffusion'' of effects: a researcher attempts to associate an exposure with a result, but actually measures the effect of a third—sometimes unnoticed—factor, that is, a confounding variable. This bias can be diminished, but only if the confounding factor is anticipated and the relevant data are collected to allow proper adjustment.

Confounding factors can be controlled in various ways: restriction, matching, stratification, standardization, and multivariable techniques. All of these approaches are focused on achieving homogeneity between study groups,[1] and in recent years multivariable regression models (MRMs) such as linear, logistic, Poisson, or Cox regression have become popular and very frequently used.[2] A review of research based on Canada National Health Survey data found that nearly 80% of the studies used some type of MRM, predominantly logistical modeling.[3] A systematic review of studies published by 10 prestigious journals in epidemiology and general medicine showed that almost 95% used MRMs, in addition to other techniques, as the adjustment methodology.[4] The frequency with which any statistical method is applied is often determined by the available software and computational capacity; therefore, this high rate of MRM usage could be due to major advances in computational capabilities with increased availability of data, but also to the ease with which these techniques can now be applied using standard statistical software.[5]

An advantage of MRM analysis is that it allows the control of more confounding factors, compared to stratification, and a simultaneous evaluation of the relationship between several exposure factors and response variables of different types (continuous, dichotomous, count, or time-dependent events).[2,6] The estimated effect of each variable reflects its association with the outcome, taking into account the contribution of the rest of the variables introduced into the model. However, modification effect is not identifiable by simple inclusion of the variable in the regression model; the interaction terms between exposure and effect modification, or the confounding variable, must also be included.

Moreover, MRMs assume probability distributions that include underlying assumptions (e.g., assumptions of normality, homoscedasticity, independence of errors, etc.). In addition, parameter estimation could be inefficient if there is multicollinearity between 2 or more variables, which affects convergence in the inference process, among other potential problems.[5,6]

Regression models produce nonbiased results for each variable of interest if the model is correctly specified and all potential confounding factors are included and correctly measured.[7] Furthermore, if not all confounders are included or the model is not properly specified, the consequences are residual confounding and biased estimates.[8,9] Although the underlying ''true'' model is seldom known, specification errors and residual confounding can be minimized by testing the formal assumptions of the selected model.[6,10] Specific statistical tools are available to evaluate whether all necessary conditions have been met to apply a particular type of adjusted modeling and the appropriateness of the model that was finally selected.[6,10] In addition, given that MRMs are usually sensitive to model specification, it is desirable to carry out more than 1 adjustment strategy to evaluate the stability of the estimated effects of different settings.[11] All these measures, together with a sensitivity analysis (variation by subgroups) and interaction assessment, lead to more consistency in evaluating the adjusted measures of association and increase their validity and level of evidence.[12,13]

Various studies have described the statistical methodology used in published biomedical research.[3,4,14,15] Strasak et al[15] showed that inappropriate use of some statistical tests is one of the most common errors. In 2008, Groenwold et al[4] carried out a systematic review of observational studies published in general medical and epidemiology journals with a high impact factor and reported finding poor quality in the adjustment methods used. More recently, in 2014, another systematic review of the use and application of generalized linear mixed models showed their increased use and, at the same time, room for improvement in reporting quality.[14] However, there is a lack of evidence on the quality of reporting or the validation procedures used when MRMs are applied in observational studies. Therefore, the objective of the present study was to review the quality of statistical reporting when the most commonly used MRMs (logistic, linear, and Cox regression) were applied in analytical observational studies published between 2003 and 2014 by journals indexed in MEDLINE.

MATERIALS AND METHODS

We reviewed a representative random sample of articles indexed by MEDLINE using the PubMed search engine. The search was specifically designed to identify original studies with an analytical observational design that stated their use of logistic, Cox, or linear MRMs focused on confirmatory analysis (i.e., to assess the effect of exposure) (Supplementary Table S1, http://links.lww.com/MD/A979 of the Appendix). The search was limited to studies in humans that were published in English between January 1, 2003 and February 16, 2014. Clinical trials, editorials, commentaries, and case reports were excluded. This strategy retrieved 71,519 references, from which a simple random sample of 500 articles was selected. A sample size of 500 randomly selected papers was calculated to allow estimation with 95% confidence and a precision of ±5% units, a population percentage considered to be of 50%. We assumed a 50% prevalence to maximize the sample size. A replacement rate of 20% was anticipated. Exclusion criteria removed 72 references, including those that proved to be focused on diagnosis, prognosis, or other analytical approaches. Therefore, 428 papers were finally reviewed (Figure 1).

Items Reviewed in Full-Text Analysis

Based on the literature,[2,11,16,17] a list of aspects related to the application of MRMs was specified, including testing formal assumptions, goodness of fit: interactions, and sensitivity analysis of the adjustment models (Table 1). An initial review of 10 articles served as a pilot test for the entire research team to define the list of items to be included, establish precise definitions, and improve interrater homogeneity. Finally, the definitive set of MRM-related items to be verified in each relevant section of the manuscript was established (Table 1). Each item was classified according to whether it would likely be stated in the methods section or was mainly involved in the communication of findings and would appear in the results section. If an item was reported in any section, the paper was considered to meet the criteria.

Review Procedure

The selected articles were randomly distributed among the 3 designated reviewers on the research team. Any doubts were shared and resolved by consensus. In addition, 12 articles were randomly selected for reviewing by all 3 reviewers, for blinded evaluation of interrater agreement. No significant differences in outcomes between reviewers were observed (Supplementary Table S2, http://links.lww.com/MD/A979); there was high interrater agreement (Kappa index > 0.73) and intraclass correlation coefficient of the number of completed items (0.88). The Kappa index measures the agreement between reviewers for compliance with each of the items separately (dummy variables) and the intraclass correlation coefficient quantifies the correlation of the number of completed items (numerical variable) between reviewers. A detailed analysis of intra- and interrater agreement is shown in the Appendix, Supplementary Table S2, http://links.lww.com/MD/A979.

Statistical Analysis

For each item specified for review, prevalence estimates and 95% confidence interval (CI) were obtained. We also calculated mean and standard deviation (SD) for the total number of review items fulfilled. CIs were computed using normal approximation. All analyses were stratified in groups according to the impact factor of the journal in the year of publication ($\leq$2, 2–4, $>$4), sample size ($<$500, 500–1500, $\geq$1500), design (cross-sectional, cohort, and case–control), data source (ad hoc, clinical/administrative records, both, or ''mixed''), and type of MRM (logistic, linear, and Cox). Pearson χ^2 and trend tests were used to assess the association between prevalence of the items of interest and the categorical secondary variables. Mann–Whitney *U* test was used to examine the relationship between prevalence of items, sample size, and the journal's impact factor. We computed the 2-sided criteria for all variables and 1-sided criteria for the impact factor level because the higher the impact factor, greater rigor is required in reporting the use of MRMs in the scientific literature. To assess and control for possible interactions, the analysis was again stratified by impact factor, sample size, design, and type of modeling. Significance level was set at $\alpha = 0.05$. All analysis was carried out using the SPSS statistical software, version 18.0. (PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.).

Ethical statement: None required the approval of the Ethics Committee because the primary source was secondary data from published scientific articles.

RESULTS

Of 428 articles reviewed, published in 313 journals (mean of impact factor = 3.38), 49.5% were cohort studies, with data primarily collected using questionnaires specifically designed

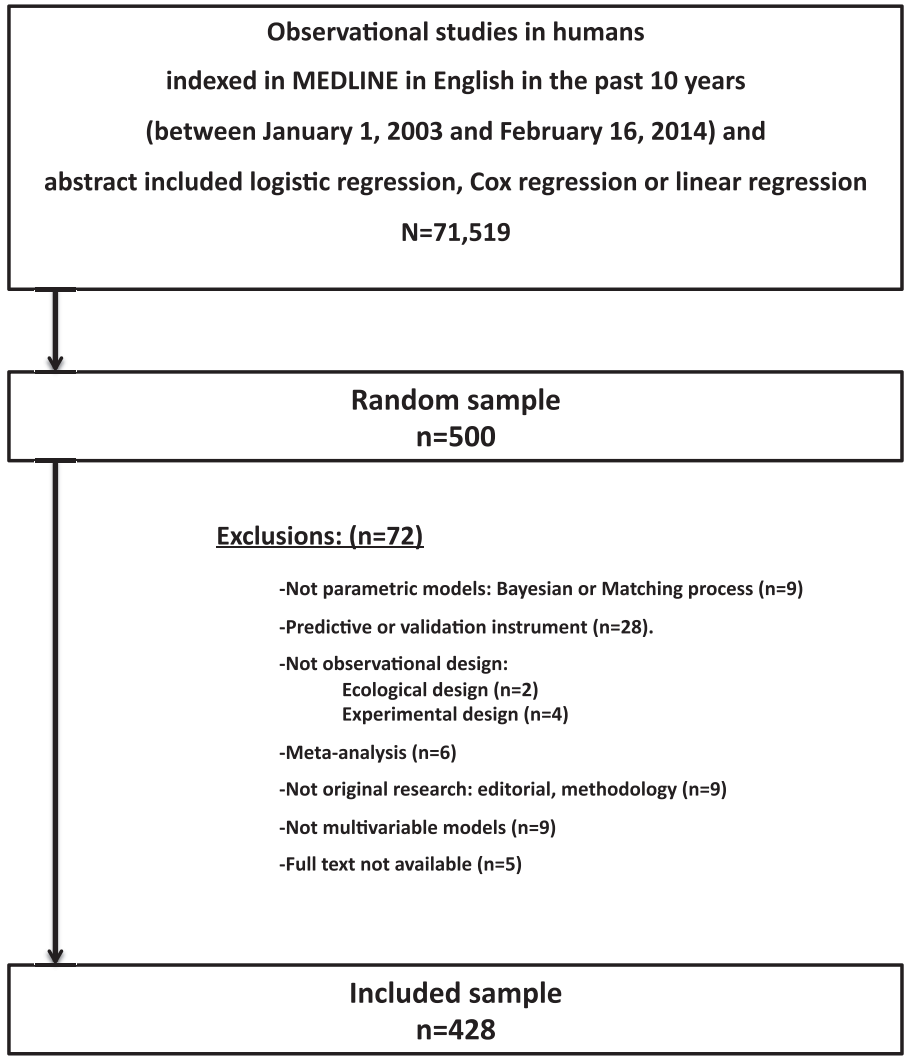


FIGURE 1. Flowchart of articles included.

to address the research objective (45.6%). The most frequently used type of modeling was logistic regression (67.5%), followed by Cox (22.9%) and linear (18%) regression. Nearly half (48.8%) of the articles reviewed were published during the last 3 years of our study period (2010–2013). Only 4% of the articles referenced any other publication that expanded on the methodology used in the study.

Table 2 shows the overall percentage observed for each of the items reviewed, and in relation to all selected variables. The major item that was reported most often (33.4%; 95% CI: 28.9–37.8%) was ''crude and adjusted effect'' (item 4, Table 2), followed by sensitivity analysis (32.7%; 95% CI: 28.3–37.1%). The least-reported item was interaction analysis (18.5%, 95% CI: 14.8%–22.1%). Testing the assumptions of the model and fitting more than 1 model were reported in 26.2% and 25.7% of the articles, respectively (items 1 and 5, Table 2).

The percentages observed for all of the items analyzed were higher in studies published in journals with a moderate or high impact factor (Table 2). The assessment of model adjustment criteria (item 1) was primarily observed in articles published in journals with a moderate impact factor and in studies that used linear models. The criteria referring to interactions, sensitivity analysis, and testing more than 1 model (items 2, 3, and 5, respectively) were also significantly and directly associated with sample size (Table 2).

The mean number of items identified in the articles reviewed was 1.36 (SD = 1.17), and increased with sample size and impact factor ($P < 0.001$). Both factors act independently of the mean number of items: there was no observed interaction between impact factor and sample size. Figure 2 shows how the frequency of each item increased with impact factor (Figure 2A), independently of sample size (Figure 2B), study design (Figure 2C), and type of MRM used (Figure 2D).

DISCUSSION AND CONCLUSIONS

Our study shows very low reporting of MRM validation in observational studies indexed in MEDLINE, being higher in studies with larger sample sizes published in journals with a higher impact factor. Only 26.2% of the articles reviewed described their validation analysis of assumptions or goodness-of-fit for the MRM used, 33.4% showed both the crude

TABLE 1. Primary Items Reviewed in Manuscripts of Observational Studies That Used Multivariable Methods (Logistic Regression, Cox Regression, or Linear Regression)

Item	Issues Reviewed in the Manuscript (Yes/No) Detail and Justify the Item Reviewed	Method Section	Results Section
1	Model assumptions and goodness-of-fit*: Normality, linearity/log-linearity, homoscedasticity, proportional hazards assumption (Cox models) or goodness-of-fit. Is the functional form of the selected model correct? How far away from the data is the selected model?	x	X
2	Interaction analysis: Some interaction term was evaluated in the models. Is there any potential variable that can modify the estimated effect?	x	x
3	Sensitivity analysis: Sensitivity analysis of the models was performed with subsamples Are the findings sufficiently robust, considering the process used to obtain them?	x	x
4	Crude and adjusted effect estimate: Report of crude measures of association in addition to those adjusted according to the model used (Odds ratio, hazards ratio, etc.). How much does the studied effect change when other variables are taken into account?		x
5	More than one adjusted model specified: For each response variable, more than 1 adjusted model with different combinations of variables was shown. Does the estimated effect differ between the different adjusted models, settings, specifications, etc.?		x

*Specific statistical methods: Kolmogorov–Smirnov about residuals, Q–Q plots; Hosmer–Lemeshow test for logistic regression, Schoenfeld residuals for Cox Regression, R^2, receiver operating characteristic (ROC) curve.

and adjusted effects, and 32.7% described any sensitivity analysis. Interaction analysis was only observed in 18.5% of the articles reviewed.

Our results are consistent with previous scientific evidence.[4,14,18] Müllner et al[18] showed that journals with a higher impact factor had better statistical reporting, perhaps because their editorial process specifically includes statistical review. In our study, the percentages observed for all of the items analyzed were higher in studies published in journals with a higher impact factor. A systematic review by Casals et al[14] of 108 articles that applied generalized linear mixed models, without discriminating between type of design or research objective, found that validation of the model and testing for goodness-of-fit were reported in 6.5% and 15.7%, respectively, of the articles. In contrast, our results showed a higher prevalence of this item (17.7%–33.7%, depending on the impact factor of the journal). This difference could be explained because our review is based on a random sample that included methodologies whose use is much more widespread.[19] Another systematic review found a lack of attention to adjustment methods in analytical observational studies,[4] in contrast with diagnostic, prognostic, or predictive validation studies in which combinations of variables were modeled with greater precision.[16,20–23] In the latter types of studies, calibration, discriminatory power, goodness-of-fit, and validation of the statistical model are considered essential 1st steps before selecting the final adjusted model.[16] The recent Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement[24] provides guidelines that highlight the essential aspects of developing and validating a predictive multivariable model. Requirements of this guide include, among others, the need for using internal validation methods to evaluate model's performance and to compare multiple models.[24]

On the other hand, published guidelines provide specific recommendations on the reporting of the scientific results of clinical trials (CONSORT),[25] observational studies (STROBE),[26] or statistical analysis in general scientific literature (SAMPL).[27] These guidelines were developed to provide more complete and precise information about key aspects of research studies, and some have been incorporated into the author guidelines of major scientific journals. Nonetheless, even though the STROBE guidelines highlight the control of confounders as a crucial aspect of observational studies and the SAMPL and TRIPOD guidelines broaden the standards for the scrutiny of statistical methods, there is still a void in requiring or assessing multivariable methodology in observational designs. Notably, the Guide for Authors and Editors (Manual of Style for the American Medical Association) includes the need to report model diagnostics and proportion of variance explained by both individual variables and the complete model.[28] In this sense, even though the data analysis may be correct, inadequate reporting makes it impossible for the reader to assess whether the data were processed appropriately.[18]

In observational research, best practice includes avoiding bias in the study design, adjusting for possible bias in the data analysis if it is not possible to avoid bias entirely in the design, and quantifying and analyzing the effects of residual bias on the study results.[7] Nonetheless, if the model was not properly selected, there may be major residual confounding even after MRM adjustment,[29–31] which leads to bias in the associations studied. For example, Liang et al[8] recently published the results of a simulation study, concluding that ''even when all confounding factors are known and controlled for using conventional multivariable analysis, the observed association between exposure and outcome can still be dominated by residual confounding effects.''

TABLE 2. Frequency of Items Related to the Application of Statistical Models, Based on Study Characteristics in Articles Reviewed

			Percentage					Reporting at Least Item
Variable	**Category**	**N**	**it1**	**it2**	**it3**	**it4**	**it5**	**≥1**
Overall		428	26.2%	18.5%	32.7%	33.4%	25.7%	71.5%
	95% CI		22.0–30.3	14.8–22.1	28.3–37.1	28.9–37.8	21.6–29.8	67.2–75.8
Design*								
	Cross-sectional	108	31.5%	12.0%	33.3%	30.6%	17.6%	68.5%
	Cohort	212	26.9%	20.8%	32.1%	36.3%	26.4%	72.6%
	Case–control	72	16.7%	25.0%	43.1%	23.6%	37.5%	75.0%
	P-value†		0.083	0.065	0.229	0.123	0.011	0.603
Data source*								
	Ad hoc	195	27.2%	16.4%	31.8%	33.8%	25.1%	69.2%
	Clinical record	106	24.5%	16.0%	30.2%	32.1%	20.8%	69.8%
	Mixed	114	25.4%	26.3%	37.7%	34.2%	31.6%	78.9%
	P-value†		0.870	0.067	0.437	0.936	0.179	0.155
Sample size								
	<=500	205	26.8%	12.2%	21.5%	30.2%	18.5%	62.0%
	501–1500	90	25.6%	21.1%	35.6%	43.3%	31.1%	75.6%
	1501+	133	25.6%	26.3%	48.1%	31.6%	33.1%	83.5%
	P-value†		0.956	0.004	<0.001	0.078	0.005	<0.001
	P-value‡		0.786	0.001	<0.001	0.634	0.002	<0.001
	P-value§		0.780	<0.001	<0.001	0.543	0.002	<0.001
Impact factor of journal								
	<=2.00	147	17.7%	8.8%	17.7%	29.3%	17.7%	55.1%
	2.01–4.00	166	33.7%	18.1%	34.3%	34.3%	25.9%	76.5%
	4.01+	115	26.1%	31.3%	49.6%	37.4%	35.7%	85.2%
	P-value†,‖		0.003	<0.001	<0.001	0.185	0.002	<0.001
	P-value‡,‖		0.040	<0.001	<0.001	0.0.08	<0.001	<0.001
	P-value§,‖		0.033	<0.001	0.001	0.229	<0.001	<0.001
Model¶								
	Logistic	289	14.9%	14.2%	26.6%	28.7%	22.1%	67.8%
	Linear	77	39.0%	14.3%	32.5%	22.1%	18.2%	87.0%
	Cox	98	29.6%	19.4%	22.4%	32.7%	20.4%	72.4%
	P-value†		<0.001	0.343	0.081	0.462	0.987	0.006

Description of items 1–5: it1 = model assumptions and fit, it2 = interaction analysis, it3 = sensitivity analysis, it4 = crude and adjusted effect estimates, it5 = more than one adjusted model. CI = confidence interval.

*The category with undefined information is not included (36 design manuscripts and 13 of data source).

†*P*-value computed with χ^2 Pearson test.

‡*P*-value computed with χ^2-trend test.

§*P*-value computed with Mann–Whitney *U* test.

‖*P*-value computed with unilateral test.

¶Statistical test excluded 35 manuscripts with multiple models.

In this sense, overall goodness-of-fit, along with graphical validation analysis, can allow researchers to evaluate possible conflicts between the models and alert them to possible specification problems, even without ensuring that the model is completely correct.[32] Properly fitting the model may require additional adjustment variables, their transformation, inclusion of interactions, or the choice of other adjustment techniques that are less sensitive to the selection of a particular model, such as stratified analysis, matching techniques,[33] or other flexible modeling approaches.[34]

Strengths and Limitations

One of the limitations of this analysis is that the primary source of data was the abstracts accessed in MEDLINE by the PubMed search engine. Therefore, the universe of potential studies for analysis was limited to that repository and the sensitivity of the search strategy used. In an effort to minimize missed records, we designed a highly specific search-term strategy. During our review we only had to discard 14.4% of the manuscripts for failing to meet at least 1 inclusion criterion.

Another limitation was that the quality or transparency of the methodological reporting could be affected by the word limitations imposed by a journal's guidelines. Nonetheless, it is now usually possible to complement an article with online supplementary information or to disseminate the methodological details and protocols in a separate manuscript that provides greater detail about the more technical aspects. However, only 4% of the manuscripts included in the present review contained any reference to a separate article detailing the methodology used.

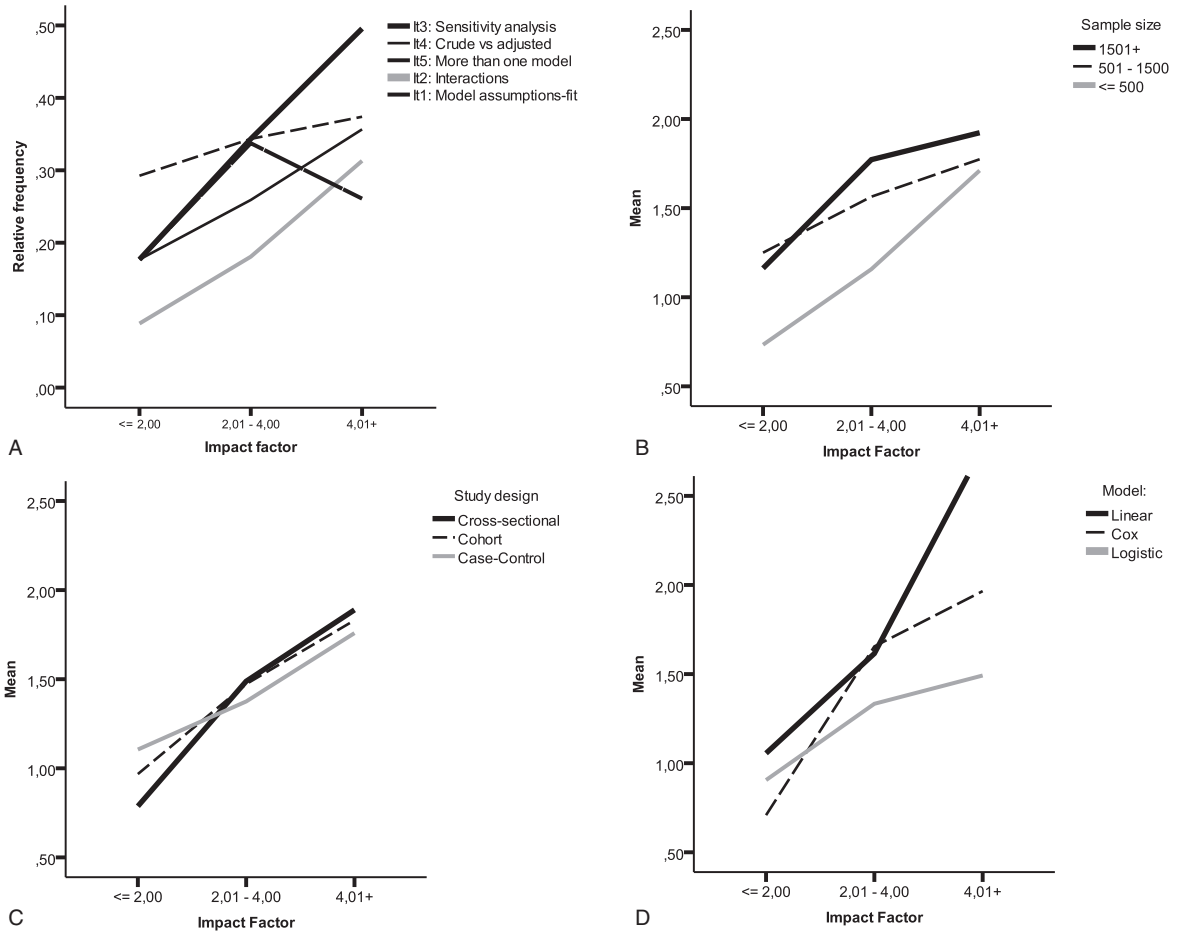


FIGURE 2. Relative frequency for each item searched (A), mean number of items per article (i.e., application of a multivariable regression model, stratified by impact factor and by sample size) (B), study design, (C) and type of model used (D).

We are aware that the items we reviewed need not have the same relevance and weight—and are not even always necessary. Assessing their interactions is not always justified, especially in small samples, and there are studies on medical interventions in which confounding could be considered negligible. Furthermore, there are other important aspects that would affect the quality of the analysis and results (e.g., the model was prespecified prior to undertaking the data analysis, or the research team had insufficient statistical background and knowledge).

Finally, our review was not paired and there could be a certain interrater variability. We attempted to minimize this potential limitation in 2 ways: very detailed specification of each item to be reviewed, all of which were easily identifiable; and prior training of reviewers with a pilot test. In addition, testing for agreement after completing the review showed a high level of intra- and interrater agreement (Supplementary Table S2, http://links.lww.com/MD/A979).

CONCLUSIONS

Statistical adjustment using MRM is a powerful tool for isolating the actual effect of exposure factors on potential confounders. However, the use of these techniques is not free of potential errors because they have strong underlying assumptions that must be tested. Our study showed that, despite the availability of known statistical tools that allow the evaluation of how well the models meet the conditions for their application, only a troublingly low percentage of published articles report information about model validation or measures to ensure the rigorous application of MRMs as an adjustment method. Given the importance of these statistical methods to the final conclusions, biomedical journals should require greater rigor in reporting the assumptions of the MRMs in the methods and results of observational studies.

ACKNOWLEDGMENTS

The authors thank Lluís Alvarez for his role in the management of full manuscripts; Gisela Galindo, and Inés Cruz for their valuable advice on the latest versions; and Elaine Lilly for English editing, supported by IDIAP-Jordi Gol.

REFERENCES

1. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet.* 2002;359:248–252.
2. Bender R. Introduction to the use of regression models in epidemiology. In: Mukesh Verma, ed. *Methods in Molecular Biology, Cancer Epidemiology. Vol. 471.* United States: Springer Science; 2009:179–195.

3. Yergens DW, Dutton DJ, Patten SB. An overview of the statistical methods reported by studies using the Canadian community health survey. *BMC Med Res Methodol.* 2014;14:1.
4. Groenwold RH, Van Deursen AM, Hoes AW, et al. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. *Ann Epidemiol.* 2008;18:746–751.
5. Gentle JE, Härdle WK, Mori Y. How computational statistics became the backbone of modern data science. In: Handbook of Computational Statistics. Berlin Heidelberg: Springer; 2012:3–16.
6. Dobson AJ. An Introduction to Generalized Linear Models 2nd ed. United States of America: Chapman and Hall; 2001.
7. Gerhard T. Bias: considerations for research practice. *Am J Health Syst Pharm.* 2008;65:2159–2168.
8. Liang W, Zhao Y, Lee AH. An investigation of the significance of residual confounding effect. *Biomed Res Int.* 2014;2014:658056.
9. Groenwold RH, Klungel OH, Altman DG, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. *CMAJ.* 2013;185:401–406.
10. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat Med.* 1995;14:1707–1723.
11. Vittinghoff E, Shiboski S, McCulloch CE. Regression Methods in Biostatistics.Springer; 2005.
12. Oxman AD. GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ.* 2004;328:1490–1494.
13. Wu R, Glen P, Ramsay T, et al. Reporting quality of statistical methods in surgical observational studies: protocol for systematic review. *Syst Rev.* 2014;3:70.
14. Casals M, Girabent-Farres M, Carrasco JL. Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): a systematic review. *PLoS One.* 2014;9:e112653.
15. Strasak AM, Zaman Q, Pfeiffer KP, et al. Statistical errors in medical research-a review of common pitfalls. *Swiss Med Wkly.* 2007;137 (3/4):44.
16. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ.* 2009;338:b605.
17. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* John Wiley & Sons; 2004. http://books.google.es/books?id=Po0RLQ7USIMC. Accessed June 4, 2015.
18. Müllner M, Matthews H, Altman D. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Ann Intern Med.* 2002;136:122–126.
19. Real J, Cleries R, Forne C, et al. Use of multiple regression models in observational studies (1970–2013) and requirements of the STROBE guidelines in Spanish scientific journals. *Semergen.* 2015;11:S1138–S3593.
20. Lee YH, Hsu CY, Hsia CY, et al. A prognostic model for patients with hepatocellular carcinoma within the Milan criteria undergoing non-transplant therapies, based on 1106 patients. *Aliment Pharmacol Ther.* 2012;36:551–559.
21. Chen S, Huang L, Liu Y, et al. The predictive and prognostic significance of pre- and post-treatment topoisomerase IIalpha in anthracycline-based neoadjuvant chemotherapy for local advanced breast cancer. *Eur J Surg Oncol.* 2013;39:619–626.
22. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ.* 2009;338:b606.
23. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925–1931.
24. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ.* 2015;350:g7594.
25. Schulz KF, Altman DG, Moher D. CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med.* 2010;152:726–732.
26. Noah N. The STROBE initiative: STrengthening the reporting of OBservational studies in epidemiology (STROBE). *Epidemiol Infect.* 2008;136:865.
27. Lang T, Altman D. Statistical analyses and methods in the published literature: The SAMPL guidelines. *Guidelines for Reporting Health Research.* 2014:264–274.
28. Evans R. AMA manual of style-A guide for authors and editors. *Nurs Stand.* 2007;21:31–131.
29. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol.* 2007;166:646–655.
30. Ho KM. Residual confounding in observational studies. *Anesthesiology.* 2009;110:430author reply 430.
31. Sainani K. The limitations of statistical adjustment. *PM R.* 2011;3:868–872.
32. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Stat Med.* 2013;32:67–80.
33. Ho DE, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal.* 2007;15:199–236.
34. Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. *Stat Med.* 2004;23:3781–3801.

Annex

TABLE S1. PubMed Search Strategy and List of Descriptors Used to Select Articles Published in English Between January 1, 2003 and February 16, 2014*

Search criteria in PubMed (n=71 519)	Fields in PubMed			
Term or keyword used	MeSH	ptyp	tiab	text
Observational studies (n=1 150 728)				
evaluation studies		+		
evaluation studies as topic	+			
evaluation study			+	
evaluation studies			+	
intervention studies	+			
intervention study			+	
intervention studies			+	
case-control studies	+			
case-control		+		
cohort studies	+			
cohort			+	
longitudinal studies	+			
longitudinal			+	
longitudinally			+	
prospective			+	
prospectively			+	
retrospective studies	+			
retrospective			+	
follow up			+	
comparative study		+	+	
observational			+	
Human subjects	+			
Exclusion of clinical trials, editorials, commentaries or case reports (-)				
editorial / letter / comment / case reports		-		
case report / case series			-	
clinical trial		-		
Logistical model (n=61 739)				
logistic models				+
logistic regression				+
logistic regressions				+
Linear regression (n=22 965)				
linear models				+
linear regression				+
linear regressions				+
Cox regression model (n=28 853)				
proportional hazards models				+
Cox regression				+
Cox regressions				+
Cox models				+

MeSH= MeSH terms, ptyp= publication type, tiab= title or abstract, text= full-text access, += included term, -= excluded term,
*= search was executed on April 16, 2014.

TABLE S2 (a). Inter-rater Agreement on the 12 Articles Evaluated in Common. Overall Agreement (%) and 2x2 Kappa

Item description	Comparisons						Overall Agreement
	Reviewer 1 vs 2		Reviewer 1 vs 3		Reviewer 2 vs 3		
	%	Kappa	%	Kappa	%	Kappa	
Exclusions (3/12): 100% agreement							
1.Model assumptions and goodness-of-fit	100.0%	1.00	100.0%	1.00	100.0%	1.00	100%
2.Interaction analysis:	100.0%	1.00	100.0%	1.00	100.0%	1.00	100%
3.Sensitivity analysis:	88.9%	0.73	88.9%	0.73	100.0%	1.00	89%
4.Crude and adjusted effect	100.0%	1.00	88.9%	0.78	88.9%	0.78	89%
5.More than one adjusted model	100.0%	1.00	88.9%	0.73	88.9%	0.73	89%
Number of items							
Intraclass correlation coefficient:	0.964		0.860		0.899		0.885

TABLE S2 (b). Intra-rater Agreement on Each Item: Comparison of Percentage of Agreement

Item description		Reviewer 1 n=136		Reviewer 2 n=140		Reviewer 3 n=143	
	P value[a]	n	%	n	%	n	%
1. Model assumptions and goodness-of-fit	0.530	33	24.3%	34	24.3%	42	29.4%
2. Interaction analysis	0.095	33	24.3%	21	15.0%	23	16.1%
3. Sensitivity analysis	0.148	46	33.8%	37	26.4%	53	37.1%
4. Crude and adjusted effect	0.343	52	38.2%	46	32.9%	43	30.1%
5. More than one adjusted model	0.984	35	25.7%	35	25.0%	37	25.9%

a= p value computed with chi square test.

6.3 Estudio SIMULACIÓN: Trabajo presentado en la SEE

Comparación del error tipo I y sesgo entre métodos *matching* y modelos paramétricos en presencia de confusión: estudio de simulación. (Manuscrito).

Título: Comparación del error tipo I entre los métodos matching y los modelos paramétricos: estudio de simulación

Autores

Jordi Real (1,2), Carles Forné (3,4), Jose M Martínez-Sánchez (2,5)

(1) Institut Universitari d'Investigació en Atenció Primària Jordi Gol, Barcelona.
(2) Universitat Intenacional de Catalunya, Sant Cugat.
(3) IRBLLEIDA - Universitat de Lleida, Lleida.
(4) Oblikue Consulting, Barcelona
(5) Grup de Prevenció i Control del Càncer, IDIBELL, L'Hospitalet de Llobregat

RESUMEN

Antecedentes/Objetivos: En los estudios observacionales es habitual utilizar las técnicas estadísticas multivariables como herramienta de ajuste para controlar el potencial sesgo de confusión. Sin embargo, pese al creciente interés en alternativas como los métodos *matching)* en el contexto del propensity score analysis, el ajuste multivariable GLM, como la regresión logística, siguen siendo muy utilizado. El objetivo del presente estudio es comparar el sesgo de confusión residual entre de 2 aproximaciones distintas (3 Algoritmos matching, versus GLMs) en escenarios donde la relación confusor-respuesta no es lineal.

Métodos: Se simularon conjuntos de datos de 10.000 observaciones de 3 variables: una variable *respuesta Y* Binaria Y (0,1), una variable exposición dicotómica (X) que también puede ser 0,1, generada totalmente independiente de la respuesta (Y), y una variable confusora continua (Z) relacionada con la exposición (X) y con la respuesta (Y). La respuesta se generó mediante una distribución binomial condicionada a 8 formas no lineales de Z, y 2 niveles de correlación entre la exposición y la variable confusora (X-Z), r=0,5 y 0,3. En total se simularon 7500 muestras en 16 escenarios distintos. Se emplearon 7 estrategias de análisis (3 algoritmos Matching vs 4 GLMs) para estimar el efecto nulo de X sobre Y (asumiendo un error de tipo I nominal α=0,05): 3 métodos *matching* (exact, subclasification y *nearest neighbour*) y 4 especificaciones del modelo de regresión logística (lineal, categorizando el confusor , polinomial, y un modelo GAM con una función *smoothing spline*). Se calculó el error de tipo I empírico y el error cuadrático medio de la estimación del efecto nulo.

Resultados: Los métodos *matching* y el modelo GAM mostraron errores tipo I empíricos más cercanos al nominal que los métodos paramétricos en las 8 formas funcionales consideradas. El método *matching* exacto presentó la menor tasa de error tipo I, inferior al nivel teórico en los dos niveles de correlación X-Z (α=0,041 y 0,039, para r=0,5 y 0,3 respectivamente), mientras que el método de *subclasificación* fue la estrategia *matching* que mostró un mayor error en ambos escenarios de asociación (α=0,084 y 0,061). Con el modelo GAM se observaron errores α=0,053 y 0,056, y con regresión logística errores entre α=0,085 y 0,093 categorizando el confusor, y α=0,176 y 0,474 en el modelo lineal.

Conclusiones: Los métodos *matching* con reducción de muestra proporcionan mayor credibilidad a los resultados en comparación a la regresión logística multivariable independientemente de la relación funcional entre el confusor y la respuesta.

INTRODUCCIÓN:

Las técnicas estadísticas multivariantes son ampliamente conocidas y utilizadas, ya que permiten evaluar las relaciones entre distintos factores de exposición e indicadores de salud de diversa naturaleza (dicotómicos, continuos, eventos dependientes del tiempo o recuentos)(1). Además, en los estudios observacionales es habitual utilizarlos como herramienta de ajuste debido al potencial sesgo de confusión existente en este tipo de diseño. Aunque existen otros métodos de control de la confusión, los métodos de regresión multivariables tradicionales (como regresión logística, lineal y de Cox) siguen siendo muy utilizados(2), quizás gracias a la accesibilidad a su uso desde su inclusión en la mayoría de software estadístico estándar(3) y a su potencia estadística en estudios con muestras pequeñas(1).

A pesar de su potencial, si la especificación del modelo de regresión es incorrecta, el efecto estudiado puede estar sesgado aunque se tengan en cuenta todos los confusores(4, 5). Para hacer frente a la mala especificación del modelo, la aproximación semi-paramétrica con la inclusión de funciones de suavizado tipo smothing splines (también llamados modelos aditivos generalizados (GAM)), ha demostrado mejores resultados en comparación con las técnicas de regresión paramétricas más habituales(6). Otras alternativas no paramétricas con un interés creciente son los métodos matching, ampliamente utilizados en el ámbito de la farmacoepidemiología(7-9) en la inferencia causal, cuando el investigador desea inferir efectos de tratamientos mediante diseños observacionales.

Los métodos matching consisten en seleccionar una submuestra de observaciones de forma que los grupos de comparación no contengan desequilibrios según las potenciales variables de confusión, y después estimar el efecto de interés. Existen distintos algoritmos en el pre proceso de datos para la elección de las observaciones y formación de grupos (Exact, subclassification, nearest-neighbour(N-N), optimal, genetic matching, Coarsened Exact Matching, etc…). La mayoría de estos algoritmos están incorporados en paquetes estadísticos como STATA, SAS, o R.

La gran mayoría de estos enfoques se basan en la metodología del propensity score (PS). El PS es la probabilidad de recibir un tratamiento activo condicionado a las covariables basales (10, 11), y permite la formación de grupos similares y/o ajustar la estimación del efecto estudiado, en relación a las covariables observadas. Hay varias maneras de usar el PS para reducir la confusión: estratificación por PS (ST-PS), Inverse probability of treatment weighting using PS (IPTW-PS), ajuste por PS y PS matching (PS-M)(10, 12, 13). A diferencia del resto de alternativas, el PS-M permite descartar observaciones para equilibrar los grupos. El resto de métodos aprovechan todas las observaciones y la puntuación de equilibrio PS se utiliza como medida de ajuste. Todas las metodologías basadas en el PS ofrecen la promesa de realizar inferencias más robustas en comparación a los modelos paramétricos tradicionales(10, 12) . Sin embargo dos trabajos de revisión de estudios que comparaban la efectividad de tratamientos mediante el ajuste por PS vs regresión logística multivariable (RegLogit) encontravan estimaciones bastante similares(14, 15). En cambio, Kurth T et al (16) en un estudio analítico, compararon distintos métodos de ajuste (IPTW-PS, Logistic regression model adjusted for the propensity score (as a linear term and as decile categories) , PS-M y RegLogit) para evaluar el efecto del

activador del plasminógeno tisular sobre la mortalidad de pacientes con ictus isquémico, y hallaron una fuerte dependencia de la técnica de ajuste sobre la medida de asociación estudiada. Posteriormente, Martens et al. (17) con un estudio de simulación encontraron estimaciones más cercanas al efecto real mediante un análisis con ST-PS que empleando regresión logística (RegLogit). Mediante estudios de simulación, Austin PC (18-20) analizó el rendimiento de varios métodos basados en PS sobre la estimación de distintas medidas de asociación (Odds ratio, riesgo relativo y hazard ratios). En sus resultados, PS-M y IPTW-PS produjeron estimaciones con menor sesgo en comparación con otras técnicas (ST-PS y ajuste por PS), mientras que ST-PS proporcionó estimaciones con mayor precisión. Aunque hay varios trabajos de simulación que comparan las propiedades de varias estrategias basadas en PS, la mayoría de los que incluyen PS-M se basan únicamente en el algoritmo nearest-neigbourg matching, que ha mostrado un óptimo rendimiento en comparación con otras alternativas comunes (21). Además, en los estudios publicados no se consideran otras situaciones posibles, como que la relación confusor-respuesta pueda no ser lineal, circunstancia habitual en estudios epidemiológicos (22), y que si es ignorada o no anticipada por parte del investigador, el modelo paramétrico produce estimaciones incorrectas (4, 6). En este sentido no hemos encontrado estudios que examinen el rendimiento de métodos de ajuste en tales circunstancias. Por este motivo seria de interés evaluar el impacto de la mala especificación de distintas formas no lineales de un confusor Z sobre una respuesta binaria (Y), en la estimación de una exposición X sobre Y en modelos paramétricos comparado con métodos no paramétricos PS-M tratando de controlar la confusión.

Por lo tanto, el objetivo del presente estudio fue examinar el rendimiento y robustez de distintos métodos de ajuste mediante análisis PS-M en comparación con regresión logística multivariante. Particularmente examinamos tres algoritmos PS-M (exact, subclasification y nearest-neighbour) mediante un estudio de simulación, induciendo distintas formas no lineales de la relación confusor-respuesta.

METODOLOGIA

Generación de datos: El estudio de simulación consistió en generar conjuntos de datos de 10000 observaciones de 3 variables: una variable *outcome* binaria (Y), una variable exposición dicotómica (X) y una variable confusora continua (Z).

Exposición (X) y relación con el confusor (Z): La variable exposición, o tratamiento X, se generó de forma totalmente independiente del *outcome* (Y) a partir de una distribución Binomial con P(X=1)=0.3.

Los valores de la variable confusora Z se generaron mediante una distribución Normal truncada condicionada a la variable X, $Z \sim N(mu, sd_z | X=x, Z \in [a, b])$, si X=0 y desplazada 5 unidades si X=1. Los dos niveles de asociación entre X y Z (r_{xz}) se predeterminaron modificando dos desviaciones típicas de la distribución de Z (sd_z=5, y 10), que corresponden a niveles de correlación entre X y Z de $r_{xz} \approx 0.5$ y ≈0.3, o a diferencias medias estandarizadas (Cohen's d index) de d=0,70 y d=1,27 respectivamente. En el anexo 1 se puede ver el código R de la simulación de los datos y un diagrama *box plot* de una muestra generada para los dos niveles de asociación prefijada, así como las semillas de aleatorización usadas en la simulación. Las magnitudes de asociación X-Z establecidas (d=0,7 y d=1,27) son consideradas moderadas-grandes, pero se encuentran en trabajos publicados (23-25). Por ejemplo en el estudio de Muñoz et al. (24) donde examinan factores de control de riesgo cardiovascular sobre episodios de hearth failure existen grandes diferencias de edades entre los grupos de estudio, y en el estudio de Gonzalez-de-Paz et al. (23) sobre cuidadores informales, hay un alto porcentaje de mujeres en comparación con el grupo control (23-25).

Dado que la exposición (X) se generó de forma totalmente independiente del outcome (Y), su efecto sobre la respuesta es nulo (β=0, OR=1 o RR=1).

Relación del outcome (Y) con el confusor (Z): El outcome (Y) se generó con dependencia de Z mediante una distribución binomial condicionada a una función de Z, $Y|G(Z=z) \sim$binomial. Así pues, $Logit(Y=1) = B_0 + G(Z)$, donde la G(Z) tomó las siguientes formas inspiradas en el estudio de Benedetti et al.(6): (a) Linear, (b) Quadratic Threshold, (c) Cubic Asymmetric, (d) Plateau Threshold, (e) Gausiana Asymetric, (f) Asymetric U Threshold, (g) "Hump", (h) Double-Hump (Figura 1).

La forma más simple de la relación exposición-riesgo es considerar un incremento del riesgo proporcional al factor de exposición como es el efecto lineal (a). Sin embargo, existen otras formas de relación exposición-respuesta (Z-Y), típicas en distintas áreas de la epidemiologia: Formas polinomiales, o J-Shape (b,c,d), son características de los efectos adversos en salud, como en la relación del daño pulmonar por presión en función del volumen de aire inducido mediante ventilación mecánica(26); El efecto umbral (b y f) se caracteriza por efectos constantes a partir o por debajo de un cierto valor, como el efecto de la exposición acumulada a hexavalent Chromium sobre el desarrollo de cáncer de pulmón (22); Las formas Gaussianas o hump (g, e) se caracterizan por efecto nulo en dosis extremas y efectos máximos en rangos intermedios, como la toxicidad por vitamina A que ha demostrado ser perjudicial tanto en dosis demasiado bajas como o en dosis excesivas en mujeres embarazadas; Además las relaciones Gaussianas o hump pueden ocurrir de manera distinta en dos subpoblaciones como representaría el escenario Doble-Hump (h) (22).

Simulaciones de Monte Carlo: En total se simularon 7500 muestras en cada uno de los 16 escenarios considerados (8 relaciones Y-Z, y 2 tipos de correlación exposición(X)-confusor(Z)). En cada muestra simulada se obtuvo el efecto de X (β_x=0) sobre el logit(Y) y el p-valor con 7 estrategias de análisis.

Estrategias de análisis: Se emplearon 3 métodos basados en matching (PS-M: exact, nearest-neighbour, subclassification), todos ellos implementados con la libreria MatchIt de R (versión 3.0.2) (27), y combinados con un análisis univariado posterior (univariate logistic regression), 3 especificaciones distintas de análisis multivariable mediante regresión logística (confusor continuo, categorizado en quintiles y términos polinomiales), y 1 modelo semiparamétrico GAM (Smothing spline) (28). En la tabla 1 se muestra una descripción de los métodos comparados y las librerías y funciones de R utilizadas.

Brevemente detallamos las características principales de las 7 estrategias de análisis utilizadas:

1)Exact: Metodo exact matching categorizando el confusor en deciles. Algoritmo que empareja cada unidad tratada (X=1) con todas las posibles unidades del grupo control (X=0) de manera que ambos grupos contengan exactamente los mismos valores según las covariables especificadas (Z). El confusor Z se categorizó mediante deciles ya que encontrar grupos idénticos en presencia de variables continuas, como es el caso de Z, no es posible.

2)Nearest: Nearest-neighbour con la opción calliper (caliper=0.1). Este algoritmo selecciona los mejores controles emparejados para cada individuo del grupo tratamiento (X=1). En cada etapa del matching se elige la observación del grupo control que esté lo más cercana a la observación tratada según una distancia especificada, habitualmente el logit(X). Si múltiples sujetos controles se encuentran a la misma distancia del sujeto tratado, se selecciona aleatoriamente a uno de estos. La opción caliper (número de desviaciones estándar de la medida de la distancia) establece una distancia máxima entre tratados y no tratados (la anchura del caliper).

3) Subclass: Subclassification matching con la opción *discard* en ambos grupos. Algoritmo que forma estratos, en función de la distribución del PS estimado, de tal manera que asegura la igualdad de distribuciones dentro de cada estrato. La opción *discard* elimina observaciones para mejorar la igualdad de distribuciones dentro de cada estrato.

4-7) Modelo lineal generalizado con link logit (GLM). Generalización del modelo de regresión logística multivariante con 4 especificaciones de ajuste distintas: 4) LogitLinear, incluyendo el confusor Z como un término lineal; 5) LogitNolinear, incluyendo términos cuadráticos o polinomiales de 3^r orden para la variable Z ($z + z^2 + z^3$); y 6) LogitCat, incluyendo en el modelo el confusor Z categorizado en quintiles. $\text{logit}(Y)= \beta_0 + X\beta_x + Z1\beta_{z1} + Z2\beta_{z2} + Z3\beta_{z3} + Z4\beta_{z4}$, donde Z1, Z2, Z3, Z4 son variables dicotómicas que correspondientes a los quintiles 1, 2, 3 y 4. 7) LogitGAM: Modelo aditivo generalizado: Un modelo aditivo generalizado (GAM) es un modelo lineal general (GLM) que combina una parte paramétrica con otra no paramétrica. La parte no paramétrica fue dada por una función de suavizado "*Smothing splines*" de la covariable (Z) con un parámetro de suavizado automático, añadida a una componente paramétrica convencional con el predictor (X). Por lo tanto el modelo ajustado fue $\text{logit}(Y)= \beta_0 + X\beta_x + s(Z)$.

Estimación: Se calculó el error de tipo I empírico como la proporción de estimaciones de β_x en las que su p-valor fue inferior a 0.05. Además, se calculó el error cuadrático medio (ECM) de βx.

RESULTADOS

En la tabla 2 se pueden observar los errores de tipo I empíricos en función de la estrategia de análisis utilizada con los distintos escenarios considerados. El gradiente de color representa, de claro a más oscuro, la distancia del error empírico al nivel de significación habitualmente establecido ($\alpha = 0.05$). Se pueden observar que en las muestras generadas con la forma Z-Y lineal la mayoría de métodos de ajuste obtienen errores de tipo I cercanos al 0.05, a excepción de logitCat, y el método Subclass con alta asociación XZ reportando una proporción de estimaciones significativas del 13,6 y el 9% respectivamente. En cambio en las muestras generadas con relaciones no lineales, los ajustes GLM presentan una frecuencia de errores tipo I muy alejados del real llegando al 25-30% en las muestras Quadratic T, Cubic o Gausiana en los escenarios de alta asociación XZ, y el 70-98% cuando la asociación XZ es baja. La aproximación polinomial (LogNolineal) presenta un error de tipo I más cercano al 5%, excepto en las muestras generadas con relación Gausiana, donde el error empírico se sitúa entre el 30% y 52% para los dos niveles de asociación XZ simulados (r_{xz}=0.5 y 0.3). Los análisis mediante PS-M, presentan un error tipo I estimado más cercano al nominal al igual que el ajuste por GAM en comparación a las estrategias de regresión paramétrica. Entre las tres estrategias de matching el método exacto es el que presenta una menor proporción de errores, incluso por debajo del nivel teórico (0.041 y 0.039). El método *subclass* es la estrategia *matching* que obtiene un mayor error en ambos escenarios de asociación X-Z (0.5 y 0.3).

La tabla 3 muestra el ECM según el método de ajuste y la forma de la relación Z-Y generada. El gradiente de color representa valores del ECM mayores. Se puede observar como los métodos de ajuste GLM paramétricos, obtienen mayores ECMs en referencia al resto de métodos.

Las figura 2 y 3 muestran la media de las estimaciones de Beta ± una desviación típica, para cada método de ajuste, y para cada forma funcional Z-Y generada, para cada nivel de asociación X-Z, respectivamente. La línea de puntos horizontal en 0 representa el valor real de B_x. Con las muestras generadas según una forma de la relación Y-Z lineal (Figura 1a), todos los métodos obtienen estimaciones insesgadas o muy cercanas al 0, a excepción del método logitCat. En cambio en los otros escenarios presentan una mayor variabilidad del método de ajuste sobre la estimación de B_x siendo los métodos GLM (principalmente paramétricos LogLinear, LogitCat y LogNoLinear) los que en general obtienen estimaciones con mayor sesgo. La forma de la relación Z-Y también afectó a la variabilidad de las estimaciones con desviaciones típicas (DT) mayores en las formas cuadráticas (Fig 2b y Fig3b).

Finalmente notar que de las 10.000 observaciones iniciales generadas por cada muestra, en las 3 estrategias de matching se eliminaron en promedio de un 52%/42%, 29%/20% y 50%/42% observaciones con los métodos *exact* con el escenario alta/baja correlación, *subclass* escenario alta/baja y *nearest* escenario alta/baja correlación respectivamente.

DISCUSIÓN

Este trabajo ha examinado el rendimiento de 2 aproximaciones distintas de ajuste (PS-M *versus* regresión logística multivariable) en cuanto a eliminación del sesgo de confusión, induciendo escenarios donde la relación del confusor-respuesta no es lineal, con muestras grandes (N=10000). En este contexto la representación paramétrica no ha llegado a eliminar por completo el sesgo de confusión existente en los datos ya que la estimación de la asociación de interés en determinados escenarios aun presentaba un substancial sesgo después de ajustar por el confusor. En cambio los métodos matching examinados, y la representación no paramétrica (modelo GAM) ha mostrado un mejor comportamiento presentado menores tasas de error de tipo I, y un menor sesgo residual en comparación con una técnica 100% paramétrica, como es la regresión logística multivariable.

El ajuste mediante modelos estadísticos es una herramienta muy poderosa para estimar el efecto real aislando los potenciales factores de confusión. Pero los modelos paramétricos demandan asunciones muy estrictas sobre el ajuste de los datos (linealidad de los predictores, normalidad, homocedasticidad, incolinealidad, etc.) cuyo incumplimiento puede invalidar el modelo y la inferencia realizada(29, 30). La consecuencia más *dramática* de la mala especificación del modelo es la confusión residual (4, 31) produciendo estimaciones aun sesgadas después del ajuste. Prueba de ello, son los resultados de nuestro estudio, donde en condiciones ideales (sin errores de medición ni confusores no medidos), con grandes muestras, y por lo tanto alta potencia estadística, una técnica de análisis multivariable como es la regresión logística puede producir estimaciones muy alejadas del parámetro real y una alta tasa de error de tipo I en circunstancias de efecto nulo.

Cuando un investigador analiza los resultados de un estudio, la asociación verdadera y la correcta forma funcional del modelo es raramente conocida. Los errores de especificación de los modelos utilizados se pueden minimizar realizando la comprobación de las asunciones formales del modelo finalmente seleccionado (29, 30). Sin embargo, algunos estudios previos ponen en duda que se verifique la validación de las técnicas de regresión utilizadas o que se utilicen medidas que aseguren la rigurosa aplicación de los métodos multivariables (32, 33). Casals et al. (33) en una revisión sistemática de 108 artículos que utilizaban modelos lineales mixtos, encontraron que la validación del modelo y reporte de medidas de bondad de ajuste tan solo de un 6,5% y un 15,7%, respectivamente. Un estudio más reciente(33) muestra que en únicamente el 15% de estudios observacionales indexados en MEDLINE se reportan pruebas de validación y/o diagnóstico de los modelos de regresión logística.

En nuestro estudio, los modelos GAM con una representación no paramétrica del confusor continuo (suavizado Smothing spline, con parámetro de suavizado automático) presenta mejores resultados en términos de reducción del sesgo de confusión, en comparación con cualquiera de las representaciones paramétricas evaluadas con regresión logística. En este sentido nuestros resultados son consistentes con los reportados por el estudio de Benedetti et al (6). Este tipo de aproximación se adapta mucho mejor a la forma funcional de los datos y consigue un mejor ajuste. Sin embargo, esta mayor flexibilidad también puede generar algunos

problemas, como el sesgo en la estimación de los errores estándar en caso de sobreajuste en presencia de concurbidity (alta correlación entre dos o más covariables)(34).

Respecto a los métodos basados en el PS-M, el trabajo de Austin et al (10) señala varias razones prácticas para preferir el uso de estos métodos sobre el ajuste de regresión habitual. Primero, es más simple determinar si el modelo ha sido adecuadamente especificado (solo con verificar la homogeneidad de los grupos después del matching o creación del PS). En contraste, es mucho más difícil determinar si el modelo de regresión relativo a la variable a evaluar, más los confusores basales sobre el *outcome* han sido correctamente especificados (35). Segundo, estos métodos permiten separar el diseño del estudio del análisis, de la misma forma que lo hace un ensayo clínico aleatorizado (RCT), donde solamente se puede estimar el efecto del tratamiento estudiado una vez el estudio ha finalizado. Cuando se usan las técnicas de matching, o métodos basados en el PS, la construcción de los grupos se puede realizar sin ninguna referencia al outcome. Solamente una vez se acepta el equilibrio de los grupos se realiza la estimación del efecto del tratamiento sobre el outcome. Sin embargo, cuando utilizamos técnicas de regresión, el *outcome* siempre está a la vista, y el investigador se enfrenta continuamente a la sutil tentación de modificar el modelo de regresión hasta que se alcanza el resultado esperado. Tercero, se puede examinar explícitamente el grado de solapamiento o superposición de la distribución de las covariables entre los grupos. Si hay substanciales diferencias de las covariables entre grupos, estas serán evidentes, dado el pequeño número de sujetos emparejados. Este hecho es importante ya que, por ejemplo, si casi la totalidad de fumadores están en uno de los grupos y la mayoría de los no fumadores están en el otro, ajustar por tabaquismo no eliminará la confusión causada por esta variable (36). En el caso de que haya problemas de solapamiento, se generarán insuficientes sujetos en los grupos, el analista o investigador puede concluir que los grupos son tan diferentes que una comparativa de la respuesta no tiene sentido y por lo tanto no será plausible el análisis. En cambio, al utilizar enfoques basados en la regresión, puede ser difícil evaluar el grado de solapamiento entre grupos. Y en un contexto en el que exista una fuerte separación entre los grupos, se puede cometer el error de proceder a un análisis basado en la regresión sin ser consciente que el modelo de regresión ajustado está interpolado en dos poblaciones distintas. El problema de escasez de datos es menos probable que se produzca en estudios con muestras grandes.

La informatización de la historia clínica en la era del Big Data ha proporcionado un gran potencial para la investigación clínica ya que a partir de grandes bases de datos existentes se pueden realizar estudios epidemiológicos con grandes muestras a un coste muy inferior a los estudios convencionales(37). La capacidad de que una investigación encuentre una relación significativa viene reflejada por la potencia estadística: 1 - β (uno menos la tasa de error de Tipo II), y esta se incrementa notablemente con el tamaño muestral del estudio. Pero los estudios con grandes muestras no están libres de errores. En este contexto, los métodos matching pueden ser una alternativa válida, ya que una reducción de la muestra no se traduce en una pérdida de potencia estadística determinante, y en cambio puede mejorar la calidad de la inferencia realizada sobre las asociaciones examinadas.

Nuestros hallazgos se basan en simulaciones de Monte Carlo, lo cual implica ciertas limitaciones propias de este tipo de estudios. Mientras que los resultados analíticos son deseables no siempre es posible obtenerlos ya que no existen otras formas de inducir relaciones conociendo a priori la verdadera asociación(38). La generalización de nuestros resultados se limita a los escenarios planteados, los cuales pueden ser considerados relativamente simples: estudios con grandes muestras, con tres únicas variables, dos niveles de asociación exposición-respuesta sobre una respuesta binaria. Los fenómenos de interés binarios son tratados de forma habitual en la literatura médica (1, 39) y asumiendo que la distribución del confusor proviene de una distribución normal truncada. Esta es precisamente una aportación del presente estudio, pues el truncamiento de una variable se da habitualmente en estudios epidemiológicos debido a los criterios de inclusión previamente establecidos.

Este trabajo se ha enfocado en examinar 3 algoritmos matching más conocidos y utilizados (40) (Exact, nearest– neighbour y subclassification). Otros algoritmos existentes, tales como el *optimal, o genetic* etc.(12, 16) no se han evaluado debido a que la alta intensidad computacional requerida en los escenarios de grandes muestras. En cuanto a los modelos GAM, no se ha examinado otras funciones de suavizado existentes tales como lowess, b-splines, o más flexibles representaciones paramétricas como polinomios fraccionales (6). Por lo tanto sería deseable examinar otros métodos de ajuste, más combinaciones de variables, la sensibilidad de los métodos de análisis (por ejemplo evaluar el análisis estándar versus análisis condicional).

Finalmente, también somos conscientes que una de las limitaciones más importantes de los estudios observacionales es que es imposible ajustar por aquellas variables no medidas o desconocidas (41). Por lo tanto, en la práctica incluso aunque la técnica estadística se utilice de manera correcta y no haya errores de medición de las variables, siempre cabe la posibilidad de que la asociación estimada permanezca confundida. Por lo tanto la aplicación correcta del método de ajuste no garantiza la eliminación completa del sesgo.

Conclusiones

En resumen, los métodos matching en estudios de muestras grandes presentan una mayor robustez en comparación con una técnica de regresión convencional como es la logística multivariable (principalmente paramétrica), ya que eliminan mejor el sesgo de confusión y presentan un error de tipo I más cercano al nominal. Así pues, los métodos no paramétricos matching con reducción de muestra proporcionan mayor credibilidad a los resultados en comparación a la regresión logística multivariable independientemente de la relación funcional entre el confusor y la respuesta.

References

1. Bender R. Introduction to the use of regression models in epidemiology. In: Mukesh Verma, editor. Methods in Molecular Biology, Cancer Epidemiology. United States: Springer Science; 2009. p. 179-95.

2. Real J, Cleries R, Forne C, Roso-Llorach A, Martinez-Sanchez JM. Use of multiple regression models in observational studies (1970-2013) and requirements of the STROBE guidelines in Spanish scientific journals. Semergen. 2015 Nov 5.

3. Gentle JE, Härdle WK, Mori Y. How computational statistics became the backbone of modern data science. In: Handbook of Computational Statistics. Springer; 2012. p. 3-16.

4. Groenwold RH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KG, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. CMAJ. 2013 Mar 19;185(5):401-6.

5. Ho KM. Residual confounding in observational studies. Anesthesiology. 2009 Feb;110(2):430; author reply 430.

6. Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. Stat Med. 2004 Dec 30;23(24):3781-801.

7. Gan HL, Zhang JQ, Bo P, Wang SX, Lu CS. Statins decrease adverse outcomes in coronary artery bypass for extensive coronary artery disease as well as left main coronary stenosis. Cardiovasc Ther. 2010 Apr;28(2):70-9.

8. Chaux A, Peskoe SB, Gonzalez-Roibon N, Schultz L, Albadine R, Hicks J, et al. Loss of PTEN expression is associated with increased risk of recurrence after prostatectomy for clinically localized prostate cancer. Mod Pathol. 2012 Nov;25(11):1543-9.

9. Stamou SC, Hill PC, Haile E, Prince S, Mack MJ, Corso PJ. Clinical outcomes of nonelective coronary revascularization with and without cardiopulmonary bypass. J Thorac Cardiovasc Surg. 2006 Jan;131(1):28-33.

10. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011;46(3):399-424.

11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

12. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis. 2007;15(3):199-236.

13. Pattanayak CW, Rubin DB, Zell ER. Propensity score methods for creating covariate balance in observational studies. Rev Esp Cardiol. 2011 Oct;64(10):897-903.

14. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006 May;59(5):437-47.

15. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005;58(6):550-9.

16. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol. 2006 Feb 1;163(3):262-70.

17. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol. 2008 Oct;37(5):1142-7.

18. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med. 2013;32(16):2837-49.

19. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078-94.

20. Austin PC. The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol. 2008;61(6):537-45.

21. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011 Mar-Apr;10(2):150-61.

22. May S, Bigelow C. Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges. Nonlinearity in Biology, Toxicology, Medicine. 2005;3(4):dose,response. 003.04. 004.

23. Gonzalez-de Paz L, Real J, Borras-Santos A, Martinez-Sanchez JM, Rodrigo-Banos V, Dolores Navarro-Rubio M. Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study. J Public Health Policy. 2016 May;37(2):173-89.

24. Munoz MA, Real J, Del Val JL, Vinyoles E, Mundet X, Domingo M, et al. Impact of the sustained control of cardiovascular risk factors on first episode heart failure: The relevant role of primary care. Eur J Gen Pract. 2015;21(4):224-30.

25. Wu J, Han Y, Zhao FL, Zhou J, Chen Z, Sun H. Validation and comparison of EuroQoL-5 dimension (EQ-5D) and Short Form-6 dimension (SF-6D) among stable angina patients. Health Qual Life Outcomes. 2014 Oct 25;12:156,014-0156-6.

26. López-Aguilar J, Villagrá A, Bernabé F, Murias G, Piacentini E, Real J, et al. Massive brain injury enhances lung damage in an isolated lung model of ventilator-induced lung injury*. Crit Care Med. 2005;33(5):1077-83.

27. Ho D, Imai K, Imai MK. Package 'MatchIt'. Retrieved June. 2013;23:2014.

28. Wood S. Generalized additive models: an introduction with R. CRC press; 2006.

29. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med. 1995 Aug 15;14(15):1707-23.

30. Dobson AJ. An Introduction to Generalized Linear Models. 2nd ed. Annette J. Dobson, editor. United States of America: Chapman and Hall; 2001.

31. Liang W, Zhao Y, Lee AH. An investigation of the significance of residual confounding effect. Biomed Res Int. 2014;2014:658056.

32. Casals M, Girabent-Farres M, Carrasco JL. Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000-2012): a systematic review. PLoS One. 2014 Nov 18;9(11):e112653.

33. Real J, Forne C, Roso-Llorach A, Martinez-Sanchez JM. Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine (Baltimore). 2016 May;95(20):e3653.

34. Ramsay TO, Burnett RT, Krewski D. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. Epidemiology. 2003;14(1):18-23.

35. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. Stat Med. 2013;32(1):67-80.

36. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. Cambridge university press; 2011.

37. Bolibar B, Fina Aviles F, Morros R, Garcia-Gil Mdel M, Hermosilla E, Ramos R, et al. SIDIAP database: electronic clinical records in primary care as a source of information for epidemiologic research. Med Clin (Barc). 2012 May 19;138(14):617-21.

38. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol. 2007 Sep 15;166(6):646-55.

39. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol. 2001 Oct;54(10):979-85.

40. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. The International Journal of Biostatistics. 2009;5(1).

41. Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2002;359(9302):248-52.

Tabla 1. Descripción de los métodos de ajuste utilizados

Method	Description	Confusor	Librería R	Función
GLM	Ajuste mediante modelo de regression logística multivariable			
LogLinear	Modelo multivariable lineal	z	glm	glm()
LogNolinear	Modelo multivariable polinomial con términos cuadraticos	z + z^2 + z^3	glm	
LogitCat	Modelo multivariable con confusor categorizado	Quintiles de z	glm	
LogitGAM	Generalized Additive Model: Semiparametric adjusted	s(z)	mgcv	gam()
Matching	Matching seguido de regresión logística estándar			
Exact	Exact Matching categorizando el confusor z (N-N con la opcion exact)	Deciles de z	MatchIt	matchit()
Nearest	Nearest - Neigbourg con la opción *caliper* (caliper=0.1)	z	MatchIt	matchit()
Subclass	Subclassification con la opción *discard* en ambos grupos	z	MatchIt	matchit()

Tabla 2. Error de tipo I empírico de X en condiciones de relación nula entre X e Y, en función de la forma de la asociación ZY generada

Correlacion X-Z		GLM			Matching		
Forma de la relación Z-Y	LogLinear	LogitCat	LogNolineal	GAM	Exact	Subclass	Nearest
Alta (SD=5; r=0.5)							
Linear	0,050	0,136	0,051	0,050	0,050	0,090	0,056
Quadratic T	0,254	0,096	0,054	0,049	0,046	0,068	0,050
Cubic Asymetric	0,084	0,082	0,050	0,049	0,047	0,119	0,046
Plateau	0,051	0,083	0,048	0,048	0,044	0,073	0,055
Gaussian	0,313	0,057	0,301	0,072	0,024	0,074	0,023
Asymetric U T	0,310	0,050	0,128	0,048	0,044	0,090	0,047
Hump	0,264	0,068	0,088	0,052	0,039	0,080	0,044
Double Hump	0,083	0,106	0,091	0,057	0,037	0,079	0,126
Total	0,176	0,085	0,101	0,053	0,041	0,084	0,056
Baja (SD=10; r=0.3)							
Linear	0,051	0,079	0,052	0,052	0,045	0,059	0,053
Quadratic T	0,701	0,066	0,064	0,051	0,051	0,052	0,052
Cubic Asymetric	0,980	0,324	0,050	0,053	0,033	0,105	0,038
Plateau	0,066	0,051	0,060	0,050	0,042	0,054	0,043
Gaussian	0,945	0,055	0,525	0,082	0,029	0,051	0,032
Asymetric U T	0,142	0,052	0,074	0,051	0,046	0,057	0,052
Hump	0,050	0,052	0,051	0,050	0,041	0,055	0,048
Double Hump	0,802	0,058	0,049	0,054	0,027	0,051	0,031
Total	0,474	0,093	0,117	0,056	0,039	0,061	0,044

Color de fondo condicionado a la magnitud del error de tipo I empírico: Cuanto más oscuro más alejado de 0,05.

Tabla 3. Error cuadrático medio del efecto de X en condiciones de no relación entre X e Y, en función del tipo de generación de datos (relación Z-Y)

Correlation X-Z		GLM				Matching		
Forma de la relación Z-Y	LogLinear	LogitCat	LogNolineal	GAM	Exact	Subclass	Nearest
Alto (SD=5; corr=0.5)							
Linear	0,003	0,005	0,003	0,003	0,003	0,003	0,004
Quadratic T	0,027	0,016	0,012	0,012	0,015	0,013	0,015
Cubic Asymetric	0,008	0,008	0,007	0,007	0,008	0,010	0,008
Plateau	0,003	0,004	0,003	0,003	0,004	0,003	0,004
Gaussian	0,010	0,004	0,010	0,005	0,003	0,004	0,003
Asymetric U T	0,009	0,003	0,005	0,003	0,004	0,004	0,004
Hump	0,010	0,004	0,005	0,004	0,004	0,005	0,004
Double Hump	0,004	0,005	0,004	0,004	0,003	0,004	0,006
Total	0,009	0,006	0,006	0,005	0,006	0,006	0,006
Bajo (SD=10; r=0.3)							
ECM							
Linear	0,002	0,003	0,003	0,003	0,003	0,003	0,003
Quadratic T	0,068	0,013	0,012	0,011	0,014	0,011	0,015
Cubic Asymetric	0,052	0,014	0,005	0,005	0,004	0,006	0,004
Plateau	0,003	0,003	0,003	0,003	0,003	0,003	0,003
Gaussian	0,041	0,004	0,015	0,004	0,003	0,003	0,003
Asymetric U T	0,004	0,003	0,003	0,003	0,003	0,003	0,003
Hump	0,003	0,004	0,003	0,004	0,004	0,004	0,004
Double Hump	0,022	0,003	0,003	0,003	0,002	0,003	0,003
Total	0,025	0,006	0,006	0,004	0,005	0,005	0,005

Color de fondo condicionado a la magnitud del valor ECM: Color oscuro significa mayor error cuadrático medio y más alejado de 0

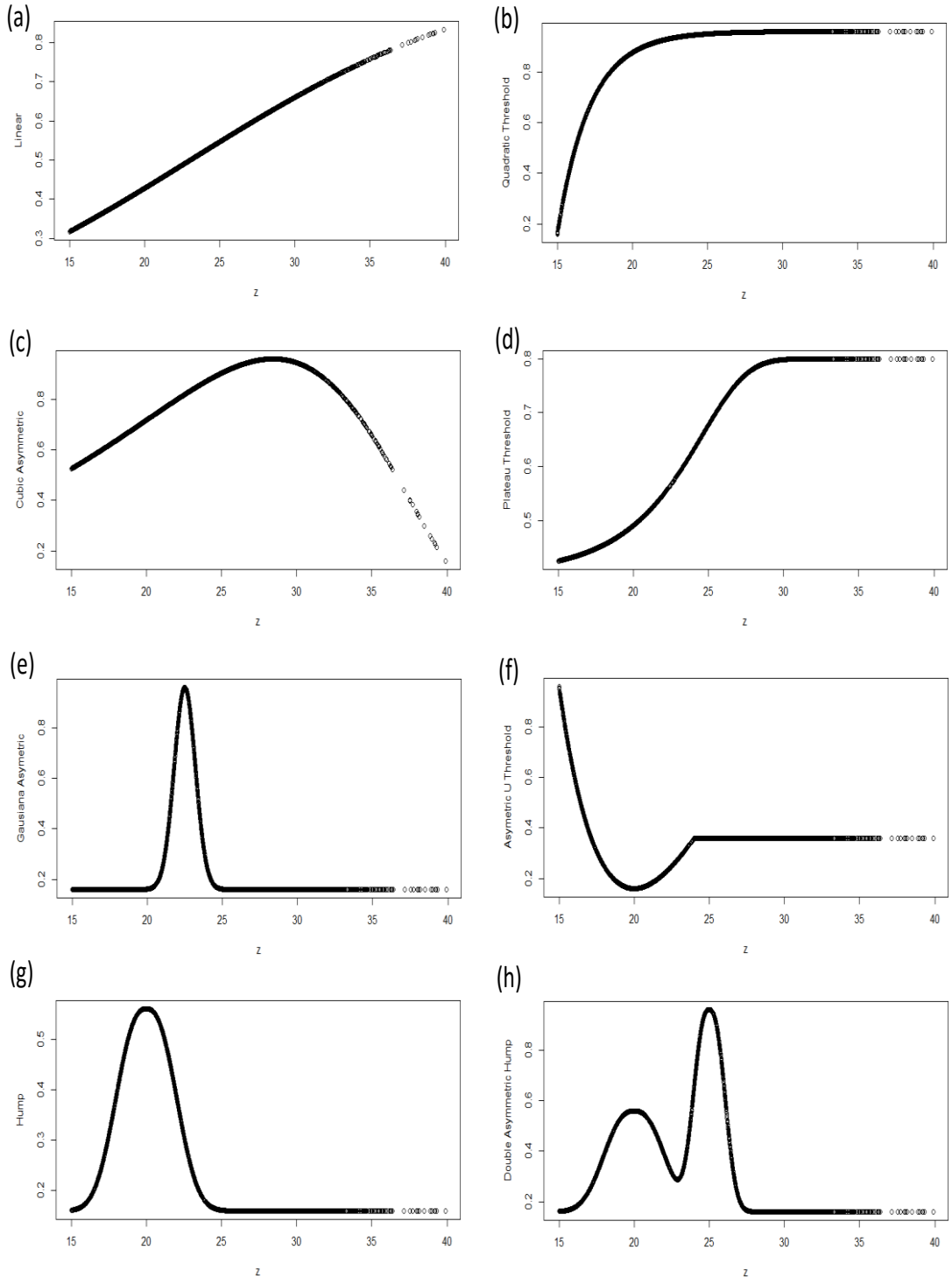


Figura 1. Relaciónes generadas de Z-P(Y): (a) Linear, (b) Quadratic Threshold, (c) Cubic Asymmetric, (d) Plateau Threshold, (e) Gausiana Asymetric , (f) Asymetric U Threshold, (g) "Hump", (h) Double Hump.

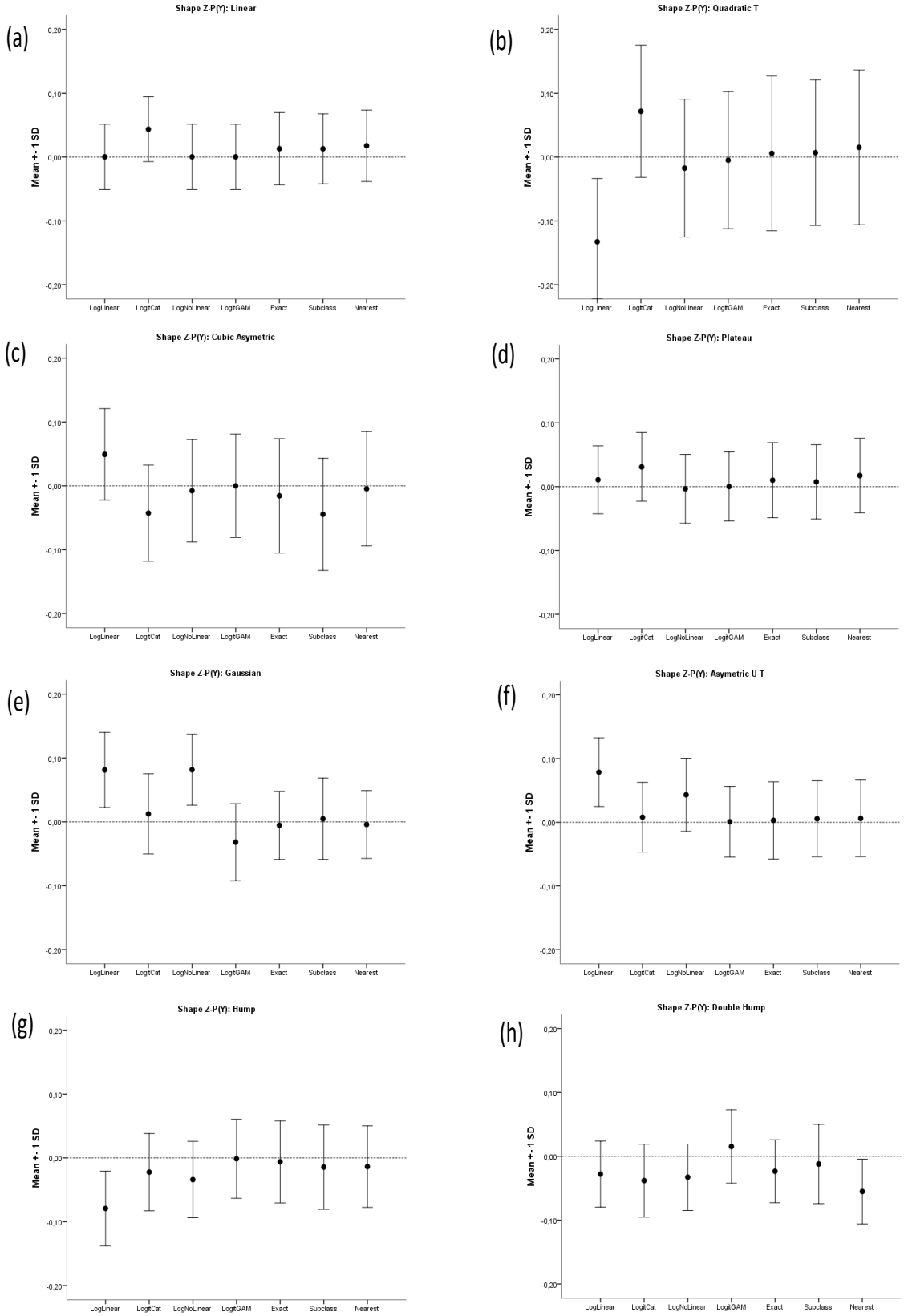


Figura 2. Media de Beta +- DT según método de ajuste para las formas de relación Z-Y (r XZ=0.5;SD=5)

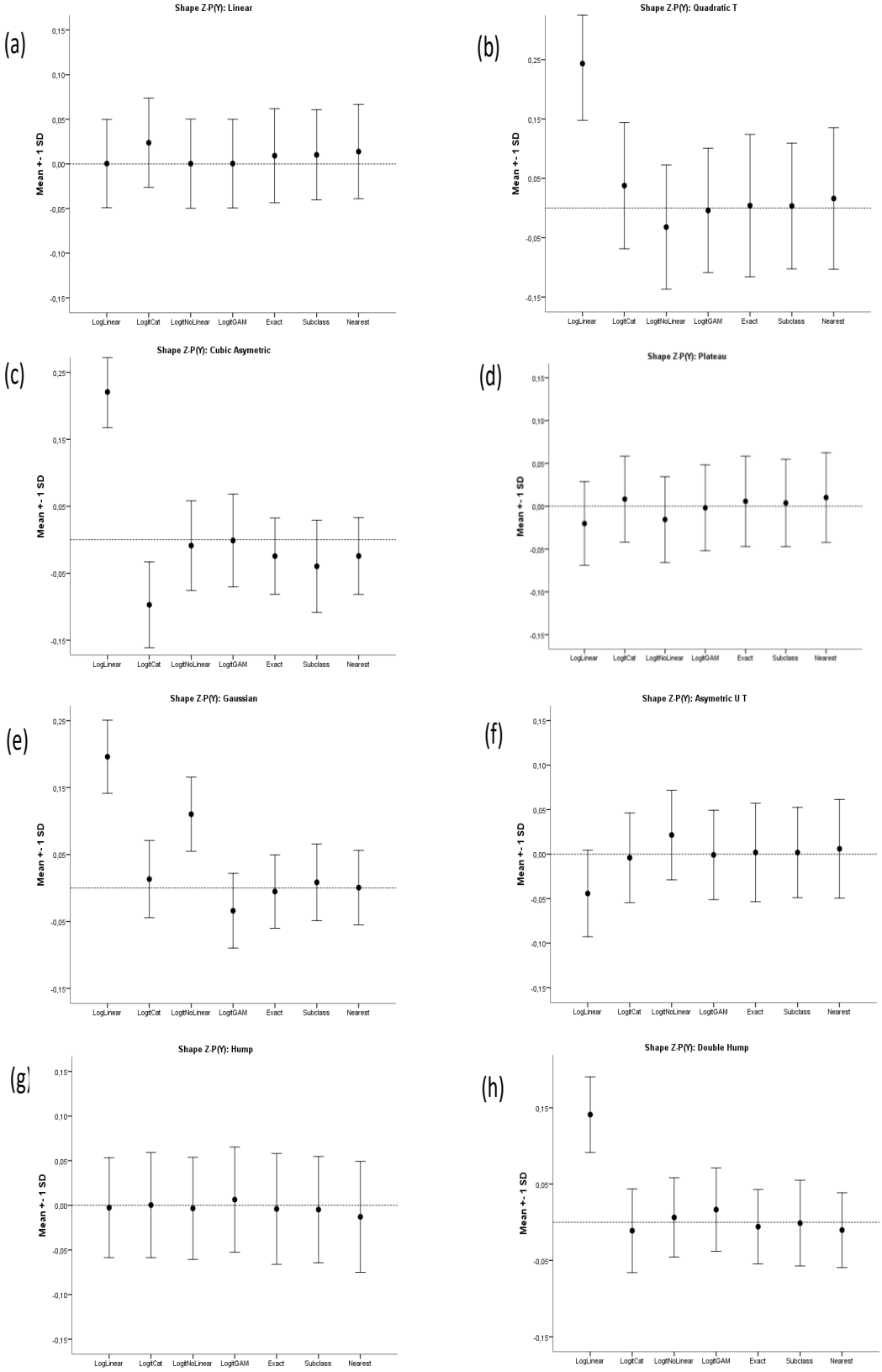


Figure 3. Media de Beta+-DT según método de ajuste para las formas de relación Z-Y (r xz=0.3;SD=10)

Apéndice 1. Datos técnicos de la simulación

1. Código en R de simulación de 10000 observaciones según una X binomial (0.3), y Z normal truncada, para dos niveles de asociación entre X y Z.

```
library("truncnorm")
set.seed(1)                 #####       Semillas de 1 hasta 3750 ; 10001 hasta 13750
nObs<-10000                 #####       Muestra de tamaño 10000
x<-rbinom(nObs,1,0.3)       #####       Genero binomial (10000, 0.3) independiente de Y

DT<-10                      #####       DT=10
z<-20+5*x+rtruncnorm(n=nObs,a=-5,b=90,mean=0,sd=DT)    # Genero Z función de X
corr(x,z)
par(c(1,2))
boxplot(z~x)
set.seed(1)

DT<-5                       #####       DT=5
z<-20+5*x+rtruncnorm(n=nObs,a=-5,b=90,mean=0,sd=DT)    # Genero Z función de X

boxplot(z~x)
```

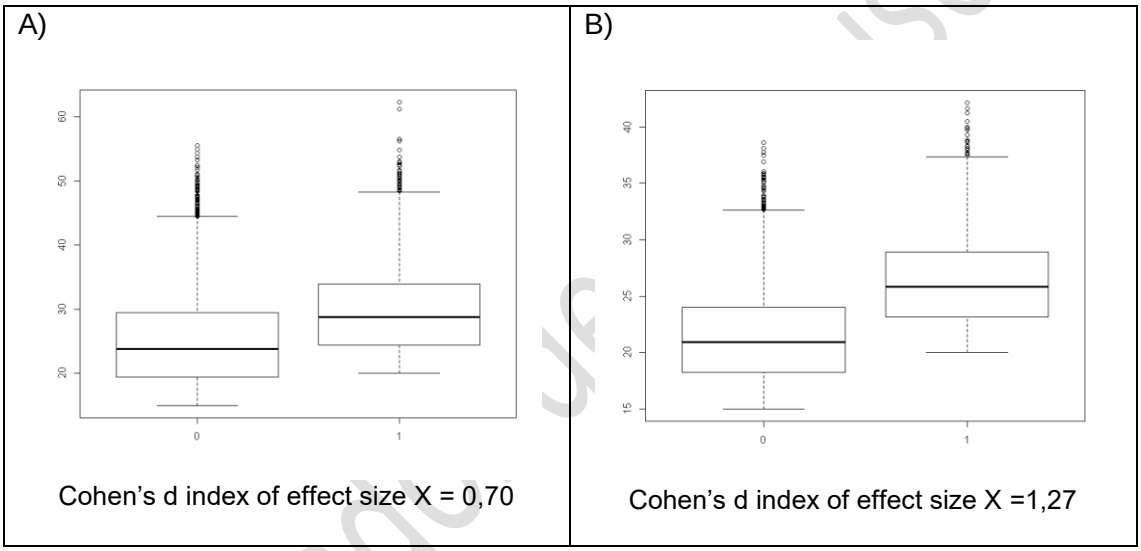


Figura 1. Box plot de muestra de datos simulada de Z en relación a valor de X para r_{xz}≈0.3 (panel A) y r_{xz}≈0.5 (panel B).

2. Código de R de las funciones que simularon las muestras de las observaciones de la probabilidad de Y condicionado a las distintas formas nolineales. P(Y)=g(Z)

```
#######  Funciones  ##########
jocdedadesredux=list(
  ##########                  a) Linear                              ######
  Plineal=function(x) {
    #### ODDSRATIOz<-1.1
    Beta1<-log(1.1)
    Beta0<-log(0.10/(1-0.10))
    klineal<- Beta0 + (Beta1*x)
    Plineal<-exp(klineal) / (exp(klineal)+1)
  },
  #
  ##########                  b) PquadraticV (Quadratic V)           ######
  PquadraticV=function(x) {
    klineal<- x^0.8
    p<-exp(klineal) / (exp(klineal)+1)
    p<-(((p-min(p))/(max(p)-min(p)))+0.2)*0.8
  },
  #########                   c) Cubic Asymetric                     ######
  PcubicV=function(x) {
    Beta1<-log(1.05)
    Beta0<-log(0.001/(1-0.001))
    multi<-5
    klineal<- Beta0 + (Beta1*x)+((Beta1*x)^multi)-((Beta1*x)^(multi+0.1))
    p<-exp(klineal) / (exp(klineal)+1)
    p<-(((p-min(p))/(max(p)-min(p)))+0.2)*0.8
  },
  ##########                  d) Plateau Threshold                   #######
  PQuadraticT=function(x) {
    klineal<- 1.3^(x-mean(x))
    p<-exp(klineal) / (exp(klineal)+1)
    p<-p*0.8
    p<-ifelse(is.nan(p),0.8,p)
  },
  ##########                  e) Gausiana asymetric                  #######
  PGaus=function(x) {
    p<-exp(-(x-22.5)^2)
    klineal<-log(p)/(1-log(p))
    p<-(p-min(p))/(max(p)-min(p))
    p<-(p+0.2)*0.8
  },
  ##########                  f) Asymetric U Threshold               ########
  PJinvAsiT=function(x) {
    klineal<-((x-20)/x)^2
    p<-exp(klineal) / (exp(klineal)+1)
    limit<-24
    p[x> limit]<-exp(((limit-20)/limit)^2) / (exp(((limit-20)/limit)^2)+1)
    p<-(p-min(p))/(max(p)-min(p))
    p<-(p+0.2)*0.8
  },
  ##########                  g) Hump                                #######
  Phump=function(x) {
    k1<-0.1; k2<-0.5  ### Altura
    p<-0.5*exp(-k1*abs(x-20)^2.5) +0*exp(-k2*abs(x-25)^2.5)
    p<-(p+0.2)*0.8
  },
  ##########                  h) Double Asymmetric Hump              #######
  PDhumpII =function(x) {
    k1<-0.1; k2<-0.5  ### Altura
    p<-0.5*exp(-k1*abs(x-20)^2.5) +1*exp(-k2*abs(x-25)^2.5)
    p<-(p-min(p))/(max(p)-min(p))
    p<-(p+0.2)*0.8
  },

)
```

3. **Características descriptivas de una muestra simulada (n=10000) y relación Y-Z según los 16 escenarios considerados. Numero de eventos (Y=1), Media de z para Y=1, y para Y=0, diferencia media estandarizada de Z (d) entre grupos (Y=0 vs Y=0) , y correlación(r).**

Relación	SD=10	Cor=low		Effect size		SD=5	Cor=high		Effect size	
P(Y)-Z	N	Mean z Y=1	Mean z Y=0	d	r	N	Mean z Y=1	Mean z Y=0	d	r
Linear	5718	27,7	23,1	0,61	0,29	4882	23,5	21,5	0,44	0,22
Quadratic T	8934	26,2	18,3	0,93	0,28	8647	23,0	17,7	1,01	0,33
Cubic Asymetric	8071	25,4	26,4	0,43	-0,17	8238	23,0	20,2	0,53	0,20
Plateau	5890	27,9	23,0	0,58	0,28	5885	23,7	21,1	0,54	0,26
Gaussian	2392	23,1	26,4	0,20	-0,08	2814	22,5	22,1	0,06	-0,03
Asymetric U T	3359	26,2	25,0	0,06	0,03	3171	23,3	22,2	0,00	0,00
Hump	2495	22,0	26,5	0,37	-0,16	2913	20,9	23,4	0,38	-0,17
Double Hump	3456	24,2	27,7	0,40	-0,19	4123	22,6	22,3	0,09	-0,05

N: Número de eventos; d=diferencia media estandarizada de Z

6.4 Estudio OSTEOFRAC: Trabajo publicado en PLoS ONE

Use of oral bisphosphonates in primary prevention of fractures in postmenopausal women: a population-based cohort study. PLoS One. 2015;10(4):e0118178.

PLOS ONE

RESEARCH ARTICLE

Use of Oral Bisphosphonates in Primary Prevention of Fractures in Postmenopausal Women: A Population-Based Cohort Study

Jordi Real[1,2*‡], Gisela Galindo[1,3‡], Leonardo Galván[4‡], María Antonia Lafarga[5‡], María Dolors Rodrigo[5‡], Marta Ortega[6‡]

1 Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Lleida, Spain, **2** School of Medicine and Health Sciences, Universitat Internacional de Catalunya, Sant Cugat del Valles, Spain, **3** Primer de Maig Center, Institut Català de la Salut, Lleida, Spain, **4** Catalan Health Departament, Lleida, Spain, **5** Bordeta-Magraners Center, Institut Català de la Salut, Lleida, Spain, **6** Cappont Center, Institut Català de la Salut, Lleida, Spain

‡ JR and GG contributed equally to this work. LG, MAL, MDR and MO also contributed equally to this work.
* jreal.lleida.ics@gencat.cat (JR)

Citation: Real J, Galindo G, Galván L, Lafarga MA, Rodrigo MD, Ortega M (2015) Use of Oral Bisphosphonates in Primary Prevention of Fractures in Postmenopausal Women: A Population-Based Cohort Study. PLoS ONE 10(4): e0118178. doi:10.1371/journal.pone.0118178

Academic Editor: Alfonso Carvajal, Universidad de Valladolid, SPAIN

Received: July 14, 2014

Accepted: January 7, 2015

Published: April 10, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Oral bisphosphonates are first-line drugs in the treatment of osteoporosis under most guidelines, and have been shown to decrease risk of first fracture only in asymptomatic vertebral fractures and in clinical trial populations that are generally very different from the general population.

Objective

To compare incidence of first osteoporotic fracture in two cohorts of postmenopausal women, one treated with bisphosphonates and the other only with calcium and vitamin D.

Design

Retrospective population cohort study with paired matching based on data from electronic health records.

Setting

Women aged 60 years and older in 2005, from 21 primary care centers in a healthcare region of Spain.

Participants

Two groups of women aged 60 years and older (n = 1208), prescribed either calcium and vitamin D (CalVitD) or bisphosphonates (BIPHOS) with or without calcium and vitamin D, were compared for the end point of first recorded osteoporotic-related fracture, with 5-years follow-up.

Main Outcome Measure

Incidence of first fracture: Vertebral fracture, osteoporosis with pathological fracture, fracture of the upper humeral epiphysis, fracture of the lower radial epiphysis, or femur fracture.

Results

Estimated 10-year risk of fracture was 11.4% (95% confidence interval: 9.6 to 13.2), 11.8% (9.2 to 14.3) in the BIPHOS group and 11.1% (8.6 to 13.6) in the CalVitD group. No significant differences were found between groups in total fractures (Hazard ratio = 0.934 (0.67 to 1.31)) or location (vertebral, femoral, radial or humeral).

Conclusions

In postmenopausal women, bisphosphonates have not been shown to better decrease risk of first fracture compared with calcium and vitamin D therapy alone.

Introduction

Osteoporosis is clinically characterized by a loss of bone mass and changes in bone structure that cause fragility and contribute to the appearance of fractures, mainly of the vertebrae, femoral neck, and wrist [1]. The condition began to be defined in the 1990s, coinciding with the development of densitometry, and since then has been classified as a disease [2].

In 1994, a World Health Organization report classified women as healthy or diseased according to their bone mineral density (BMD) value, comparing them with an average 30-year-old woman [3]. This led to classify many healthy women as having osteoporosis and starting drug therapies in women who were not at risk of future fractures.[4] At present, a decline in BMD is considered a risk factor, not an indication of the disease, and patients whose only symptom is low BMD, determined by computed tomography (CT) scan, are not labelled as having osteoporosis [2].

In clinical practice, it is important to identify patients with a high risk of fracture and decide who should be treated and how [5,6]. In daily practice, however, decision-making is difficult because of many uncertainties, heterogeneity in clinical guidelines published by the various scientific societies [7], and even differences among doctors in the same country and medical specialty [8]. To decrease this variability, tools have been introduced to estimate the risk of future fractures, taking into account the various risk factors; the two main scales are FRAX [9], and QFRACTURE [10,11]. Both scales incorporate history of fracture, family history of hip fracture, underweight (BMI<18.5 kg/m2), smoking, alcohol consumption, and glucocorticoid treatment.

Of the available treatment options, bisphosphonates have the longest track record, have been the most studied, and are the least expensive drug choice. Meta-analysis of the different bisphosphonates has repeatedly shown a decline in new fractures among postmenopausal women in secondary prevention, defined as women with previous fracture and women without fractures and at least 2 SD values below the peak bone mass or older than 62 years when these data were not available. However, no treatment benefit has been observed in primary prevention except in the case of asymptomatic morphometric spinal fractures in women taking alendronate [12–14].

In the general population and in our setting, few studies have analysed the impact of osteoporosis treatments. One of these, an ecological study in Galicia by Guerra-García, observed that the number of units of anti-resorptive agents dispensed by pharmacies nearly doubled from 2004 to 2008 but there was no decline in the number of femoral fractures, which are the most serious osteoporotic fractures and have the worst consequences for patients [15]. Another ecological study using data from all the Spanish public health system detected a slight decrease between 2002–2008 years of adjusted hip fracture rates in women over 50 years (from 4.1 to 3.91 per 1000) in contrast with the sharp increase in the bisphosphonates consumption, multiplied by 5, in the same time period [16]. The 10-year cost of avoiding one hip fracture ranges from 54,134 to 84,287 euros with alendronate and 67,853 to 173,748 euros with risedronate treatment [17].

In daily clinical practice, anti-resorptive therapy is often prescribed as primary prevention in women younger than 60 years. Sanfélix-Gimeno commented on "the peculiar panorama" of osteoporosis management in our country, where excessive diagnostic testing is ordered and the treatment pattern is to prescribe anti-resorptive drugs and calcium plus vitamin D supplements more frequently for younger women with low risk than for older patients with high risk [18].

In 2009, a meta-analysis of oral bisphosphonates use in women older than 65 years showed a 24% reduction in osteoporotic fracture risk, a lower benefit than has been indicated in some clinical trials and highly associated with treatment adherence [19]. Another study identified an increased risk of atypical femoral fractures with this therapy [20]. Nonetheless, the results of these studies cannot be generalized to the population treated in usual clinical practice. It has been estimated that only 20% of the participants in randomized clinical trials are representative of the general population with osteoporosis [21].

Although oral bisphosphonates are first-line drugs for osteoporosis treatment under most guidelines [22,23], it is not clear that the associated reduction in the risk of first osteoporotic fracture is cost-effective in the general population. Maintaining long-term anti-resorptive therapy when its effectiveness is doubtful is a waste of resources. In addition, one should be very cautious in primary prevention because interventions have secondary effects; these must have highly conclusive evidence of effectiveness and long-term safety because they target large segments of the population and healthy individuals [24].

Despite the high social and healthcare impact of osteoporosis, the efficiency of the drugs most commonly prescribed in Spain for the prevention of osteoporotic fractures has not been sufficiently evaluated [25]. The aim of the present study was to estimate the incidence of first osteoporotic fracture in a cohort of postmenopausal women treated with bisphosphonates, compared with women treated only with calcium and vitamin D, using a population database of retrospective clinical records and 5-year follow-up.

Material and Methods

Study population

The study was carried out in a cohort of women aged 60 years and older assigned to any of the 21 healthcare centres in the Health Region of Lleida (HRL) belonging to the Spanish National Health Service, with universal coverage during the studied period. In 2005, the HRL covered a total population of 360,489, of which 42,234 were women aged 60 years and older.

Design

We designed a retrospective population cohort study with 5-year follow-up, matching two cohorts by clinical characteristics and drugs taken, based on the HRL database of electronic health records. The research potential these data provide for population studies has been previously

described [26]. All patients registered in the HRL database who were at least 60 years old at the time of inclusion and taking calcium, vitamin D, and/or bisphosphonates under their doctor's prescription were included in the study. The date of treatment initiation was considered the date that a pharmacy dispensed the first prescription (index date), according to the official pharmaceutical database (Fig. 1).

Exclusion criteria (Table 1 and Fig. 2) included previous treatment (before index date at 2005) with drugs that modify bone metabolism (bisphosphonates and/or calcium, vitamin D, oestrogens, calcitonin, parathyroid hormone, strontium ranelate, or raloxifene); known history of osteoporotic fracture, kidney failure, Paget disease, or multiple myeloma; enrolment in the HRL database after 2002; and lack of contact with their HRL doctor during the follow-up period (2005–2010).

Study participants were divided into two groups (Fig. 2): BIPHOS, consisting of the women who had retrieved a prescription from their pharmacy for bisphosphonates, with or without calcium and vitamin D, and CalVitD, the control group of women who had only taken calcium and vitamin D.

Data sources

Drug information was obtained from the HRL Pharmacy Unit, which has collected data on all HRL prescriptions dispensed by pharmacies since 2002. Primary care centres managed by the Catalan Institute of Health provide free universal healthcare to 95% of the population of this HRL; during the study period, pharmaceuticals were also provided free of charge to patients older than 65 years and at a 60% subsidy to younger patients.

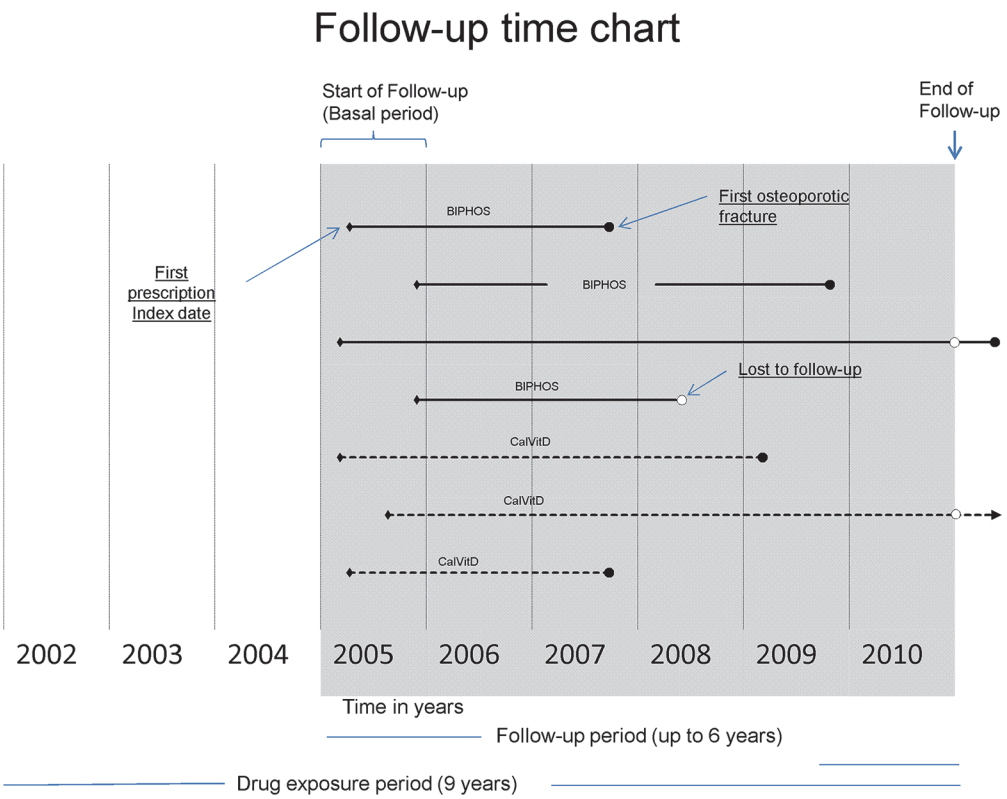


Fig 1. Follow-up time chart.

doi:10.1371/journal.pone.0118178.g001

Table 1. Study variables.

Variable type	Codes: Source	
Definition (Time)	ICD-10: Primary care	ICD-9: Hospital
Primary outcome (between 2005 and 2010)		
Vertebral fracture: fatigue / unspecified	M48.4 / T08	733.13/ 805.xx
Osteoporosis with pathological fracture	M80	733.11–733.19
Fracture of the upper humeral epiphysis	S42.2	812.xx/ 733.11
Fracture of the lower radial epiphysis	S52.5	813.42/ 813.52/733.12
Fracture of the lower radial epiphysis	S52.5	813.42/ 813.52/733.12
Femur fracture	S72	820.xx-821.xx/733.14/733.15
Comorbidity, basal at index date		
Diabetes mellitus	E10-E14	
Behavioural syndromes associated with physiological disturbances and physical factors	F50-F59	
Extrapyramidal and movement disorders	G20-G26	
Episodic and paroxysmal disorders	G40-G47	
Hypertensive diseases	I10-I15	
Ischemic heart diseases	I20-I25	
Cerebrovascular diseases	I60-I69	
Chronic lower respiratory diseases	J40-J47	
Inflammatory polyarthropathies	M05-M14	
Arthrosis	M15-M19	
Exclusion diseases (at index date)		
Renal failure	N17-N19	
Paget's diagnoses	M88	
Multiple myeloma	C90.0	
Drugs	**ATC: Pharmacy**	
Primary		
Bisphosphonates, with/without Calcium	M05BA, M05BB	
Calcium + vitamin D	A11CC A12AA A12AX	
Secondary drugs		
Antidepressants	N06AA, N06AB,N06AG,N06AX	
Proton-pump inhibitors	A02BC	
Glucocorticoids	H02AB	
Benzodiazepine	N05AH, N05AL, N05BA, N05CD NC5CD, N05CF	
Opiates	N02AA,N02AB,N02AC,N02AE,	
Exclusion (prior index date)		
Osteoporosis drugs	M05BX, H05AA, H05BA	
Other hormone therapies	H03AA, L02AE,G03XC,L02BG	
Barbiturates, antipsychotic and antiepileptic drugs	N03AA, N03AB, N03AD, N03AE, N03AF, N03AG, N03AX,N05AA, N05AB,N05AC,N05AD,N05AE, N05AF,N05AG,N05AN,N05AX,	
Antihistamines	R06AA, R06AB, R06AC, R06AE, R06AE, R06AX	
Oestrogens	G03CA, G03FA, G03FB	

ATC: Anatomical, Therapeutic, Chemical classification system; ICD-9 or ICD-10: International Statistical Classification of Diseases, version 9 or version 10.

doi:10.1371/journal.pone.0118178.t001

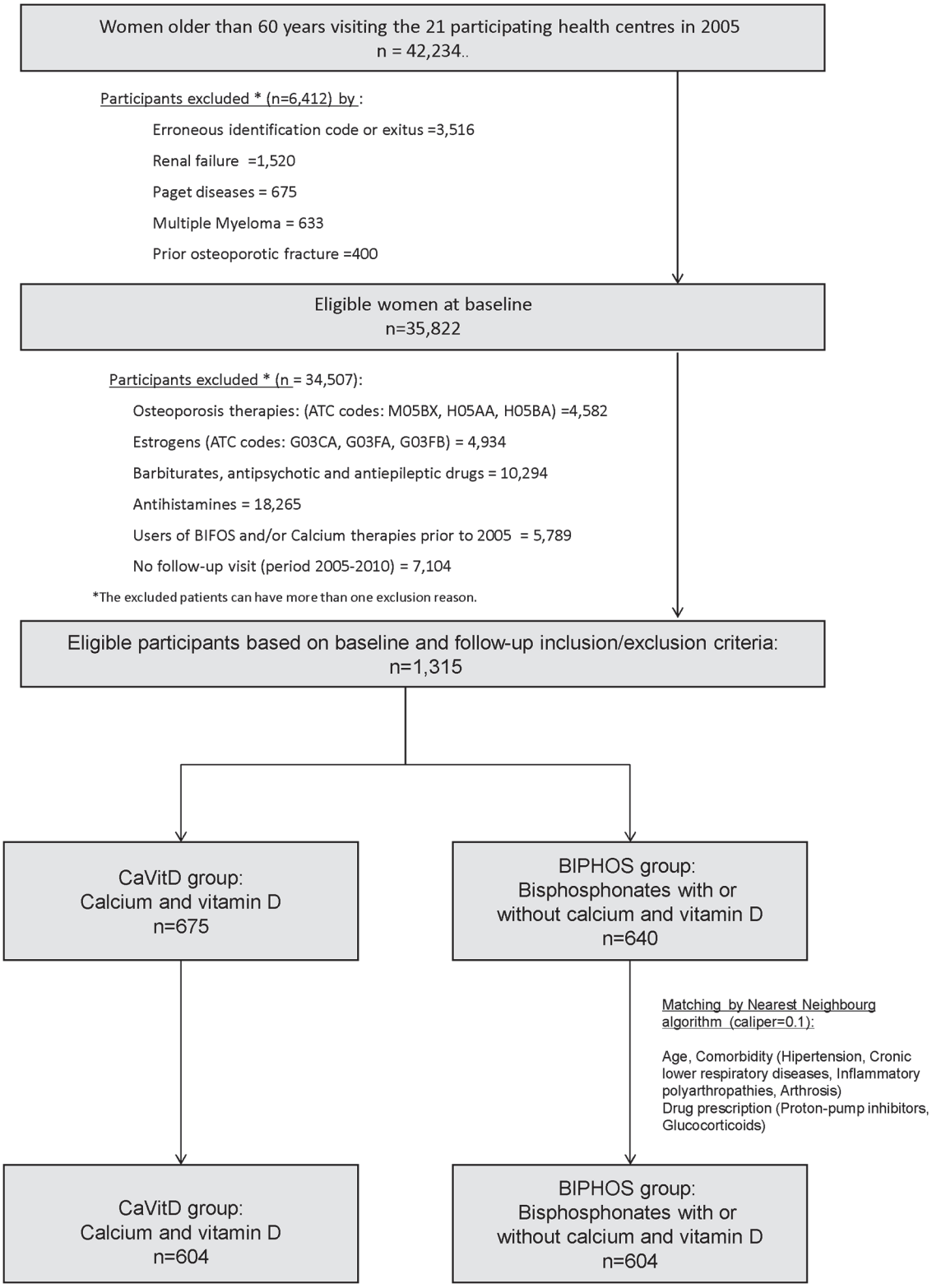


Fig 2. Flow chart showing the participants' selection process.

doi:10.1371/journal.pone.0118178.g002

Baseline information about fractures and co-morbidities was extracted from two sources: the primary care system's database of electronic health records and the hospital discharge databases of the HRL's two referral hospitals. The pharmacy database was cross-referenced with the diagnostic records of the HRL's two hospitals and the primary care centres to obtain baseline and follow-up information.

Ethical aspects

We carefully respected all the Helsinki Declaration criteria. Since this was an observational study, participants underwent no interventions other than usual clinical care. Information from clinical records was correctly anonymized before analysis in order to preserve the participants' confidentiality, in accordance with Spanish law (Ley Orgánica 15/1999, Protección de Datos de Carácter Personal). The study protocol was approved by Clinical Ethics Committee of the Primary Healthcare University Research Institute IDIAP-Jordi Gol (P11/85) (S1 Fig).

Sample, matching process, and statistical power

The two study groups were matched to ensure balance in terms of basal comorbidities, age, and use of other drugs that modify bone metabolism (Table 2, Fig. 3). Matching was done by the "Nearest Neighbour algorithm" (caliper = 0.1), using the "MatchIt" library of the R (v3.0.1) statistical package [27,28]. The Nearest-Neighbour matching algorithm was employed to find as many matches between groups based on the propensity scores to produce two balanced patient cohorts. The distance was created with the link logit according the following variables: Age, Comorbidity (Hypertension, Chronic lower respiratory diseases, Inflammatory polyarthropathies, Arthrosis), and Drugs prescription (Proton-pump inhibitors and Glucocorticoids).The final matched sample included 1208 women, 604 per group; after the matching process, the potential selection bias between the two samples (total vs. matched) was reduced by 69%. The selection bias reduction was computed according the overall difference between the matched sample and pre-matched sample regarding the sum of relative differences (between exposure groups) in the variables represented in Fig. 3. Assuming a minimum absolute risk reduction of 5% (15% vs. 10% incidence)[14] in a sample of 1208 women, we obtained a statistical power approximated of 91%, with an alpha level of 0.05 and standard deviation of random effect at cluster level of 0.9 (according our data analysis) using logistic regression test with sandwich robust standard (This approximation was performed with R simulation code done by Arnold B.F., 2011 [29].

Length of follow-up

Time free of fracture was defined as the time between the first dispensation of medication in 2005 until the first fracture recorded by the primary care doctor, or hospital admission or urgent care visit for fracture, or abandonment. Reasons for abandoning the study (lost to follow-up) were death, change of address, or final medical contact in the HRL's records before 31 December 2010 (Figs. 1 and 2).

Primary outcomes

The primary event was defined as first fracture. Osteoporosis-related fracture diagnoses were selected. They were coded according to the International Classification of Diseases (ICD-10 or ICD-9): Fracture of femur; Osteoporosis, pathological fracture, Fatigue fracture of vertebra; Fracture of lower end of radius; and Fracture of upper end of humerus (Table 1).

Table 2. Baseline characteristics of the participants according to the group.

Variable		CalVitD (n = 604)		BIPHOS (n = 604)		p value
	Category	n	(%)	n	(%)	
Age group						0.236[a]
	60–64	113	(18.7)	98	(16.2)	
	65–74	230	(38.1)	259	(42.9)	
	>74	261	(43.2)	247	(40.9)	
Age	Mean ± SD	73.5	±8.6	73.1	±7.9	0.384[b]
BMI	Mean ± SD	29.1	±4.3	28.8	±4.1	0.903[b]
Comorbidity						
	Diabetes mellitus	94	(15.6)	83	(13.7)	0.377[a]
	Behavioural syndromes associated with physiological disturbances and physical factors	6	(1.0)	4	(0.7)	0.530[a]
	Extrapyramidal and movement disorders	9	(1.5)	8	(1.3)	0.808[a]
	Episodic and paroxysmal disorders	9	(1.5)	6	(1.0)	0.409[a]
	Hypertensive diseases	277	(45.9)	273	(45.2)	0.814[a]
	Ischemic heart diseases	23	(3.8)	14	(2.3)	0.133[a]
	Cerebrovascular diseases	17	(2.8)	11	(1.8)	0.262[a]
	Chronic lower respiratory diseases	36	(6.0)	32	(5.3)	0.602[a]
	Inflammatory polyarthropathies	16	(2.6)	19	(3.1)	0.579[a]
	Arthrosis	75	(12.4)	74	(12.3)	0.927[a]
Drugs, number of boxes dispensed (mean ± SD)						
	Antidepressants	15.9	±32.5	14.7	±31.3	0.525[c]
	Proton-pump inhibitors	24.4	±31.6	25.3	±31.1	0.655[c]
	Glucocorticoids	3.4	±11.5	3.7	±11.2	0.669[c]
	Benzodiazepine	31.6	±51.5	33.6	±49.1	0.496[c]
	Opiates	6.0	±28.0	6.0	±19.0	0.995[c]

n: Frequency; SD: Standard deviation; P value computed using: Univariate logistic regression with robust standard errors (a); Mixed linear regression by pairs(b) and Mann-Whitney test (c).

doi:10.1371/journal.pone.0118178.t002

Drug exposure

For each patient, the number of boxes of medication dispensed with HRL prescriptions, mainly oral bisphosphonates, calcium, and vitamin D, was calculated from initiation to the last date of follow-up. Drugs were coded according to the Anatomical Therapeutic Chemical Classification System (ATC). Table 1 presents the remaining drug variables as well as the other co-variables analysed.

Statistical methods

Initially, basal characteristics of both groups were evaluated to establish homogeneity in age, comorbidities, and exposure to other drugs that modify bone metabolism. The incidence of fracture and accumulated risk of fracture after five years was calculated for each group (BIPHOS vs. CalVitD). To evaluate time-related incidence curves, we performed Cox regression models. Risk functions and hazard ratios (HR) with their 95% confidence interval (95% CI) were estimated to compare the BIPHOS group to the CalVitD group. The 95% confidence intervals and p values was computed with robust standard errors to account the matched sample. The models were constructed using co-variables that were clinically adjusted and/or

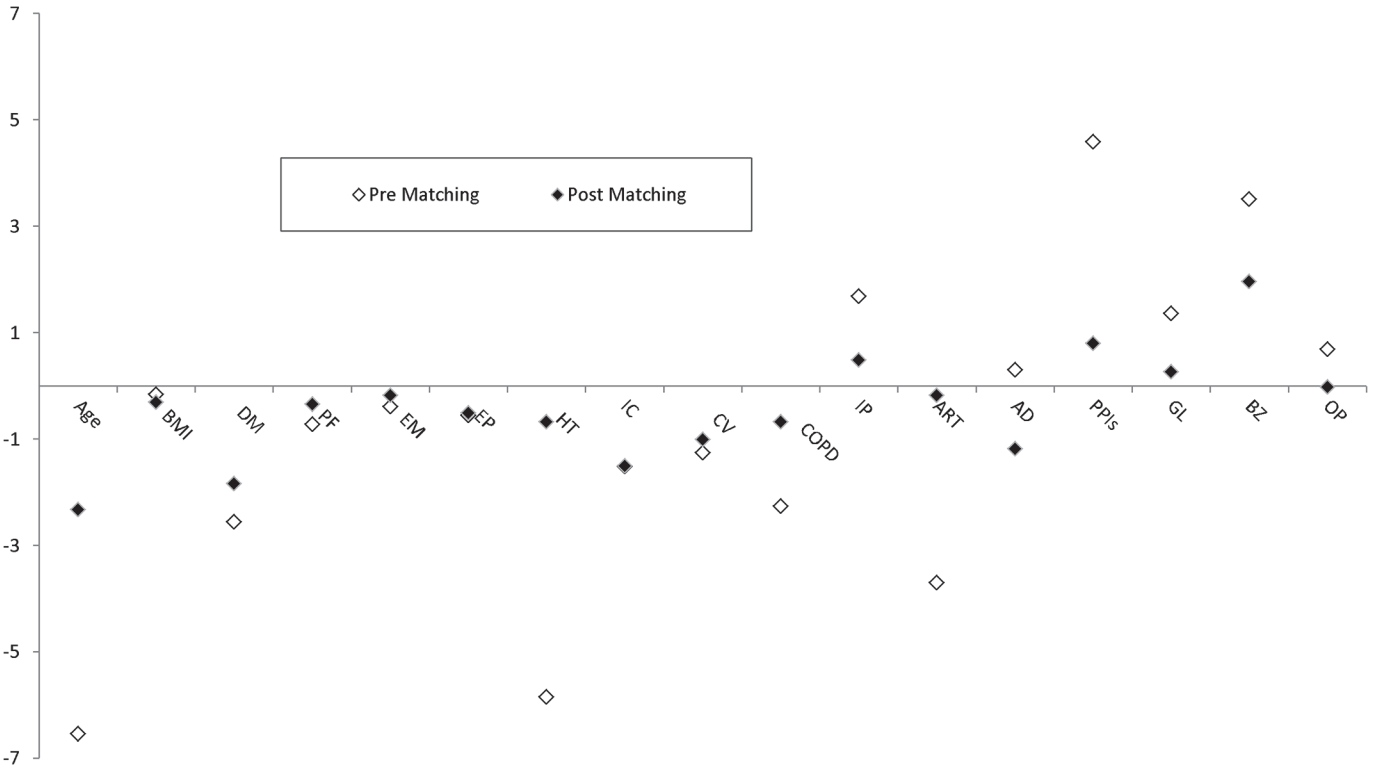


Fig 3. Absolute differences between groups (BIPHOS-CalVitD) pre and post matching. BMI: Body mass index; DM: Type II diabetes mellitus; PF: Behavioural syndromes associated with physiological disturbances and physical factors; EM: Extrapyramidal and movement disorders; EP: Episodic and paroxysmal disorders; HT: Hypertension; IC: Ischaemic heart diseases; CV: Cerebrovascular disease; COPD: Chronic lower respiratory diseases; IP: Inflammatory polyarthropathies; ART: Arthrosis; AD: Antidepressants; PPIs: Proton-pump inhibitors; GC: Glucocorticoids; BZ: Benzodiazepine; OP: Opiates.

doi:10.1371/journal.pone.0118178.g003

statistically associated with fracture risk. We evaluated goodness-of-fit and the Cox model's proportional risk assumption, as well as the interactions at different levels of exposure to each drug, using the Schoenfeld residual analysis. A secondary analysis of the sensitivity of the estimated HR for drug exposure levels was carried out, considering the number of boxes of medication collected at the pharmacy: low ($\leq$12), moderate (13–36), and high (>36). The stability and consistency of the models was evaluated using various subsamples of patients whose doctors meet high-quality standards for data entry in the medical records system (22% of the sample). This quality sample (SIDIAP-Q database) minimizes the risk of global bias in epidemiological studies and improves representativeness, as previously published in a validation study[30]. Statistical significance was established as a p-value < 0.05. Data management and analysis was done with the SPSS (v17) and STATA v11-IC statistical packages.

Results

Sociodemographic characteristics of the study participants are shown in Table 2; there were no significant differences between the two study groups. At the beginning of follow-up, the mean age of participants was 73.3 years (SD = 8.3) and body mass index (BMI) was about 29 (SD = 4.2). The most prevalent pathology was hypertension (45%), followed by diabetes (14.7%).

During a mean follow-up of 4.87 years, 138 fractures were recorded, representing an accumulated risk after 5 years of 11.4% (95% confidence interval 9.6 to 13.2%). Half of the fractures were of the femur (50.7%), followed by vertebral fractures (27.5%), unspecified osteoporotic fractures (23.9%), and fractures of the humerus or radius (19.5%).

Table 3. Frequency and fracture risk according to study group and basal comorbidity.

Variable	Risk at 5 years			
	Fractures		CI 95%	
Category	n	(%)	(L inf to L Sup)	p value[a]
Group				
CalVitD: Calcium + Vitamin D	71	(11.8)	(9.2 to 14.3)	0.710
BIPHOS: Bisphosphonate	67	(11.1)	(8.6 to 13.6)	
Exposure level: Number of boxes dispensed				
Low: ≤12				
CalVitD: Calcium + Vitamin D	60	(12.9)	(9.9 to 16.0)	0.604
BIPHOS: bisphosphonate	36	(11.7)	(8.1 to 15.3)	
Moderate: 13–36				
CalVitD: Calcium + Vitamin D	5	(5.4)	(0.8 to 10.1)	0.017
BIPHOS: bisphosphonate	21	(15.9)	(9.7 to 22.1)	
High: >36				
CalVitD: Calcium + Vitamin D	6	(12.5)	(3.1 to 21.9)	0.155
BIPHOS: bisphosphonate	10	(6.1)	(2.4 to 9.8)	
Age group at baseline				
60–64	8	(3.8)	(1.2 to 6.4)	<0.001
65–74	40	(8.2)	(5.8 to 10.6)	
>74	90	(17.7)	(14.4 to 21.0)	
Body mass index at baseline				
≤ 25.00	23	(13.1)	(8.1 to 18.0)	0.482
25.01–30.00	75	(11.9)	(9.4 to 14.4)	
30.01+	39	(9.9)	(6.9 to 12.8)	
Basal comorbidity				
Diabetes mellitus	23	(13.0)	(8.0 to 17.9)	0.465
Behavioural syndromes associated with physiological disturbances and physical factors	5	(50.0)	(19.0 to 81.0)	<0.001
Extrapyramidal and movement disorders	3	(17.6)	(-0.5 to 35.8)	0.417
Episodic and paroxysmal disorders	3	(20.0)	(-0.2 to 40.2)	0.265
Hypertensive diseases	84	(15.3)	(12.3 to 18.3)	<0.001
Ischemic heart diseases	5	(13.5)	(2.5 to 24.5)	0.688
Cerebrovascular diseases	5	(17.9)	(3.7 to 32.0)	0.282
Chronic lower respiratory diseases	9	(13.2)	(5.2 to 21.3)	0.631
Inflammatory polyarthropathies	6	(17.1)	(4.7 to 29.6)	0.297
Arthrosis	12	(8.1)	(3.7 to 12.4)	0.202

CI 95%: Confidence Interval with 95%; n: Fracture frequency

a: p value computed using univariate logistic regression with robust standard errors by clusters (pairs)

doi:10.1371/journal.pone.0118178.t003

During their fracture-free period, the CalVitD cohort received a mean 11 (SD = 20) boxes of calcium and vitamin D and the BIPHOS cohort received 11.6 (SD = 16.4) boxes, in addition to a mean 22.6 (SD = 24.4) boxes of bisphosphonates.

Table 3 shows the cumulative incidence of fracture at 5-year follow-up by study group, stratified by level of drug exposure and by the co-variables studied. Accumulated risk was 11.8% among women in the BIPHOS group and 11.1% in the CalVitD group (no significant difference between groups). Among women with moderate drug exposure, the BIPHOS group had significantly higher accumulated fracture than the controls. The remaining co-variables –older age,

Table 4. Crude risk fracture and adjusted by groups and baseline comorbidity.

Model Variable Category[b]	Hazard Ratio	(95% CI)	p value[a]
Crude			
Bisphosphonate group (Ref: CalVitD)	0.899	(0.64 to 1.26)	0.532
Adjusted			
Bisphosphonate group (Ref:CalVitD)	0.934	(0.67 to 1.30)	0.687
Age (in years)	1.075	(1.05 to 1.10)	<0.001
Body mass index	0.953	(0.91 to 1.00)	0.034
Comorbidity (Ref = No)			
Hypertensive diseases	1.705	(1.20 to 2.42)	0.003
Cerebrovascular diseases	1.323	(0.54 to 3.23)	0.538
Arthrosis	0.641	(0.34 to 1.21)	0.169
Drug therapies at baseline (1 year previous)			
Proton-pump inhibitors	0.962	(0.66 to 1.41)	0.844
Glucocorticoids	1.234	(0.74 to 2.06)	0.422
Antidepressants	1.159	(0.77 to 1.75)	0.482
Benzodiazepine	0.838	(0.58 to 1.20)	0.334
Opiates	1.396	(0.89 to 2.20)	0.149
Specific fractures			
Femur			
Crude			
Bisphosphonate group (Ref: CalVitD)	0.742	(0.46 to 1.20)	0.227
Adjusted			
Bisphosphonate group (Ref: CalVitD)	0.735	(0.45 to 1.21)	0.228
Upper humeral or lower radial epiphysis			
Crude			
Bisphosphonate group (Ref: CalVitD)	0.867	(0.40 to 1.87)	0.716
Adjusted			
Bisphosphonate group (Ref: CalVitD)	0.883	(0.41 to 1.91)	0.752
Vertebral fracture: fatigue or unspecified			
Crude			
Bisphosphonate group (Ref: CalVitD)	1.256	(0.66 to 2.38)	0.484
Adjusted			
Bisphosphonate group (Ref: CalVitD)	1.405	(0.82 to 2.42)	0.219

a: p values, and Confidence Interval according the Cox regression model with robust standard errors

b: Reference category is CalVitD group, and not presence of comorbidity; 95% CI: Confidence Interval with 95%.

doi:10.1371/journal.pone.0118178.t004

lower BMI, and a basal hypertension diagnosis–were significantly associated with increased risk of fracture.

Table 4 shows the crude and adjusted Hazard Ratio (HR) for the specific risk of femoral, radial-humeral, and vertebral fractures. None of the models detected significant differences in fracture risk between groups. The BIPHOS group had a slightly –but not significantly–lower global risk of fracture (HR_{crude} = 0.899 / HR_{adj} = 0.934). In the analysis by level of exposure, women with moderate use of bisphosphonates (13–36 units over 1–3 years) had a higher global risk of fracture (HR = 3.0; 95%CI: 1.13 to 7.9). With respect to fracture typology, the BIPHOS group had a lower risk of femoral fracture (HR_{adj} = 0.73; 0.45 to 1.21) and higher risk of

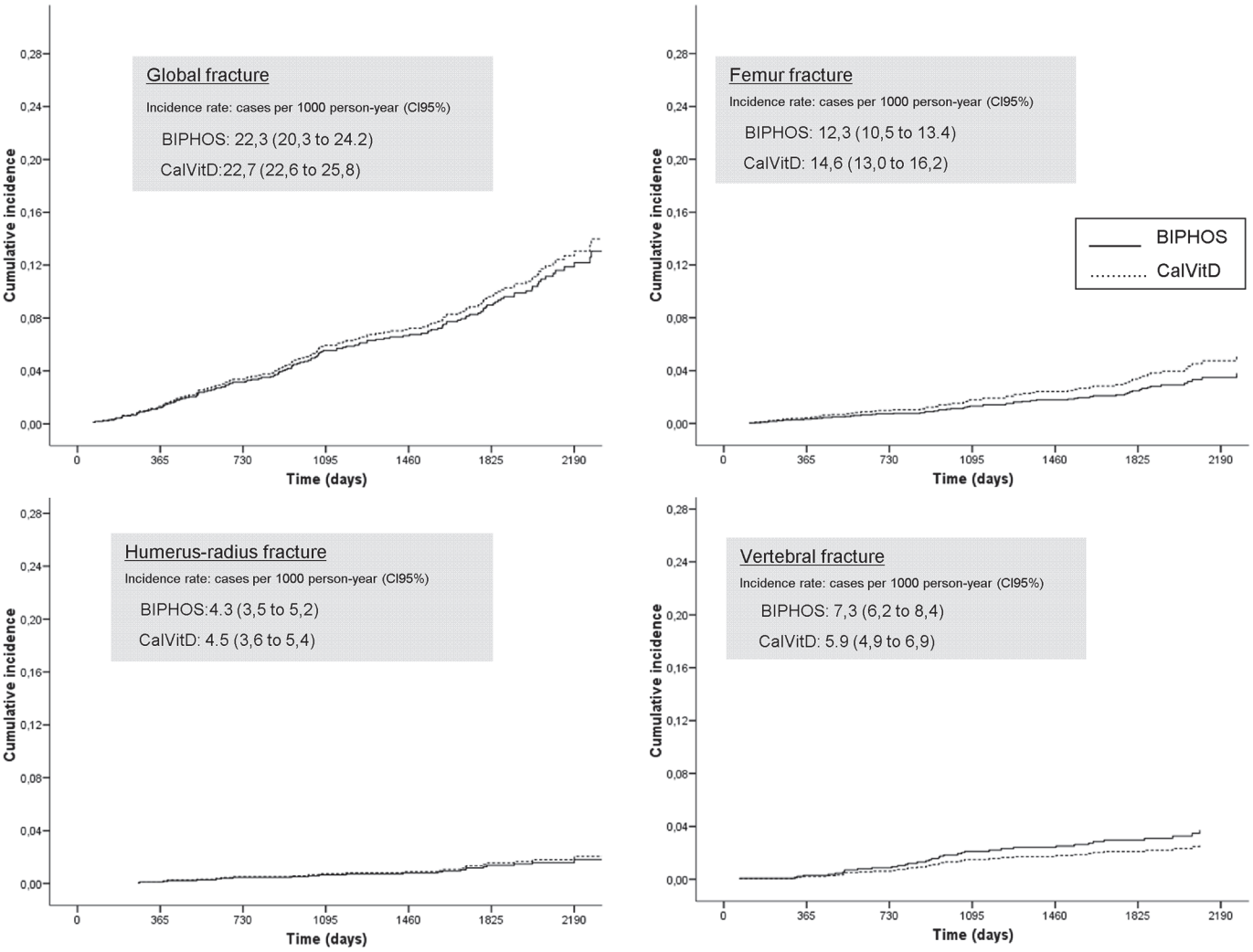


Fig 4. Incidence curves of fracture during 5 years follow-up by group (BIPHOS / CalVitD).

doi:10.1371/journal.pone.0118178.g004

vertebral fracture (HR_{adj} = 1.40; 0.82 to 2.42); again, none of these differences achieved significance (p-value > 0.05).

Fig. 4 shows the incidence curves of global and site-specific fractures, adjusted by basal characteristics, for both study groups. In humeral or radial long bones, the curves are practically superimposed; in the femur, the CalVitD curve is slightly higher (but non-significant) and in vertebral fractures the BIPHOS curve is higher.

Discussion

In this study, which included 1208 women aged 60 years and older without previous fracture who were receiving drug therapy as primary prevention, the cumulative incidence of fracture during five years of follow-up was 11.4% (9.6 to 13.2%). One in 10 participants presented with some type of fracture during the study period, half of them femoral fractures; this corresponds to a 10-year incidence of 22 osteoporosis-related fractures per 100 women.

The two cohorts were selected so that the only distinguishing feature was whether or not they were taking bisphosphonates. They were of comparable age, BMI, comorbidities associated with

fracture risk, and use of other prescription drugs. Under these conditions, we did not find any decreased risk of fracture between the groups during the five years of follow-up.

The appearance of a first fracture was associated with advanced age (the main risk factor described in the literature), lower BMI, and a history of hypertension[23]. The first two are well known, particularly in secondary prevention, and are included in the FRAX and QFracture risk equations[9,10]. The hypertension association could be due to the relationship between use of hypertension medications and increased falls, especially at treatment initiation and in elderly patients [31]. We did not find any risk reduction with respect to the site of fractures (femoral, vertebral, humeral neck or head), as shown in Fig. 4.

The FIT-2 study, one of the few clinical trials with a large group of postmenopausal women (n = 4432, mean age, 68 years), compared alendronate with placebo in primary prevention. At four years of follow-up, differences in clinical (vertebral and non-vertebral) and hip fractures did not reach significance. However, a significantly lower incidence of radiological vertebral fractures was observed in the group treated with alendronate [32].

A high percentage of asymptomatic spinal fractures, also called morphometric fractures, occur; these are usually found during exploratory exams and have limited impact on quality of life [31]. Under real-life conditions, asymptomatic fractures are likely under-diagnosed and under-reported in the electronic health records, which would help to explain why no differences were observed in our study.

Given the findings about bisphosphonates treatment in primary prevention, can we justify the investment in preventing only one type of fractures that are asymptomatic and have limited impact on quality of life? We cannot forget the secondary effects of any preventive effort, especially if drug therapy is involved. The use of bisphosphonates has been associated with increased risk of mandibular osteonecrosis, osteoarticular pain, atrial fibrillation, sub-trochanteric and diaphyseal fractures of the femur, and esophagitis [33].

In concordance with the present results, the Catalan Agency for Healthcare Quality and Assessment (*Agencia per la Qualitat i Avaluació Sanitàries Catalana*) advises against the use of bisphosphonates in postmenopausal women with low risk of fracture because the benefits do not exceed the risks[34]. In addition, the 2010 report from Spain's Agency for the Evaluation of Healthcare Technologies (AETS, *Instituto de Salud Carlos III*) concluded that, assuming partial adherence to treatment, none of the drug interventions evaluated in comparison with calcium plus vitamin D or placebo obtained acceptable cost-utility outcomes if treatment was initiated before 69 years of age [25].

In 2010, the United Kingdom's Secretary of State issued a report that analysed variations in drug uses in 14 countries. Spain was first on the list in use of osteoporosis medications, even though it had one of the lowest levels of osteoporotic fracture risk [35].

Various scientific societies agree on eradicating low-value clinical practices, such as systematic primary prevention CT scans of postmenopausal women without other risk factors (*Compromiso por la calidad de las Sociedades Científicas*, Ministry of Health, Social Policy and Equality, Spain), [36] in women younger than 65 years without other risk factors (Choosing Wisely, American Academy of Family Physicians) [37], or without evaluating risk using the FRAX or QFracture equations (Do not do, UK National Institute of Health and Care Excellence (NICE) [38].

With the current trend of population aging, the incidence of fractures associated with osteoporosis is expected to increase. Prevention is an objective for all healthcare systems, focused mainly on hip fractures and followed by clinical vertebral fractures because of their implications for morbidity, mortality, and quality of life. In any case, any preventive interventions must be undertaken with caution because, as stated above, they target healthy individuals and large segments of the population [22].

Strengths and limitations

We are aware that our study has several potential limitations that preclude us from providing definitive evidence of the absence of benefit or the association between bisphosphonates use and the risk of fracture in comparison with calcium and vitamin D. Clinical records from a large population database bring with them implicit biases related to under-reporting. To minimize this limitation, in addition to the fractures included in the primary care electronic records we included those that caused a visit to urgent care or a hospital admission in the region's two reference hospitals. The availability of this large population database (Fig. 2) allowed us to set very high standards for the final participant selection, eliminating potential selection biases and allowing the inclusion of homogeneous cohorts. We included only current users of the healthcare system (i.e., active records), evidenced by follow-up visits, and without a known history of fracture or previous use of osteoporosis treatment or oestrogens. Finally, strict matching was performed to construct balanced study groups (Table 2 and Fig. 3), which reduced the potential selection bias by 69%.

The total sample used for analysis consisted of 1208 women. This sample size might not be sufficient to detect actual reductions in the incidence of fractures observed in our study (<7% over 5 years: HR = 0.934), but it has sufficient power to detect a reduction of 20% or more in the risk of fracture, as reported in the literature [32].

We know that collecting drugs from a pharmacy with a doctor's prescription does not provide a precise measurement of the use of the medication, although we assumed that this was evidence of treatment adherence.

An important strength of our study is that the evaluation of the potential effectiveness of bisphosphonates use was done under actual conditions of daily clinical practice, unlike the structure of clinical trials [21]. Furthermore, the availability of a subsample of electronic health records in a database that has been validated for high-quality coding by clinicians reinforces the validity of our results.

Conclusions

In our study, postmenopausal women obtained no benefit from primary prevention with bisphosphonates in reducing their 5-year risk of first osteoporosis-related fracture, compared to treatment only with calcium plus vitamin D. There was also no risk reduction according to fracture site (femur, vertebrae, or humeral head or neck).

If bisphosphonates use is not shown to have better outcomes than calcium plus vitamin D, primary prevention strategies shall be reconsidered and one should stop doing what is not effective. Medical societies must work together to unify their criteria, reduce the inappropriate use of tests and treatments [18], and use BMD scores appropriately [33]. The availability of an objective measurement simplifies clinical decision-making, which is probably the reason BMD has been used as the measure indicating the prescription of osteoporosis therapy and as a diagnostic method in women younger than 60 years.

In clinical practice, emphasis should be focused to improvement secondary prevention and to identification of those patients at high risk and, therefore, would benefit from a primary prevention therapy. It may be stressed that BMD, per se, is simply a risk factor to be considered, not the identifier by which patients should be selected for treatment.

Supporting Information

S1 Dataset. Data set in Stata format (dta). Data set from 1208 records including: Group, Follow-up time, Fracture, Fracture femur, Fracture of humeral/radial, Fracture vertebral, Propensity distance group, Age, BMI, Hypertensive diseases, Cerebrovascular diseases, Arthrosis,

Proton-pump inhibitors, Glucocorticoids, Antidepressants, Benzodiazepine, Opiates, BMI Group, Age group, Diabetes mellitus, Behavioural syndromes associated with physiological disturbances, Extrapyramidal and movement disorders, Episodic and paroxysmal disorders, Ischaemic heart diseases, Chronic lower respiratory diseases, Inflammatory polyarthropathies, Exposure level.
(DTA)

S1 Fig. Statement from the clinical investigation ethics.
(PDF)

S1 Table. STROBE Statement—Checklist of items that should be included in reports of cohort studies.
(DOCX)

Acknowledgments

We thank Montserrat Rue for their helpful comments on an earlier version of this manuscript.

Author Contributions

Conceived and designed the experiments: GG JR MAL. Analyzed the data: JR LG. Wrote the paper: MO MDR GG JR.

References

1. Pages-Castella A, Prieto Alhambra D. Degenerative osteoarthritis, osteoporosis and fractures: Controversies and evidences. Med Clin (Barc). 2013; 141: 217–220. doi: 10.1016/j.medcli.2013.01.036 PMID: 23540390
2. Vazquez M. Osteoporosis: The crisis of a paradigm. Med Clin (Barc). 2010; 134: 206–207. doi: 10.1016/j.medcli.2009.10.009 PMID: 19939415
3. WHO Study Group. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. report of a WHO study group. World Health Organ Tech Rep Ser. 1994; 843: 1–129. PMID: 7941614
4. Pressman A, Forsyth B, Ettinger B, Tosteson AN. Initiation of osteoporosis treatment after bone mineral density testing. Osteoporos Int. 2001; 12: 337–342. doi: 10.1007/s001980170099 PMID: 11444079
5. Azagra R, Roca G, Encabo G, Aguye A, Zwart M, et al. FRAX(R) tool, the WHO algorithm to predict osteoporotic fractures: The first analysis of its discriminative and predictive ability in the spanish FRIDEX cohort. BMC Musculoskelet Disord. 2012; 13: 204-2474-13-204. doi: 10.1186/1471-2474-13-204
6. Kanis JA, Johnell O, Oden A, Sembo I, Redlund-Johnell I, et al. Long-term risk of osteoporotic fracture in malmo. Osteoporos Int. 2000; 11: 669–674. PMID: 11095169
7. Sanfelix-Genoves J, Catala-Lopez F, Sanfelix-Gimeno G, Hurtado I, Baixauli C, et al. Variability in the recommendations for the clinical management of osteoporosis. Med Clin (Barc). 2014; 142: 15–22. doi: 10.1016/j.medcli.2012.10.025 PMID: 23332628
8. Casado E, Caamano M, Sanchez-Burson J, Salas E, Malouf J, et al. Management of the patient with a high risk of fracture in clinical practice. results from a survey of 174 spanish rheumatologists (OSTEOPAR project). Reumatol Clin. 2011; 7: 305–313. doi: 10.1016/j.reuma.2010.12.008 PMID: 21925446
9. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int. 2008; 19: 385–397. doi: 10.1007/s00198-007-0543-5 PMID: 18292978
10. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in england and wales: Prospective derivation and validation of QFractureScores. BMJ. 2009; 339: b4229. doi: 10.1136/bmj.b4229 PMID: 19926696
11. Gonzalez Lopez-Valcarcel B, Sosa Henriquez M. Estimate of the 10-year risk of osteoporotic fractures in the spanish population. Med Clin (Barc). 2013; 140: 104–109. doi: 10.1016/j.medcli.2011.11.030 PMID: 22401729

12. Wells GA, Cranney A, Peterson J, Boucher M, Shea B, et al. Etidronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database Syst Rev. 2008; (1:):CD003376. doi: CD003376. doi: 10.1002/14651858.CD003376.pub3 PMID: 18254018

13. Wells G, Cranney A, Peterson J, Boucher M, Shea B, et al. Risedronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database Syst Rev. 2008; (1:):CD004523. doi: CD004523. doi: 10.1002/14651858.CD004523.pub3 PMID: 18254053

14. Wells GA, Cranney A, Peterson J, Boucher M, Shea B, et al. Alendronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database Syst Rev. 2008; (1:):CD001155. doi: CD001155. doi: 10.1002/14651858.CD001155.pub2 PMID: 18253985

15. Guerra-Garcia MM, Rodriguez-Fernandez JB, Puga-Sarmiento E, Charle-Crespo MA, Gomes-Carvalho CS, et al. Incidence of hip fractures due to osteoporosis in relation to the prescription of drugs for their prevention and treatment in galicia, spain. Aten Primaria. 2011; 43: 82–88. doi: 10.1016/j.aprim.2010.04.010 PMID: 20554353

16. Arias LM, Treceno C, Garcia-Ortega P, Rodríguez-Paredes J, Escudero A, et al. Hip fracture rates and bisphosphonate consumption in spain. an ecologic study. Eur J Clin Pharmacol. 2013; 69: 559–564. doi: 10.1007/s00228-012-1337-z PMID: 22821192

17. Álvarez Rodríguez E. Optimización del tratamiento con alendronate en osteoporosis [Improvement of treatment with alendronate in osteoporosis]. Universidad Complutense de Madrid. 2009; Available: http://eprints.ucm.es/8891/1/T30917.pdf. Accessed 04 April 2014

18. Sanfelix-Gimeno G. Opportunities for improvement in the management of osteoporosis. time to tackle the essential. Med Clin (Barc). 2013; 141: 527–528. doi: 10.1016/j.medcli.2013.09.013 PMID: 24210981

19. Imaz I, Zegarra P, Gonzalez-Enriquez J, Rubio B, Alcazar R, et al. Poor bisphosphonate adherence for treatment of osteoporosis increases fracture risk: Systematic review and meta-analysis. Osteoporos Int. 2010; 21: 1943–1951. doi: 10.1007/s00198-009-1134-4 PMID: 19967338

20. Park-Wyllie LY, Mamdani MM, Juurlink DN, Hawker GA, Gunraj N, et al. Bisphosphonate use and the risk of subtrochanteric or femoral shaft fractures in older women. JAMA. 2011; 305: 783–789. doi: 10.1001/jama.2011.190 PMID: 21343577

21. Giannini S, Varenna M. Observational studies in osteoporosis treatment. Reumatismo. 2009; 61 Suppl 2: 2–10. PMID: 19999185

22. Migliore A, Broccoli S, Massafra U, Cassol M, Frediani B. Ranking antireabsorptive agents to prevent vertebral fractures in postmenopausal osteoporosis by mixed treatment comparison meta-analysis. Eur Rev Med Pharmacol Sci. 2013; 17: 658–667. 3443 [pii]. PMID: 23543450

23. Casado E. Nuevos datos sobre el tratamiento con bifosfonatos: ¿Son aconsejables unas vacaciones terapeúticas? [New data on bisphosphonate therapy: Is a therapeutic advisable vacation?]. Reumatol Clin. 2011; 7: S28–S33. doi: 10.1016/j.reuma.2011.10.004 PMID: 22152287

24. Agirrezabala J, Aizpurua I, Albizuri M, Iciar A, Armendáriz M, et al. Osteoporosis postmenopáusica: ¿estamos previniendo las fracturas? [Postmenopausal osteoporosis: Are we preventing fractures?]. Infac. 2006; 14: 43–48.

25. Imaz Iglesia I, Rubio González B, López Delgado M, Amate Blanco J, Gómez Pajuelo P, et al. Análisis coste-utilidad de los tratamientos farmacológicos para la prevención de fracturas en mujeres con osteoporosis en España. [cost-utility analysis of the prevention drug treatment of fractures in osteoporosis women in Spain]. Madrid: Agencia de Evaluación de Tecnologías Sanitarias—Instituto de Salud Carlos III. 2010; Available: http://gesdoc.isciii.es/gesdoccontroller?action = download&id=14/09/2012-3fdd17b5be. Accessed 04 April 2014.

26. Bolibar B, Fina Aviles F, Morros R, Garcia-Gil Mdel M, Hermosilla E, et al. SIDIAP database: Electronic clinical records in primary care as a source of information for epidemiologic research. Med Clin (Barc). 2012; 138: 617–621. doi: 10.1016/j.medcli.2012.01.020 PMID: 22444996

27. Daniel H, Imai K, King G, Stuart E. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis. 2007; 15: 199–236.

28. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011; 10: 150–161. doi: 10.1002/pst.433 PMID: 20925139

29. Arnold BF, Hogan DR, Colford JM Jr,Hubbard AE. Simulation methods to estimate design power: An overview for applied research. BMC Med Res Methodol. 2011; 11: 94-2288-11-94. doi: 10.1186/1471-2288-11-94 PMID: 21689447

30. Garcia-Gil Mdel M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care. 2011; 19: 135–145. PMID: 22688222

31. Butt DA, Mamdani M, Austin PC, Tu K, Gomes T, et al. The risk of falls on initiation of antihypertensive drugs in the elderly. Osteoporos Int. 2013; 24: 2649–2657. doi: 10.1007/s00198-013-2369-7 PMID: 23612794

32. Cummings SR, Black DM, Thompson DE, Applegate WB, Barrett-Connor E, et al. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: Results from the fracture intervention trial. JAMA. 1998; 280: 2077–2082. joc80627 [pii]. PMID: 9875874

33. Jamart Sáncheza L, Herrero Hernándeza S, Barreda Velázqueza C. ¿Está justificado el gasto en fármacos contra la osteoporosis? [Is it justified spending on drugs for osteoporosis?]. MC Form Med Contin Aten Prim. 2011; 18: 317.

34. Departament de Salut de la Generalitat de Catalunya. Bifosfonats en dones postmenopàusiques amb risc baix de fractures [Bisphosphonates in postmenopausal women with low risk fracture]. 2013; Available: http://www20.gencat.cat. Accessed Abril 2014.

35. Richards M. A report for the secretary of state for health by professor sir mike richards CBE. COI Crown. 2010; Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216249/dh_117977.pdf. Accessed 04 April 2014.

36. Brotons Muntó F, Cerecedo Pérez MJ, González González A, Lázaro Gómez MJ, León Vázquez F, et al. Grupo de trabajo de la semFYC para el proyecto recomendaciones «No hacer» [Working group on recommendations semFYC project "not do"]. e-documentos semFYC. 2014; Available: http://e-documentossemfyc.es/recomendacion-para-no-hacer-de-la-sociedad-espanola-de-medicina-de-familia-y-comunitaria/. Accessed 24 Abril 2014.

37. American Academy of Family Physicians. Fifteen things physicians and patiens should question. ABIM Foundation. 2014; Available: http://www.choosingwisely.org/doctor-patient-lists/american-academy-of-family-physicians/. Accessed 04 February 2014.

38. Barry P, Aspray T, Briers K, Collins G, Compston J, et al. 'Do not do' recommendations: Osteoporosis: Assessing the risk of fragility fracture: NICE guideline. 2014; Available: http://www.nice.org.uk/usingguidance/donotdorecommendations/detail.jsp?action = details&dndid=1112. Accessed 20 June 2014.

6.5 Estudio CUIDADORES: Trabajo publicado en J Public Health Policy

Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study. J Public Health Policy.2016 May;37(2):173-89

Original Article

Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study

Luís González-de Paz[a,b,*], Jordi Real[a,c], Alicia Borrás-Santos[a,d,e], José M. Martínez-Sánchez[f,g,h], Virginia Rodrigo-Baños[b], and María Dolores Navarro-Rubio[a,i,j]

[a]Facultat de Medicina i Ciencies de la Salut, Public Health Unit, School of Medicine and Health Sciences, Universitat Internacional de Catalunya, C. Doctor Trueta S/N, Sant Cugat del Vallès, Barcelona 08195, Spain.

[b]Centre d'Atenció Primària Les Corts. Transverse Group for Research in Primary Care, IDIBAPS, Barcelona, Spain.

[c]Institut d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol). USR-Lleida, Lleida, Spain.

[d]Centro de Investigación Biomédica en Red Enfermedades Respiratorias (CIBERES), Madrid, Spain.

[e]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

[f]Biostatistics Unit, School of Medicine and Health Sciences, Universitat Internacional de Catalunya, Sant Cugat del Vallès Barcelona, Spain.

[g]Tobacco Control Unit, Cancer Prevention and Control Programme, Catalan Institute of Oncology-ICO, Hospitalet de Llobregat Barcelona, Spain.

[h]Cancer Control and Prevention Group, Bellvitge Biomedical Research Institute-IDIBELL, Hospitalet de Llobregat Barcelona, Spain.

[i]Albert Jovell Institute for Public Health and Patients. Universitat Internacional de Catalunya, Sant Cugat del Vallès Barcelona, Spain.

[j]Spanish Patient's Forum, Barcelona, Spain.

*Corresponding author. E-mail: gonzalezdepaz@hotmail.com

Abstract This population-based study using 2011–2012 Spanish National Health Survey data aimed to measure the impact of disease, health-related habits, and risk factors associated with informal caregiving. We included and matched self-reported informal caregivers [ICs] with controls (1:4) from the same survey. For each outcome, we analyzed associations between ICs and controls using linear regression or logistic regression models. ICs had 3.4 per cent more depression (OR: 1.33, 95 per cent confidence intervals [CI]:1.06, 1.68). ICs had lower social support (95 per cent CI: 1.64, 3.28), they did more housework alone (OR:3.6,

95 per cent CI:2.65, 4.89), and had greater stress (95 per cent CI:0.13, 0.83). Women ICs caring alone had more anxiety than other groups. We found no statistical association between caregivers and worse health-related habits or increased risk factors (less physical activity, smoking, drinking, and cholesterol). Our results provide evidence that health-care professionals and organizations should recognize the importance of caring for those who care.
Journal of Public Health Policy (2016) **37**, 173–189. doi:10.1057/jphp.2016.3; published online 11 February 2016

Keywords: caregivers; risk factors; burden; social support; quality of life

Introduction

Aging is associated with an increase in chronic and non-communicable diseases, and reduced functional and social independence.[1] When family members require care because of functional limitations, household roles change and relatives often become informal caregivers (ICs). They aid home maintenance and the activities of daily living, such as bathing, dressing, and eating.[2] The well-being of ICs influences the control of disease symptoms[3] and the quality of life of care receivers,[4] and is crucial to health budgets,[5,6] that could not afford nursing home care for all who might benefit from it. Examining epidemiologically the consequences of informal caregiving is relevant, as the regional office for Europe of the World Health Organization forecasts that, by 2050, people aged $\geqslant 65$ years will form >25 per cent of the population. They will, at some point, require a caregiver.[1]

Informal caregiving has social and health consequences. ICs often abandon social relationships, consider changing jobs or work schedules, and may suffer psychological distress and physical disease.[7,8] The negative outcomes, resulting from ICs feelings of inability to manage care receivers' demands, have been called the *caregiver's burden*.[9] Factors such as sex, age, socioeconomic status, educational attainment, ethnicity, social support, household organization, and time spent caring may also have an influence.[2,10,11] Current guidelines on dementia, degenerative diseases, and severe mental impairment include charters and information to reduce negative health-related outcomes for caregivers.[12–14] Effective support actions include assessing living conditions to educate ICs on environmental changes in the home, psycho-educational interventions to help ICs assimilate a more clinical belief set about their

role, and enhancement of coping skills.[3] These have been shown to lessen caregivers' burdens and improve the quality of life of ICs.[15]

Interventions with ICs aim to avoid disease and symptoms as ICs suffer more depression, anxiety, gastric disease, and a worsening in chronic conditions, plus health-related habits (smoking, drinking, and little physical activity) as compared with the general population. These differences may occur because ICs reduce their leisure time and abandon social relationships, which in turn affect health-related habits.[7,8,15,16] Most studies have concluded that the association between the disease burden and informal caregiving warrants further epidemiological study and analysis.[2]

Most studies have enrolled ICs by looking at groups of patients with specific conditions (Alzheimer's, the terminally ill, disabled children, and so on). Little is known about ICs in general.[10,17,18] Strong evidence of links between informal caregiving and disease would permit more effective interventions by health-care professionals aimed at protecting ICs – disease prevention and promoting a healthy household.

Does informal caregiving cause mental health problems, physical disease, and even increased mortality? Probably, but measures of the association with disease are not proven and come from studies with small sample sizes in special groups of caregivers.[2] The primary objective of our study was to learn whether informal caregiving was associated with disease and, secondarily, to examine differences in risk factors and health-related habits.

Methods

Study design

We undertook a matched cross-sectional study using a representative survey of Spanish households to describe the association between informal caregiving and health outcomes for the caregivers.

Hypothesis

On the basis of previous studies, we expected that ICs would have more depression and anxiety, worse health status, and would report fewer desirable health-related habits, compared with matched controls.

We expected that perceived social support would mediate the effect of depression and anxiety.

Data

The 2011–2012 Spanish National Health Survey, a survey conducted by the National Statistics Institute, collected information on the determinants and opinions of citizens about their health. It comprised three questionnaires: for household residents, for adults, and for children,[19] and provided a representative sample ($N = 21{,}007$) of Spain's non-institutionalized home-dwelling population. The units of analysis are Spanish households. A multistage cluster method with proportional random selection of primary and secondary sampling units (regions, towns, and census tracts, respectively) creates the sample, with the final sample selected by sex- and age-based quotas. Trained interviewers conducted the survey with face-to face interviews in homes. Methodological details have been published elsewhere.[19]

Participants and matching criteria

We used three criteria to identify eligible participants aged >15 years:

(1) having an individual with any functional limitation or disability, unable to self-care, in need of a caregiver (excluding normally-developing children) in the household;
(2) reporting being the sole or shared caregiver of the disabled resident;
(3) caregiving for >1 year (thus avoiding any disease bias with respect to the start of caring).

Once selected, each completed the Spanish National Health Survey health questionnaire. We selected four controls who did not report caring for anybody in the household for each IC among individuals completing the survey.

We matched ICs and controls using a propensity score, with the nearest neighbor method (caliper = 0.2, $r = 1{:}4$) to reduce bias because of confounding variables at the individual level and to produce a data set closer to that which would result from a block, randomized design.[20] We matched groups by age, sex, size of municipality, household composition, educational level, and household net monthly income.

Web Appendix I shows that the overall mean difference in the selected variables was reduced to almost 0 from the pre-matched sample to the final matched sample; each case was matched with a mean of four controls from the same Spanish region, and, therefore, comparability and representativeness was maximized. Web Appendix II shows the factors used to match ICs with controls: Percentages were similar in all variables and categories.

Variables

We defined four main outcomes:

(1) diagnosis of chronic depression;
(2) anxiety;
(3) perceived personal social support; and
(4) degree of psychological well-being.

To collect diagnosis of chronic depression or anxiety we used the question: 'Has your doctor ever told you that you suffer from (disease)?'.[19] To collect the perceived personal social support information, we used the Spanish adaptation of the DUKE-UNC Functional Social Support Questionnaire.[21] This is a self-administered 11-item questionnaire, with each item scored from 1 to 5. It assesses the degree of an individual's social interaction and subjective support. To measure the degree of psychological well-being, we used the Goldberg GHQ 12, [22]a 12-item scale, to identify the severity of an individual's psychological distress. It is used to assess mental health, covering disorders or patterns of adjustment associated with distress. Each scale item has four responses from 'better than usual' to 'much less than usual.'

Secondary outcomes were:

- Quality of life, measured using the EQ-5D-5L questionnaire – a widely used instrument that assesses five self-perceived and health-related quality of life states (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression)[23] plus other health problems including mental problems (not depression or anxiety), migraine or frequent headaches, chronic constipation, ulcers (stomach or duodenum), and chronic back pain (lumbar or cervical).
- Drug use during the previous 2 weeks (of tranquilizers, muscle relaxants, sleeping drugs, antidepressants, stimulants) using the question:

'Have you taken this kind of drug in the past two weeks and, if so, what drugs have been prescribed by a doctor?'. This variable had two levels: drugs prescribed and drugs consumed.

- Health-related habits based on six risk factors targeted by the World Health Organization were assessed using National Health Survey questionnaire data:[24] medical diagnoses of diabetes (types 2 and 1), high blood pressure, and elevated total cholesterol; smoking (*current smoker, ex-smoker, never smoker*); physical activity (*low, medium, high*) was measured using the IPAQ questionnaire; chronic heavy drinking was determined by the mean units of alcohol per week (men $\geqslant$ 27, women $\geqslant$ 14).[25]
- Other variables. We studied type of care (caring alone or sharing care with others); the degree of participation in housework – measured using the question 'Who mainly does the housework such as washing, cooking, ironing, etc.?' answers classified as: *No housework (others do it)*, *shared with another person*, or *alone;* current employment (*yes, no*).
- Stress and work satisfaction measured using a 7-point Likert scale.

Statistical methods

Continuous variables were expressed as means and standard deviation (SD) or medians plus 25th and 75th percentiles; categorical variables as percentages. The prevalence of household-dwellers with functional limitations or disability and of ICs in Spanish households were computed using the whole National Health Survey sample. We calculated the error and 95 per cent confidence intervals (CI) using weighting coefficients.

For all outcomes, we analyzed associations between ICs and controls using a linear regression model or a logistic regression model with clustered standard errors. We report results as a linear coefficient or an adjusted odds ratios (OR) with 95 per cent CI. We examined differences in stress and satisfaction at work in employed ICs. We studied the impact of disease in ICs according to sex for the type of caregiving: alone or shared. We reported the results as the prevalence ratio (PR) and 95 per cent CI. We studied the impact of depression and anxiety using a linear regression model, and reported results as a linear coefficient with 95 per cent CI.

We tested mediation effects (indirect) using bootstrap methods and adjusting by sex. We chose this method to prevent type I error, because

bootstrapping does not require the assumption of normality of the sampling distribution. It allows the indirect effect to be estimated using repeated samples from the data set, and produces bias-corrected accelerated confidence intervals for the indirect effect. After examining the bootstrap confidence interval, we calculated the ratio of the indirect effect to the total effect (*mediation ratio*).[26] In all analyses, the level of statistical significance was set at $\alpha = 0.05$. For the analysis we used SPSS v. 22th (IBM Corp) and R v. 3.1.2 statistical packages.[27,28]

Ethical considerations

Our study used information from a public data set. The Spanish National Statistics Institute asserted that all permissions and ethical concerns were guaranteed and all participants had consented to their information being published anonymously. These public data files contain anonymous information and assure participant confidentiality.[19] For these reasons, Ethics Committee approval was not required under Spanish legislation and European Union (EU) rules.

Results

An estimated 6.60 per cent (95 per cent CI: 5.90, 7.40) of Spanish households had a disabled resident requiring informal care. Five hundred fifteen (2.45 per cent) of ICs fulfilled our inclusion criteria. Women made up 63.1 per cent of these ICs, with a mean age of 58.01 years ($SD = 16.12$, range 17–96). Four hundred forty-three (86.01 per cent) were caring for one adult and 48 (9.32 per cent) were caring for children with limitations or disabilities (Table 1).

Table 2 shows the primary outcomes. ICs were more likely to have received a diagnosis of depression (3.4 per cent, $P<0.05$) and less social support ($P<0.001$) than controls. ICs had worse psychological well-being and more diagnoses of anxiety, although the results were not significantly different. We found a significant mediation effect on depression ($P<0.001$) from functional social support – 26.37 per cent of the total effect. The direct effect between caregiving and depression fell ($P = 0.484$). The same result was found for anxiety: functional social support explained 35.45 per cent of the total effect of anxiety in caregivers (Table 2).

Table 1: Caregiver characteristics

Characteristics	*(n = 515)*
Who provides care	
Caregiving alone	302 (58.06%)
Sharing care	213 (41.40%)
Mean years providing care (SD)	17.96 (25.57)
25th–50th–75th percentiles	3–7–20
Mean care in hours, total week (SD)	55.37 (*58.18*)
25th–50th–75th percentiles	16–42–64
Reference person (main household provider)	231 (44.9%)
Age of care receiver by household	
Adult	467 (90.70%)
Child	47 (9.1%)
Adult and child	1 (0.20%)
Caregiver relationship with care receiver (N = 540)	
Spouse or partner	156 (28.90%)
Parents or in-laws	141 (26.1%)
Grandparents	3 (0.60%)
Son, daughter or son-in-law	170 (31.50%)
Other relative or unrelated	71 (13.10%)
Occupational status	
Working	127 (24.70%)
Unemployed	73 (14.205%)
Housework	110 (21.40%)
Retired	170 (3.35%)
Other	35 (6.90%)
Marital Status n (%)	
Single	115 (22.30%)
Married	320 (62.10%)
Widowed	45 (8.70%)
Legally separated or divorced	35 (6.80%)
High risk drinking	
Women	10 (3.1%)
Men	17 (8.9%)
Physical activity	
Low	160 (46.90%)
Moderate	110 (32.30%)
Vigorous	71 (20.80%)

Secondary outcomes (Table 3) showed that quality of life was very similar in the two groups, except for the anxiety and depression dimensions in the questionnaire ($P<0.001$). Unexpectedly, there were

Table 2: Results of primary outcomes.

Primary outcomes	*Informal caregivers n = 515*	*Controls n = 2053*	*P-value*	*Measure of association*	*95% CI*
Functional social support, Mean (SD)	45.68 (9.48)	48.14 (8.14)	<0.001	2.46[a]	1.64, 3.28
Psychological well-being (Goldberg GHQ 12)	3.50 (11.66)	3.40 (12.92)	0.871	0.10[a]	−1.34, 1.14
Medical diagnosis (yes)					
Chronic depression	81 (15.7%)	252 (12.3%)	0.012	1.334[b]	1.06, 1.68
Chronic anxiety	70 (13.6%)	230 (11.2%)	0.094	1.247[b]	0.96, 1.62
Functional social support as a mediator of depression[c]					
Effect of caregiving (direct effect)	–	–	0.484	1.32	1.00, 1.74[d]
Effect of functional social support (indirect effect)	–	–	<0.001	2.08	1.03, 1.13[d]
Functional social support as a mediator of anxiety[c]					
Effect of caregiving (direct effect)	–	–	0.161	1.23	0.85, 1.64
Effect of functional social support (indirect effect)	–	–	<0.001	1.07	1.04, 1.13

[a]Regression coefficient.
[b]Odds ratio.
[c]Mediating analyses were carried out by adjusting the outcome (depression and anxiety) by sex.
[d]Bootstrap 95 per cent CI with 1000 resamples.

non-significant results in outcomes previously reported to affect ICs – migraine, digestive disorders, and chronic back pain. ICs were more likely to report doing housework alone (53.20 per cent ICs versus 30 per cent of controls). Working ICs had more stress at work than controls ($P = 0.003$), but satisfaction was similar in both groups. There were no significant differences in the proportions of the two groups with diabetes diagnoses, high blood pressure, elevated cholesterol, smoking, risky drinking, or low physical activity (Table 4).

Women caregivers had a higher prevalence of depression (PR:3.36; 95 per cent CI:1.87, 6.04) and anxiety (PR:3.51; 95 per cent CI: 1.84, 6.69) than men. Caring alone had significant indirect effect on anxiety in women ICs (OR: 1.02, 95 per cent CI: 1.01, 1.04). Caring alone accounted for 17.91 per cent of the likelihood of anxiety ($P = 0.01$). Caring alone had a non-significant indirect effect on depression (OR: 0.99, 95 per cent CI: 0.97, 1.01, $P = 0.591$). Hours of informal care per week were not associated with the likelihood of depression (regression coefficient: 0.002, 95 per cent CI: −0.00, 0.01, $P = 0.525$), but a statistical trend was found in the likelihood of anxiety

Table 3: Results of the secondary outcomes

Secondary outcomes	*Informal caregivers n = 515*	*Controls n = 2053*	*P-value*	*Measure of association*	*95per cent CI*
Euroqol-Quality of Life: Mean (SD)					
Mobility	1.38 (0.80)	1.41 (0.85)	0.444	0.034[a]	0.12, 0.05
Self-care	1.13 (0.49)	1.17 (0.61)	0.058	0.04[a]	−0.09, 0.0
Usual activities	1.28 (0.70)	1.31 (0.76)	0.437	−0.03[a]	−0.10, 0.04
Pain, discomfort	1.63 (0.94)	1.61 (0.92)	0.919	0[a]	−0.07, 0.07
Anxiety, depression	1.45 (0.87)	1.31 (0.78)	*<0.001*	0.14[a]	0.06, 0.23
Medical diagnosis (yes)					
Other mental problems	11 (2.1%)	34 (1.7%)	0.418	1.296[b]	0.68, 2.46
Migraine or frequent headaches	66 (12.8%)	244 (11.9%)	0.569	1.090[b]	0.81, 1.47
Ulcers (stomach or duodenum)	34 (6.6%)	132 (6.4%)	0.344	1.164[b]	0.84, 1.60
Chronic back pain (lumbar or cervical)	160 (31.1%)	712 (34.7%)	0.238	1.090[b]	0.94, 1.26
Drug use and suitability					
Tranquilizers, muscle relaxants, and sleeping drugs.					
Consumed	91 (17.7%)	332 (16.2%)	0.317	1.112[b]	0.90, 1.37
Prescribed	91 (17.7%)	323 (15.7%)	0.427	1.141[b]	0.82, 1.59
Antidepressants and stimulants					
Consumed	45 (8.7%)	159 (7.7%)	0.427	1.141[b]	0.82, 1.59
Prescribed	45 (8.7%)	159 (7.7%)	0.427	1.141[b]	0.82, 1.59
Housework					
No (others do it)	86 (16.70%)	695 (33.90%)	–	–	–
Shared with another person	155 (30.10%)	743 (36.20%)	*0.008*	1.686[b]	1.14, 2.50
Alone	274 (53.20%)	615 (30.00%)	*<0.001*	3.600[b]	2.65, 4.89
	n = 123	*n = 633*	–	–	–
Stress at work	4.58 (1.37)	4.13 (1.79)	*0.007*	0.48[a]	0.13, 0.83
Satisfaction at work	5.76 (1.37)	5.60 (1.36)	0.366	0.105[a]	−0.12, 0.33

[a]Regression coefficient.
[b]Odds ratio.

(regression coefficient: 0.005, 95 per cent CI: −6.604, 0.011, $P = 0.053$). Years providing care were not correlated with depression (regression coefficient: 0.01, 95 per cent CI:−0.01, 0.02, $P = 0.323$) or anxiety (regression coefficient: 0.004, 95 per cent CI −0.01, 0.02, $P = 0.408$).

Discussion

Our findings showed that ICs were more likely to have chronic depression than matched controls. Prevalence of disease in ICs was mediated by two factors: sex and caring alone. The latter also affected the likelihood of anxiety. The time spent caring did not have a significant effect.

Table 4: Differences in health-related habits and risk factors between ICs and control group

Health-related habits and risk factors	*Informal caregivers n = 515*	*Controls. n = 2053.*	*P-value*	*Measure of association*	*95% CI*
Diabetes (Type 2 and 1)	67 (13.0%)	227 (11.1%)	0.171	1.203[a]	0.92, 1.57
High blood pressure	160 (31.10%)	713 (34.13%)	0.099	0.847[a]	0.70, 1.03
Elevated cholesterol	143 (27.80%)	550 (26.80%)	0.648	1.050[a]	0.85, 1.30
Smoking					
No	298 (57.90%)	1228 (59.80%)	–	–	–
Current smoker	115 (22.30%)	424 (20.70%)	0.409	1.118[a]	0.86, 1.46
Ex-smoker	102 (19.80%)	401 (19.50%)	0.791	1.048[a]	0.74, 1.49
Physical activity					
Low	160 (46.90%)	642 (47.90%)	–	–	–
Medium	110 (32.30%)	430 (32.10%)	0.801	1.026[a]	0.84, 1.26
High	71 (20.80%)	26p (20.10%)	0.708	1.059[a]	0.78, 1.43
Risk drinking	27 (5.20%)	97 (4.70%)	0.624	0.896[a]	0.58, 1.39

[a]Odds ratio.

Other factors affecting ICs were lower perceived functional social support, greater stress at work, and doing more housework.

ICs were more likely to have chronic depression than controls (OR 1.3). In a systematic review of studies of caregivers for dementia patients, the relative risk of depression ranged from 2.80 to 38.68.[29] The most recent Spanish study reported that major depressive symptoms affected 8.9 per cent of ICs.[30] The variations in impact may be due to the criteria used to establish a diagnosis of depression and the groups of ICs studied. We found that social support was a full mediator between caregiving and impact of depression. A similar effect was found for anxiety. Thus we believe that social support is essential to explain impact of depression and anxiety in caregivers, as reported.[11] The negative consequences of caring worsens if caregivers lack family support in the household, and have to undertake all the housework.[31,32] Links between the household organization and housework were also reported in a recent study in 12 European countries.[33] Interestingly, in our study, the time spent on informal caring was not associated with self-reported poor health, as found by other studies.[34]

We found none of the expected differences in risk factors and health-related habits between caregivers and controls that have been reported by other studies in which physical activity or dietary habits affected cholesterol levels or diabetes.[15,17,18,35] We found no significant

differences between caregivers and matched controls in diabetes diagnoses, high blood pressure, elevated cholesterol, or health habits. The lack of control groups in previous studies, and associations based on the Zarit Burden Inventory scores – a widely used self-report instrument designed to reflect the negative experience of caregivers – may explain these differences. The Zarit inventory overlaps in several dimensions, and was not designed to detect diseases as coded in the International Disease Classification.[2]

Our results suggest that health-care professionals should focus on caregivers' perceptions of social isolation and the symptoms of depression and anxiety. In dementia and cancer guidelines, the British National Institute for Health and Care Excellence advises health professionals to determine the central role of ICs and their need for support.[12,13] Health-care professionals should seek to identify systematically any psychological distress or psychosocial impact associated with caring. Health-care professionals might explore how to involve other family members in care plus the caregiver's needs. Health-care organizations should arrange links with other organizations to develop and deliver services for ICs and offer support and information. The evidence on health-care interventions to reduce stress among ICs, such as meditation or cognitive therapy, might be evaluated, although the current level of evidence is low.[36]

Given the crucial tasks and social relevance of ICs, support for ICs should be available when recipients of care have any conditions that require assistance with activities of daily living. Currently, schemes of support for ICs are in place in the 27 EU countries.[37] Almost all offer counseling to reduce caregivers' stress, information plans, respite care programs, and training delivered through health systems. The most common are respite programs that facilitate temporary breaks for ICs by admitting the recipient of care to nursing homes.[38] A systematic review of respite programs for caregivers where the care recipient has dementia, found no significant effects of these programs. Additional research is needed to address the range of interventions and conditions for which caregivers may provide care.[39] While ICs needs are considered, a recent review of policies across EU countries noted that there is no evaluation system to measure the effect of these programs.[37] The financial support offered to ICs is not as uniform as health-care support. In Spain, IC support varies based on the level of disability of care recipients. Budget cuts during the recent financial crisis have

resulted in ICs receiving financial support only when care recipients are severely dependent.[40]

Study limitations and strengths

The survey method may result in bias. Diseases and diagnoses were self-reported to trained professionals, who, to maximize reliability, used a computer-assisted personal interviewing method. Our cross-sectional design meant we were unable to determine whether diseases developed after or before informal caregiving began. This potential bias affected both ICs and controls. Other results supported the findings: depression-anxiety scores from the quality of life questionnaire were worse compared with controls. The mediation analysis may be affected by temporal precedence: if the ICs had less social support before becoming a caregivers, this would suggest a directional link. In analysis of mediation, there was theoretical and empirical evidence of the mediation relationship between social support and the anxiety and depression.

The strengths of the study were the sample size and the heterogeneity of ICs that permitted a view not restricted to a specific disease. The data came from a national survey. This allows correct inference and generalization of results, and ICs matched with controls allowed us to minimize potential confounding issues. These methods closely emulated an experimental design with random selection.

Policy conclusions

In this study, ICs had a greater prevalence of depression after adjusting for social and demographic determinants of health. Women caring alone had a greater likelihood of anxiety and depression, independent of the time spent caring. Health-care professionals should take the symptoms and signs of depression among ICs into account. ICs may require more health-care attention and social support. More scientific evidence, program development, and evaluation are needed to provide adequate health care to all ICs. Health-care organizations and clinical guidelines should make support for ICs more explicit.

Policymakers might address two important issues: establishment of standard welfare benefits according to the health status of caregivers – not only the care receiver – and a set of indicators to assess the effectiveness of current programs.

Acknowledgements

The authors thank Dr. Belchin Kostov, Transverse Group for Research in Primary Care, IDIBAPS, Barcelona, Spain, for his valuable comments, Dr Montse Neira León, Deputy Director of Health Information and Innovation, Spanish Ministry of Health and Social Policy, Madrid, Spain, and David Buss for technical help.

About the Authors

Luis Gonzáles-de Paz, PhD, MSc, R.N. is a researcher in the Transverse Group for Research in Primary Care, Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), Barcelona, Spain. He lectures in the Universitat Internacional de Catalunya, Sant Cugat del Vallès (Barcelona), Spain.

Jordi Real is a statistician in Institut d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Lleida, Spain. He lectures in the Universitat Internacional de Catalunya, Sant Cugat del Vallès (Barcelona), Spain.

Alicia Borrás-Santos is a researcher in the Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain. She lectures in the Universitat Internacional de Catalunya, Sant Cugat del Vallès (Barcelona), Spain.

José M. Martínez-Sánchez is a researcher in the Cancer Control and Prevention Group, Bellvitge Biomedical Research Institute-IDIBELL, Hospitalet de Llobregat (Barcelona), Spain. He is the head of the Biostatistics Unit in the Universitat Internacional de Catalunya, Sant Cugat del Vallès (Barcelona), Spain.

Virginia Rodrigo-Baños is a community and family nurse resident in Les Corts Primary Health Care Center, Barcelona, Spain.

María Dolores Navarro-Rubio is the head of the public health and epidemiology department. She is the director of the Albert Jovell Institute for Public Health and Patients in the Universitat Internacional de Catalunya, Sant Cugat del Vallès, (Barcelona), Spain.

References

1. World Health Organization Europe. (2013) The European health report 2012: Charting the way to well-being, http://www.euro.who.int/en/data-and-evidence/european-health-report-2012, accessed 28 October 2015.
2. Adelman, R.D., Tmanova, L.L., Delgado, D., Dion, S. and Lachs, M.S. (2014) Caregiver burden: A clinical review. *The Journal of American Medical Association* 311(10): 1052–1060.
3. Sorensen, S., Pinquart, M. and Duberstein, P. (2002) How effective are interventions with caregivers? An updated meta-analysis. *The Gerontologist* 42(3): 356–372.
4. Hartmann, M., Bazner, E., Wild, B., Eisler, I. and Herzog, W. (2010) Effects of interventions involving the family in the treatment of adult patients with chronic physical diseases: A meta-analysis. *Psychotherapy and Psychosomatics* 79(3): 136–148.
5. Oliva, J., Vilaplana, C. and Osuna, R. (2011) The social value of informal care provided to elderly dependent people in Spain. *Gaceta Sanitaria* 25(Suppl 2): 108–114.
6. Hurd, M.D., Martorell, P., Delavande, A., Mullen, K.J. and Langa, K.M. (2013) Monetary costs of dementia in the United States. *The New England Journal of Medicine* 368(14): 1326–1334.
7. Borghi, A.C., de Castro, V.C., Marcon, S.S. and Carreira, L. (2013) Overload of families taking care of elderly people with Alzheimer's disease: A comparative study. *Revista Latino-Americana de Enfermagem* 21(4): 876–883.
8. Etters, L., Goodall, D. and Harrison, B.E. (2008) Caregiver burden among dementia patient caregivers: A review of the literature. *Journal of the American Academy of Nurse Practitioners* 20(8): 423–428.
9. Zarit, S.H., Todd, P.A. and Zarit, J.M. (1986) Subjective burden of husbands and wives as caregivers: A longitudinal study. *The Gerontologist* 26(3): 260–266.
10. Northouse, L., Williams, A.L., Given, B. and McCorkle, R. (2012) Psychosocial care for family caregivers of patients with cancer. *Journal of Clinical Oncology* 30(11): 1227–1234.
11. Thielemann, P.A. and Conner, N.E. (2009) Social support as a mediator of depression in caregivers of patients with end-stage disease. *Journal of Hospice and Palliative Nursing* 11(2): 82–90.
12. National Institute for Health and Clinical Excellence. (2004) The NICE-GSGSP guidance on cancer services: Improving supportive and palliative care for adults with cancer. The manual (online), https://www.nice.org.uk/guidance/csgsp, accessed 28 October 2015.
13. National Institute for Health and Clinical Excellence. (2012) The NICE-SCIE guideline on supporting people with Dementia and their carers in health and social care (online), https://www.nice.org.uk/guidance/CG42, accessed 28 October 2015.
14. Ngo, J. and Holroyd-Leduc, J.M. (2015) Systematic review of recent dementia practice guidelines. *Age and Ageing* 44(1): 25–33.
15. Guedes, A.C. and Pereira Mda, G. (2013) Burden, coping, physical symptoms and psychological morbidity in caregivers of functionally dependent family members. *Revista Latino-Americana de Enfermagem* 21(4): 935–940.
16. Cabral, L., Duarte, J., Ferreira, M. and dos Santos, C. (2014) Anxiety, stress and depression in family caregivers of the mentally ill. *Atención Primaria* 46(Suppl): 5176–179.
17. Badia Llach, X., Lara Surinach, N. and Roset Gamisans, M. (2004) Quality of life, time commitment and burden perceived by the principal informal caregiver of Alzheimer's patients. *Atención Primaria* 34(4): 170–177.
18. Pena-Longobardo, L.M. and Oliva-Moreno, J. (2015) Caregiver burden in Alzheimer's disease patients in Spain. *Journal of Alzheimer's Disease* 43(4): 1293–1302.
19. Instituto Nacional de Estadística de España. (2015) National health survey. General methodology, http://www.ine.es/en/metodologia/t15/t15304191 2_en.pdf, accessed 28 October 2015.

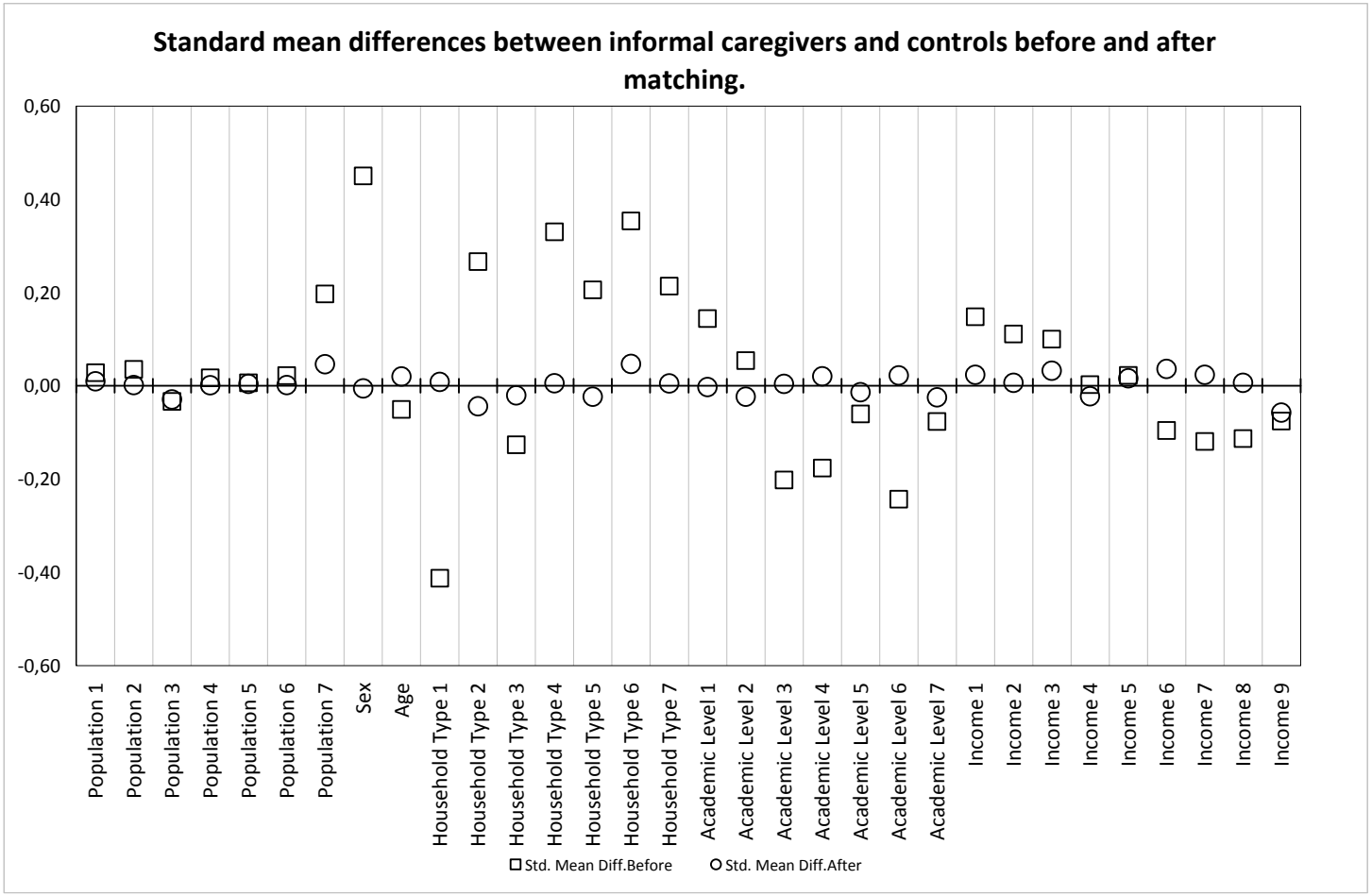
20. Stuart, E.A. (2010) Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1): 1–21.
21. Bellon Saameno, J.A., Delgado Sanchez, A., Luna del Castillo, J.D. and Lardelli Claret, P. (1996) Validity and reliability of the Duke-UNC-11 questionnaire of functional social support. *Atención Primaria* 18(4): 153–156, 158–163.
22. Sanchez-Lopez Mdel, P. and Dresch, V. (2008) The 12-item general health questionnaire (GHQ-12): Reliability, external validity and factor structure in the Spanish population. *Psicothema* 20(4): 839–843.
23. Herdman, M. *et al* (2011) Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research* 20(10): 1727–1736.
24. World Health Organization. (2014) Global status report on noncommunicable diseases 2014, http://www.who.int/nmh/publications/ncd-status-report-2014/en/, accessed 15 May 2015.
25. Valencia Martin, J.L., Gonzalez, M.J. and Galan, I. (2014) Methodological issues in the measurement of alcohol consumption: The importance of drinking patterns. *Revista Española de Salud Pública* 88(4): 433–446.
26. Preacher, K.J. and Kelley, K. (2011) Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods* 16(2): 93–115.
27. R Core Team. (2015) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
28. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014) Mediation: R package for causal mediation analysis. *Journal of Statistical Software* 59(5): 1–38.
29. Cuijpers, P. (2005) Depressive disorders in caregivers of dementia patients: A systematic review. *Aging & Mental Health* 9(4): 325–330.
30. Torres, Á., Blanco, V., Vázquez, F., Díaz, O., Otero, P. and Hermida, E. (2015) Prevalence of major depressive episodes in non-professional caregivers. *Psychiatry Research* 226(1): 333–339.
31. Adams, R.N., Mosher, C.E., Cannady, R.S., Lucette, A. and Kim, Y. (2014) Caregiving experiences predict changes in spiritual well-being among family caregivers of cancer patients. *Psycho-oncology* 23(10): 1178–1184.
32. McLennon, S.M., Bakas, T., Jessup, N.M., Habermann, B. and Weaver, M.T. (2014) Task difficulty and life changes among stroke family caregivers: Relationship to depressive symptoms. *Archives of Physical Medicine and Rehabilitation* 95(12): 2484–2490.
33. Schneider, U. and Kleindienst, J. (2015) Monetising the provision of informal long-term care by elderly people: Estimates for European out-of-home caregivers based on the well-being valuation method. *Health & Social Care in the Community*, advance online publication May 4, doi:10.1111/hsc.12250.
34. Legg, L., Weir, C.J., Langhorne, P., Smith, L.N. and Stott, D.J. (2013) Is informal caregiving independently associated with poor health? A population-based study. *Journal of Epidemiology and Community Health* 67(1): 95–97.
35. Molinuevo, J.L. and Hernandez, B. Grupo de Trabajo del Estudio IMPACT. (2011) Profile of the informal carer associated with the clinical management of the Alzheimer's disease patient refractory to symptomatic treatment of the disease. *Neurologia (Barcelona, Spain)* 26(9): 518–527.
36. Yesufu-Udechuku, A. *et al* (2015) Interventions to improve the experience of caring for people with severe mental illness: systematic review and meta-analysis. *The British Journal of Psychiatry* 206(4): 268–274.
37. Courtin, E., Jemiai, N. and Mossialos, E. (2014) Mapping support policies for informal carers across the European Union. *Health Policy (Amsterdam, Netherlands)* 118(1): 84–94.

38. Hoffmann, F. and Rodrigues, R. (2010) Informal carers: Who takes care of them? *Policy Briefs*, 13 May.
39. Maayan, N., Soares-Weiser, K. and Lee, H. (2014) Respite care for people with dementia and their carers. *The Cochrane Database of Systematic Reviews*, 1: CD004396, doi:10.1002/14651858.CD004396.pub3.
40. Gallo, P. and Gene-Badia, J. (2013) Cuts drive health system reforms in Spain. *Health Policy (Amsterdam, Netherlands)* 113(1–2): 1–7.

Supplementary information accompanies this article on the *Journal of Public Health Policy* website (www.palgrave-journals.com/jphp)

Appendix I. Characteristics used to match the samples of ICs and controls. After matching the two groups were almost the same

	Informal caregivers n=515	**Controls n=2053**
Age		
Mean (SD)	58.01 (16.12)	57.66 (18.01)
Population of municipality.		
> 10,000 inhabitants.	127 (24.7%)	484 (23.6%)
10,000 to 20,000 inhabitants.	58 (11.3%)	226 (11.0%
20,000 to 50,000 inhabitants.	79 (15.3%)	346 (16.9%)
50,000 to 100,000.	70 (7.8%)	188 (9.2%)
>100,000 to 500,000 inhabitants.	47 (9.1%)	196 (9.5%)
>500,000 inhabitants	47 (9.1%)	165 (8.0%)
Provincial capital	117 (22.7%)	448 (21.8%)
Sex.		
Female	325 (63.1%)	1257 (61.2%)
Household composition.		
Couple alone	113 (21.9%)	445 (21.7%)
Couple with children under 25 years old.	75 (16.6%)	304 (14.8%)
Couple with children over 25 years old.	87 (16.9%)	376 (18.3%)
Single parent with children under/over 25 years old	101 (19.6%)	382 (18.6%)
Couples, single parent with child under 25 years and others living in the household	48 (9.3%)	205 (10.0%)
Other type.	91 (17.7%)	341 (16.61%)
Educational Level.		
Primary education or less	207 (40.17%)	827 (40.3%)
Secondary education	213 (41.36%)	871 (42.43%)
Professional certificate	50 (9.71%)	186 (9.06%)
College or equivalent	45 (8.7%)	169 (8.2%)
Monthly income of the entire household.		
550 € or less	14 (2.70%)	77 (3.80%)
551-1050 €	135 (26,21	557 (27.13%)
1051-1550 €	129 (25.05%)	471 (22.94%)
1551-2250 €	79 (15.34%)	309 (15.05%)
2251-2700 €	19 (3.70%)	68 (3.30%)
>2701 €	25 (4.85%)	107 (5.21%)
NA's	114 (22.10%)	464 (22.60%)
Social class, based on occupation of main household provider.		
I	26 (5.00%)	132 (6.40%)
II	32 (6.20%)	124 (6.00%)
III	98 (19.00%)	352 (17.10%)
IV	68 (13.20%)	285 (13.90%)
V	194 (37.70%)	738 (35.90%)
VI - Workers/unskilled/not stated	73 (14.20%)	342 (16.70%)
NAs	24 (4.70%)	80 (3.90%)



Appendix II. The overall mean difference of selected variables used to match ICs with control was reduced to almost 0 from the pre-matched sample to the final matched sample.

6.6 Estudio DAMOCLES:Trabajo publicado en The European Journal of General Practice

Impact of the sustained control of cardiovascular risk factors on first episode heart failure: The relevant role of primary care. E J Gen Pract. 2015;21(4):224-30

European Journal of General Practice, 2015; Early Online: 1–7

Original Article

Impact of the sustained control of cardiovascular risk factors on first episode heart failure: The relevant role of primary care

Miguel-Angel Muñoz[1,2,3], *Jordi Real*[2,4], *José-Luis del Val*[1,2], *Ernest Vinyoles*[1,2], *Xavier Mundet*[1,2,3], *Mar Domingo*[1,2], *Cristina Enjuanes*[5,6], *José-Maria Verdú-Rotellar*[1,2,3]

[1]Institut Català de la Salut, Barcelona, Spain, [2]Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona, Spain, [3]School of Medicine, Universitat Autònoma de Barcelona, Bellaterra, Spain, [4]School of Medicine and Health Sciences, Universitat Internacional de Catalunya, Sant Cugat del Valles, Spain, [5]Heart Failure Programme, Department of Cardiology, Hospital del Mar, Barcelona, Spain, [6]Heart Diseases Biomedical Research Group, Programme of Research in Inflammatory and Cardiovascular Disorders, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

KEY MESSAGE:

- Since most patients at high cardiovascular risk are attended in primary care, the role of general practitioners is crucial in preventing the development of heart failure.
- Control of cardiovascular risk factors has shown to be determinant to prevent a first heart failure hospitalization.

ABSTRACT

Background: The role of cardiovascular risk factor control in the development of heart failure (HF) has not yet been clearly established.

Objective: To determine the effect of cardiovascular risk factor control on the occurrence of a first episode of hospital admission for HF.

Methods: A case-control study using propensity score-matching was carried out to analyse the occurrence of first hospital admission for HF taking into account the degree of cardiovascular risk factor control over the previous 24 months. All patients admitted to the cardiology unit of the Hospital del Mar between 2008 and 2011 because of a first episode of HF were considered cases. Controls were selected from the population in the hospital catchment area who were using primary care services. Cardiovascular risk factor measurements in the primary healthcare electronic medical records prior to the first HF episode were analysed.

Results: After the matching process, 645 participants were analysed (129 HF cases and 516 controls). Patients suffering a first HF episode had modest increments in body mass index and blood pressure levels during the previous two years. Adjusted odds ratio for experiencing a first HF hospital admission episode according to systolic blood pressure levels and body mass index was (OR: 1.031, 95% CI: 1.001–1.04), and (OR: 1.09, 95% CI: 1.03–1.15), respectively.

Conclusion: Increased levels of body mass index and systolic blood pressure during the previous 24 months may determine a higher risk of having a first HF hospital admission episode.

Keywords: Heart failure, hospital admission, primary care, cardiovascular risk factors

INTRODUCTION

Worldwide, cardiovascular diseases are the greatest cause of mortality (1). In 2013, the American Heart Association reported that 28 million Americans suffered from cardiovascular diseases and, among these, five million had experienced heart failure (HF) (2). HF represents a growing problem due to general population aging and to the prolongation of patients' lives as a result of improvements in medical treatment and procedures (3). From a clinical point of view, it is noteworthy that most HF cases can be attributed to modifiable risk factors and, consequently, preventable (4).

Correspondence: Miguel-Angel Muñoz, Sardenya 375, entlo, 08025, Barcelona, Spain. Fax: + 34 93 457 7133. E-mail: mamunoz.bcn.ics@gencat.cat

(Received 1 August 2014; revised 20 April 2015; accepted 24 April 2015)

ISSN 1381-4788 print/ISSN 1751-1402 online © 2015 Informa Healthcare
DOI: 10.3109/13814788.2015.1049154

The relationship between several cardiovascular risk factors (CVRF) such as hypertension, high blood cholesterol, diabetes, obesity, and smoking and HF incidence has been well-established (5–8). It has also been shown that minor reductions in the prevalence of some CVRF can decrease HF incidence (9). Nevertheless, prior to the publication of the 2009 ARIC Study there was insufficient evidence concerning the impact of CVRF control on HF incidence.

While adequate CVRF management has been reported to result in lower HF incidence, most studies have only measured them at the commencement of the cohort follow-up, without considering an intermediate analysis (10). In addition, the optimum high-risk cut-off remains to be determined.

Although several cohort studies have analysed the effect of CVRF in a long-term follow-up, little is known about the control of these factors prior to onset, particularly at the period close to the HF episode. This study aimed to determine the effect of good CVRF control on the occurrence of first episode HF hospitalization.

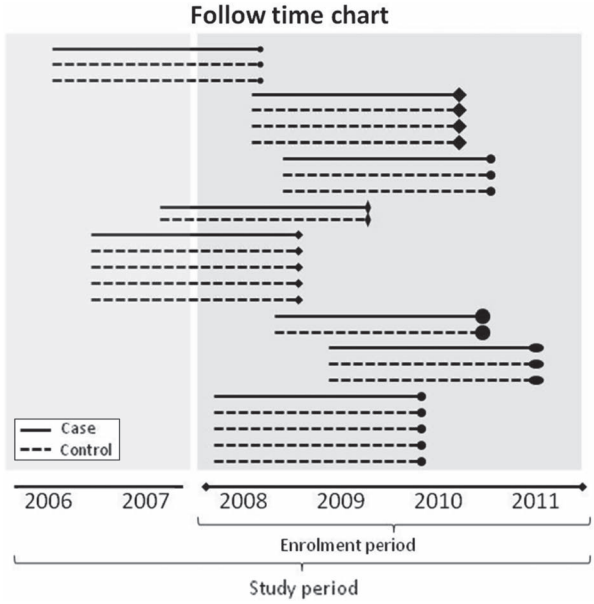


Figure 1. Follow-up chart showing the enrolment period of all participants according to inclusion date.

METHODS

Study design

A retrospective, observational study was carried out comparing two groups, one with and the other without HF. Information used in the analysis came from a database located at the Heart Failure Unit of the Hospital del Mar, which covers the catchment area of the study (to identify the cases), and the SIDIAP (system information for the development of research in primary care) for the controls. The SIDIAP is a clinical database of anonymized patient records containing information for almost six million people coming from 274 primary care practices in Catalonia.

Since a number of factors, such as variability in the quality of the primary healthcare record registration, could have influenced the effect of CVRF control on HF occurrence, we designed a case-control study using propensity score-matching methodology. This allowed us to assemble two balanced groups of patients sharing the same co-variable profile, with the exception of those of interest, in order to minimize the effect of potential bias (11) (Figure 1).

All CVRF measurements obtained in a primary care setting during the two years prior to the first HF hospital admission were analysed. Four controls were taken for every case. Time was divided as follows: 24 to 13 months prior to the first HF episode; from 12 months to the day of the first HF episode; and the whole period of 24 months prior to the HF episode.

Study population and inclusion criteria

Primary healthcare in this area is provided by 10 large centres with 168 general practitioners (GPs). All the centres share the same information technology system for recording clinical information electronically. Hospital reports are also available from this system and can be consulted by GPs. Clinical and laboratory test data corresponding to the two years prior to study inclusion were drawn from the primary healthcare records of all participants.

Cases. All patients were admitted because of a first HF episode between 2008 and 2011 to a hospital covering a population of 176 659 inhabitants living in Barcelona (Spain). The diagnoses were retrieved from the hospital discharge reports, which included the International Classification of Diseases (ICD 9: 428).

Controls. Initially, the whole adult population >30 years old residing in the catchment area was selected, who had consulted their GP in the study period, and who were free from HF on the date of inclusion (ICD 10: I50). ICD 10 was employed because it is the classification system used in primary healthcare records in Catalonia.

Matching process. Groups were paired by age, gender, CVRF, treatment for cardiovascular prevention, and the presence of coronary heart disease.

The process was carried out with the 'nearest neighbour (calliper $=0.2$),' 'MatchIt' procedure, which improved the quality of matching (12,13). R Statistical package was used to analyse the results (14). Because of the high number of available controls, it was possible to achieve a very similar co-variable distribution in both groups. As a result, no more than one control for each case was needed to have statistical power to test the main hypothesis.

Variables

The day of the incident HF episode was proposed as the date of study inclusion for both cases and matched controls. Incident HF was defined as a first hospitalization that, on the discharge report, included ICD 9: 428 as a first diagnosis since ICD 9 is the classification system used by hospitals. For each matched potential control, information was collected according to the date of hospital admission of the respective case. This procedure allowed us to obtain the same period of analysis for cases and controls.

Variables used for the matching procedures. Cases and controls were paired for multiple variables that could influence the effect of CVRF on HF occurrence. In addition to socio-demographic variables (gender and age), the following were included in the matching process:

- Cardiovascular co-morbidity: hypertension (ICD 10: I10–I15), dyslipidaemia (ICD10: E78), coronary heart disease (ICD 10: I20–I25), stroke (ICD 10: I61, I63, I64), and diabetes mellitus (ICD 10: E10–E14).
- Drug therapy for CVRF: angiotensin-converting enzyme inhibitors, angiotensin-receptor blocker, dosazoxine, statins, diuretics, calcium channel blockers, beta-blockers, aspirin, and glucose-lowering drugs.

Independent variables considered for analysing the effect of CVRF control on HF occurrence. Values for systolic and diastolic blood pressure, total cholesterol, LDL and HDL cholesterol, heart rate, creatinine, fasting blood glucose, triglycerides, glycosylated haemoglobin, and body mass index were also gathered from medical records. For the purpose of the analysis, measurements were summarized in three groups: the mean of all variable measurements taken the year prior to the inclusion date; the mean of all variable measurements taken in the period 24 to 13 months prior to inclusion; and, and the mean of the whole period (the two years) prior to the inclusion date.

In 2013, European guidelines set a goal to achieve blood pressure levels below 150–140/90 mmHg (15). Lipids were considered well controlled when LDL-cholesterol was lower than 115 mg/dl and glucose control when glycosylated haemoglobin was lower than 7% (16).

Statistical analysis

An initial, descriptive comparison between groups of all variables was performed to evaluate their balance after the matching process. Statistical significance was assessed by the chi-square test when the frequencies of comorbidity and demographic variables were compared among groups. The average determinations between one year and two years prior to the inclusion date were compared by student's *t*-test. This analysis was conducted with complete cases (Tables 1 and 2).

Different logistic regression models (ENTER method) were developed to analyse the relationship among the values of previous determinations of risk factors and the occurrence of HF. Models were adjusted by variables that were statistically significant at bivariate analysis or were clinically associated with HF. In all models, the goodness of fit hypothesis was tested by the Hosmer–Lemeshow test.

Multiple imputation procedure was used to handle missing data in the multivariate analysis using Stata 12 (Stata Statstical Software: Release 12; StataCorp LP, College Station, TX). The estimates of the parameters for each imputed data set were combined using Rubin's rules. Estimates of the risk differences and odds ratios are reported with corresponding 95% confidence intervals and *P* values. In all tests, a level < 0.05 was considered statistically significant.

Ethics

All Helsinki Declaration ethical criteria were respected. Since it was an observational study, participants underwent no interventions other than the usual clinical care. Information from clinical records was correctly anonymized before analysis in order to preserve participants' confidentiality. The study protocol was approved by the Primary Healthcare University Research Institute IDIAP-Jordi Gol.

RESULTS

From the potentially eligible 152 cases and 176 090 controls, 645 participants were finally selected (129 cases and 516 controls) after the matching process, which reduced the probability of a selection bias by 98% (Figure 2).

Table 1. Differences in cardiovascular comorbidity between individuals admitted to hospital because of a first heart failure episode and those without heart failure, after the matching process.

	Cases (heart failure) *n* = 129 *n* (%)	Controls (free from heart failure) *n* = 516 *n* (%)	*P* value
Comorbidity			
Smoker	20 (15.5)	58 (11.2)	0.184
Dyslipidaemia	57 (44.2)	230 (44.6)	0.937
Hypertension	87 (67.4)	363 (70.3)	0.52
Coronary heart disease	23 (17.8)	72 (14.0)	0.267
Diabetes	47 (36.4)	215 (41.7)	0.279
Stroke	11 (8.5)	47 (9.1)	0.836
Sex			
Male	68 (52.7)	268 (51.9)	0.875
Female	61 (47.3)	248 (48.1)	

SD, standard deviation.

Table 2. Differences in cardiovascular risk factor levels and their control between individuals admitted to hospital because of a first heart failure episode (cases) and those free from heart failure (controls), according to the time elapsed before the date of inclusion. (The inclusion date was determined by the hospital admission of the corresponding cases.).

	Last 12 months before the date of inclusion					Between 24 and 13 months before the date of inclusion				
	Cases (heart failure episode) $n = 129$		Controls (free from heart failure) $n = 516$			Cases (heart failure episode) $n = 129$		Controls (free from heart failure) $n = 516$		
Variable	Mean	SD	Mean	SD	*P* value	Mean	SD	Mean	SD	*P* value
Body mass index	33.0	6.2	29.8	5.4	0.001	33.4	6.9	30.0	5.1	<0.001
Heart rate	76.1	10.9	73.6	11.6	0.133	74.9	11.8	73.0	10.9	0.279
Systolic blood pressure	142.2	20.6	135.4	14.4	0.001	138.6	17.8	136.7	16.3	0.413
Diastolic blood pressure	76.2	10.0	73.0	8.9	0.009	75.9	8.0	74.6	9.2	0.331
Blood pressure uncontrolled (%)[a]	44.0%		34.0%		0.12	45.6%		42.9%		0.713
Glycosylated haemoglobin	7.3	1.6	6.9	1.3	0.173	7.3	2.1	6.5	1.2	0.011
Glycosylated haemoglobin ≥ 7 (%)	58.0%		36.5%		0.029	38.1%		24.7%		0.209
Fasting blood glucose	133.6	53.3	128.7	41.4	0.491	153.9	58.0	125.2	33.3	<0.001
Total cholesterol	186.8	42.8	198.2	39.9	0.059	191.7	43.0	199.0	40.2	0.322
HDL cholesterol	46.5	10.5	52.7	14.0	0.004	47.0	12.5	53.8	14.3	0.016
LDL cholesterol	114.7	33.9	120.8	33.7	0.264	111.0	35.0	121.1	34.3	0.154
Lipids uncontrolled (%) LDL ≥ 115)	50.0%		52.7%		0.734	39.3%		54.6%		0.131
Triglycerides	145.2	71.8	128.2	58.6	0.072	197.8	150.8	126.9	61.7	<0.001
Renal function (creatinine)	1.0	0.3	1.0	0.4	0.246	1.0	0.3	1.0	0.3	0.406

HDL, high density lipoprotein; LDL, low density lipoprotein.
[a]Systolic blood pressure ≥ 140 and/or diastolic blood pressure ≥ 90 mmHg.

Population characteristics

The average age was 73.0 (SD: 11.8) years, and 47. 9 % were women. Among cases, 49.6% had HF with preserved ejection fraction. The most frequent CVRF present in both groups was hypertension (69.8%), followed by hypercholesterolaemia (44.5%), type 2 diabetes (40.6%), and smoking (12.1%). Since the two groups were paired for all these factors, differences between them after the matching process were not significant. Mean age was 71.9 (SD: 11.2) in the group of cases and 73.4 (SD: 11.9) in the control group ($P = 0.187$). Other variables are presented in Table 1.

With respect to the number of CVRF present at inclusion, 84.7% had at least one CVRF, and 57.5% had two or more. There were no differences between cases and controls with respect to the number of CVRF present at inclusion (*P* trend $= 0.474$).

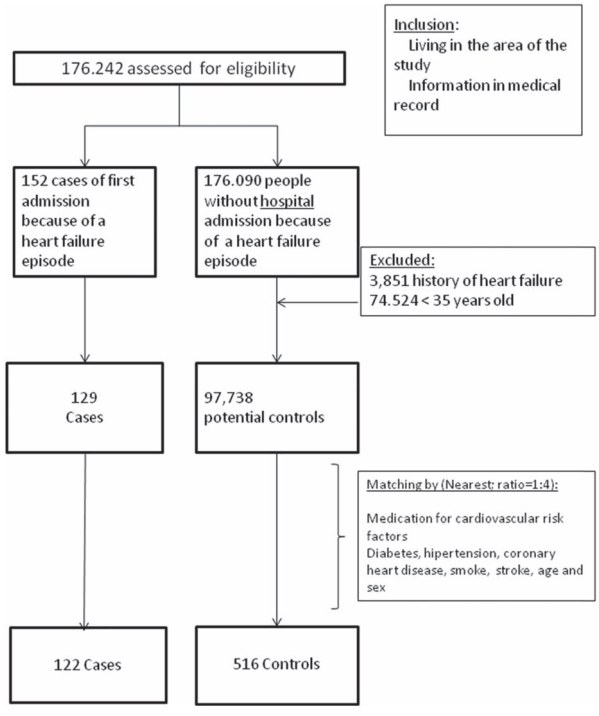


Figure 2. Flow chart showing the participants' selection process.

Cardiovascular risk factors

Regarding the time elapsed before the first HF episode, body mass index was found to be significantly higher in both the previous one and two years (Table 2). Systolic blood pressure was higher during the whole period, particularly in the 12 months prior to the first HF episode. Blood pressure control (systolic blood pressure < 140 mmHg and/or diastolic blood pressure < 90 mmHg) was worse during the whole study period without reaching a statistical significance. Fasting blood glucose levels appeared to be related to HF occurrence. Total cholesterol and low-density lipoprotein cholesterol did not show statistical differences between cases and controls although high-density lipoprotein cholesterol levels were significantly lower in the former.

When analysing all blood pressure measurements taken during the whole period of the study, it was found that 11.0% (mean 2.03) and 6. 0 % (mean 1.2) of the total ($P = 0.001$) were badly controlled in cases and controls, respectively. No differences were found between cases and controls regarding the percentage of well-controlled glycosylated haemoglobin and LDL measurements.

Multivariable analysis, after adjusting for potential confounding factors, showed that the probability of having an HF episode when systolic blood pressure was greater during the year immediately before diagnosis (OR: 1.03 95% CI: 1.01–1.04) (Table 3).

DISCUSSION

Main findings

In this study was found that even moderately increased levels of body mass index and systolic blood pressure determined a higher risk of having a first HF episode. This risk was especially significant in the case of systolic blood pressure where the possibility of experiencing an HF was 1.3 times higher for every 5 mmHg increment during the year prior to the first HF hospitalization episode. Although fasting blood glucose tended to be high in all HF patients, when diabetic patients were considered separately glycosylated haemoglobin did not affect the risk. Regarding lipid profile, only high-density lipoproteins were related to HF occurrence.

Hypertension

Hypertension is the principal risk factor involved in HF occurrence (17). Aggressive lowering of blood pressure has been reported to play a key role in reducing the risk of cardiovascular diseases and help prevent HF development (18). However, the extent to which these levels must be lowered to achieve maximum benefit remains controversial (15). A community based cohort study found a continuous positive association between systolic blood pressure and HF risk in the elderly for levels as low as 115 mmHg (19). Nevertheless, a recent meta-analysis did not observe any benefits in reaching blood pressure levels below 150/80 mmHg (20). Regarding the elderly, two Japanese studies reported that reductions of systolic blood pressure under 142 mmHg did not add any benefit to cardiovascular risk (21,22).

The results are in accordance with those recommending a more exhaustive blood pressure control (19).

Body mass index

In concurrence with other studies, a greater proportion of HF in both subjects with a higher body mass index and individuals with worse fasting blood glucose levels was observed (23–25). However, and in contrast to some other studies, we did not find a higher frequency of HF in diabetic patients with poorly controlled glycosylated haemoglobin, which may have been due to our small sample size (26).

With respect to cholesterol and in agreement with a review published by Kannel in 2000, no relationship was observed between total cholesterol and a higher occurrence of HF (3).

Strengths and limitations

The cases included in the study, had a validated HF diagnosis based on discharge reports. More than 170 000 potential controls (all living in the area and complying with the inclusion criteria) were initially retrieved from

Table 3. Crude and adjusted odds ratio for the first heart failure hospital admission episode according to previous clinical determinations. ORs have been calculated according to the variation for each unit of the corresponding variables.

Variable	OR adjusted	95% CI	*P* value	OR crude	95% CI	*P* value
Time: 12 months before heart failure episode						
Body mass index (kg/m^2)	1.06	(0.99–1.13)	0.072	1.09	(1.03–1.15)	0.005
Heart rate (beat/min)	1.02	(1.00–1.05)	0.074	1.02	(1.00–1.05)	0.055
Systolic blood pressure (mm/Hg)	1.03	(1.01–1.04)	0.006	1.03	(1.01–1.04)	<0.001
Diastolic blood pressure (mm/Hg)	1.02	(0.98–1.05)	0.351	1.04	(1.01–1.07)	0.002
HDL cholesterol (mg/dl)	0.95	(0.92–0.98)	0.005	0.95	(0.92–0.98)	0.001
LDL cholesterol (mg/dl)	1.00	(0.99–1.01)	0.469	0.99	(0.99–1.00)	0.226
Time: between 24 and 13 months before heart failure episode						
Body mass index (kg/m^2)	1.08	(1.01–1.15)	0.022	1.09	(1.04–1.16)	0.001
Heart rate (beat/min)	1.02	(0.99–1.05)	0.159	1.02	(1.00–1.05)	0.086
Systolic blood pressure (mm/Hg)	1.01	(0.99–1.03)	0.274	1.01	(0.99–1.03)	0.263
Diastolic blood pressure (mm/Hg)	1.00	(0.97–1.04)	0.798	1.02	(1.00–1.05)	0.1
HDL cholesterol (mg/dl)	0.97	(0.94–1.00)	0.064	0.97	(0.94–1.00)	0.028
LDL cholesterol (mg/dl)	1.00	(0.99–1.01)	0.578	0.99	(0.99–1.00)	0.280

OR adjusted, odds ratio adjusted by logistic model by presented variables; OR crude, odds ratio without adjustment; 95% CI, 95% confidence interval; HDL, high-density lipoprotein; LDL, low-density lipoprotein

primary healthcare records, which signified a completely external validation of our population. A considerable number of the controls were originally from other European, African, Asian, and Latin American countries. Thus, the population was sufficiently heterogeneous to ensure external validity to many areas with similar population characteristics.

Propensity score methodology allowed us almost to eliminate the potential bias between cases and controls and to introduce a high number of co-variables into the matching process.

Unlike other studies, data from primary healthcare records were used. GPs have a privileged role in the care of a population at high cardiovascular risk and with varying chronic conditions (27,28) because of the possibility of carrying out a long-term follow-up of these patients. The asymptomatic left ventricular systolic dysfunction has been recently reported in approximately one out of every 20 at-risk medical inpatients with at least one HF risk factor (29).

Although HF still has a poor prognosis, the widespread use of evidence-based treatment has improved outcomes (30). Since a substantial number of patients with HF have preserved left ventricular ejection fraction, and that this condition is particularly frequent in older, female, obese, and hypertensive patients, we should be especially careful in treating these subjects and applying preventive measures, which have been proven effective (31).

Since this was an observational retrospective analysis, there could have been some variability in registration and quality of CVRF measurements. Laboratory tests in clinical practice are, however, well standardized and, in this study, centralized in two laboratories. Blood pressure measure procedure has a clear protocol in the primary care setting and it is usually taken with periodically calibrated automatic devices.

We have only employed data from the two years prior to the first HF episode, which means that the long-term effect of CVRF cannot be evaluated.

The lengthy period of analysis (2006–2011) could also have influenced variability in the quality of available data (during this period administrative changes were carried out to improve the register of information in the medical records). The issue was resolved by matching the data from each case with a control at the same period.

Anti-hypertensive medication was included in the matching process to minimize its effect on both CVRF control and the probability of having an HF. Nevertheless, we cannot be sure whether subjects taking more anti-hypertensive drugs had a worse prognosis, in spite of reaching lower blood pressure levels, due to the harmful effects of these medications. The issue could be raised as a future research question.

Implications for clinical practice

These results should encourage GPs to fill the preventive gap that occurs with many patients due to their not displaying a high-risk profile. The results indicate the relevance of sustained, well-controlled CVRF. Moreover, we would like to stress that the continuity and longitudinality of care provided by primary healthcare professionals are crucial in order to prevent a first episode of HF in a population at risk.

Conclusion

Increased levels of body mass index and systolic blood pressure during the previous 24 months may determine a higher risk of having a first HF hospitalization episode.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES

1. World Health Organization. The 10 leading causes of death in the world, 2000 and 2012. Fact sheet No. 310.Updated May 2014. Available at: http://who.int/mediacentre/factsheets/fs310/en/ (accessed 29 January 2015).
2. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Borden WB, et al. American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Executive summary: Heart disease and stroke statistics—2013 update: A report from the American Heart Association. Circulation 2013;127:143–52.
3. Kannel WB. Incidence and epidemiology of heart failure. Heart Fail Rev. 2000;5: 167–73.
4. Kalogeropoulos A, Georgiopoulou V, Kritchevsky SB, Psaty BM, Smith NL, Newman AB, et al. Epidemiology of incident heart failure in a contemporary elderly cohort: The health, aging, and body composition study. Arch Intern Med. 2009;169:708–15.
5. Cowie MR, Wood DA, Coats AJ, Gibbs JS, Underwood SR, Turner RM, et al. Coronary artery disease as the cause of incident heart failure in the population. Eur Heart J. 2001;22:228–36.
6. Mosterd A, Hoes AW. Clinical epidemiology of heart failure. Heart 2007;93:1137–46.
7. Nichols GA, Gullion CM, Koro CE, Ephross SA, Brown JB. The incidence of congestive heart failure in type 2 diabetes: An update. Diabetes Care 2004;27:1879–84.
8. Gopal DM, Kalogeropoulos AP, Georgiopoulou VV, Smith AL, Bauer DC, Newman AB, et al. Cigarette smoking exposure and heart failure risk in older adults: The health, aging, and body composition study. Am Heart J. 2012;164:236–42
9. Avery CL, Loehr LR, Baggett C, Chang PP, Kucharska-Newton AM, Matsushita K, et al. The population burden of heart failure attributable to modifiable risk factors: The ARIC (atherosclerosis risk in communities) study. J Am Coll Cardiol. 2012;60:1640–6.
10. Folsom AR, Yamagishi K, Hozawa A, Chambless LE. Absolute and attributable risks of heart failure incidence in relation to optimal risk factors. Circ Heart Fail. 2009;2:11–7.
11. D'Agostino RB Jr. Propensity scores in cardiovascular research. Circulation 2007;115:2340–3.

12. Ho D, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 2007;15:199–236.
13. Austin PC. Optimal calliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10:150–61.
14. R Development Core Team. R: A language and environment for statistical computing. Vienna, R Foundation for Statistical Computing: 2008. Available at: http://www.R-project.org (accessed 4 May 2015).
15. Mancia G, Fagard R, Narkiewicz K, Redon J, Zanchetti A, Böhm M, et al. Task Force for the Management of Arterial Hypertension of the European Society of Hypertension and the European Society of Cardiology. 2013 ESH/ESC practice guidelines for the management of arterial hypertension. Blood Press 2014;23:3–16.
16. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren M, et al. European Association for Cardiovascular Prevention & Rehabilitation (EACPR); ESC Committee for Practice Guidelines (CPG). European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). Eur Heart J. 2012;33:1635–701.
17. Tocci G, Sciarretta S, Volpe M. Development of heart failure in recent hypertension trials. J Hypertens. 2008;26:1477–86.
18. Black HR. The burden of cardiovascular disease: Following the link from hypertension to myocardial infarction and heart failure. Am J Hypertens. 2003;16(Suppl.2):4–6.
19. Butler J, Kalogeropoulos AP, Georgiopoulou VV, Bibbins-Domingo K, Najjar SS, Sutton-Tyrrell KC, et al. Systolic blood pressure and incident heart failure in the elderly. The cardiovascular health study and the health, ageing and body composition study. Heart 2011;97:1304–11.
20. Briasoulis A, Agarwal V, Tousoulis D, Stefanadis C. Effects of antihypertensive treatment in patients over 65 years of age: A meta-analysis of randomised controlled studies. Heart 2014;100:317–23.
21. JATOS Study Group. Principal results of the Japanese trial to assess optimal systolic blood pressure in elderly hypertensive patients (JATOS). Hypertens Res. 2008;31:2115–27.
22. Ogihara T, Saruta T, Rakugi H, Matsuoka H, Shimamoto K, Shimada K, et al. Target blood pressure for treatment of isolated systolic hypertension in the elderly: Valsartan in elderly isolated systolic hypertension study. Hypertension 2010;56:196–202.
23. Loehr LR, Rosamond WD, Poole C, McNeill AM, Chang PP, Folsom AR, et al. Association of multiple anthropometrics of overweight and obesity with incident heart failure: The atherosclerosis risk in communities study. Circ Heart Fail. 2009;2:18–24.
24. Dunlay SM, Weston SA, Jacobsen SJ, Roger VL. Risk factors for heart failure: A population-based case-control study. Am J Med. 2009;122:1023–8.
25. Nichols GA, Gullion CM, Koro CE, Ephross SA, Brown JB. The incidence of congestive heart failure in type 2 diabetes: An update. Diabetes Care 2004;27:1879–8.
26. Lind M, Olsson M, Rosengren A, Svensson AM, Bounias I, Gudbjörnsdottir S. The relationship between glycaemic control and heart failure in 83 021 patients with type 2 diabetes. Diabetologia 2012;55:2946–5.
27. Macinko J, Starfield B, Shi L. The contribution of primary care systems to health outcomes within Organization for Economic Cooperation and Development (OECD) countries, 1970–1998. Health Serv Res. 2003;38:831–65.
28. Cancian M, Battaggia A, Celebrano M, Del Zotti F, Novelletto BF, Michieli R, et al. The care for chronic heart failure by general practitioners. Results from a clinical audit in Italy. Eur J Gen Pract. 2013;19:3–10.
29. Martin LD, Mathews S, Ziegelstein RC, Martire C, Howell EE, Hellmann DB, et al. Prevalence of asymptomatic left ventricular systolic dysfunction in at-risk medical inpatients. Am J Med. 2013;126:68–73.
30. Maggioni AP, Dahlström U, Filippatos G, Chioncel O, Crespo Leiro M, Drozdz J, et al. Heart Failure Association of the European Society of Cardiology (HFA). EUR Observational Research Programme: Regional differences and 1-year follow-up results of the Heart Failure Pilot Survey (ESC-HF Pilot). Eur J Heart Fail. 2013; 15:808–17.
31. Gurwitz JH, Magid DJ, Smith DH, Goldberg RJ, McManus DD, Allen LA, et al. Contemporary prevalence and correlates of incident heart failure with preserved ejection fraction. Am J Med. 2013;126:393–400.

7. Discusión conjunta de los artículos

El primer artículo de esta tesis doctoral(6) ha constatado el extensivo y creciente uso de técnicas estadísticas de regresión multivariable en los artículos de investigación médica publicados en revistas indexadas en PubMed, especialmente los modelos de regresión logística. El incremento observado coincide con el incremento de la capacidad computacional de los ordenadores y la cada vez mayor disponibilidad de estos métodos en paquetes estadísticos estándares y la mayor facilidad para realizar estos análisis (7). Otras revisiones previas de metodología estadística empleada en artículos publicados en campos específicos de ciencias de la salud, también muestran un mayor y creciente frecuencia de utilización de los modelos de regresión, predominando el modelo logístico (4,53).

A pesar de ello, el uso de estas técnicas no está libre de potenciales errores y no siempre su uso es apropiado, pues sobre ellas descasan fuertes asunciones que no siempre se cumplen. En este sentido, existen herramientas estadísticas que permiten evaluar si los modelos cumplen las condiciones de aplicación(9,11) y así poder evitar errores de especificación de los modelos. En esta línea, el segundo artículo de la presente tesis publicado en la revista Medicine (54) mostró el preocupante bajo reporte en la verificación de las hipótesis de estas técnicas de ajuste en trabajos publicados e indexados en el repositorio PubMed. En concreto, los resultados mostraron que solo 1 de cada 4 trabajos que utiliza MRM, informa de la realización de alguna técnica diagnóstica de los modelos o presenta estadísticos de bondad de ajuste de los modelos utilizados en su análisis. Otros aspectos revisados, como la evaluación de las interacciones o el reporte de estimaciones crudas y ajustadas, también se encontró en un porcentaje bajo de los artículos revisados. A pesar de ello, encontramos que los estudios publicados en revistas con un mayor factor de impacto presentaron con mayor frecuencia información más completa de la utilización de las técnicas de validación de los modelos de regresión. Nuestros resultados fueron consistentes con los estudios previos (5,15,55). En el año 2002 el estudio de Müllner M et al. (55) mostró que las revistas con mayor factor de impacto presentaban un más completo reporte estadístico de los MRM, quizás gracias a que su proceso editorial, incluye de forma específica una revisión estadística. Una revisión sistemática realizada por Casals et al. (15) de 108 artículos que aplicaban modelos mixtos lineales generalizados, sin discriminar entre tipos de diseños, encontró que la validación del modelo y/o pruebas de bondad de ajuste únicamente se reportaba en el 15,7% de los

artículos. Otro trabajo, realizado por Groenwold, R.H et al. (5) encontró de forma general una falta de atención a los métodos de ajuste en los estudios observacionales analíticos, hecho que contrasta con los estudios sobre modelos predictivos, o evaluación de pruebas diagnósticas, donde las combinaciones de variables son modeladas con mayor precisión (56-60). En este sentido la guía TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) (61) proporciona directrices y señala los aspectos esenciales para desarrollar y validar un modelo predictivo. Entre los puntos esenciales que esta guía incluye están: realizar una validación interna y externa, evaluar el rendimiento del modelo y realizar un análisis de sensibilidad del modelo finalmente seleccionado. Sin embargo, otras guías de recomendaciones, específicas para estudios observacionales analíticos (como la STROBE(1)), o sobre aspectos más estadísticos (como SAMPL(62)), no incluyen entre sus puntos esenciales la validación del modelo estadístico empleado. La necesidad de evaluar críticamente la calidad metodológica de los estudios ha puesto de manifiesto graves deficiencias en los artículos de investigación y las guías de recomendaciones son herramientas desarrolladas para facilitar información más exacta de los aspectos clave de los estudios de investigación(63) . Sin embargo, estas guías todavía siguen estando incompletas en cuanto a aspectos sobre metodología estadística multivariable.

Los MRMs son herramientas muy poderosas y utilizadas para estimar el efecto real, aislando los potenciales factores de confusión, pero el incumplimiento de las asunciones formales de los modelos puede invalidar los resultados que se derivan del estudio, produciendo importantes sesgos de las estimaciones(13). Con el fin de evitar estas circunstancias, la construcción de un modelo estadístico, además de basarse en hipótesis previas, requiere de la correcta especificación del modelo y realizar un buen diagnóstico del modelo verificando o examinando la bondad de ajuste a los datos o examinando la calibración del mismo (8). Paradójicamente, sorprende que a pesar de la importancia de estas comprobaciones formales en los resultados inferenciales en los estudios analíticos, se encuentre un porcentaje tan elevado (≈75%)(64) de trabajos que ni siquiera presentan una sola medida de bondad de ajuste de los métodos estadísticos utilizados (ni declaración de intenciones en la sección de métodos del artículo). Ésta práctica habitual en la investigación médica publicada, puede reflejar la poca atención o relevancia que se le da a la metodología estadística (o al propio uso). Las consecuencias de la incorrecta especificación del modelo de regresión es que, aunque se tengan en cuenta todos los confusores aún puede permanecer lo que se conoce como confusión residual(14,65). Sin embargo, el

investigador puede no ser consciente de este riesgo potencial debido a que normalmente se desconoce la verdadera medida de asociación.

En este sentido el tercer trabajo de la presente tesis (manuscrito) (66) mediante un estudio de simulación, muestra empíricamente el sesgo residual existente y la elevada tasa de falsos positivos que pueden ocurrir, si una técnica de regresión multivariable (como es la regresión logística) es inapropiada (situación inapreciable si la verificación de las asunciones del modelo es ignorada).

Para hacer frente a la mala especificación de los modelos paramétricos, la aproximación semi-paramétrica con la inclusión de funciones de suavizado tipo *smothing splines* de los confusores continuos (como los modelos GAM), mostró mejores resultados (estimaciones con menor sesgo) en comparación con la representación 100% paramétrica, incluso sin predeterminar la forma funcional del confusor. Los modelos GAM, son una extensión de los modelos tradicionales (GLM) donde los predictores lineales son reemplazados con ajustes de funciones no-paramétricas para estimar la relación entre una variable respuesta y los predictores independientes no lineales.

Los resultados del estudio de simulación de la tesis doctoral (66) son consistentes con los aportados por Benedetti A et al.(67). Este estudio también investigó el rendimiento de los modelos GAM, aunque solo centrándose en evaluar estrategias de análisis con funciones no paramétricas que involucraban la elección del mejor parámetro de suavizado (gl) en 3 estrategias (fijando gl=4, gl según el criterio de información de Akaike (AIC) y gl según Bayes (BIC)). Sin embargo, en nuestro estudio, se comparó la estimación con GAM, estableciendo el grado de suavizado de forma automático, como parte del modelo ajustado, frente a otras técnicas de ajuste. Además, a diferencia del realizado por Benedetti A. et al.(67) y otros trabajos de simulación, que anteriormente han examinado técnicas de ajuste(48-50), el trabajo que se presenta considera un escenario de efecto nulo y el confusor continuo es generado según una distribución asimétrica y truncada (circunstancia también común en estudios clínicos, por ejemplo derivada de los criterios de inclusión de las covariables continuas incluidas). Además, a pesar de los óptimos resultados, los modelos GAM también tienen sus riesgos potenciales y limitaciones como son el sesgo de los errores estándar en presencia de concurbidity (alta correlación entre dos o más covariables) (68) y la medida de asociación estimada de la parte no paramétrica sólo se puede apreciar mediante una figura de la relación no lineal, dificultando la interpretación de sus coeficientes.

Pero el estudio de simulación de la presente tesis doctoral (66) también examina el comportamiento de 3 métodos *matching* (PS-M), con estimación posterior, como alternativa a los GLM (modelo logístico). Los resultados muestran como las 3 estrategias matching evaluadas en los escenarios planteados obtuvieron mejores resultados (menor sesgo (ECM) y menor $\hat{\alpha}$), generando estimaciones más creíbles del efecto nulo, en comparación con el ajuste mediante un MRM y por lo tanto demostrando ser un método más robusto. Además este trabajo evidenció un mayor beneficio de PS-M sobre todo, en comparación con las distintas especificaciones de regresión (regresión logística paramétrica), donde los MRM pueden alcanzar altas tasas de falsos positivos en determinados escenarios.

Anteriormente la serie estudios de Austin PC.(48-50) había examinado, también mediante simulación, el rendimiento de varios métodos basados en el PS sobre la estimación de distintas medidas de asociación (odds ratio, riesgo relativo y hazard ratio). Estas investigaciones mostraron que los métodos PS-M producían estimaciones con menor sesgo, junto con el método IPTW-PS (Inverse probability of treatment weighting using PS), en comparación con otras técnicas (*subclas* por PS o ajuste por PS), si bien el método *subclas* proporcionó estimaciones con mayor precisión. Estos estudios se centraron solamente en el algoritmo matching Nearest-Neighbour(N-N), que también había mostrado un buen rendimiento frente a otras alternativas comunes (47). Pero a diferencia del trabajo que se presenta en esta tesis, en anteriores estudios (47-50), no se consideraron escenarios donde fuera probable una mala especificación de los MRMs, a causa, por ejemplo de existir relaciones no lineales difíciles de modelar mediante MRM, circunstancia habitual en estudios epidemiológicos (69). Por lo tanto el trabajo de simulación de esta tesis añade a los trabajos previos (47-50) la comparativa de 3 estrategias M-PS en escenarios donde, según nuestro conocimiento, no habían sido todavía evaluados.

Entre los tres algoritmos matching examinados, *exact* y N-N obtuvieron parecidos errores cuadráticos medios (ECM), mientras que el método *subclas* presentó el mayor error de tipo I empírico ($\hat{\alpha}$) de los tres métodos matching examinados en la mayoría de los escenarios considerados. Una posible explicación es que el método *subclas* elimina menos muestra (20% versus 40% en promedio) en comparación con el resto, debido a que el algoritmo *subclas* solo requiere que las observaciones dentro de los estratos se parezcan. En esencia, el método *subclas* con un confusor sería similar a la estrategia GLM categorizando el confusor (*Logitcat*) mejorado por la eliminación de

observaciones. Esta última estrategia matching (subclas) es la que presentó peores resultados ($\hat{\alpha}$ =0.061, y $\hat{\alpha}$=0,084 en promedio) permaneciendo un ligero sesgo residual en algunas situaciones. Por otra parte, el método *exact* es el que obtuvo un menor $\hat{\alpha}$ entre los tres métodos matching evaluados. Esto es debido a que la estrategia *exact* empareja las observaciones de forma que la distribución de covariables es exactamente igual entre los grupos, por lo tanto se asimilaría al método *fullmatching* (o al diseño experimental completamente bloqueado). Ésta característica hace que además del sesgo, también se reduzca la varianza de la estimación y por lo tanto en términos de eficiencia, resulte más recomendable en contra de los métodos PS-M aproximados (N-N, o subclas basados en PS), tal como también señala recientemente Gary King (70).

Por otro lado, cuando la relación confusor-*outcome* se simuló de forma lineal (por lo tanto, teóricamente sin apenas mala especificación en los MRMs), el análisis mediante métodos matching obtuvo una estimación del α ligeramente inferior al valor nominal ($\hat{\alpha} < \alpha = 0{,}05$). Una posible explicación es que la estrategia de análisis utilizada no tuvo en cuenta la naturaleza emparejada de la muestra. En este sentido, existe cierto debate y controversia sobre si el análisis del *outcome* después del matching requiere tener en cuenta la naturaleza emparejada de los datos (23,39-43). Recientemente Neil Pearce (43) discute sobre este asunto, particularmente en diseños de casos y controles, y señala que el análisis estándar (no condicional, sin tener en cuenta el emparejamiento de la muestra) después del matching puede infraestimar la varianza del estimador hecho que podría explicar nuestros resultados en cuanto a las diferencias encontradas entre el error de tipo I empírico y el nominal. Parte de las diferencias encontras ($\hat{\alpha} \neq \alpha$) también pueden ser debidas a que la distribución del confusor no proviene de una distribución normal simétrica y, además, siempre hay que considerar la presencia del error de estimación muestral, el cual con 7500 réplicas estaría próximo a ±0,5% (para un α=0,05).

Sin embargo, en el contexto de los estudios basados en registros clínicos, la estimación de la varianza debido al error muestral (y por lo tanto la potencia estadística), no es una limitación preocupante, ya que la característica común de estos estudios es que cuentan con gran cantidad de información y con grandes muestras poblacionales. Si bien tradicionalmente el emparejamiento se ha usado en estudios de casos y controles, con muestras más pequeñas ad-hoc y sobre todo en el marco de la inferencia causal. En esta tesis doctoral se presentan tres artículos científicos (36,71,72) basados en información clínica de grandes bases de datos con 3 diseños

epidemiológicos distintos (transversal, cohortes y, de casos y controles) utilizando la metodología matching explicada anteriormente y que se ha demostrado empíricamente según el estudio de simulación presentado, que tienen un mejor control del sesgo de confusión(66).

Los principales resultados de estos tres trabajos clínicos donde se utilizó la metodología matching son: 1) los bifosfonatos pueden no ser eficaces en población general para prevenir las fracturas osteoporoticas(71); 2) la carga atribuida a ser cuidador informal está asociada a tener más frecuencia de depresión y un menor soporte social (36); y 3) el buen control del índice de masa corporal (IMC) y la presión arterial sistólica se asocia a un menor riesgo de sufrir un episodio de cardiopatía isquémica (72).

Las fuentes primarias de los tres estudios aplicados son, o bien registros clínicos provenientes de la práctica clínica habitual (46), o bien datos de una gran encuesta nacional como es la Encuesta Nacional de Salud (73). Por lo tanto, para la realización de estos estudios retrospectivos, no se necesitó un trabajo de campo adicional, ni un seguimiento posterior, lo cual posibilitó la realización de estos estudios a un costo muy inferior al de los estudios prospectivos ad-hoc. No obstante, una de las dificultades de estos estudios, es que su viabilidad depende de la disponibilidad de suficiente información y/o la calidad de ésta (por ejemplo: los registros clínicos conllevan sesgos implícitos relativos al infra registro). Sin embargo, gracias a la disponibilidad de grandes muestras, también es posible configurar un muy alto estándar de calidad en la selección del participante final, eliminando potenciales sesgos de selección, permitiendo la inclusión de grupos más homogéneos. Por ejemplo, en el estudio OSTEOFRAC (71), a pesar de tener una muestra potencial inicial de 42.234 mujeres, solamente se incluyeron usuarias activas en el sistema de salud, con evidencia de seguimiento y sin dispensación de otros fármacos modificadores óseos que pudieran interferir en la visión del posible efecto estudiado (ver figura 2 y metodología del cuarto artículo de la tesis doctoral (71)).

En sendos estudios observacionales con hipótesis analíticas(36,71,72), existían importantes factores de confusión conocidos. En el estudio CUIDADORES(36), antes del tratamiento de datos, era conocido el papel de variables como la edad, sexo, y factores socioeconómicos, como factores a tener en cuenta (confusores potenciales) asociados a ser cuidador informal y a los resultados en salud a evaluar. Por lo tanto, la estrategia analítica debía incluir el uso de alguna de las herramientas para controlar o minimizar el posible sesgo de confusión.

A modo de ejemplo en el anexo II (Modelo 3) de esta tesis se puede ver los resultados del análisis de uno de los *outcomes* (Depresión si/no) del trabajo CUIDADORES (36), mediante un análisis de regresión logística estandard, utilizando toda la muestra (n=20.029; 515 cuidadores informales (CI) y 19.514 controles) y ajustado por todas las variables potencialmente confusoras (Las mismas incluidas en la aproximación matching utilizada en el artículo publicado). Según estos resultados la estimación por MRM el OR de depresión de los cuidadores seria de 1,57 (IC95%:1,21-2,04; p=0,001), en cambio el análisis mediante la estrategia matching (muestra n=2.568; 515 CI y 2053 controles) obtuvo un OR estimado de 1,33 (IC95%:1,06-1,68; p=0,02). Como vemos existen substanciales diferencias entre las dos aproximaciones sobre la estimación de la misma medida de asociación: La estimación vía GLM, obtiene un OR de 1,57 lo que supone un incremento que un 57% del odds (o riesgo), y la misma estimación con la aproximación M-PS queda reducida, a un incremento del odds (o riesgo) de depresión del 33%. Como se ha visto en anteriores investigaciones (52), existe una fuerte dependencia del método de ajuste sobre la medida de asociación estimada. La primera aproximación (MRM), utiliza toda la muestra (n=20.029) pero el modelo que se utiliza para realizar la inferencia no se ajusta a los datos, ya que por ejemplo asume un incremento quasilineal de la frecuencia de depresión en relación a la edad, hecho que no queda reflejado por los datos (anexo II; figura A4), y además la hipótesis de bondad de ajuste de H&L es rechazada (p<0,001) (anexo II bondad de ajuste del modelo 3). Por lo tanto, aunque mediante la aproximación GLM se aprovechan todas las observaciones (estrategia aparentemente ideal en términos de potencia estadística) los resultados no son tan verosímiles, a pesar de haber realizado el ajuste por regresión. En cambio la aproximación matching con solamente un 13% (n=2.568) de la muestra respecto la inicial, se obtiene una estimación de menor magnitud (aun significativa, incluso con mayor precisión), pero probablemente más próxima a la realidad. Estas diferencias aún podrían tener una mayor trascendencia y gravedad si la confusión residual afectase a un efecto nulo (como los escenarios planteados en el estudio de simulación), que implican resultados falsos positivos.

En la actualidad, hay una creciente preocupación sobre sí los resultados de las investigaciones publicadas pueden ser falsos. En este sentido, John P.A. Ianndidis(74) afirmaba que “cuanto mayor es la flexibilidad en los diseños, las definiciones, los resultados y los modos de análisis en un campo científico, es menos probable que los resultados de la investigación sean verdad”. Como hemos visto en los resultados de esta tesis la estimación mediante el uso de estrategias matching, reduce el

desequilibrio de los datos, la sensibilidad al método de análisis, la dependencia al modelo, la discreción del investigador y el sesgo. Por lo tanto, en el contexto de estudios basados en registros clínicos y muestras grandes, la metodología matching tiene un gran potencial, ya que la eliminación de observaciones puede no ser determinante, en términos de eficiencia y poder estadístico, y además el beneficio en validez interna es considerable.

8. Limitaciones de la tesis doctoral

Los resultados y artículos de esta tesis doctoral presentan ciertas limitaciones ya comentadas y consideradas en cada uno de los artículos presentados. A continuación se exponen y resumen algunas de las limitaciones:

- El incremento del uso de los modelos paramétricos detectado, se basó en el motor de búsqueda de PubMed y la estrategia de búsqueda utilizada. Además, los resultados de la búsqueda dependen de que los autores hayan mencionado las técnicas de regresión usadas en el resumen. Este tipo de limitación es común en los estudios bibliométricos basados en motores de búsqueda por palabras como PubMed (53). Sin embargo, teniendo en cuenta que a partir del año 1990 los descriptores para los métodos de regresión ya estaban incorporados en PubMed y que nos hemos basado en todo su repositorio, en términos de evolución, el efecto de esta limitación puede ser reducido.

- Ésta característica también afecta a la revisión metodológica de artículos, donde el universo potencial de artículos que usaban técnicas multivariables, fue limitada a la sensibilidad de la estrategia de búsqueda y a su repositorio. En un esfuerzo para no tener que descartar muchos artículos durante la revisión manual, se diseñó una estrategia altamente específica.

- Por otro lado, no todos los aspectos revisados sobre el reporte de los MRM tienen la misma relevancia y sentido. Por ejemplo, la evaluación de las interacciones no siempre se justifica, especialmente en muestras pequeñas, o hay estudios sobre intervenciones médicas en las que la confusión podría considerarse irrelevante. Además hay otros aspectos importantes que podrían afectar a la calidad de los análisis y resultados (por ejemplo, si el modelo se especificó previamente antes de realizar el análisis de datos, o el grado de conocimiento estadístico del equipo investigador).

- En cuanto a la evaluación diagnostica de los modelos, estamos de acuerdo que no existe una prueba de bondad de ajuste analítica que garantice la certeza de la correcta especificación de un modelo. Por ejemplo, un R^2 bajo no garantiza el incumplimiento de las hipótesis del modelo, pero un R^2 bajo puede ser un síntoma de que el modelo está incorrectamente especificado. Además las pruebas analíticas de bondad de ajuste son sensibles al tamaño de la muestra por lo tanto en los estudios con muestras grandes, pequeñas desviaciones producirán significación estadística rechazándose la hipótesis de bondad de ajuste(44). Por lo tanto para evaluar las posibles discrepancias entre los datos y el modelo y la necesidad de remodelar, pueden ser más apropiadas pruebas de validación descriptiva o gráfica.

- La generalización de los resultados de la simulación se limita a los escenarios planteados, que pueden ser considerados relativamente simples. Sin embargo, los fenómenos de interés binarios son tratados de forma habitual en la literatura médica(10,75).

- La librería MatchIt de R y la metodología matching incorpora otros algoritmos que no han sido examinados (*Optimal, Genetic, Fullmatching Coarseted-Matching*). Además estos métodos permiten múltiples configuraciones como el uso de la distancia de Mahalanobis en vez del PS, u otras opciones (caliper≠0,1; ratio tratado-control≠1). Desafortunadamente no se han podido realizar un análisis de sensibilidad de más algoritmos y opciones por la elevada intensidad computacional que hubiera supuesto, y nos hemos centrado en solo los 3 algoritmos matching más familiares y conocidos (*Exact, N-N y subclas*), con las especificaciones habitualmente utilizadas (Caliper=0,1; ratio tratado-control=1). En este sentido futuras extensiones de este trabajo serían deseables.

9. Conclusiones

- Los modelos de regresión multivariables (logística, lineal, Cox y Poisson) son cada vez más utilizados en estudios observacionales publicados e indexados en MEDLINE en revistas internacionales a nivel global como en revistas publicadas en lengua española. Debido al aumento de la utilización, y la popularización de los métodos multivariables, parece necesario establecer estándares que garanticen la correcta aplicación y buen uso de estas técnicas.

- Un bajo porcentaje de los trabajos observacionales indexados en MEDLINE que utilizan técnicas multivariables como herramienta de ajuste (Lineal, logístico y Cox), proporcionan información que asegure la rigurosa aplicación del modelo de ajuste seleccionado. Dada la importancia de estos métodos en los resultados y las conclusiones finales, especialmente en estudios observacionales analíticos, sería deseable un mayor rigor en la aplicación y presentación de los MRM en la literatura científica. En este sentido, sería recomendable que los procesos editoriales y/o las directrices o guías en que se basan las revistas científicas biomédicas ampliasen a metodología multivariable.

- Según el estudio de simulación de la tesis doctoral se puede concluir que:

    a. La estimación paramétrica con GLM (Regresión logística), como método de ajuste en presencia de confusión, puede resultar muy sesgada si la comprobación de las asunciones del modelo seleccionado es ignorada.

    b. La estimación mediante un modelo GAM (Link logít) con una representación no paramétrica del confusor continuo presenta mejores resultados que los GLM paramétricos en términos de reducción del sesgo de confusión.

    c. La estimación aplicando metodología matching presentan una mayor robustez en comparación con una técnica de regresión paramétrica (Modelo logístico), ya que eliminan mejor el sesgo de confusión y presenta una menor tasa de resultados falsos positivos en la estimación de un efecto nulo.

- En un contexto de estudios observacionales basados en la práctica clínica o, con grandes muestras disponibles (*Real world data*) se recomienda la aplicación de los métodos matching ya que estos proporcionan mayor credibilidad a los resultados en comparación con los métodos de regresión tradicionales.

10. Bibliografía

(1) Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 2007;45(4):247-251.

(2) Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet 2002;359(9302):248-252.

(3) Vandenbroucke JP, Von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Gac Sanit 2009 Mar-Apr;23(2):158.

(4) Yergens DW, Dutton DJ, Patten SB. An overview of the statistical methods reported by studies using the Canadian community health survey. BMC Med Res Methodol 2014 Jan 25;14:15-2288-14-15.

(5) Groenwold RH, Van Deursen AM, Hoes AW, Hak E. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. Ann Epidemiol 2008 Oct;18(10):746-751.

(6) Real J, Cleries R, Forne C, Roso-Llorach A, Martinez-Sanchez JM. Use of multiple regression models in observational studies (1970-2013) and requirements of the STROBE guidelines in Spanish scientific journals. Semergen 2015 Nov 5.

(7) Gentle JE, Härdle WK, Mori Y. How computational statistics became the backbone of modern data science. Handbook of Computational Statistics: Springer; 2012. p. 3-16.

(8) Núñez E, Steyerberg EW, Núñez J. Estrategias para la elaboración de modelos estadísticos de regresión. Revista Española de Cardiología 2011;64(6):501-507.

(9) Dobson AJ. An Introduction to Generalized Linear Models. 2nd ed. United States of America: Chapman and Hall; 2001.

(10) Bender R. Introduction to the use of regression models in epidemiology. In: Mukesh Verma, editor. Methods in Molecular Biology, Cancer Epidemiology United States: Springer Science; 2009. p. 179-195.

(11) Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med 1995 Aug 15;14(15):1707-1723.

(12) Cobo E. Multivariate models in biomedical research: criteria for the inclusion of variables. Med Clin (Barc) 2002 Jul 13;119(6):230-237.

(13) Liang W, Zhao Y, Lee AH. An investigation of the significance of residual confounding effect. Biomed Res Int 2014;2014:658056.

(14) Groenwold RH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KG, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. CMAJ 2013 Mar 19;185(5):401-406.

(15) Casals M, Girabent-Farres M, Carrasco JL. Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000-2012): a systematic review. PLoS One 2014 Nov 18;9(11):e112653.

(16) Strasak AM, Zaman Q, Pfeiffer KP, Gobel G, Ulmer H. Statistical errors in medical research-a review of common pitfalls. Swiss medical weekly 2007;137(3/4):44.

(17) Groenwold RH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. J Clin Epidemiol 2009 Jan;62(1):22-28.

(18) Gan HL, Zhang JQ, Bo P, Wang SX, Lu CS. Statins decrease adverse outcomes in coronary artery bypass for extensive coronary artery disease as well as left main coronary stenosis. Cardiovasc Ther 2010 Apr;28(2):70-79.

(19) Chaux A, Peskoe SB, Gonzalez-Roibon N, Schultz L, Albadine R, Hicks J, et al. Loss of PTEN expression is associated with increased risk of recurrence after prostatectomy for clinically localized prostate cancer. Mod Pathol 2012 Nov;25(11):1543-1549.

(20) Stamou SC, Hill PC, Haile E, Prince S, Mack MJ, Corso PJ. Clinical outcomes of nonelective coronary revascularization with and without cardiopulmonary bypass. J Thorac Cardiovasc Surg 2006 Jan;131(1):28-33.

(21) Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. : Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

(22) Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis 2007;15(3):199-236.

(23) Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research 2011;46(3):399-424.

(24) Greenwood E. Experimental sociology: A study in method. : King's crown Press; 1945.

(25) Chapin FS. Experimental Designs in. Sociological Research.New York: Harper and Brothers 1947.

(26) Cochran WG, Rubin DB. Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A 1973:417-446.

(27) Althauser RP, Rubin D. The computerized construction of a matched sample. American Journal of Sociology 1970:325-346.

(28) Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 1985;39(1):33-38.

(29) Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci 2010 Feb 1;25(1):1-21.

(30) Ho D, Imai K, King G, Stuart E. MatchIt: MatchIt: Nonparametric Preprocessing for Parametric Casual Inference. R package version 2006:2.2-11.

(31) Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996:249-264.

(32) King G, Nielsen R, Coberley C, Pope JE, Wells A. Comparative effectiveness of matching methods for causal inference. Unpublished manuscript 2011;15.

(33) King G, Lucas C, Nielsen R, King G, Pan J, Roberts M, et al. The Balance-Sample Size Frontier in Matching Methods for Causal Inference}. PS: Political Science and Politics} 2014;42:S11-S22.

(34) Agodini R, Dynarski M. Are experiments the only option? A look at dropout prevention programs. Rev Econ Stat 2004;86(1):180-194.

(35) Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. The International Journal of Biostatistics 2009;5(1).

(36) Gonzalez-de Paz L, Real J, Borras-Santos A, Martinez-Sanchez JM, Rodrigo-Banos V, Dolores Navarro-Rubio M. Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study. J Public Health Policy 2016 May;37(2):173-189.

(37) Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica 2006;74(1):235-267.

(38) Glazerman S, Levy DM, Myers D. Nonexperimental versus experimental estimates of earnings impacts. Ann Am Acad Pol Soc Sci 2003;589(1):63-93.

(39) Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. Stat Med 2006;25(13):2230-2256.

(40) Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis 2007;15(3):199-236.

(41) Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods 2008;13(4):279.

(42) Stuart EA. Developing practical recommendations for the use of propensity scores: Discussion of'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003'by Peter. Stat Med 2008;27(12):2062-2065.

(43) Pearce N. Analysis of matched case-control studies. BMJ 2016 Feb 25;352:i969.

(44) Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. Stat Med 2013;32(1):67-80.

(45) Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. : Cambridge university press; 2011.

(46) Bolibar B, Fina Aviles F, Morros R, Garcia-Gil Mdel M, Hermosilla E, Ramos R, et al. SIDIAP database: electronic clinical records in primary care as a source of information for epidemiologic research. Med Clin (Barc) 2012 May 19;138(14):617-621.

(47) Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat 2011 Mar-Apr;10(2):150-161.

(48) Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med 2013;32(16):2837-2849.

(49) Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 2007;26(16):3078-3094.

(50) Austin PC. The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol 2008;61(6):537-545.

(51) Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol 2008 Oct;37(5):1142-1147.

(52) Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006 Feb 1;163(3):262-270.

(53) Scotch M, Duggal M, Brandt C, Lin Z, Shiffman R. Use of statistical analysis in the biomedical informatics literature. J Am Med Inform Assoc 2010 Jan-Feb;17(1):3-5.

(54) Real J, Forne C, Roso-Llorach A, Martinez-Sanchez JM. Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine (Baltimore) 2016 May;95(20):e3653.

(55) Müllner M, Matthews H, Altman D. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. Ann Intern Med 2002;136(2):122-6.

(56) Lee YH, Hsu CY, Hsia CY, Huang YH, Su CW, Lin HC, et al. A prognostic model for patients with hepatocellular carcinoma within the Milan criteria undergoing non-transplant therapies, based on 1106 patients. Aliment Pharmacol Ther 2012 Sep;36(6):551-559.

(57) Chen S, Huang L, Liu Y, Chen CM, Wu J, Shao ZM. The predictive and prognostic significance of pre- and post-treatment topoisomerase IIalpha in anthracycline-based neoadjuvant chemotherapy for local advanced breast cancer. Eur J Surg Oncol 2013 Jun;39(6):619-626.

(58) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009 Jun 4;338:b606.

(59) Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014 Aug 1;35(29):1925-1931.

(60) Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009 May 28;338:b605.

(61) Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 7;350:g7594.

(62) Lang T AD. Statistical Analyses and Methods in the Published Literature: the SAMPL Guidelines. 2013.

(63) Altman D, Hoey J, Marušić A, Moher D, Schulz KF. EQUATOR Network.Enhancing the QUAlity and Transparency Of health Research. 2014; Available at: http://www.equator-network.org. Accessed 05/11, 2014.

(64) Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM. Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine 2016;95(20):e3653.

(65) Ho KM. Residual confounding in observational studies. Anesthesiology 2009 Feb;110(2):430; author reply 430.

(66) Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM. Error tipo I entre métodos matching y modelos paramétricos (en presencia de confusión): Estudio de simulación. 2016.

(67) Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. Stat Med 2004 Dec 30;23(24):3781-3801.

(68) Ramsay TO, Burnett RT, Krewski D. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. Epidemiology 2003;14(1):18-23.

(69) May S, Bigelow C. Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges. Nonlinearity in Biology, Toxicology, Medicine 2005;3(4):dose-response. 003.04. 004.

(70) King G, Nielsen R. Why propensity scores should not be used for matching. Copy at http://j.mp/1sexgVw Download Citation BibTex Tagged XML Download Paper 2016;378.

(71) Real J, Galindo G, Galván L, Lafarga MA, Rodrigo MD, Ortega M. Use of oral bisphosphonates in primary prevention of fractures in postmenopausal women: a population-based cohort study. PloS one 2015;10(4):e0118178.

(72) Munoz MA, Real J, Del Val JL, Vinyoles E, Mundet X, Domingo M, et al. Impact of the sustained control of cardiovascular risk factors on first episode heart failure: The relevant role of primary care. Eur J Gen Pract 2015;21(4):224-230.

(73) Instituto Nacional de Estadística de España. (2015). National health survey. General methodology. . Accessed October/28, 2015.

(74) Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124.

(75) Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 2001 Oct;54(10):979-985.

Anexos

Anexo I: Ajuste de un modelo lineal

En el estudio CUIDADORES(36), se evaluó el impacto sobre la salud asociada de ser cuidador informal(CI), sobre una escala continua de salud percibida (con rango de 0 a 100). Un modelo de regresión lineal multivariable podría ser el siguiente:

ESalud= 97,7 - 3.5 * CI (1 si es CI/ 0 si no lo es) – 0,41 * Edad

Donde ESalud es el estado de salud percibido (escala cuantitativa), CI es ser o no ser cuidador informal y la edad medida en años. Este modelo matemático de regresión lineal asume: 1) linealidad: El efecto de la edad es proporcional en todos los grupos de edad (0,41 unidades menos por año de edad); y 2) Inexistencia de interacción: La escala de salud percibida disminuye en 3,5 puntos en los cuidadores en comparación con el resto y se considera constante en todo el rango de edad (Figura A1). La violación de estas asunciones invalidaría la certeza del modelo.

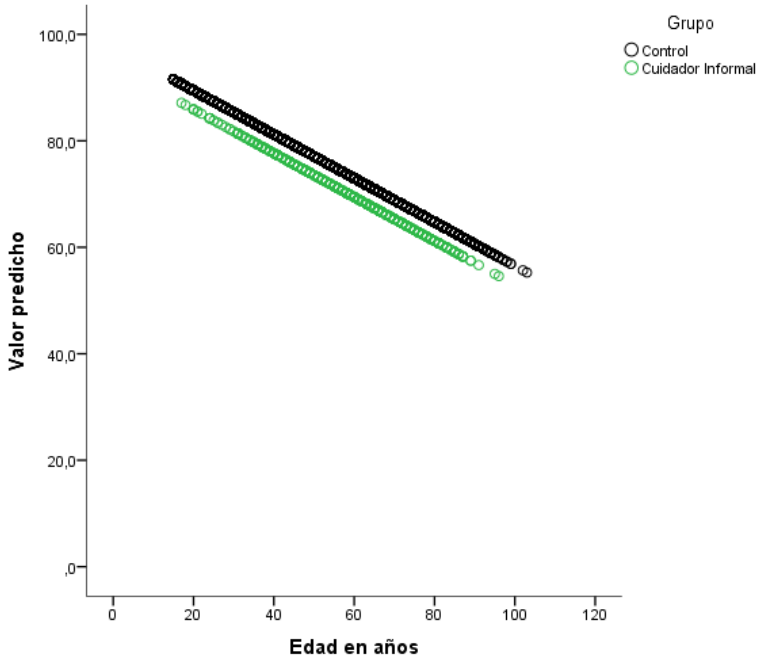


Figura A1. Valor pronosticado de una escala de estado de salud percibido en función de la edad según el grupo (Cuidador informal versus control) estimado según un modelo de regresión lineal múltiple.

Anexo II: Ajuste de un modelo logístico

- **Modelo 1:** Modelo logístico sobre depresión con una variable explicativa (Edad en años)

En el estudio CUIDADORES (36) incluido en esta tesis, se analizaba la frecuencia de depresión y su relación con ser cuidador informal teniendo en cuenta también su edad. En la figura A2 se presentan dos gráficos de líneas: probabilidades pronosticadas de depresión según un modelo logístico en función de la edad (Figura A2 panel a); y la frecuencia relativa observada de depresión en función de la edad (Figura A2 panel b). Según las probabilidades pronosticadas por el modelo , la probabilidad de depresión siempre crece con la edad de una forma cuasi lineal (En realidad exponencial). En cambio las frecuencias observadas (Panel b) muestran que a partir de un rango de edad (>61-72 años) la frecuencia de depresión deja de crecer con la edad. Por lo tanto el modelo estimado no reflejaría exactamente lo que ocurre con las frecuencias observadas.

a) b)

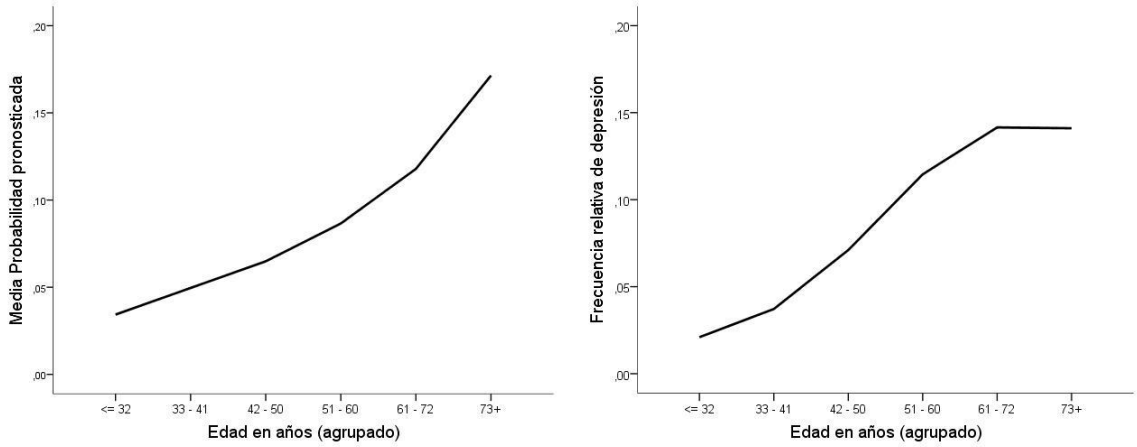


Figura A2. Gráficos de líneas de la frecuencia de depresión por edad según un modelo de regresión logística (Panel a) y frecuencia de depresión (eje y), en función de la edad (Panel b)

El modelo ajustado toma la siguiente forma funcional (Modelo 1):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Edad$$

$$\mathrm{p} = \frac{Exp(\beta_0 + \beta_1 Edad)}{Exp(\beta_0 + \beta_1 Edad) + 1}$$

Donde p es la probabilidad de depresión y β_1 es el coeficiente que acompaña a la variable edad, introducida de forma aditiva en el modelo. La estimación del efecto($\hat{\beta}_1$) que compaña a la variable edad según el modelo ajustado a los datos fue de $\hat{\beta}_1$ =0,018, que corresponde a un OR=1,02 (exp($\hat{\beta}$)), lo cual se significaría que el odds de depresión (≈Riesgo de depresión) se incrementaría un 2% por cada año de edad.

- **Modelo 2:** Modelos logístico con dos variables explicativas: Edad y cuidador (Si/No)

En la figura A3 se presentan los gráficos de líneas de: Las probabilidades pronosticadas de depresión (Figura A3 a) según un modelo Logístico (Modelo 2) que incluye además de la edad, el grupo (1. Cuidador / 0. Control) y; las frecuencias observadas de depresión según las dos variables analizadas, edad y grupo (Figura A3 b). Según el modelo ajustado la frecuencia pronosticada de depresión es mayor siempre (en todos el rango de edad) en el grupo cuidador informal, en cambio las frecuencias observadas, muestran que no en todos los grupos de edad, se observa mayor frecuencia de depresión en los cuidadores. Lo cual sugiere que puede existir una interacción no tenida en cuenta según el modelo ajustado (modelo 2).

a) b)

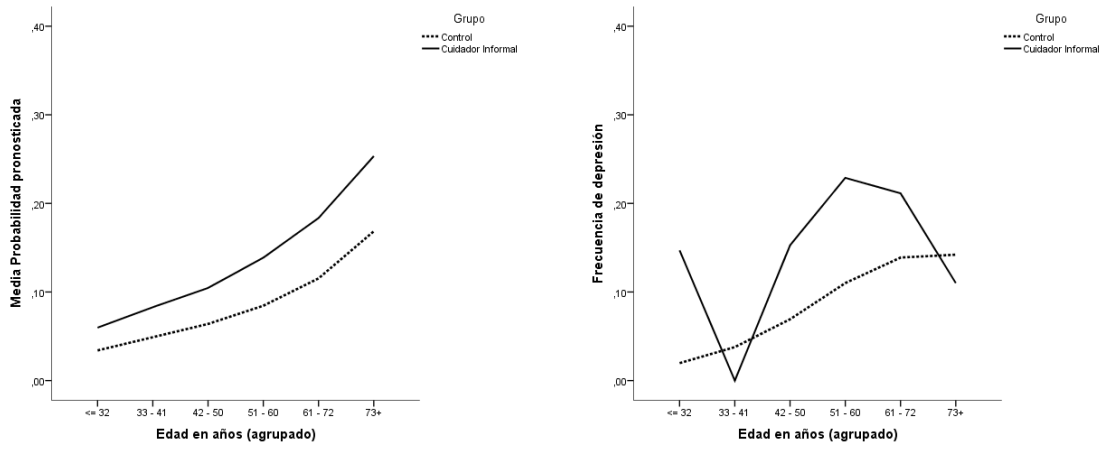


Figura A3. Gráficos de líneas de la frecuencia de depresión (eje y) pronosticada y observada, en función de la edad (eje x), según el grupo (cuidador informal o control). Panel a): probabilidades predichas según modelo 2; Panel b): frecuencia relativa observada de depresión por grupo

El modelo ajustado toma la siguiente forma funcional (Modelo 2):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Edad + \beta_2 CI$$

Donde CI toma valores 1 si es del grupo cuidador y 0 si no, y ahora β_2 es el coeficiente que acompañan a la variable CI. En esta ocasión la estimación del efecto $\hat{\beta}_2$ que compaña a la variable CI según el modelo ajustado a los datos fue de $\hat{\beta}_2$ =0,54, que corresponde a un OR de

1,71 (Exp($\hat{\beta}_2$)) lo cual significaría que las personas que del grupo CI, incrementarían las posibilidades de depresión un 71% (En realidad odds de depresión), independientemente de su edad.

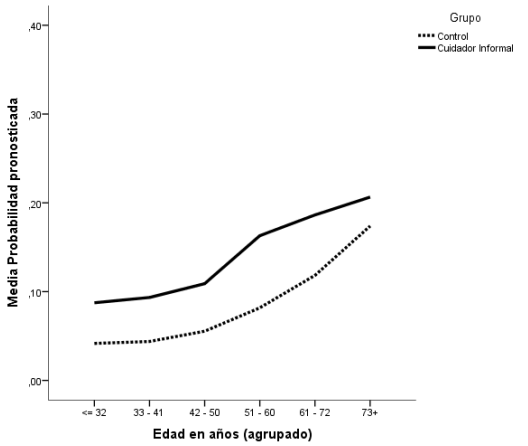
- **Modelo 3:** Modelo logístico sobre depresión según variable: Cuidador (Si/No), edad, sexo y variables socioeconómicas y demográficas

Las variables incluidas en el siguiente modelo ajustado son: edad, estrato social, estrato al que pertenece el hogar (por tamaño de municipio), composición del hogar, nivel de estudios, clase social basada en la ocupación de la persona de referencia, intervalo de ingreso mensual neto de todo el hogar. En la siguiente tabla presentamos los coeficientes del modelo ajustado:

Variables en la ecuación		p valor	Exp(B)	95% C.I. para EXP(B)	
	Categoria		OR	Inferior	Superior
Grupo					
	Cuidador informal	0,001	1,574	(1,21-	2,04)
Edad					
	años	<0,001	1,018	(1,01-	1,02)
Sexo					
	Mujeres	<0,001	2,774	(2,46-	3,13)
Municipio					
	Municipios de más de 500.000 habitantes	0,005			
	Municipio capital de provincia (excepto los anteriores)	0,758	1,031	(0,85-	1,25)
	Municipios con más de 100.000 habitantes (excepto los anteriores)	0,014	1,335	(1,06-	1,68)
	Municipios de 50.000 a 100.000 habitantes (excepto los anteriores)	0,417	1,103	(0,87-	1,40)
	Municipios de 20.000 a 50.000 habitantes (excepto los anteriores)	0,204	1,143	(0,93-	1,40)
	Municipios de 10.000 a 20.000 habitantes	0,063	1,228	(0,99-	1,53)
	Municipios con menos de 10.000 habitantes	0,467	0,931	(0,77-	1,13)
Composición del hogar					
	Hogar unipersonal	<0,001			
	Pareja sola	0,024	0,846	(0,73-	0,98)
	Pareja con algún hijo menor de 25 años	<0,001	0,582	(0,48-	0,71)
	Pareja con todos los hijos mayores de 25 años	0,737	1,039	(0,83-	1,30)
	Padre o madre solo, con algún hijo menor de 25 años	0,510	0,914	(0,70-	1,19)
	Padre o madre solo, con todos los hijos mayores de 25 años	0,048	1,249	(1,00-	1,56)
	Pareja, padre o madre solo con hijo menor de 25 años y otras personas vivie	0,713	0,947	(0,71-	1,27)
	Otro tipo de hogar	0,116	0,808	(0,62-	1,05)
Nivel de estudios					
	No sabe leer o escribir	<0,001			
	Ha asistido menos de 5 años a la escuela	0,080	1,296	(0,97-	1,73)
	No llegó al último curso de la enseñanza obligatoria	0,030	1,386	(1,03-	1,86)
	Enseñanza Secundaria de Primera etapa	0,129	1,256	(0,94-	1,69)
	Estudios de Bachillerato	0,324	0,840	(0,60-	1,19)
	Enseñanzas profesionales de grado medio o equivalentes	0,435	1,152	(0,81-	1,64)
	Enseñanzas profesionales de grado superior o equivalentes	0,097	0,695	(0,45-	1,07)
	Estudios universitarios o equivalentes	0,025	0,654	(0,45-	0,95)
Clase social					
	Clase Social I	0,220			
	Clase Social II	0,454	1,125	(0,83-	1,53)
	Clase Social III	0,534	1,087	(0,84-	1,42)
	Clase Social IV	0,503	1,099	(0,83-	1,45)
	Clase Social V	0,429	1,110	(0,86-	1,44)
	Clase Social VI - Trabajadores/as no cualificados/as	0,122	1,240	(0,94-	1,63)
	No consta	0,439	0,876	(0,63-	1,23)
Intervalo de ingreso mensual neto de todo el hogar					
	550 euros o menos	<0,001			
	De 551 a 800 euros	<0,001	0,597	(0,49-	0,73)
	De 801 a 1050 euros	<0,001	0,610	(0,49-	0,76)
	De 1051 a 1300 euros	<0,001	0,497	(0,39-	0,63)
	De 1301 a 1550 euros	<0,001	0,475	(0,36-	0,62)
	De 1551 a 1850 euros	<0,001	0,369	(0,27-	0,50)
	De 1851 a 2250 euros	<0,001	0,371	(0,27-	0,51)
	De 2251 a 2700 euros	<0,001	0,336	(0,23-	0,48)
	De 2701 a 3450 euros	<0,001	0,353	(0,23-	0,53)
	Más de 3450 euros	<0,001	0,269	(0,16-	0,45)
	NC	<0,001	0,395	(0,32-	0,49)
Constante		<0,001	0,032		

Según el modelo 3 ajustado la estimación del OR de depresión es de 1,57, lo cual se significa que las personas que están en una situación de cuidadores incrementarían las posibilidades de depresión un 57% (En realidad odds de depresión), independientemente de su edad, estrato social, estrato al que pertenece el hogar (por tamaño de municipio), composición del hogar, nivel de estudios, clase social e ingresos mensuales neto de todo. En la figura A4 podemos ver las probabilidades predichas versus las observadas.

a)



b)

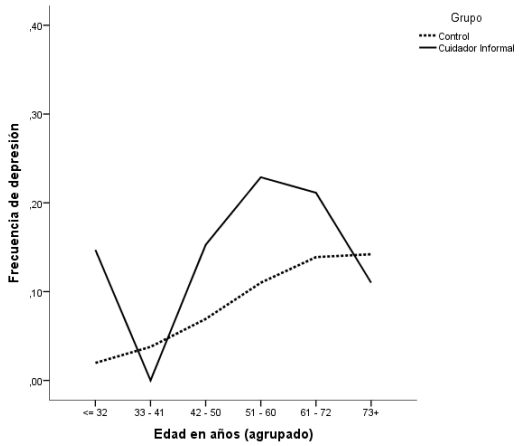


Figura A4. Gráficos de líneas de la frecuencia de depresión (eje y) pronosticada y observada, en función de la edad (eje x), según el grupo (cuidador informal o control). Panel a): probabilidades predichas según modelo 3; Panel b): frecuencia relativa observada de depresión por grupo

- **Bondad de ajuste**

En las siguientes resultados podemos ver los *outputs* de los estadísticos resumen del ajuste del **modelo 3**, junto con las pruebas de bondad de ajuste.

Tabla de contingencia para la prueba de Hosmer y Lemeshow del modelo 3

		depresión = No		depresión = SI		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	1985	1976,827	18	26,173	2003
	2	1976	1958,970	27	44,030	2003
	3	1920	1942,217	83	60,783	2003
	4	1933	1921,432	70	81,568	2003
	5	1897	1897,191	106	105,809	2003
	6	1879	1867,554	124	135,446	2003
	7	1820	1826,733	183	176,267	2003
	8	1745	1764,049	258	238,951	2003
	9	1624	1663,510	379	339,490	2003
	10	1526	1486,519	476	515,481	2002

Resumen del modelo 3

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	10477,153[a]	,062	,139

a. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.

Prueba de Hosmer y Lemeshow del modelo 3

Escalón	Chi-cuadrado	gl	Sig.
1	32,059	8	,000

En los ejemplos mostrados (modelos 1 2 o 3) en los 3 modelos ajustados se rechazaría la hipótesis de bondad de ajuste de H&L ya que el resultado de sendas pruebas de significación presentan p-valores inferiores a 0,05 (p-valor< 0,001). Esto significa que la hipótesis nula de que los datos se ajustan a los modelos se rechazaría en todos los casos.

Anexo III: Distribución de variables antes y después del matching

En la figura A5 podemos observar un diagrama de cajas con datos del estudio CUIDADORES(36) de la variable edad (potencial confusor) por grupos antes y después del matching. En ella se puede apreciar como en la muestra post matching la distribución de la variable edad queda más equilibrada entre grupos, presentado medianas parecidas.

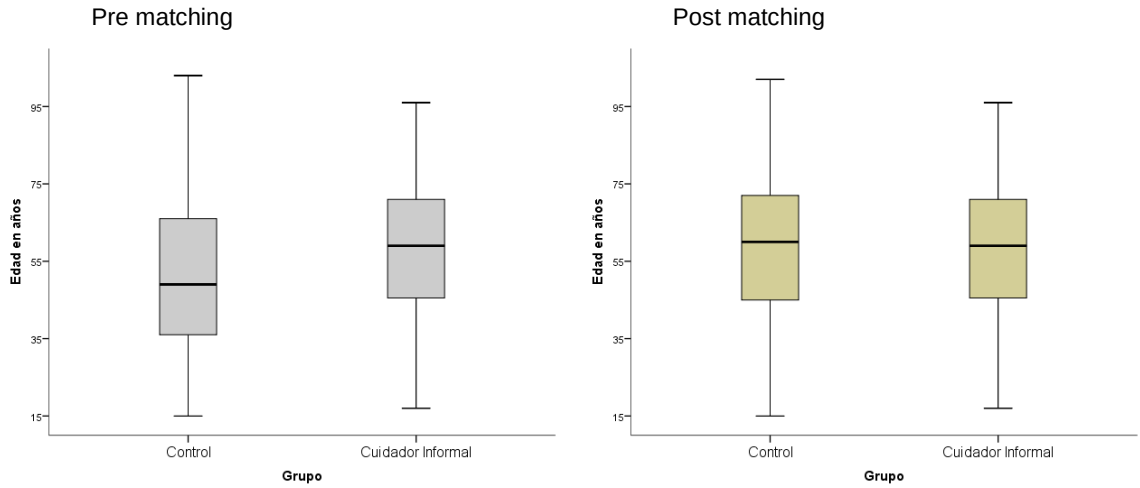


Figura A5. Diagrama de cajas de la distribución de la edad antes y después del matching en el estudio CUIDADORES

En la siguiente figura (Figura A6) se puede ver un ejemplo de Gary King (33) donde se representa un gráfico de dispersión de dos variables, nivel educativo y edad, antes y después del proceso de matching. En ella podemos ver como después de la eliminación de observaciones, en la muestra post matching, solo permanecen aquellas observaciones control (C) más cercanas a las tratadas (T), según su edad (Age) y nivel educativo (Education).

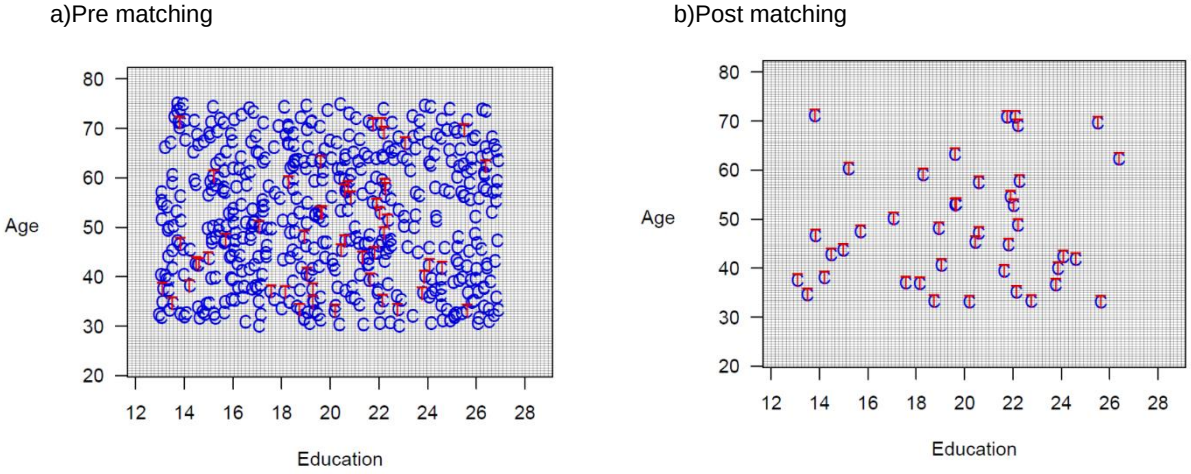


Figura A6. Gráfico de dispersión de educación versus edad de una muestra pre matching (Panel a) y post matching (Panel b). Ejemplo de Gary King & Richard Nielsen. MacMillan-CSAP Workshop on Quantitative Research Methods, Yale University, 10/3/2016

Anexo IV: Diagnóstico del matching

Diferencia de medias estandarizada de cada covariable entre grupos antes y después del matching.

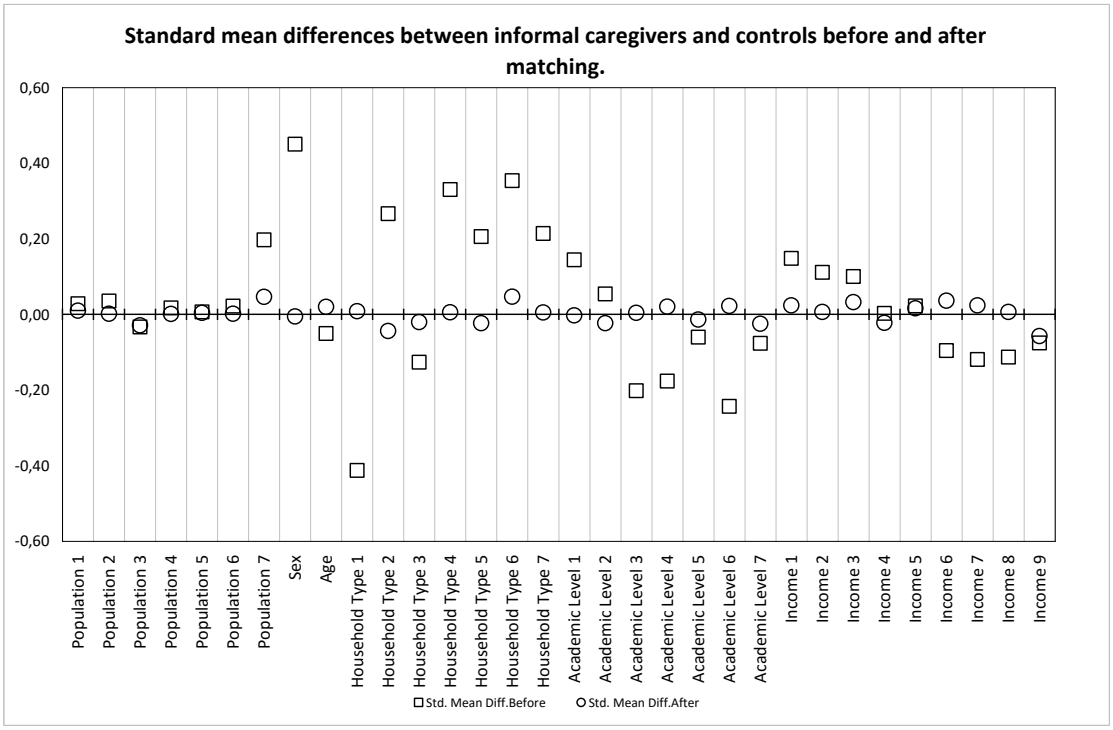


Figura A7. Apendice II del artículo publicado en J Public Health Policy 2016 May;37(2):173-189. The overall mean difference of selected variables used to match ICs with control was reduced to almost 0 from the pre-matched sample to the final matched sample

Anexo V: Cuestionario de recogida de datos del trabajo REVISION

Qüestionari i guia pel revisor:

ID: _____ Revisor: _____ Data de revisió: __/__/____

Article

PMID: _____ 1e autor:_____ Any: _____

1.ABSTRACT

1.1. Compleix algun dels següents criteris d'exclusió (Si / No)? ____

- No us de models paramètrics clàssics como mètode principal de d'ajust : ___
  - Models no-paramètrics (Models GAM) / Models bayesians
  - Mètodes d'estandardització de taxes / Mètodes de Matching
- Estudis de validació d'un instrument
- Estudis experimentals
- Estudis ecològics. Series temporals o geo estadístics "Tipus diseases mapping"
- Series de casos
- Revisions sistemàtiques – Meta anàlisis

Seguir en al següent apartat, si no compleix cap criteri d'exclusió:

1.2. Inspeccionar el resum:

-Mida de la mostra: _____
-Font de dades (1. Adhoc; 2. Registro clínic o administratiu; 3. Mixte; 9.Indefinit) :___
-Disseny de l'estudi (1. Transversal / 2. Cohorts / 3. Cas-Control / 9. Indefinit): _____
-S'identifica l'ús de models paramètrics (1. Logística / 2. Lineal / 3. Reg Cox; 9. Indefinit): ____
-Reporta mesures d'associació tipus OR , HR o Ef (1.Si/0.No): _____
-Inclou Intervals de confiança de les mesures d'associació IC95% (1.Si/0.No): ___

2. Inspeccionar text complet:

TEXT COMPLET: METODOLOGIA

*2.1. Identificar **apartat de metodologia** (preferentment anàlisis estadístic): Revisar els següents ítems:*

(Només en cas de no constar en ABSTRACT)*

-Mida de la mostra: _____
-Font de dades (1. Adhoc; 2. Registre clínic o administratiu; 3. Mixte; 9.Indefinit) : __
-Disseny de l'estudi (1. Transversal / 2. Cohorts / 3. Cas-Control / 9. Indefinit): _____
-S'identifica l'ús de models paramètrics (1. Logística / 2. Lineal / 3. Reg Cox; 9. Indefinit): __
(Pot ser multi resposta)

Revisar la següent llista d'ítems de verificació:

Ítem	Descripció	Logística	Lineal	Cox
Im1	Valida interna			
Im3	Interaccions			
Im4	Sensibilitat			
Im7	Procés(*)			
Im8	Validació externa			

Descripció d'ítems:
Im1: Especifica que es realitza proba de validació del model
Im3: Especifica avaluació de les interaccions
Im4: Especifica que es realitzarà un anàlisis de sensibilitat dels models
Im7: Descriu el procés de selecció de variables per arribar al model final. (e.g., ENTER*; forward-stepwise; best subset)
*El mètode ENTER es considerat implícit si descriu totes las variables que utilitzarà per construir el model final.
Im8: Especifica que es realitza una validació externa amb submostra reserva o externa.

L'apartat de mètode conte una cita referent a la metodologia o protocol referent a aquesta investigació (S/N)? ___

RESULTATS I TAULES /FIGURES

*2.2 Identificar **Resultats i llista de taules (o figures**):*

Ítem	**Descripció**	**Logística[1]**			**Lineal[2]**			**Cox[3]**	
		H-L	ROC	Altra	R^2	Residus	Alt	Pr HR test	Altra
Ir2	Reporta mesures de bondat d'ajust i/o validació dels models								

Descripció d'ítems:

Ir2: Reporta mesures de bondat d'ajust o validació model:

1. Regressió Logística: Test de Hosmer and Lemeshow / Àrea sota la corba ROC / Altra
2. Regressió Lineal: R2 o Test de K-S sobre residus del modelo / Test gràfic de residus Q-Q plot / Altra
3. Comprovació de Riscos proporcionals: Schoenfeld residuals

Ítem	**Descripció**	**Logística**	**Lineal**	**Cox**
Ir3	Rep Interacciones			
Ir4	Anàlisis Sub mostres			
Ir5	Cru vs Ajustat			
Ir6	+ 1 model*			
Ir8	Validació externa			
Ir9	Inclou IC95% Mes Ass			

**Nombre de models màxim per "Outcome"*

Descripció d'ítems:

Ir3 Reporta avaluació de interaccions dels models

Ir4 Reporta anàlisis del models amb sub mostres (Exemple: Model separat per sexes o grups d'edat)

Ir5 Reporta mesures d'associació: crues versus ajustades (Mesures d'associació no ajustades versus ajustades segons model)

Ir6 Descriu mes d'un model ajustat per diferents variables (Nombre de models màxim per outcome). S'exclou el model cru – no ajustat

Ir8 Mostra resultats d'una proba de validació externa amb submostra amb la qual no s'ha elaborat el model

Ir9 Inclou Intervals de confiança de les mesures d'associació reportades referents en els models (OR/HR/Ef)

Tots els ítems han sigut clarament identificats ? (1.S/0.N) : __

Si la resposta anterior es un No : pendent de revisió amb la resta dels investigadors.

Anexo VI: Reunión científica de la sociedad española de epidemiologia (SEE)

Resultados presentados en la XXXIV reunión científica de la SEE (15/09/2016):

Título *(100 caracteres con espacios)***:** Comparación del error tipo I entre los métodos matching y los modelos paramétricos: estudio de simulación.

Autores: Jordi Real (1,2), Carles Forné (3,4), Albert Roso-Llorach (1), Jose M Martínez-Sánchez (2,5).

(1) Institut Universitari d'Investigació en Atenció Primària Jordi Gol, Barcelona; (2) Universitat Intenacional de Catalunya, Sant Cugat; (3) Universitat de Lleida, Lleida; (4) Oblikue Consulting, Barcelona; (5) Grup de Prevenció i Control del Càncer, IDIBELL, L'Hospitalet de Llobregat.

Resumen *(2.600 caracteres con espacios):*

Antecedentes/Objetivos: En los estudios observacionales es habitual utilizar las técnicas estadísticas multivariables como herramienta de ajuste para controlar el potencial sesgo de confusión. Sin embargo, pese al creciente interés en alternativas como los métodos de emparejamiento (*matching)* y/o, modelos aditivos generalizados (GAM), los modelos multivariables paramétricos, como la regresión logística, siguen siendo muy utilizados. El objetivo del presente estudio es comparar el sesgo de confusión residual de tres métodos de ajuste distintos (*matching*, GAM y regresión logística) en escenarios donde la relación confusor-respuesta no es lineal.

Métodos: Se simularon conjuntos de datos de 10.000 observaciones de 3 variables: una variable *respuesta* binaria (Y), una variable exposición dicotómica (X) independiente de la respuesta (Y), y una variable confusora continua (Z) relacionada con la exposición (X) y con la respuesta (Y). La respuesta se generó mediante una distribución binomial condicionada a 8 formas no lineales de Z, y 2 niveles de correlación entre la exposición y la variable confusora (X-Z), r=0,5 y 0,3. En total se simularon 7500 muestras en 16 escenarios distintos descartando aquellas muestras donde alguna de las estrategias de análisis no convergió. Se han comparado 7 estrategias de análisis para estimar el efecto nulo de X sobre Y (asumiendo un error de tipo I nominal α=0,05): 3 métodos *matching* (exacto, subclasificación y *nearest-neighbour*), 3 especificaciones del modelo de regresión logística (lineal, categorizando el confusor y polinomial), y un modelo GAM. Se calculó el error de tipo I empírico y el error cuadrático medio de la estimación del efecto nulo.

Resultados: Los métodos *matching* y el modelo GAM mostraron errores tipo I empíricos más cercanos al nominal que los métodos paramétricos en las 8 formas funcionales consideradas. El método *matching* exacto presentó la menor tasa de error tipo I, inferior al nivel teórico en los dos niveles de correlación X-Z (α=0,041 y 0,039, para r=0,5 y 0,3 respectivamente), mientras que el método de subclasificación fue la estrategia *matching* que mostró un mayor error en ambos escenarios de asociación (α=0,084 y 0,061). Con el modelo GAM se observaron errores

α=0,053 y 0,056, y con regresión logística errores entre α=0,085 y 0,093 categorizando el confusor, y α=0,176 y 0,474 en el modelo lineal.

Conclusiones: Los métodos *matching* con reducción de muestra proporcionan mayor credibilidad a los resultados en comparación a la regresión logística multivariable independientemente de la relación funcional entre el confusor y la respuesta.

Presentación Power point

Error tipo I entre métodos matching y modelos paramétricos (en presencia de confusión):

Estudio de simulación

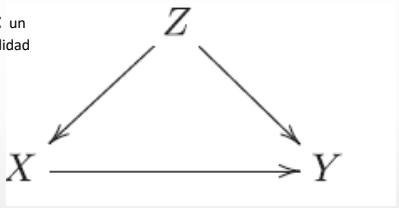
Jordi Real (1,2), Carles Forné (3,4), Albert Roso-Llorach (1), Jose M Martínez-Sánchez (2,5)
(1) Institut Universitari d'Investigació en Atenció Primària Jordi Gol, Barcelona.
(2) Universitat Intenacional de Catalunya, Sant Cugat.
(3) IRBLLEIDA. Universitat de Lleida, Lleida.
(4) Oblikue Consulting, Barcelona
(5) Grup de Prevenció i Control del Càncer, IDIBELL, L'Hospitalet de Llobregat

1. Introducción
   1. Confusión en estudios observacionales
   2. Métodos de control de la confusión
2. Objetivos
3. Métodos
4. Resultados
5. Discusión / conclusiones

Sesgo de confusión en estudios observacionales

En los estudios observacionales es habitual el sesgo de confusión

Confusión: "mezcla" o "difuminación" de efectos: un investigador trata de relacionar una exposición a un resultado, pero en realidad mide el efecto de un tercer factor (la variable de confusión)

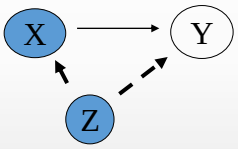


Una **variable confusión** es una variable que distorsiona la medida de la asociación entre otras dos variables.

El resultado de la presencia de una variable de confusión puede ser la observación de un efecto donde en realidad no existe o la exageración de una asociación real (confusión positiva) o, por el contrario, la atenuación de una asociación real e incluso una inversión del sentido de una asociación real (confusión negativa).

Sesgo de confusión en estudios observacionales (Ejemplos)

La edad y el sexo son "sospechosas habituales" a ser variables de confusión ya que a menudo van ligadas a resultados en salud a la vez que la exposición

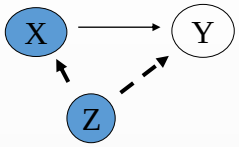


- Comparación de tasas de mortalidad entre países con estructuras de edad distintas.
- Exposición a tabaquismo, dieta y ejercicio → eventos cardiovasculares

Sesgo de confusión en estudios observacionales (Caso práctico)

Objetivo

- Impacto sobre la salud (Ansiedad, depresión etc..) y factores asociados relacionado con estar en una situación de cuidador Informal



Método:

- ENSE: Encuesta Nacional de Salud Española a 20.000 hogares
- Se identificó 515 personas consideradas Cuidadores informales (superior al año)
- Resultados en Salud: Diagnostico depresión, Ansiedad, Calidad de vida, Estado de salud percibido, soporte social

González-de Paz L, Real J, Borrás-Santos A, Martínez-Sánchez JM, Rodrigo-Baños V, Dolores Navarro-Rubio M. Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study. J Public Health Policy. 2016 May;37(2):173-89. doi: 10.1057/jphp.2016.3. Epub 2016 Feb 11. PubMed PMID: 26865318.

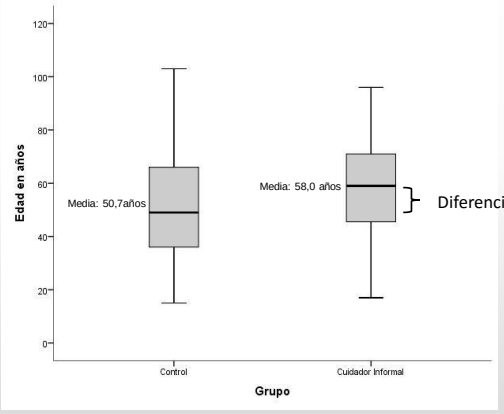
Métodos de corrección del sesgo de confusión

1. Restricción (Diseño)
2. Anticipación de confusores potenciales (Diseño)
3. Análisis
   - Estratificación por confusor/es:
     - Menor potencia estadística
     - Difícil con muchas covariables
     - Método simple
   - Métodos de regresión (Ajuste por covarianza)
     - Fuertes asunciones de los modelos a los datos (Linealidad predictores, normalidad de residuos, interacciones, etc..)
     - Potencia estadística
4. Diseño / Análisis:
   - Matching

Caso práctico: Cuidadores informales y factores asociados

Muestra	% Depresión
515 CI	15,7%
19.514 resto	8,4 %

OR crudo depresión (CI)=**2,03**



Diferencia media= 7,3 años

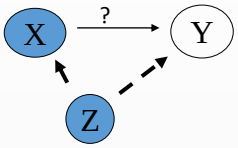
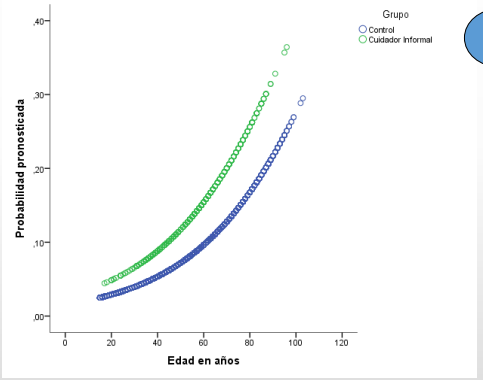


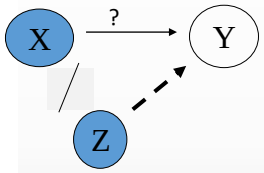
Figura 1. Box plot de la edad en función del grupo muestra inicial (n=515 vs 19.514)

Ajuste mediante modelo multivariable: Regresión logística

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 edad + \beta_2 CI$$

OR ajustado = **1,71** ; IC95%: 1,34- 2,19

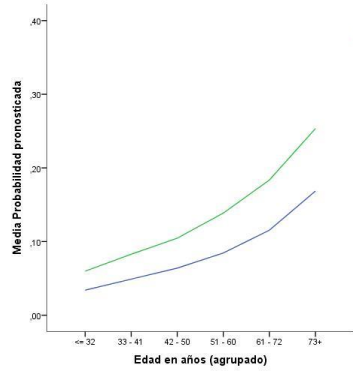




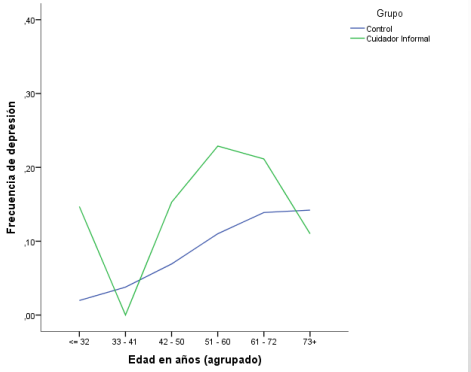
Diagnóstico: Modelo vs realidad

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 edad_1 + \beta_2 CI_2 + \text{Error}$$

Predicciones

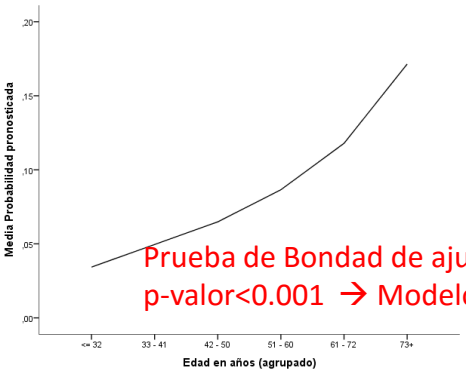


Observaciones

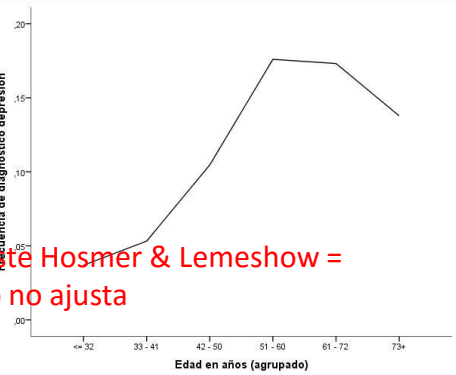


Diagnóstico: Modelo vs realidad

Expectativas según modelo



Realidad

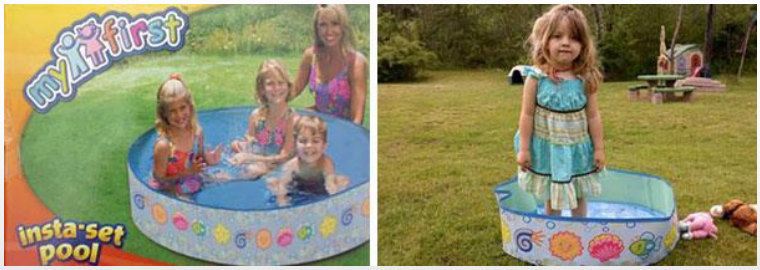


Prueba de Bondad de ajuste Hosmer & Lemeshow = p-valor<0.001 → Modelo no ajusta

Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM. Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine (Baltimore). 2016 May;95(20)

Diagnóstico: Modelo vs realidad

Expectativa vs Realidad



Fuente: Google imágenes

Prueba de satisfacción segun expectativas generadas
→ Modelo no ajusta

Métodos de corrección del sesgo de confusión

1. Restricción (Diseño)
2. Anticipación del las variables potenciales (Diseño)
3. Análisis
   - Estratificación por confusor/es:
     - Menor potencia estadística
     - Difícil si tenemos muchas covariables
     - Método sencillo
   - Métodos de regresión (Ajuste por covarianza)
     - Fuertes asunciones sobre los modelos paramétricos
     - (Linealidad predictores, normalidad de residuos, interacciones)
     - Alta potencia estadística
4. **Diseño / Análisis:**
   - **Aplicar métodos de Matching**
     - Eliminas muestra
       Potencia estadística
     - Independencia de modelo

Matching: En que consiste?

Dado una muestra N , seleccionar sub muestra n (n<N) tal que los grupos sean comparables según confusores potenciales

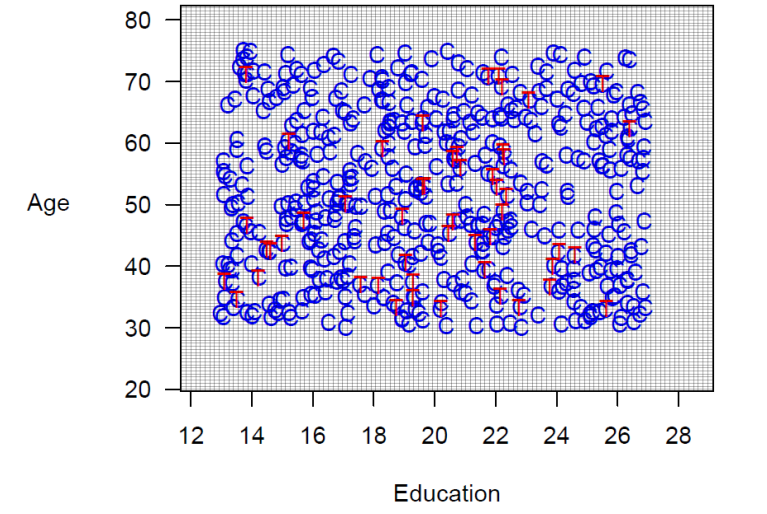
4 Fases:

- 1. Estimar distancia D (PS de estar en un grupo condicionado a Z): P(X/Z)
- 2. Aplicar algoritmo Matching que usará D para seleccionar observaciones:

  Exacto, Subclassification, Nearest-Neigbourg etc...
- 3. Evaluar el equilibrado
- 4. Estimar medida de asociación con la nueva muestra "n" :

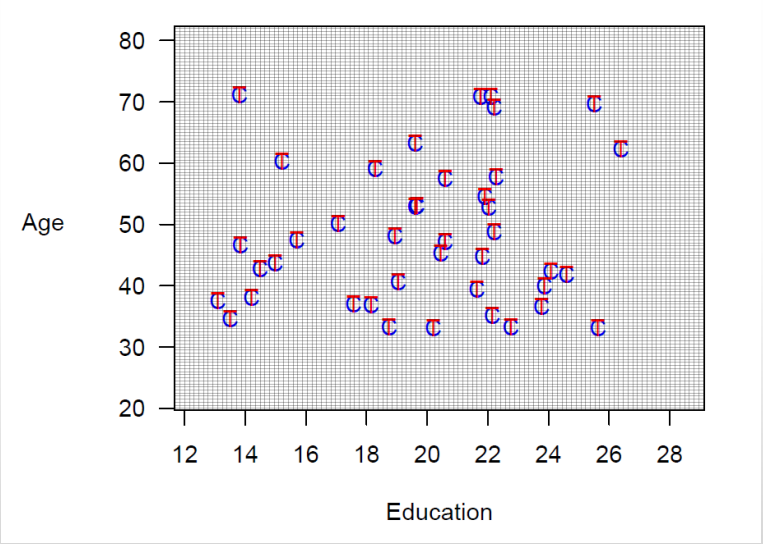
  Métodos convencionales / modelos ajustado por cada pareja etc….

Matching: En que consiste?



Gary King & Richard Nielsen. MacMillan-CSAP Workshop on Quantitative Research Methods, Yale University, 10/3/2016

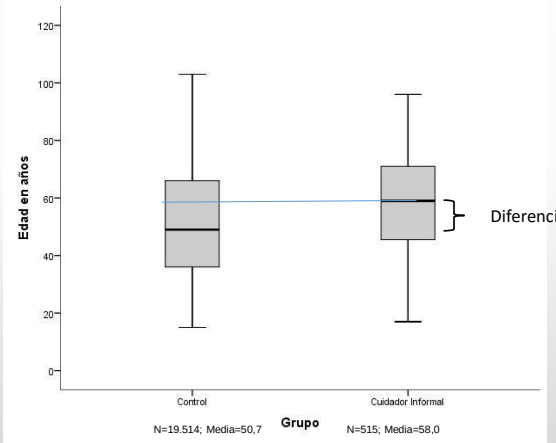
Matching: En que consiste?

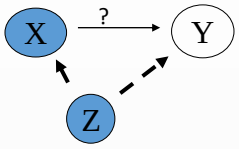


Gary King & Richard Nielsen. MacMillan-CSAP Workshop on Quantitative Research Methods, Yale University, 10/3/2016

Caso práctico: Cuidadores informales y factores asociados

Box plot de la edad en función del grupo muestra inicial (n=515 vs 19.514)

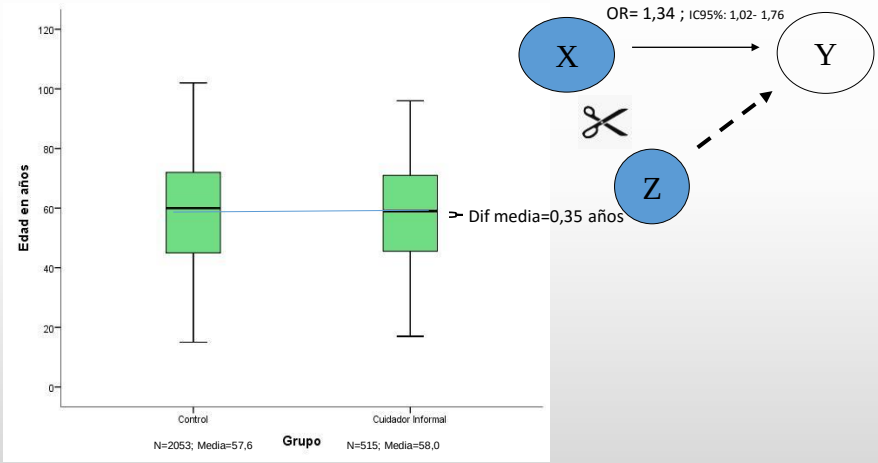




Diferencia media= 7,3 años

Caso práctico: Cuidadores informales y factores asociados

Box plot de la edad en función del grupo muestra apareada (n=515 vs 2053)



- Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol 2008 Oct;37(5):1142-1147.
- Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006 Feb 1;163(3):262-270.

2. Objetivo

Comparar métodos matching con otras técnicas de ajuste conocidas (Reg Logística) en escenarios donde la relación confusor-respuesta no sea lineal:

Y, los datos pueden no cumplir las expectativas de los modelos

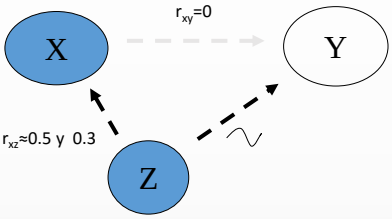
3. Métodos

Generamos muestras simuladas (n=10.000) tal que:

X ~ Bernouilli (0,3) ind Y

Z ~ Normal (mu, sd/X)

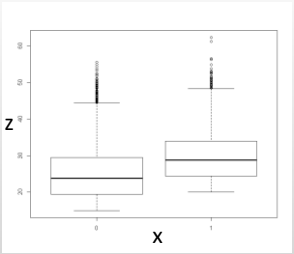
Z e X relacionadas según:



$Z \sim N(mu+5, sd_z | X=1)$

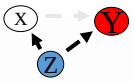
$Z \sim N(mu, sd_z | X=0)$

Dos niveles de asociación X-Z , tal que: r_{xz}≈0.5 y 0.3



3. Métodos

Relaciones confusor Z vs P(Y):P(Y/Z)~G(Z)



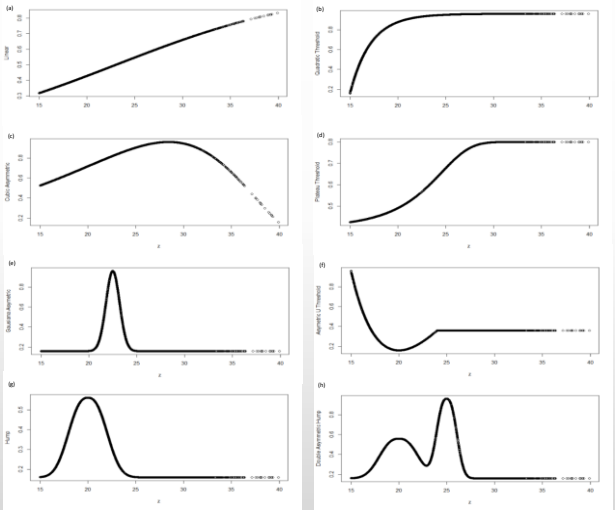
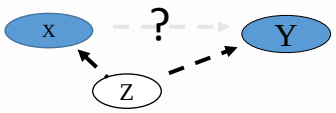


Figura 1. Relaciónes generadas de Z-P(Y): (a) Linear, (b) Quadratic Threshold, (c) Cubic Asymmetric, (d) Plateau Threshold, (e) Gausiana Asymetric , (f) Asymetric U Threshold, (g) "Hump", (h) Double Hump.

3. Métodos

Aplicamos métodos de ajuste para estimar efecto de X sobre Y



Muestra (n=1..10000)

Y

Filter

	event	x	z
1	1	0	19.73941
2	1	0	32.30784
3	1	1	29.83667
4	0	1	33.96142
5	1	1	32.56139
6	1	0	22.88538
7	0	0	18.78676
8	0	0	15.33610
9	0	1	19.98123
10	0	0	19.61898
11	1	1	33.90117

Métodos

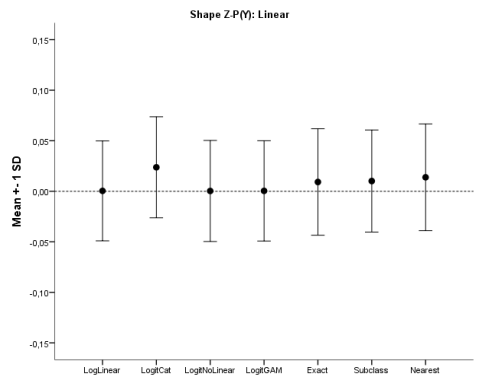
1. GLM / GAM Link Logit:
   1. Z Lineal
   2. Funciones polinómicas de Z orden 3
   3. Categorizando Z en quintiles
   4. Función no paramétrica S (Z) (GAM)
2. Métodos de matching + Univariable
   5. Exacto
   6. Neigbourg-Nearest
   7. Subclasification (con descartes)

Ho D, Imai K, King G, Stuart E. MatchIt: MatchIt: Nonparametric Preprocessing for Parametric Casual Inference. R package version 2006:2.2-11

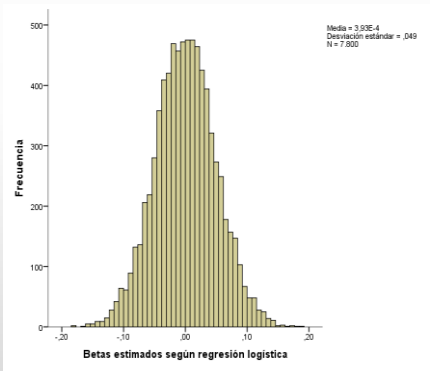
4. Resultados

Escenario: Relación Z-Y Lineal

Estimación de ßs según método de ajuste ± 1dt

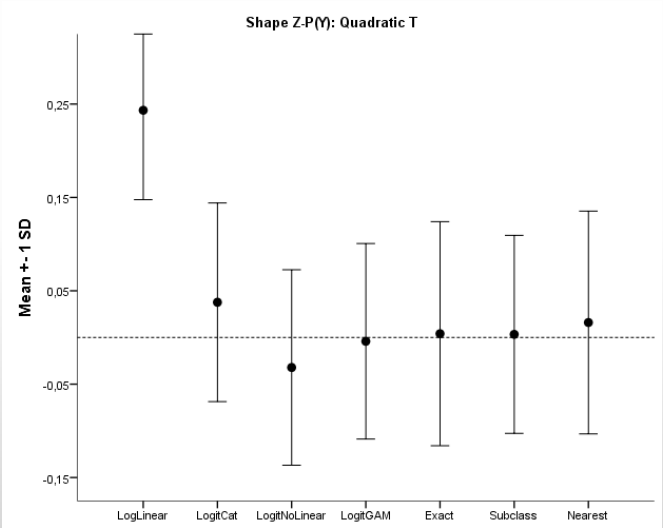


Histograma de estimaciones de ßs



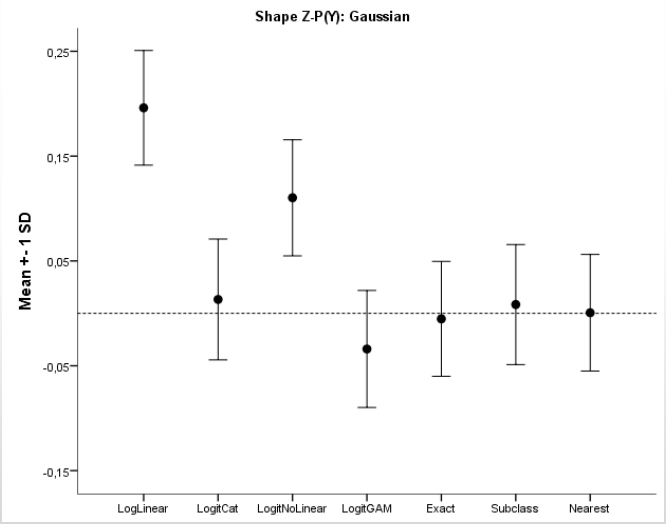
4. Resultados

Escenario: Relación Z-Y No lineal



4. Resultados

Escenario: Relación Z-Y No lineal



4. Resultados: α empírico por escenario-método

α empírico: Tasa de falsos positivos

Error de tipo I empírico de X en condiciones de relación nula entre X e Y, en función de la forma de la asociación ZY generada.

Correlation X-Z

		GLM			Matching		
Shape relation Z-Y	LogLinear	LogitCat	LogNolineal	GAM	Exact	Subclass	Nearest
Low (SD=10; r=0.3)							
Linear	0,051	0,079	0,052	0,052	0,045	0,059	0,053
Quadratic T	0,701	0,066	0,064	0,051	0,051	0,052	0,052
Cubic Asymetric	0,980	0,324	0,050	0,053	0,033	0,105	0,038
Plateau	0,066	0,051	0,060	0,050	0,042	0,054	0,043
Gaussian	0,945	0,055	0,525	0,082	0,029	0,051	0,032
Asymetric U T	0,142	0,052	0,074	0,051	0,046	0,057	0,052
Hump	0,050	0,052	0,051	0,050	0,041	0,055	0,048
Double Hump	0,802	0,058	0,049	0,054	0,027	0,051	0,031
Total	0,474	0,093	0,117	0,056	0,039	0,061	0,044

Color de fondo condicionado a la magnitud del error de tipo I empirico: Cuanto más oscuro mas alejado de 0.05

5. Discusión

Limitaciones (Escenarios planteados):

- Escenarios relativamente simples
  - Respuesta binaria, sin considerar confusores no medidos, ni existencia de interacciones , n=10000.
  - Pero relacionado no linealmente con el outcome , Z→ Normal asimétrica truncada
- Algoritmos no evaluados:
  - (Genetic, Optimal, Full etc…), Otras distancias (Mahalanobis) o otras parametrizaciones de análisis (sensibilidad caliper., n etc..).
- Costo computacional alto
  - Generación de casi 1 millón de estimaciones
  - Los algoritmos son lentos y existen muchas opciones de parametrización

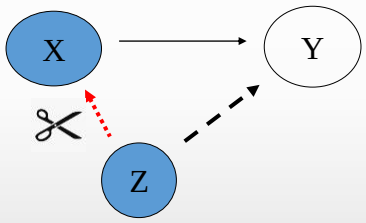
5. Conclusiones

- Los modelos paramétricos pueden generar estimaciones muy sesgadas si la comprobación de las asunciones es ignorada.
  - Aunque la estimación paramétrica con GLM basada en la selección de un buen modelo todavía es eficiente en términos de MSE (error) y puede mejorar usando algoritmos de validación cruzada (machine learning algorithms and cross validation)
- Los modelos GAM ofrecen eficientes resultados en términos de control de confusión, sin especificar a priori la forma funcional del confusor
- Los tres algoritmos *matching examinados* proporcionan mayor credibilidad a los resultados , como método de ajuste en comparación a la regresión logística multivariable

5. Conclusiones

- La principal ventaja de la metodología "matching" es que reducen en gran medida la dependencia de los modelos y sus asunciones: Sin embargo cualquier permanencia de desequilibrio debe ser tratado con ellos (Ho et al, 2007)

  Ambos métodos pueden ser complementarios si existe cierto desequilibrio después del matching



- Calidad del matching:

  Método, datos, muestra
  Requiere validación manual

5. Conclusiones

Requiere validación manual

Fuente: Google Imágenes

6. Alguna bibliografía

- Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol 2008 Oct;37(5):1142-1147.
- Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006 Feb 1;163(3):262-270.

- Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis 2007;15(3):199-236.
- Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 2007;26(16):3078-3094.
- Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci 2010 Feb 1;25(1):1-21.
- King G, Nielsen R, Coberley C, Pope JE, Wells A. Comparative effectiveness of matching methods for causal inference. Unpublished manuscript 2011;15.
- King G, Nielsen R. Why propensity scores should not be used for matching. Copy at http://j.mp/1sexgVw Download Citation BibTex Tagged XML Download Paper 2016;378.
- King G, Lucas C, Nielsen R, King G, Pan J, Roberts M, et al. The Balance-Sample Size Frontier in Matching Methods for Causal Inference}. PS: Political Science and Politics} 2014;42:S11-S22.
- Pearce N. Analysis of matched case-control studies. BMJ 2016 Feb 25;352:i969

- Real J, Forné C, Roso-Llorach A, Martínez-Sánchez JM. Quality Reporting of Multivariable Regression Models in Observational Studies: Review of a Representative Sample of Articles Published in Biomedical Journals. Medicine (Baltimore). 2016 May;95(20)
- González-de Paz L, Real J, Borrás-Santos A, Martínez-Sánchez JM, Rodrigo-Baños V, Dolores Navarro-Rubio M. Associations between informal care, disease, and risk factors: A Spanish country-wide population-based study. J Public Health Policy. 2016 May;37(2):173-89. doi: 10.1057/jphp.2016.3. Epub 2016 Feb 11. PubMed PMID: 26865318.