

# Métodos de emparejamiento (Matching)

En grandes bases de datos

Jordi Real

jreal@idiapjgol.info

IDIAP- Jordi Gol

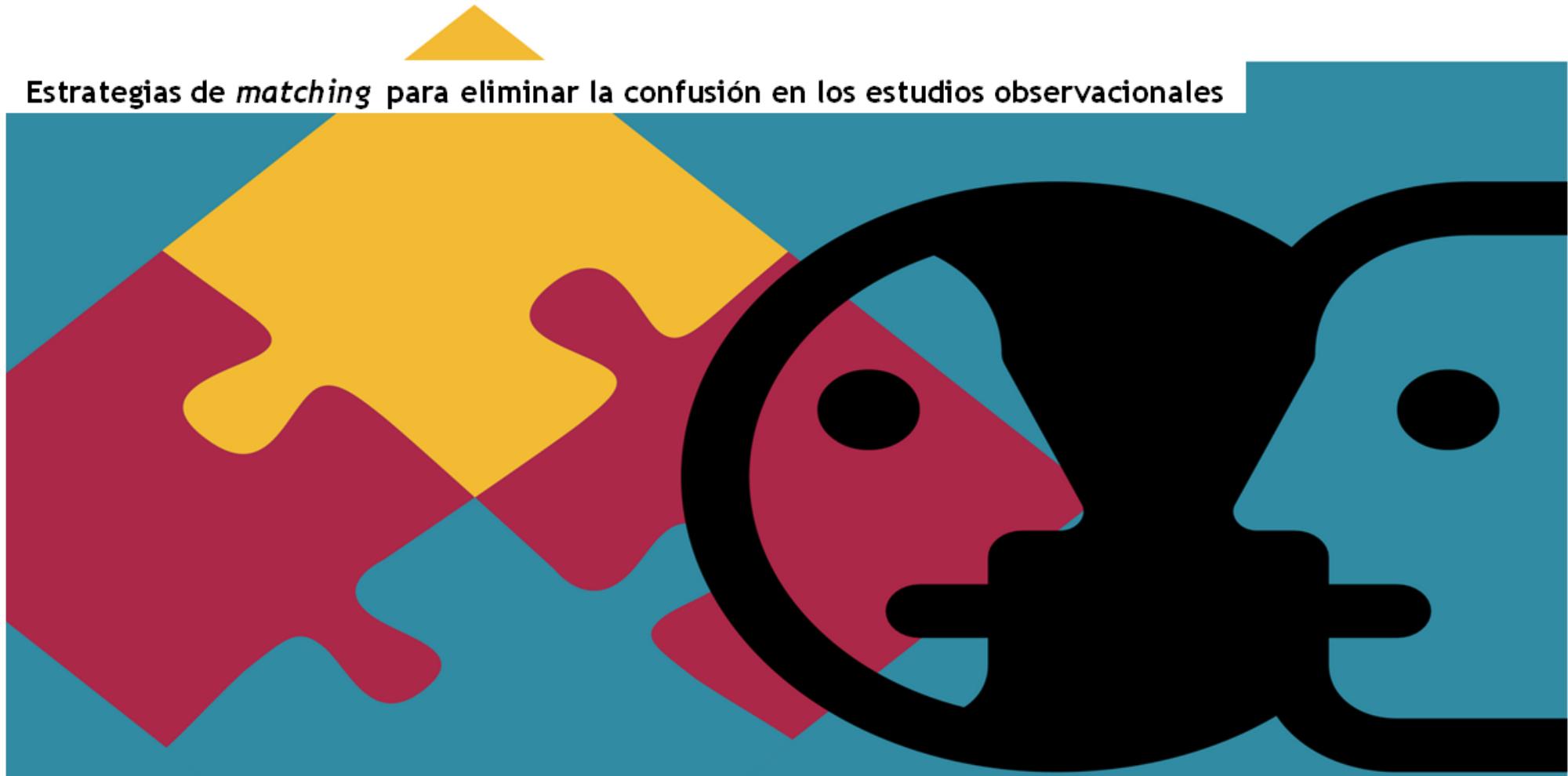
USR-BCN. Grup DAP-Cat

Métodos de Matching en estudios retrospectivos



# Métodos de Matching

Jordi Real - 2022



# Guión

## ¿Por qué? ¿Como? ¿Cuándo?

- Introducción
  - Confusión
  - Métodos de ajuste
  - Comparativa
- Métodos de matching
  - Proceso
  - Distancias, algoritmos y Validación
  - Herramientas y grados de libertad
- Conclusiones / Resumen
  - Cuándo aplicarlo
  - Resumen

# Introducción

Aspecto más importante de una investigación?



TARGET

# Introducción

Estudio CVD-Real

# Cardiovascular Events Associated With SGLT-2 Inhibitors Versus Other Glucose-Lowering Drugs

## The CVD-REAL 2 Study

Mikhail Kosiborod, MD,<sup>a</sup> Carolyn S.P. Lam, MBBS, PhD,<sup>b,c</sup> Shun Kohsaka, MD,<sup>d</sup> Dae Jung Kim, MD,<sup>e</sup> Avraham Karasik, MD,<sup>f</sup> Jonathan Shaw, MD,<sup>g</sup> Navdeep Tangri, MD, PhD,<sup>h</sup> Su-Yen Goh, MD,<sup>i</sup> Marcus Thuresson, PhD,<sup>j</sup> Hungta Chen, PhD,<sup>k</sup> Filip Surmont, MD,<sup>l</sup> Niklas Hammar, PhD,<sup>m,n</sup> Peter Fenici, MD,<sup>o</sup>  
on behalf of the CVD-REAL Investigators and Study Group

### ABSTRACT

**BACKGROUND** Randomized trials demonstrated a lower risk of cardiovascular (CV) events with sodium-glucose cotransporter-2 inhibitors (SGLT-2i) in patients with type 2 diabetes (T2D) at high CV risk. Prior real-world data suggested similar SGLT-2i effects in T2D patients with a broader risk profile, but these studies focused on heart failure and death and were limited to the United States and Europe.

**OBJECTIVES** The purpose of this study was to examine a broad range of CV outcomes in patients initiated on SGLT-2i versus other glucose-lowering drugs (oGLDs) across 6 countries in the Asia Pacific, the Middle East, and North American regions.

**METHODS** New users of SGLT-2i and oGLDs were identified via claims, medical records, and national registries in South Korea, Japan, Singapore, Israel, Australia, and Canada. Propensity scores for SGLT-2i initiation were developed in each country, with 1:1 matching. Hazard ratios (HRs) for death, hospitalization for heart failure (HHF), death or HHF, MI, and stroke were assessed by country and pooled using weighted meta-analysis.

**RESULTS** After propensity-matching, there were 235,064 episodes of treatment initiation in each group; ~27% had established CV disease. Patient characteristics were well-balanced between groups. Dapagliflozin, empagliflozin, dapagliflozin, canagliflozin, tofogliflozin, and luseogliflozin accounted for 75%, 9%, 8%, 4%, 3%, and 1% of exposure time in the SGLT-2i group, respectively. Use of SGLT-2i versus oGLDs was associated with a lower risk of death (HR: 0.51; 95% confidence interval [CI]: 0.37 to 0.70;  $p < 0.001$ ), HHF (HR: 0.64; 95% CI: 0.50 to 0.82;  $p = 0.001$ ), death or HHF (HR: 0.60; 95% CI: 0.47 to 0.76;  $p < 0.001$ ), MI (HR: 0.81; 95% CI: 0.74 to 0.88;  $p < 0.001$ ), and stroke (HR: 0.68; 95% CI: 0.55 to 0.84;  $p < 0.001$ ). Results were directionally consistent across both countries and patient subgroups, including those with and without CV disease.

**CONCLUSIONS** In this large, international study of patients with T2D from the Asia Pacific, the Middle East, and North America, initiation of SGLT-2i was associated with a lower risk of CV events across a broad range of outcomes and patient characteristics. (Comparative Effectiveness of Cardiovascular Outcomes in New Users of SGLT-2 Inhibitors [CVD-REAL]; [NCT02993614](#))  
(J Am Coll Cardiol 2018;71:2628–39) © 2018 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# Introducción

## Estudio CVD-Real Catalonia

ORIGINAL INVESTIGATION

Open Access



### Cardiovascular and mortality benefits of sodium–glucose co-transporter-2 inhibitors in patients with type 2 diabetes mellitus: CVD-Real Catalonia

Jordi Real<sup>1,2†</sup>, Bogdan Vlacho<sup>1†</sup>, Emilio Ortega<sup>3,4</sup>, Joan Antoni Vallés<sup>1,5</sup>, Manel Mata-Cases<sup>1,2,6</sup>, Esmeralda Castelblanco<sup>1,2</sup>, Eric T. Wittbrodt<sup>7</sup>, Peter Fenici<sup>8</sup>, Mikhail Kosiborod<sup>9</sup>, Dídac Mauricio<sup>1,2,10,11\*</sup> and Josep Franch-Nadal<sup>1,2,12\*</sup>

#### Abstract

**Background:** Evidence from prospective cardiovascular (CV) outcome trials in type 2 diabetes (T2DM) patients supports the use of sodium–glucose co-transporter-2 inhibitors (SGLT2i) to reduce the risk of CV events. In this study, we compared the risk of several CV outcomes between new users of SGLT2i and other glucose-lowering drugs (oGLDs) in Catalonia, Spain.

**Methods:** CVD-REAL Catalonia was a retrospective cohort study using real-world data routinely collected between 2013 and 2016. The cohorts of new users of SGLT2i and oGLDs were matched by propensity score on a 1:1 ratio. We compared the incidence rates and hazard ratio (HR) for all-cause death, hospitalization for heart failure, chronic kidney disease, and modified major adverse CV event (MACE; all-cause mortality, myocardial infarction, or stroke).

**Results:** After propensity score matching, 12,917 new users were included in each group. About 27% of users had a previous history of CV disease. In the SGLT2i group, the exposure time was 60% for dapagliflozin, 26% for empagliflozin and 14% for canagliflozin. The use of SGLT2i was associated with a lower risk of heart failure (HR: 0.59; 95% confidence interval [CI] 0.47–0.74; p < 0.001), all-cause death (HR = 0.41; 95% CI 0.31–0.54; p < 0.001), all-cause death or heart failure (HR = 0.55; 95% CI 0.47–0.63; p < 0.001), modified MACE (HR = 0.62; 95% CI 0.52–0.74; p < 0.001), and chronic kidney disease (HR = 0.66; 95% CI 0.54–0.80; p < 0.001).

**Conclusions:** In this large, retrospective observational study of patients with T2DM from a Catalonia, initiation of SGLT-2i was associated with lower risk of mortality, as well as heart failure and CKD.

**Keywords:** SGLT2i, Heart failure, All-cause mortality, Type 2 diabetes mellitus

# Introducción

## Estudio Cuidadores

---

## Original Article

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

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

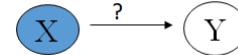
<sup>a</sup>Facultat de Medicina i Ciències de la Salut, Public Health Unit, School of Medicine and Health Sciences, Universitat Internacional de Catalunya, C. Doctor Trueta S/N, Sant Cugat del Vallès, Barcelona 08195, Spain.

<sup>b</sup>Centre d'Atenció Primària Les Corts. Transverse Group for Research in Primary Care, IDIBAPS, Barcelona, Spain.

<sup>c</sup>Institut d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol). USR-Lleida, Lleida, Spain.

# Ejemplo

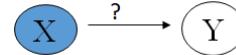
## Estudio Cuidadores



- Método:
  - ENSE: Encuesta nacional de salud española. 20.000 hogares
  - Se identificó **515** personas en situación de cuidador informal (superior al año)
  - Resultados en Salud: Diagnóstico depresión, Ansiedad, Calidad de vida, Estado de salud percibido, soporte social

# Ejemplo

## Enfoque analítico básico

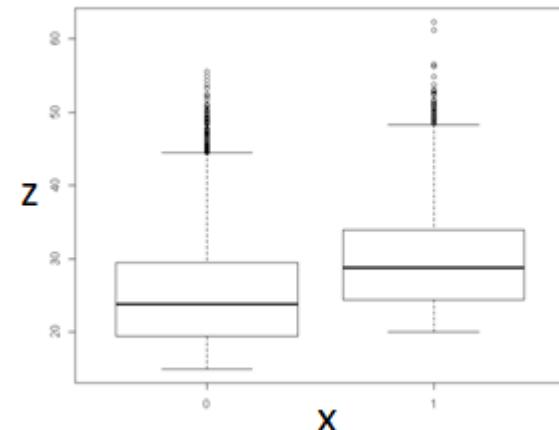


- $P(\text{Depresión} / \text{Cuidador}) = 16\% (n=515)$
- $P(\text{Depresión} / \text{Resto}) = 8.4\% (n=19.500)$
- Medida de asociación :  $OR = 2.03 (CI95: 1.7 - 2.5)$

# Introducción

## Preguntas

- Puedo comparar resultados en salud de los 515 cuidadores con el resto 19.514 no cuidadores?
- Son comparables?
- Veamos la diferencia en edad entre los cuidadores versus los no-cuidadores
- Los cuidadores 7.3 años mayores que los no cuidadores



(a) Boxplot

Métodos de Matching en estudios retrospectivos

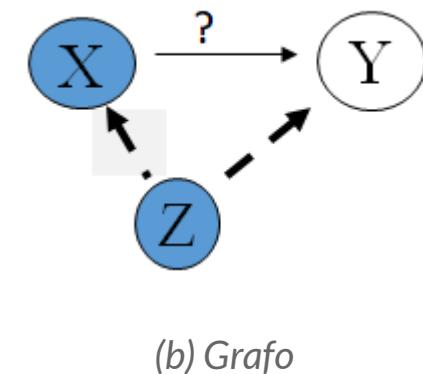
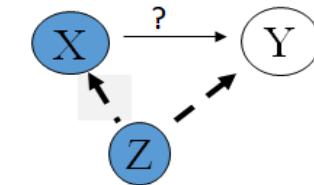


Figure 1: DAG

Métodos de Matching en estudios retrospectivos

# Sesgo de confusión

## Sesgo de confusión



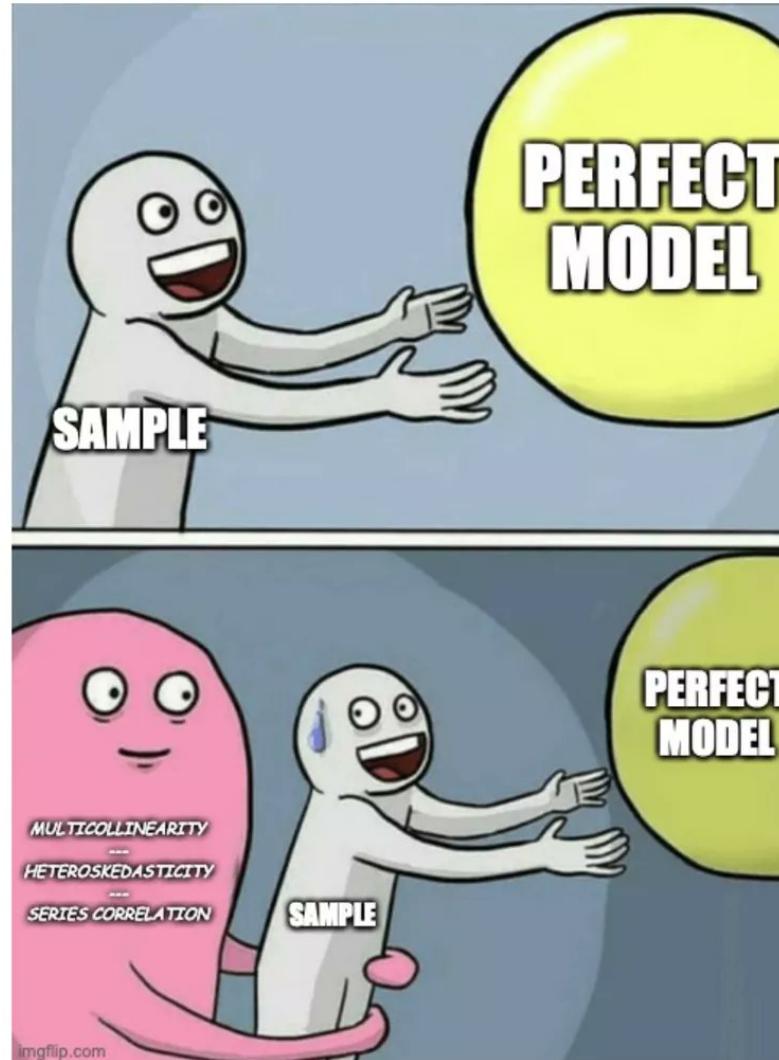
- Habitual en estudios observacionales
- Confusión: “mezcla” o “difuminación” de efectos
- Se trata de relacionar una exposición a un resultado
- En realidad mide el efecto de un tercer factor (la variable de confusión)
- Distorsiona la medida de la asociación entre otras dos variables
- El resultado en presencia de una variable de confusión puede ser la observación de:
  1. Efecto donde en realidad no existe (Asociación espuria)
  2. Exageración o atenuació d'una asociación real (confusión positiva)
  3. Inversión del sentido de una asociación real (confusión negativa)

# Métodos de ajuste

- Restricción (Diseño)
- Anticipación de confusores potenciales (Diseño)
- Estratificación por confusor/es (Análisis):
  - Simple
  - Difícil con muchas covariables
- Técnicas de estandarización (Análisis)
- Métodos de regresión (Ajuste por covarianza) (Análisis)
  - Mayor potencia estadística
  - Técnico
  - Asunciones de modelos

# Modelos de regresión

## Asunciones sobre modelos



Métodos de Matching en estudios retrospectivos

# Modelos de regressión

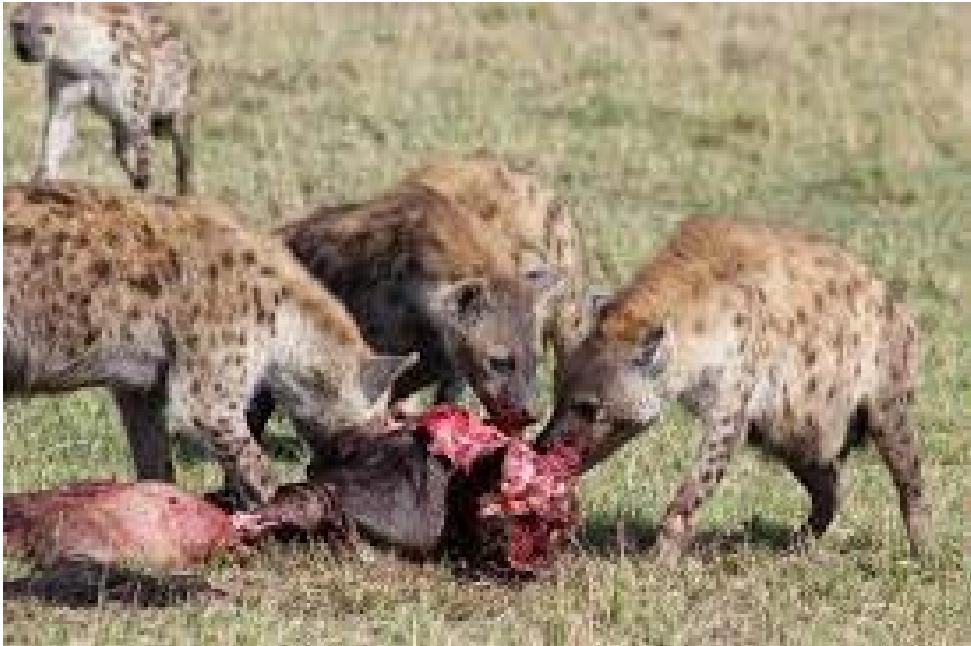
## Asunciones sobre modelos de regressión



Métodos de Matching en estudios retrospectivos

# Modelos de regressión

Asunciones sobre modelos de regression



# Modelos de regressión

## Asunciones sobre modelos de regression



**Reviewer #2: (27/09/2023)** The paper is beautifully written and considers an important topic. Since this is a modelling study, including additional details on the models is necessary. My detailed comments follow.

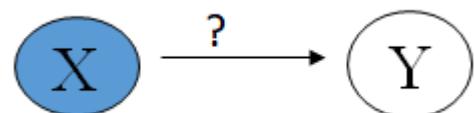
1. For the univariate and multivariate (presumably multivariable) analyses mentioned on page 5, line 14 (or 28 with the journal's numbering), please specify what kind of models were used, e.g. Poisson, Cox etc.
2. Please briefly describe what model **assumption and goodness of fit** checks were performed, and what the findings were (you could include details in supplementary material).
3. There are many variables included in the multivariable analyses - how were these selected, and what checks for **collinearity** were performed?

# Ajuste por regressión

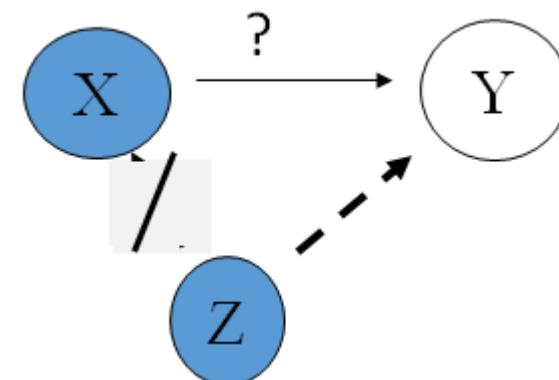
Estudio cuidadores. **Enfoque multivariable**

Ajuste multivariable mediante regressión logística

- Tener en cuenta la edad y otros factores asociados<sup>1</sup>
- Medida de asociación **no ajustada**: OR= 2.03 (IC95%: 1.7 - 2.5)
- Medida de asociación **ajustada**: OR= 1.56 (IC95%: 1.21 - 2.04)



OR: 2.03; IC95%=(1,71-2,5)



OR: 1.56; IC95%=(1.21-2.04)

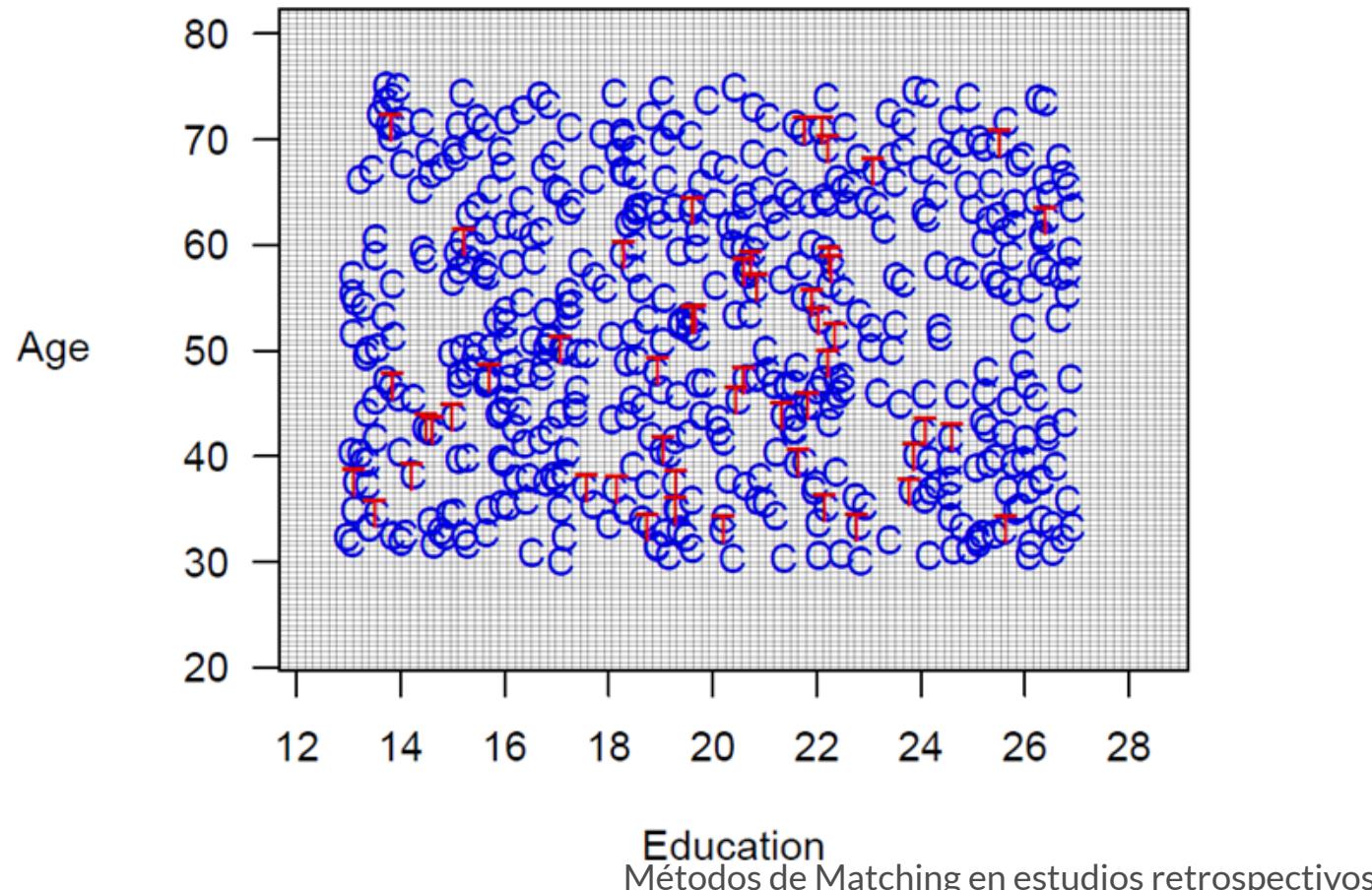
# Métodos de ajuste

- Restricción (Diseño)
- Anticipación de confusores potenciales (Diseño)
- Estratificación por confusor/es (Análisis):
  - Simple
  - Difícil con muchas covariables
- Técnicas de estandarización (Análisis)
- Métodos de regresión (Ajuste por covarianza) (Análisis)
  - Mayor potencia estadística
  - Técnico
  - Asunciones de modelos
- Matching (Diseño / Análisis)

# Matching

## En que consiste?

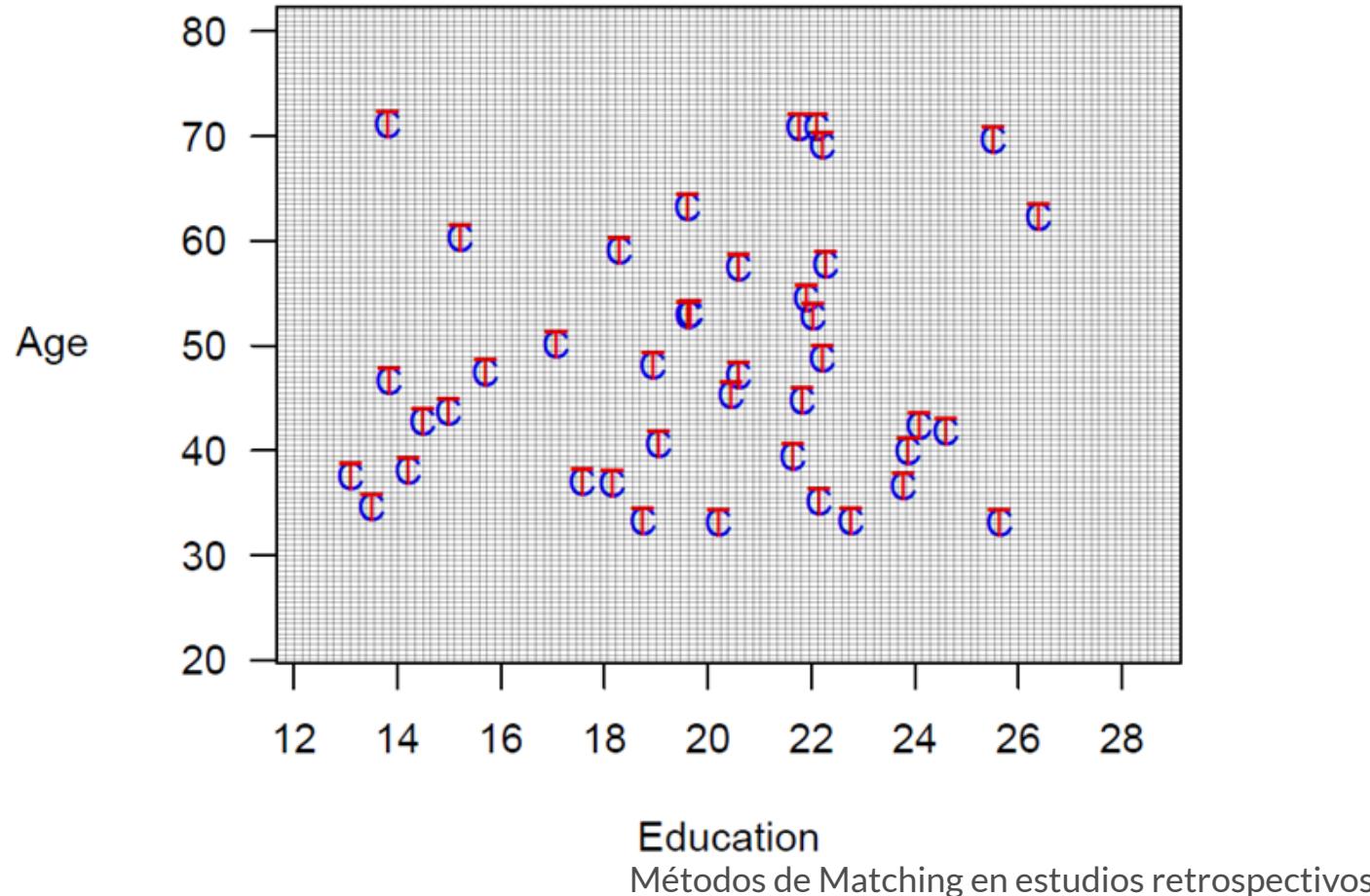
Dada N, encontrar n ( $n < N$ ), tal que los grupos sean de igual o de similares características



# Matching

En que consiste?

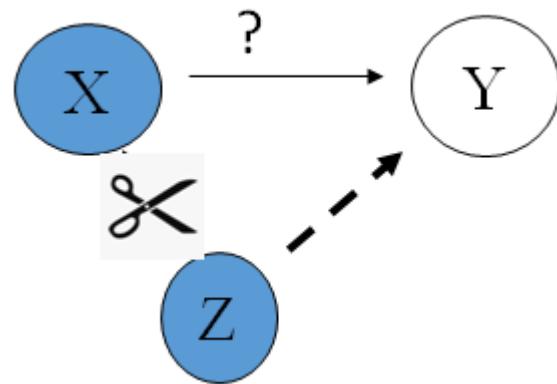
Dada N, encontrar n ( $n < N$ ), tal que los grupos sean de igual o de similares características



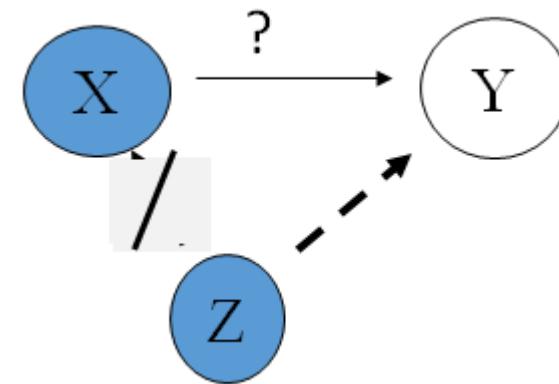
# Estudio cuidadores

## Comparativa entre métodos

- Medida de asociación **no ajustada**: OR= 2,03 (IC95%: 1,7 - 2,5)
- Medida de asociación **ajustada según regressión**: OR= 1,56 (IC95%: 1,21 - 2,04)
- Medida de asociación **según método de matching**: OR = 1,34 (IC95%: 1,02 - 1,76)



OR: 1.34; IC95%=(1,02-1,76)

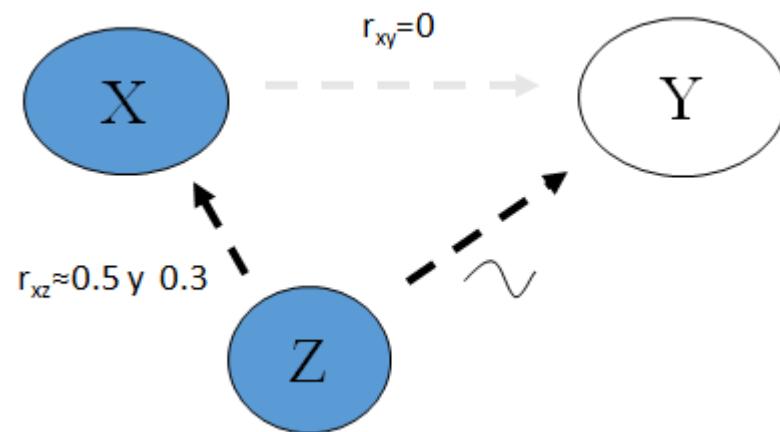


OR: 1.56; IC95%=(1.21-2.04)

# ¿Por que?

## La prueba del algodón. Estudio de simulación

- Objetivo: Comparar la corrección del **sesgo de confusión** entre distintas aproximaciones mediante un estudio de simulación
- En una situación controlada donde conocemos la realidad
- Generamos 7500 muestras simuladas de tamaño grande ( $n=10.000$ ) tal que:



# Estudio de simulación

- Distintos escenarios donde la Relación confusor Z vs P(Y):

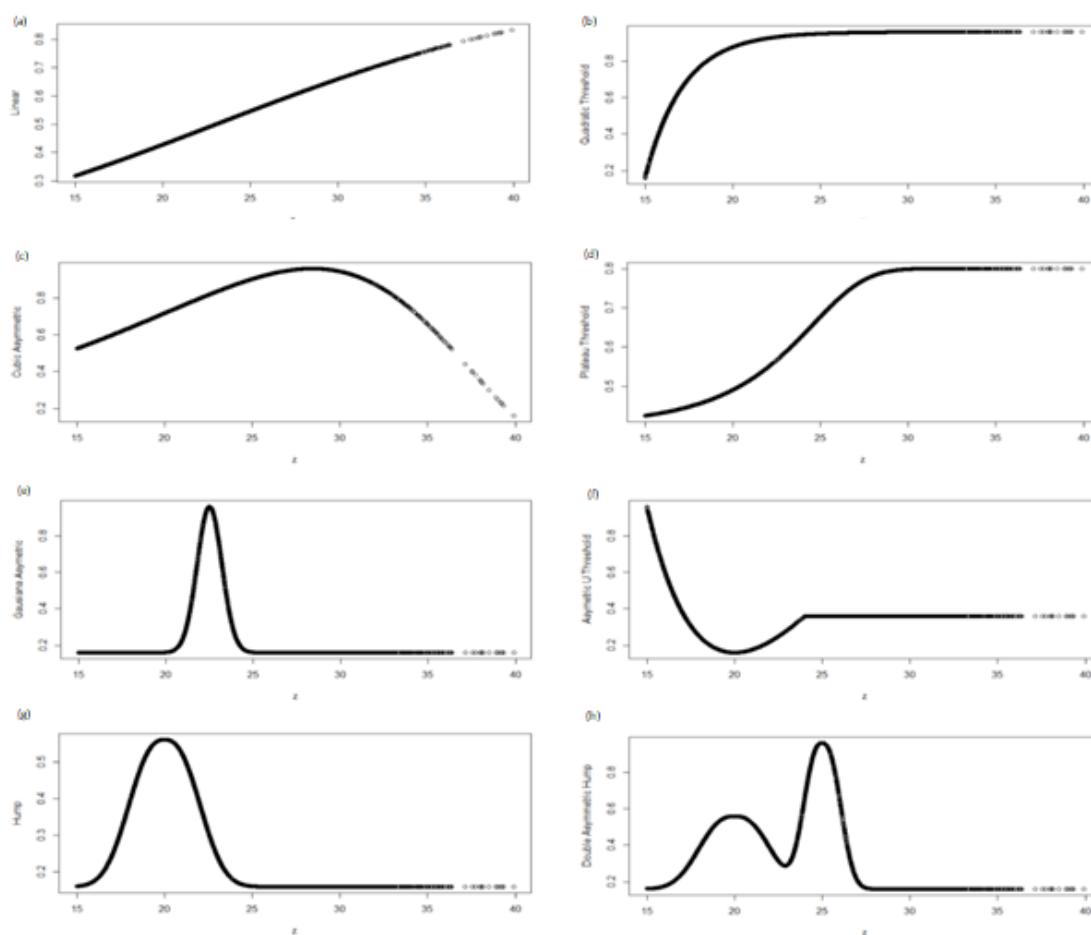
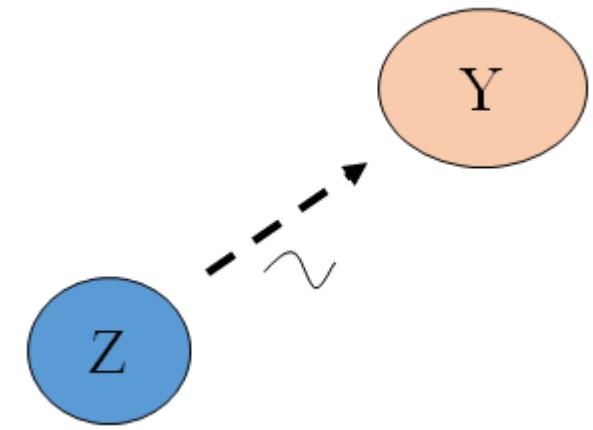


Figura 1. Relaciones generadas de Z-P(Y): (a) Linear, (b) Quadratic Threshold, (c) Cubic Asymmetric, (d) Plateau Threshold, (e) Gausiana Asymmetric, (f) Asymmetric U Threshold, (g) "Hump", (h) Double Hump.

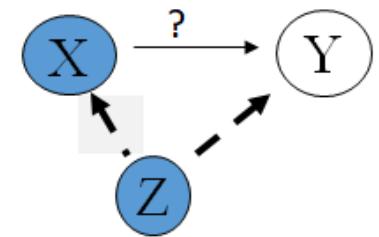


# Estudio de simulación

Estudio de simulación

Distintas aproximaciones para estimar el efecto de X sobre Y

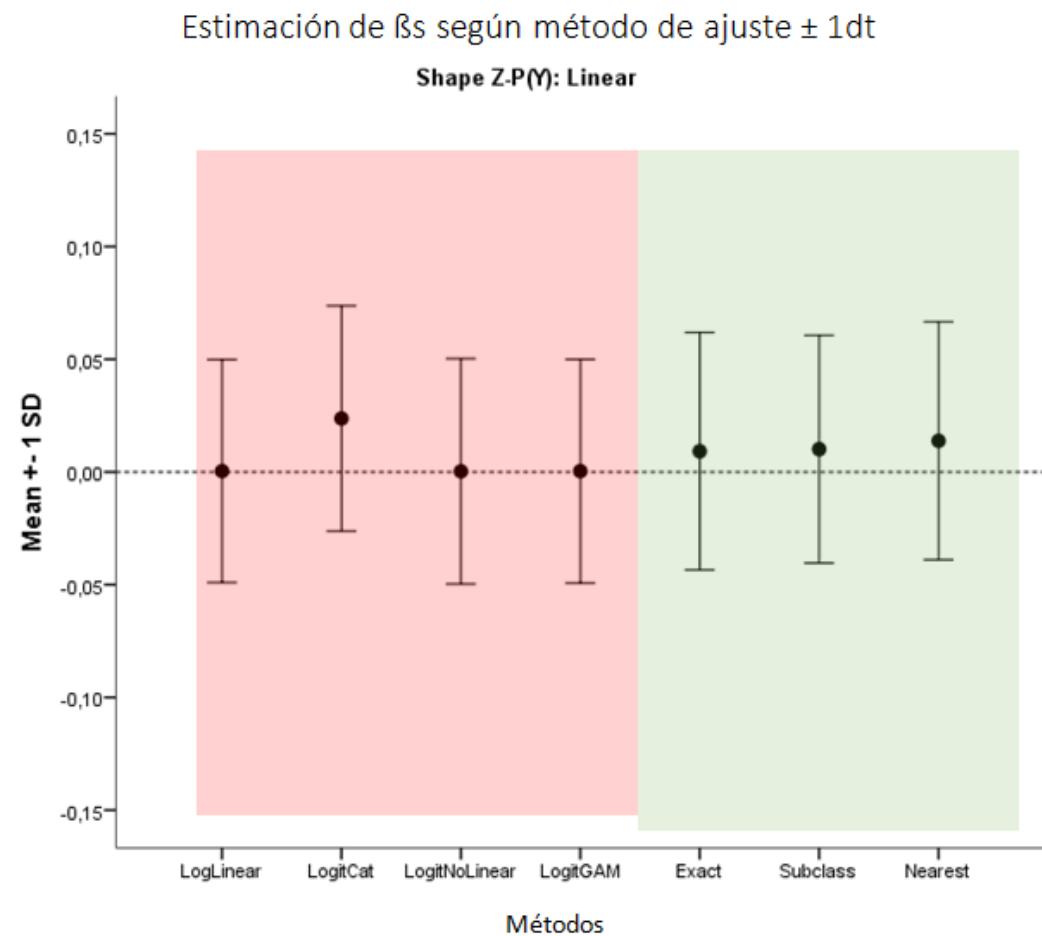
<i>Descripción</i>	Método
1. Z Lineal (LogLineal)	Regresión
2. Categorizando Z En quintiles (LogitCat)	Regresión
3. Fuciones polinómicas $Z^3$ (LogNoLineal)	Regresión
4. Función no paramétrica $s(Z)$ (GAM)	Regresión
5. Exacto	Matching exacto
6. Subclasificación con descartes	Matching
7. Nearest- Neighbour	Matching N-N



# Estudio de simulación

## Resultados escenario Lineal

Estimación de efecto nulo ( $OR=1$ ) según el método de ajuste

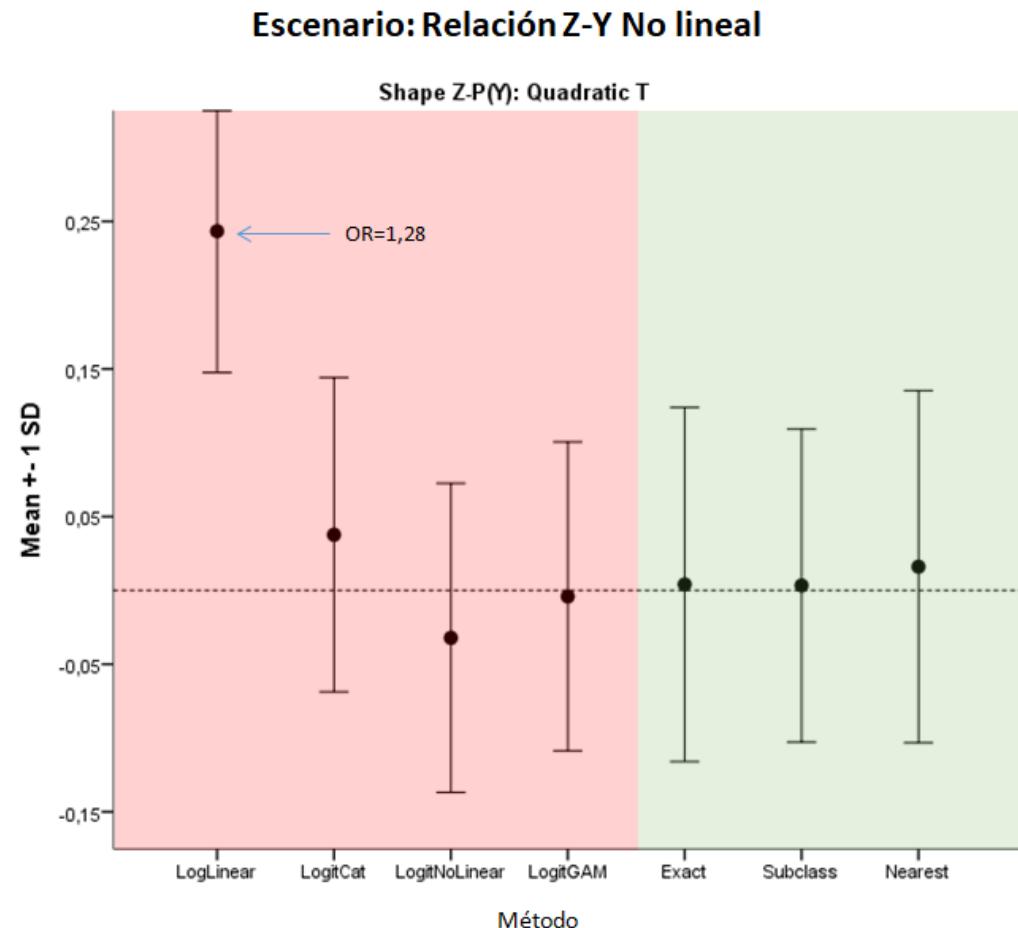


Métodos de Matching en estudios retrospectivos

# Estudio de simulación

## Resultados escenario No Lineal

Estimación de efecto nulo ( $OR=1$ ) según el método de ajuste

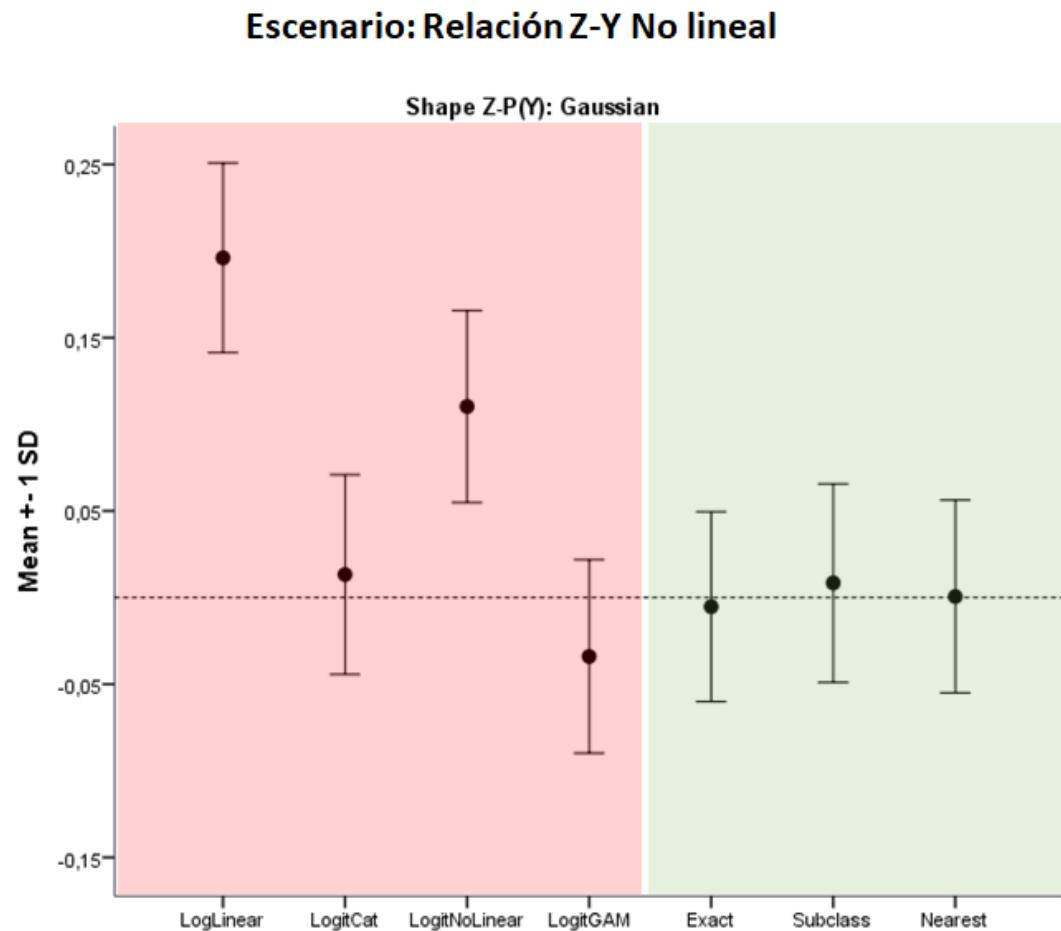


Métodos de Matching en estudios retrospectivos

# Estudio de simulación

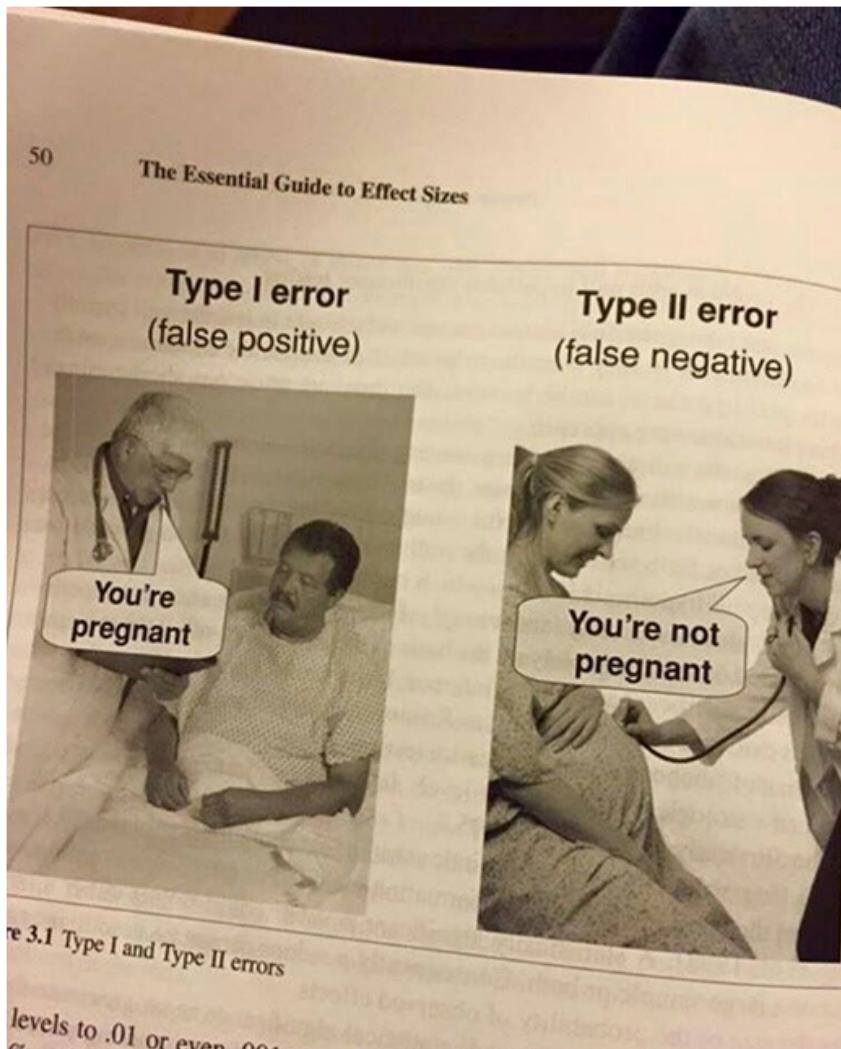
## Resultados escenario No lineal

Estimación de efecto nulo ( $OR=1$ ) según el método de ajuste



# Estudio de simulación

Error de tipo I empírico según escenario y método



# Estudio de simulación

## Tasa de falsos positivos según escenario y método

$\alpha$  empírico: Tasa de falsos positivos – escenario y método

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

Shape relation Z-Y	GLM				Matching		
	LogLinear	LogitCat	LogNolineal	GAM	Exact	Subclass	Nearest
Low (SD=10; r=0.3)							
Linear	<b>0,051</b>	<b>0,079</b>	<b>0,052</b>	<b>0,052</b>	<b>0,045</b>	<b>0,059</b>	<b>0,053</b>
Quadratic T	<b>0,701</b>	<b>0,066</b>	<b>0,064</b>	<b>0,051</b>	<b>0,051</b>	<b>0,052</b>	<b>0,052</b>
Cubic Asymmetric	<b>0,980</b>	<b>0,324</b>	<b>0,050</b>	<b>0,053</b>	<b>0,033</b>	<b>0,105</b>	<b>0,038</b>
Plateau	<b>0,066</b>	<b>0,051</b>	<b>0,060</b>	<b>0,050</b>	<b>0,042</b>	<b>0,054</b>	<b>0,043</b>
Gaussian	<b>0,945</b>	<b>0,055</b>	<b>0,525</b>	<b>0,082</b>	<b>0,029</b>	<b>0,051</b>	<b>0,032</b>
Asymmetric U T	<b>0,142</b>	<b>0,052</b>	<b>0,074</b>	<b>0,051</b>	<b>0,046</b>	<b>0,057</b>	<b>0,052</b>
Hump	<b>0,050</b>	<b>0,052</b>	<b>0,051</b>	<b>0,050</b>	<b>0,041</b>	<b>0,055</b>	<b>0,048</b>
Double Hump	<b>0,802</b>	<b>0,058</b>	<b>0,049</b>	<b>0,054</b>	<b>0,027</b>	<b>0,051</b>	<b>0,031</b>
Total	<b>0,474</b>	<b>0,093</b>	<b>0,117</b>	<b>0,056</b>	<b>0,039</b>	<b>0,061</b>	<b>0,044</b>

Color de fondo condicionado a la magnitud del error de tipo I empírico: Cuanto más oscuro mas alejado de 0.05

# ¿Por qué fallan algunos modelos?



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

# Tipos de matching

- Matching aproximado por frecuencia (según una distancia)
- Matching exacto por frecuencia
- Matching individual por densidad de incidencia

# Etapas del matching

## 4 fases

Donde las tres primeras representan la fase de “diseño” y la última la de “análisis”

### 1. Distancia

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

### 2. Algoritmo

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

### 3. Validación

- 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

# Distancias

## Métodos para calcular distancias de covariables

"euclidean" The Euclidean distance is the raw distance between units, computed as

$$d_{ij} = \sqrt{(x_i - x_j)(x_i - x_j)'} \quad \text{where } x_i \text{ and } x_j \text{ are vectors of covariates}$$

It is sensitive to the scale of the covariates, so covariates with larger scales will take higher priority.

"scaled\_euclidean" The scaled Euclidean distance is the Euclidean distance computed on the scaled (i.e., standardized) covariates. This ensures the covariates are on the same scale. The covariates are standardized using the pooled within-group standard deviations, computed by treatment group-mean centering each covariate before computing the standard deviation in the full sample.

"mahalanobis" The Mahalanobis distance is computed as

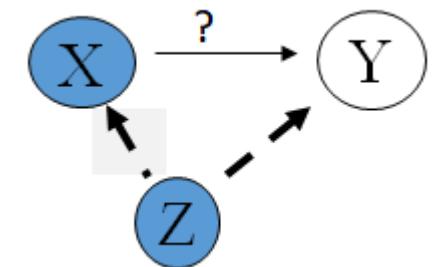
$$d_{ij} = \sqrt{(x_i - x_j)\Sigma^{-1}(x_i - x_j)'} \quad \text{where } \Sigma \text{ is the covariance matrix of covariates}$$

where  $\Sigma$  is the pooled within-group covariance matrix of the covariates, computed by treatment group-mean centering each covariate before computing the covariance in the full sample. This ensures the variables are on the same scale and accounts for the correlation between covariates.

"robust\_mahalanobis" The robust rank-based Mahalanobis distance is the Mahalanobis distance computed on the ranks of the covariates with an adjustment for ties. It is described in Rosenbaum (2010, ch. 8) as an alternative to the Mahalanobis distance that handles outliers and rare categories better than the standard Mahalanobis distance but is not affinely invariant.

# Distancias

Cual és la reina de las distancias?



$$P(X) = \text{Probabilidad}(X = 1/Z) = f(Z)$$

$$\Pr(X_i/Z_i) = \frac{e^{g(Z_i)}}{1+e^{g(Z_i)}}$$

# Algoritmos de agrupación

- **Exacto**

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. Cuándo hay muchas covariables y/o las covariables pueden tomar un amplio rango de valores, exact matching puede no ser posible.

- **Nearest Neighbour (N-N)**

Mejores controles emparejados para cada individuo tratado. Observaciones del grupo control lo más cercana a tratada según la distancia especificada. Misma distancia 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 grupos para ser seleccionados asegurando una igualdad mínima entre observaciones.

- **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.

- **Otros métodos (Optimal, Full, Genetic etc..)**

# Validación

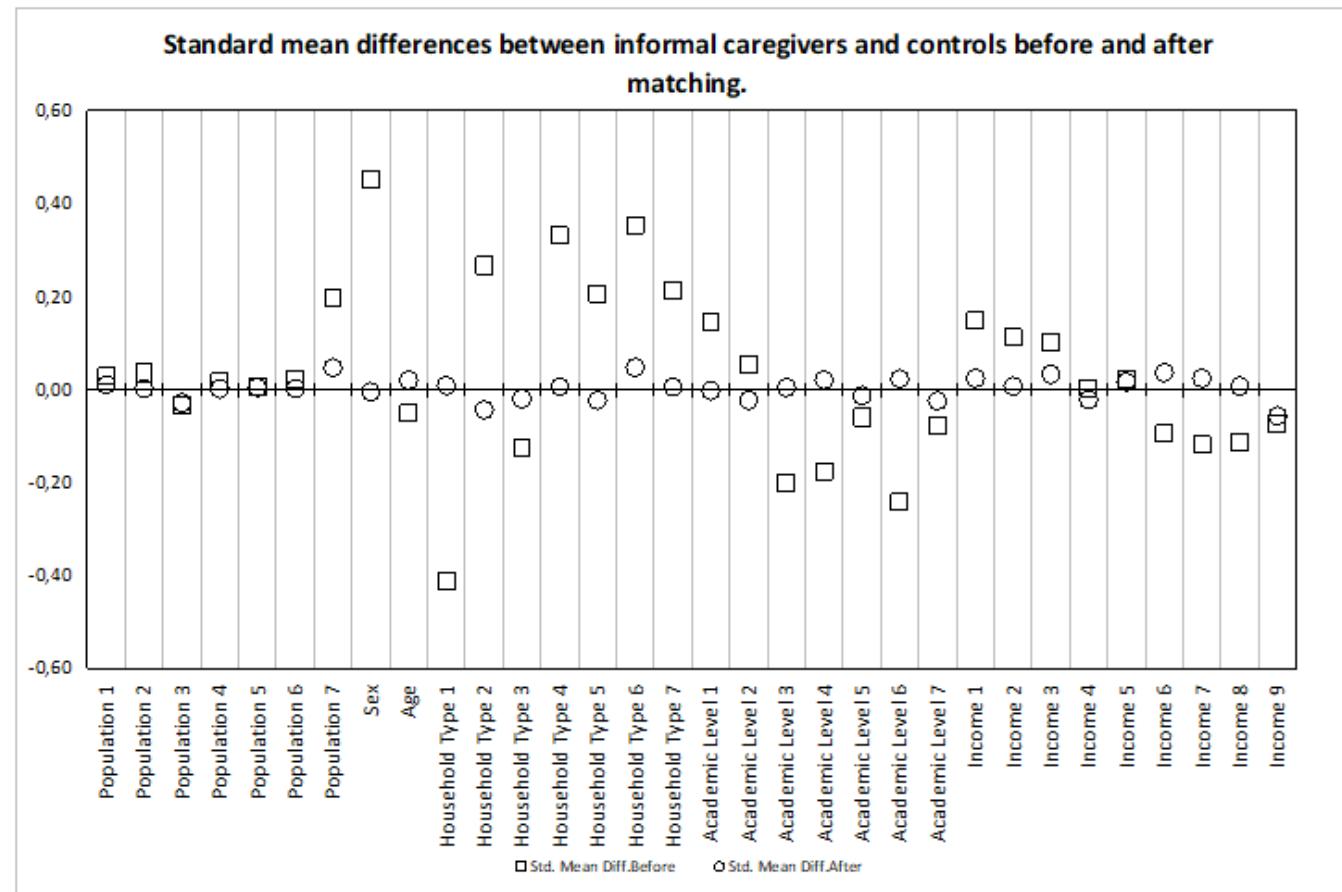
Evaluación del equilibrio (calidad del matching) de los grupos emparejados. Si la calidad del matching no se satisface se repiten las etapas 1 y 2 hasta que los grupos puedan considerarse coincidentes



# Validación

## Estudio cuidadores

Diferencias medias estandarizadas (SMD) por covariable. Se recomienda un SMD<0.1.

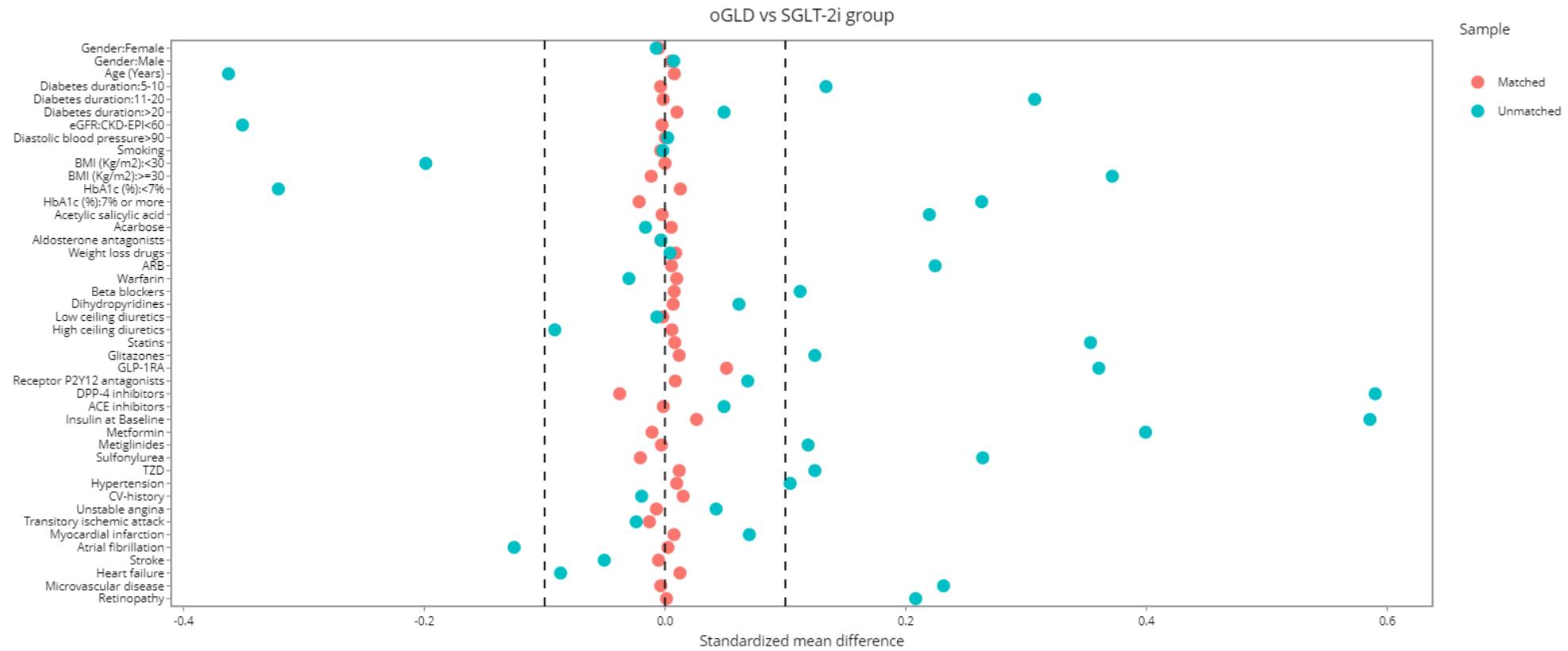


Covariate plot

Métodos de Matching en estudios retrospectivos

# Validación

Evaluación de la calidad del equilibrio de las covariables basales después del matching  
 Estudio CVD Real Catalonia



Covariate plot

# Como?

## Herramientas y grados de libertad

- Software: Stata, R, SAS
- Ratio de los grupos (1:1 hasta 1:4)
- Distancia utilizada (PS, mahalanobis, euclidiana)
- Algoritmo de agrupación (N-N, Exacto, óptimo, genético .....)
- Descartes
- Variables

# Variables

- Potenciales confusoras segun conocimiento previo
- Número?
  - No hay límite. Cuantas más, más difícil encontrar buenos grupos
- No: outcomes secundarios, ni variables subrogadas



# Herramientas

## Paquete {Matchit} de R (version 4.4.0)

### Matching for Causal Inference

#### Descripción

`matchit()` is the main function of `MatchIt` and performs pairing, subset selection, and subclassification with the aim of creating treatment and control groups balanced on included covariates. `MatchIt` implements the suggestions of Ho, Imai, King, and Stuart (2007) for improving parametric statistical models by preprocessing data with nonparametric matching methods. `MatchIt` implements a wide range of sophisticated matching methods, making it possible to greatly reduce the dependence of causal inferences on hard-to-justify, but commonly made, statistical modeling assumptions. The software also easily fits into existing research practices since, after preprocessing with `MatchIt`, researchers can use whatever parametric model they would have used without `MatchIt`, but produce inferences with substantially more robustness and less sensitivity to modeling assumptions.

#### Función:

```
matchit(formula, data = NULL, method = "nearest", distance = "glm", link = "logit", distance.options = list(), estimand = "ATT",
exact = NULL, mahvars = NULL, antiexact = NULL, discard = "none", reestimate = FALSE, s.weights = NULL, replace = FALSE,
m.order = NULL, caliper = NULL, std.caliper = TRUE, ratio = 1, verbose = FALSE,...)
```

# MatchIt

## Ejemplo

Summary descriptives table by groups of 'treat'

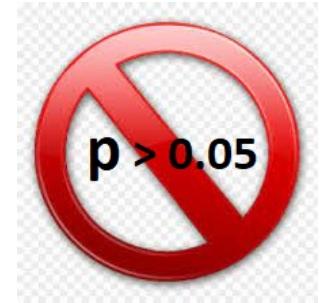
	0	1
	<i>N</i> =429	<i>N</i> =185
age	28.0 (10.8)	25.8 (7.16)
educ	10.2 (2.86)	10.3 (2.01)
race:		
black	87 (20.3%)	156 (84.3%)
hispan	61 (14.2%)	11 (5.95%)
white	281 (65.5%)	18 (9.73%)
nodegree	0.60 (0.49)	0.71 (0.46)
married	0.51 (0.50)	0.19 (0.39)
re74	5619 (6789)	2096 (4887)
re75	2466 (3292)	1532 (3219)

# MatchIt

## Ejemplo

Summary descriptives table by groups of 'treat'

	0	1	p.overall
	N=429	N=185	
age	28.0 (10.8)	25.8 (7.16)	0.003
educ	10.2 (2.86)	10.3 (2.01)	0.585
race:		<0.001	
black	87 (20.3%)	156 (84.3%)	
hispan	61 (14.2%)	11 (5.95%)	
white	281 (65.5%)	18 (9.73%)	
nodegree	0.60 (0.49)	0.71 (0.46)	0.007
married	0.51 (0.50)	0.19 (0.39)	<0.001
re74	5619 (6789)	2096 (4887)	<0.001
re75	2466 (3292)	1532 (3219)	0.001



# MatchIt

## Paquete MatchIt de R (versión 4.4.0)

```

1 dt_temp<-lalonde %>% mutate(idp=1:n())
2
3 set.seed(123)
4
5 m.out1 <- matchit(treat ~ age + educ + race + nodegree + married + re74 + re75,
6                     exact = ~ race + married,
7                     distance = "glm",
8                     method = "nearest",
9                     discard = "both",
10                    ratio = 1,
11                    caliper = .1,
12                    # unit.id="idp",
13                    data = dt_temp)

```

# MatchIt

## Paquete MatchIt de R (versión 4.4.0)

```
1 m.out1
```

A matchit object

- method: 1:1 nearest neighbor matching without replacement
- distance: Propensity score [caliper, common support]
  - estimated with logistic regression
- caliper: <distance> (0.029)
- common support: units from both groups dropped
- number of obs.: 614 (original), 212 (matched)
- target estimand: ATT
- covariates: age, educ, race, nodegree, married, re74, re75

```
1 pp<-summary(m.out1)
2 pp$nn
```

	Control	Treated
All (ESS)	429	185
All	429	185
Matched (ESS)	106	106
Matched	106	106
Unmatched	266	71
Discarded	57	8

# MatchIt

## Validación

Summary descriptives table by groups of 'treat'

	0	1
	<i>N=106</i>	<i>N=106</i>
age	25.3 (10.2)	26.1 (6.87)
educ	10.4 (2.78)	10.4 (2.06)
race:		
black	77 (72.6%)	77 (72.6%)
hispan	11 (10.4%)	11 (10.4%)
white	18 (17.0%)	18 (17.0%)
nodegree	0.62 (0.49)	0.65 (0.48)
married	0.20 (0.40)	0.20 (0.40)
re74	2578 (4476)	2139 (4137)
re75	1739 (2882)	1707 (3788)

# MatchIt

## Validación

Summary descriptives table by groups of 'treat'

	0	1	p.overall
	<i>N</i> =106	<i>N</i> =106	
age	25.3 (10.2)	26.1 (6.87)	0.509
educ	10.4 (2.78)	10.4 (2.06)	0.911
race:			1.000
black	77 (72.6%)	77 (72.6%)	
hispan	11 (10.4%)	11 (10.4%)	
white	18 (17.0%)	18 (17.0%)	
nodegree	0.62 (0.49)	0.65 (0.48)	0.670
married	0.20 (0.40)	0.20 (0.40)	1.000
re74	2578 (4476)	2139 (4137)	0.459
re75	1739 (2882)	1707 (3788)	0.945

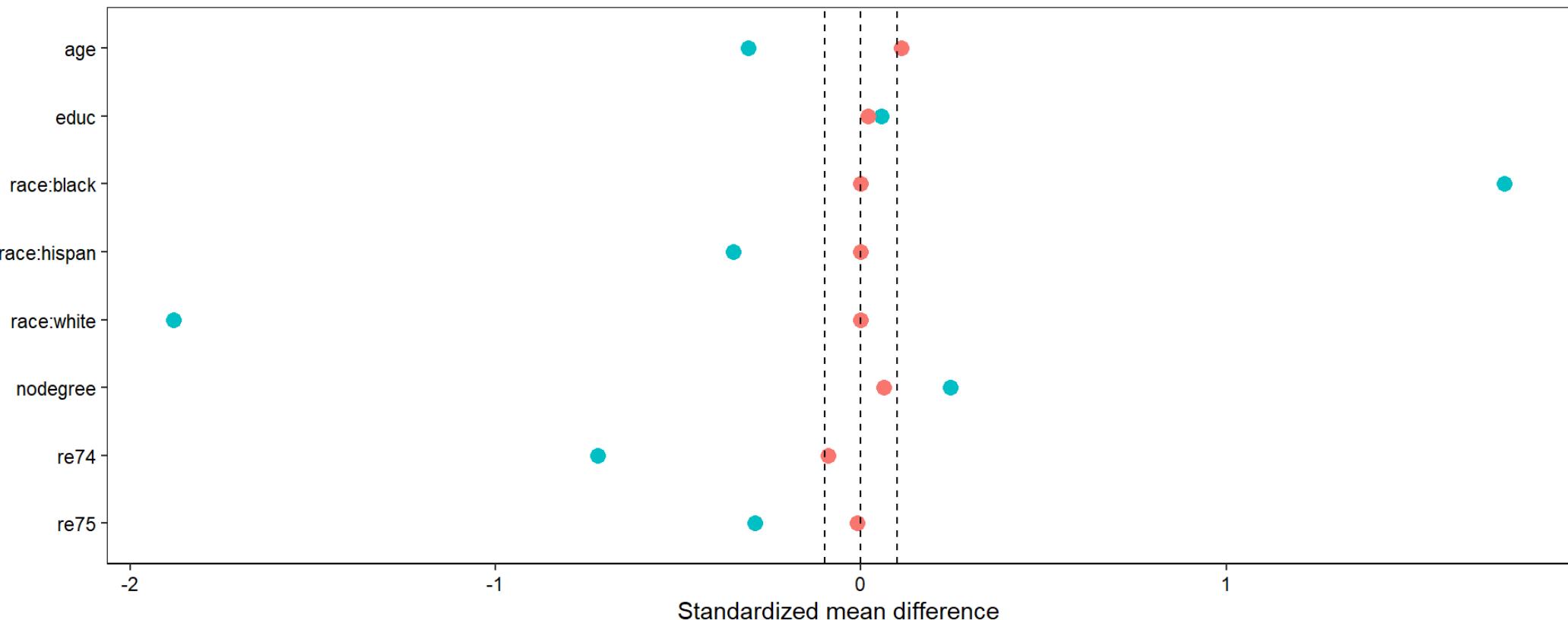


# MatchIt

## Validación

**Covariate plot**  
 Grupo intervención vs Control

Sample    ● Matched    ● Unmatched



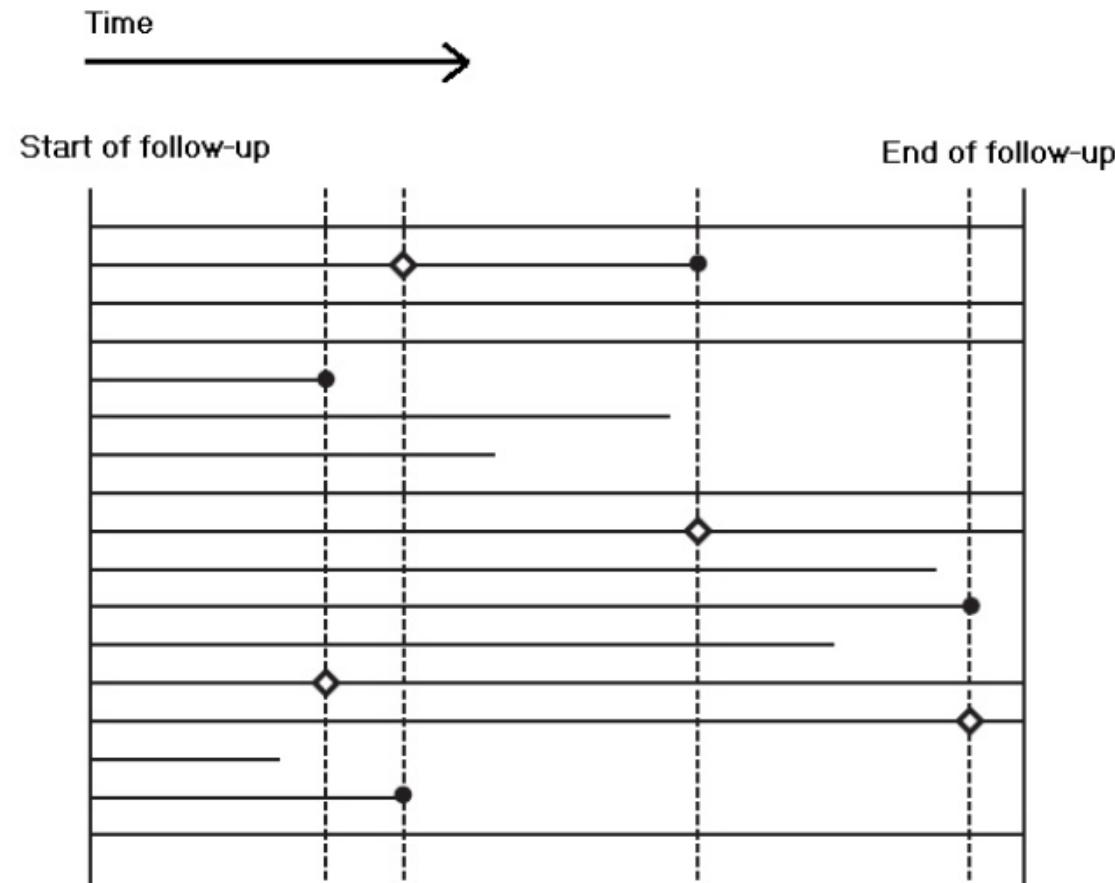
# Muestreo por densidad de incidencia

Matching individual/exacto

Apareamiento por densidad de incidencia

- Diseño: Caso-control anidado / Cohorte dinámica (o abierta)
- Los controles se seleccionan a medida que se producen los casos (Matching on time)
- Se construyen conjuntos de riesgo emparejados
- Los controles pueden ser remuestreados(más de una vez por caso)
- Los controles se pueden convertir en casos

# Muestreo por densidad de incidencia



Esquema de selección de casos

# Muestreo por densidad de incidencia

Ejemplo: Cohorte dinámica

Estudio DM\_TBC

```

1 library(Macedonia)
2 dat2<-Macedonia::match_density_incidence(dt=Macedonia::dat,
3                                              id="idp",
4                                              llistaPS=c("sex"),
5                                              eventcontrol=TRUE,
6                                              reemplacement=FALSE,
7                                              numcores=NA,
8                                              Ncontrols=1,
9                                              seed=123)
10 dat$idp %>% length()

```

[1] 55

```
1 dat2$idp %>% length()
```

[1] 18

```

1 dat2 %>% head() %>% select(-c(diabetes,heartdis,byear)) %>% kableExtra::kable() %>%
2   kableExtra::kable_classic_2() %>% kableExtra::kable_styling(font_size = 15)

```

idp	.caseid	.dtindex	.event	.n	event	sex	dtindex_case	dtindex_control
P41	1	5	1	1	1	fem	5	NA
P20	1	5	0	1	0	fem	NA	20
P42	2	8	1	1	1	fem	8	NA
P13	2	8	0	1	0	fem	NA	13
P43	3	11	1	1	1	fem	11	NA
P19	3	11	Métodos de Matching en estudios retrospectivos					19

# Muestreo por densidad de incidencia

## Ejemplo: Caso-Control

```

1 dades<-readRDS("dades_setMatch.Rds")
2 # data de censura en els casos com a molt la data de CAS
3 dades<-dados %>% mutate(dtindex_control=ifelse(event==1,dtindex_case,dtindex_control))
4 llistaPS<-c("sexe","year_DM2","year_naix")
5
6 dt_aparellada<-match_density_incidence(dades,
7                               id="id",
8                               llistaPS=llistaPS,
9                               eventcontrol = T,
10                              reemplacement=F,
11                              Ncontrols = 10,
12                              seed=131)
13 dades %>% n_distinct("id")

```

[1] 400

```
1 dt_aparellada %>% n_distinct("id")
```

[1] 30

```

1 dt_aparellada %>% select(id,.caseid,.dtindex,.event,.n, Fecha_caso=dat_cas,llistaPS) %>%
2   mutate(.dtindex=lubridate::as_date(.dtindex)) %>%
3   head(5) %>% kableExtra::kable() %>% kableExtra::kable_classic() %>% kableExtra::kable_styling(font_size = 1)

```

<b>id</b>	<b>.caseid</b>	<b>.dtindex</b>	<b>.event</b>	<b>.n</b>	<b>Fecha_caso</b>	<b>sexe</b>	<b>year_DM2</b>	<b>year_naix</b>
299	1	2012-09-13	1	2	2012-09-13	D	2004	1929
214	1	2012-09-13	0	2	2017-06-15	D	2004	1929
35	1	2012-09-13	0	2	NA	D	2004	1929
400	3	2013-04-11	1	1	2013-04-11	D	2004	1931
226	3	2013-04-11	0	1	NA	D	2004	1931

# ¿Cuándo No?

- Muestras insuficientes (“pequeñas”)
- Objetivos
  - Descriptivos
    - Multivariantes: Clustering, PCA, factorial
  - Estimar parámetros poblacionales (ej. prevalencia)
  - Predictivos
- Múltiples hipótesis / objetivos

# ¿Cuándo No?

Hipótesis / objetivos múltiples



Métodos de Matching en estudios retrospectivos

# ¿Cuándo Sí?

- Objetivos analíticos confirmatorios
- Inferencia causal
- Diseños
  - Transversal/Cohortes /Caso-control
- Variable principal (X o Y) de agrupación de naturaleza categórica (preferentemente binaria)<sup>1</sup>
- Estudio comparativo aislando factores ya conocidos

1. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006 May 6;332(7549):1080. doi: 10.1136/bmj.332.7549.1080 PMID: 16675816 DMCID: DMC1458573

# Para concluir...

## Razones

- Simplicidad
- Facilidad de evaluar su viabilidad
- Evaluar grado de solapamiento
- Conclusiones contextuales

# Para concluir...

- Modelos paramétricos funcionan, pero .....
- Modelos semiparamétricos funcionan mejor
- Robustez de los métodos Matching
- Análisis dirigido a un objetivo concreto
- Separan el diseño del análisis
  - El outcome es invisible al investigador

# Conclusiones



# Conclusiones

Sesgo de confirmación



ReactionGIFS.me

# Muchas gráciás

Último MEME



Métodos de Matching en estudios retrospectivos

# Bibliografía

- Rosenbaum, P.R. and D.B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70, 1, 41-55.
- 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.
- 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.
- 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.
- Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006 May 6;332(7549):1080. doi: 10.1136/bmj.332.7549.1080. PMID: 16675816; PMCID: PMC1458573.

# Material extra

## Propensity Score Matching for Acute Kidney Injury

vanced statistical methods such as propensity score (PS) matching, a statistical matching technique that deals with unmeasured residual confounding, is increasingly being used in large studies on PC-AKI using non-randomized observational data. In this systematic review, we give an overview of recent advanced observational studies on PC-AKI using PS matching and possible implications for guideline development.

distribution ranging from 0 (0% probability that a patient will be given a specific treatment) to 1.0 (100% probability that a patient will be given a specific treatment). Patients with a high PS for CM administration will tend to be older and have more comorbidities, which influences their likelihood of being offered IV CM in usual practice. Because propensity scores are an estimate of probabilities and not actual proportions, the treat-

and measur true rando of these co ing logistic able accor ences the t matching e it facilitate groups in t matching :

