

## Tópicos en investigación

# DISEÑO DE CASOS Y CONTROLES

María Alejandra Palacios<sup>1</sup>

1. Médico y cirujano. MSc en Epidemiología. Instructor Asociado Unidad de Investigación, Fundación Universitaria Sanitas, Bogotá D.C., Colombia.

## RESUMEN

El diseño de casos y controles es óptimo para el estudio de condiciones con baja prevalencia, y es de uso rutinario en el estudio de brotes en vigilancia epidemiológica. Sus ventajas incluyen un tiempo corto de desarrollo, bajo costo y un tamaño de muestra menor al de estudios de cohorte para condiciones infrecuentes. Sin embargo, este diseño es altamente susceptible a sesgos de selección, información y confusión, y se deben tomar medidas extensas para controlarlos. Es rol del experto clínico guiar el análisis estadístico por medio de la identificación informada de variables de confusión, el cual puede estar desprovisto de validez si no se incorpora el conocimiento previo. Por último, es también responsabilidad del experto la interpretación efectiva de las medidas de asociación obtenidas, haciendo un análisis acucioso de la significancia clínica de los resultados por encima de cualquier medida de significancia estadística.

**Palabras clave:** Estudios de casos y controles; Oportunidad relativa; Tamaño de la muestra; Sesgo.

**DOI:** 10.26852/01234250.45

## CASE CONTROL DESIGN

## ABSTRACT

Case-control study design is optimal for study of conditions with low prevalence and it is routinely used in the study of outbreaks in epidemiological surveillance. Its strengths include a brief development time, low cost, and smaller sample sizes than those in cohort studies for equally rare conditions. However, this design is highly susceptible to selection, information and confounding biases, and extensive measures should be taken to control them. It is the clinical expert's role to guide the statistical analysis by providing an informed identification of potential confounders, because may be lacked validity without incorporation of prior knowledge. Lastly, it is also expert's responsibility to effectively interpret any measures of association, by undertaking careful analysis of results clinical significance beyond any indicator of statistical significance.

**Keywords:** Case-Control studies; Odds Ratio; Sample size; Bias.

Recibido: 11 de septiembre de 2019

Aceptado: 12 de septiembre de 2019

Correspondencia: mapalaciosar@unisanitas.edu.co

## INTRODUCCIÓN

El diseño epidemiológico de casos y controles nace formalmente en la década de los cincuenta gracias a los trabajos realizados por Cornfield (1) y Mantel y Haenszel (2), quienes establecen las bases metodológicas y estadísticas para su aplicación y análisis. Otra figura reconocida es Miettinen (3), quien establece la concepción moderna de este tipo de estudio en la década de los setenta y demuestra la estrecha relación que este tipo de diseño tiene con los estudios de cohorte. Sin embargo, su aplicación empírica data del siglo XIX cuando John Snow (4), mediante una investigación realizada para determinar las causas de una epidemia de cólera, comparó personas con y sin la enfermedad respecto al área geográfica donde residían y la fuente de agua potable que utilizaban para abastecerse.

## DEFINICIÓN

El diseño de casos y controles permite realizar la comparación de dos grupos de personas en un periodo de observación definido: el primero afectado por una condición o enfermedad particular, los cuales serán conocidos como casos, y el segundo, formado por personas que no padecían la condición o enfermedad, pero estuvieron expuestos al factor que posiblemente esté asociado al desarrollo de la enfermedad, quienes se llaman controles o testigos (5,6).

Aunque este diseño permite explorar varias exposiciones asociadas a la condición de interés, se debe intentar especificar la hipótesis del estudio en el mayor grado posible. Si esto no se ha hecho, y solo se cuenta con una base de datos, se puede incurrir en una “expedición de pesca” en la que se exploran todas las asociaciones posibles buscando significancia estadística. Esta práctica ha producido múltiples falsos positivos en hallazgos de investigación y ha lastimado la credibilidad de disciplinas enteras (7).

En esencia, este diseño cuenta de tres características inamovibles: se trata de un diseño observacional dado que no se realiza ninguna intervención en la población de estudio, analítico porque busca establecer asociaciones

con un fondo causal entre variables y/o características (factores de riesgo) y la enfermedad, y retrospectivo ya que parte de población que tiene una condición o enfermedad específica y pretende estudiar las posibles exposiciones que tuvo antes desarrollarla. La unidad de análisis es el individuo y las fuentes de información pueden ser primarias (si se realiza una entrevista con los sujetos de investigación) o secundarias (si, por el contrario, se revisan historias clínicas).

## PROBLEMAS DE INVESTIGACIÓN COMPATIBLES CON ESTE DISEÑO

Este tipo de diseño se emplea en la identificación de factores de riesgo que tienen un sustento causal y se asocian con las condiciones o enfermedades que ya se diagnosticaron. Este tipo de diseño también se usa en la solución de problemas de salud poco frecuentes o en enfermedades con largos periodos de latencia (diagnóstico, pronóstico y tratamiento), en la aparición de brotes y en programas de tamizaje (6).

El estudio de enfermedades raras (baja prevalencia) se ajusta de forma óptima al diseño de casos y controles. Por un lado, es impráctico seguir una cohorte durante un largo tiempo para determinar si una exposición determinada tiene un rol en el desarrollo de la condición (infrecuente) de interés. Es posible que no se identifiquen casos en la cohorte, o que la latencia de la exposición sea tal que el estudio deba seguir a los sujetos por décadas. Frente a los ensayos clínicos, los estudios de casos y controles permiten además el estudio de exposiciones cuya asignación no sería ética (e.g. asbesto). Por último, este diseño permite estudiar múltiples exposiciones asociadas a la condición de interés, dando un panorama más completo de las posibles causas de esta.

## VENTAJAS Y DESVENTAJAS DEL DISEÑO

En comparación con otros diseños observacionales (especialmente con el de cohorte), los estudios de casos y controles tienen las siguientes ventajas (5):

1. Suelen exigir menor tiempo de realización (no es necesario esperar mucho tiempo para conocer si hay o no asociación).
2. Son menos costosos.
3. Caracterizan simultáneamente los efectos de más de un posible factor de riesgo del problema de salud que se estudia.
4. Requieren de un menor tamaño de muestra.

Sin embargo, también presentan desventajas (5,6):

1. No se puede medir directamente la incidencia o prevalencia de la enfermedad.
2. Si el problema de salud tiene una prevalencia mayor al 5%, su principal medida de asociación, el "Odds Ratio" (OR), no ofrece una estimación confiable del riesgo relativo.
3. No sirven para determinar otros posibles efectos de una exposición, porque se ocupan solamente de una enfermedad.
4. Son especialmente susceptibles a sesgos:
  - a. De selección ya que la población a riesgo a menudo no está definida y los casos seleccionados por el investigador se obtienen partir de una reserva disponible.
  - b. De información: Resulta cuando la información con respecto a la exposición se obtiene de manera diferente en los enfermos que en los sanos. Esto se puede presentar dado que la exposición se mide, se reconstruye o se cuantifica, después de la enfermedad.
  - c. De confusión: Se refiere al efecto de una variable extraña que explica toda o parcialmente el resultado aparente de una exposición en estudio o que enmascara una asociación verdadera. Tiene tres requisitos:
    - Es un factor de riesgo para la enfermedad en estudio
    - Está asociada con la exposición en estudio, pero no es consecuencia de esta (asociación estadística pero no causal)

- No es parte de la historia natural de la enfermedad o de la cadena causal de esta.

## CÁLCULO DEL TAMAÑO DE LA MUESTRA

Para determinar el tamaño muestral requerido es necesario definir un nivel de confianza ( $\alpha$ ) y un poder estadístico ( $1-\beta$ ), además de contar con al menos una aproximación de la prevalencia de exposición en cada grupo (8). De no tenerse información sobre la prevalencia de exposición en alguno de los grupos, se puede recurrir a un estimado de la magnitud que tendría el OR en el estudio puesto que los tres valores se encuentran relacionados así:

$$OR = \frac{P_1(1-P_2)}{P_2(1-P_1)}$$

Una vez se obtienen estos valores, para grupos independientes se tiene que el tamaño del grupo 1,  $n_1$ , es:

$$n_1 = \frac{\left( z_{1-\alpha/2} \sqrt{(1+\Phi)\bar{P}(1-\bar{P})} + z_{1-\beta} \sqrt{\Phi P_1(1-P_1) + P_2(1-P_2)} \right)^2}{\Phi(P_1 - P_2)^2}; n_2 = \Phi n_1$$

En donde  $\phi$  es el número de controles por cada caso, y el valor  $\bar{P}$  es:

$$\bar{P} = \frac{P_1 + \Phi P_2}{1 + \Phi}$$

Otras fórmulas similares existen para casos en los que se tienen muestras emparejadas (9).

## ANÁLISIS DE LOS RESULTADOS

"Odds ratio", Razón de disparidad o Razón de momios

En el diseño de casos y controles no se pueden estimar de manera directa las medidas de incidencia o prevalencia que tradicionalmente se obtienen de otros diseños observacionales (cohorte, transversales). Por esta razón en este tipo de estudios solo es posible estimar pseudotasas conocidas como "Odds", Momios o disparidad, que solo indican una frecuencia relativa de la exposición entre los casos y controles (8).

El “Odds” de exposición de los casos se calcula dividiendo los casos expuestos sobre los casos no expuestos. Así mismo, el “Odds” de exposición de los controles se obtiene dividiendo los controles expuestos sobre los controles no expuestos. La división de el “Odds” de casos entre el “Odds” de controles se conoce como “Odds ratio” (OR), razón de momios o razón de disparidad y es la principal medida de asociación en este diseño y señala la probabilidad de que una persona enferma (caso) haya estado expuesta al factor de estudio, con respecto a la probabilidad de que una persona sana (control) lo haya estado (Tabla 1) (5,8).

TABLA 1. TABLA DE CONTINGENCIA PARA EL ESTUDIO DE CASOS Y CONTROLES.			
	CASOS	CONTROLES	
Expuestos	A	b	a+b
No Expuestos	C	d	c+d
	a+c	b+d	a+b+c+d

$$\text{Odds de exposición}_{\text{casos}} = \frac{a}{c}$$

$$\text{Odds de exposición}_{\text{controles}} = \frac{b}{d}$$

$$\text{Odds Ratio} = \frac{a \times d}{b \times c}$$

Cuando del OR se obtiene una cifra mayor de uno, esto indica que el factor en estudio se está comportando como un factor de riesgo, en cambio cuando el resultado es menor a la unidad, se traduce en que el factor de estudio se comporta como un factor protector. Cuando el resultado sea igual a uno, la interpretación será que no existe asociación entre la característica o variable estudiada y el desarrollo de la condición o enfermedad (6,8).

### CONTROL DE LA CONFUSIÓN

Una forma común de atacar los resultados de un estudio de casos y controles es la de identificar variables de confusión que no fueron controladas.

A la vez, se ha hecho énfasis en que los estudios observacionales no pueden aproximarse a la causalidad, y esta es la raíz de una contradicción importante. Al hablar de variable de confusión necesariamente se

habla de causalidad, de modo que las conclusiones de un estudio observacional pueden atacarse como si se aproximara a la causalidad, pero no defenderse de la misma manera.

Es falso que los estudios observacionales no se puedan aproximar a la causalidad, particularmente cuando una gran parte de la literatura alrededor de factores de riesgo, exposiciones ambientales, y riesgos profesionales se ha derivado de investigación observacional.

Es provechoso realizar una evaluación de las causalidades involucradas en un problema dado. Este ejercicio permite plasmar de forma sucinta la visión que tiene el investigador del mundo del problema, permitiendo además identificar conjuntos mínimos de ajuste.

### CONSTRUCCIÓN DEL MODELO - ¿QUÉ VARIABLES DEBEN APARECER EN LA TABLA?

En la construcción de estos modelos, hay varios algoritmos automáticos que permiten identificar qué variables deben ser incluidas. Existe, por ejemplo, el criterio de Hosmer-Lemeshow, según el cual se comienza por introducir al modelo multivariado todas aquellas variables que en el análisis bivariado arrojaran valores p menores a 0,25, buscando después el modelo más parsimonioso (10). Existen también métodos automáticos, en los que se comienza de un modelo nulo y se introducen variables hasta que la inclusión de variables adicionales no incremente la varianza explicada (11). Diversos criterios de información existen para ejecutar análisis similares. Sin embargo, estos métodos automáticos pueden producir errores importantes y su uso no es recomendado en la investigación. Por ejemplo, dependiendo de la estructura causal del problema, la introducción de variables de forma automática puede llevar a la introducción de un sesgo M, o a la interpretación errónea de uno de los coeficientes (falacia de la tabla) (12,13). Esto se debe a que la introducción de más variables al modelo no “controla” la confusión de forma automática, y a que incluso si la medida de asociación entre dos variables tuviese un conjunto de ajuste correcto, esto no garantiza que las asociaciones entre la variable independiente y las otras variables en el modelo estén libres de confusión.

Es debido a los problemas descritos que la construcción del modelo no puede dejarse exclusivamente en manos de quien analiza estadísticamente los datos. El experto clínico es fundamental en la identificación de variables de confusión, y el conocimiento de cómo están relacionadas las variables estudiadas de acuerdo con investigación previa es frecuentemente parte de su arsenal diagnóstico y terapéutico. Este conocimiento se puede plasmar y compartir de forma directa y sencilla mediante gráficos acíclicos dirigidos. Esta metodología ha sido descrita de forma extensa por Hernán et al., y tiene un soporte matemático riguroso dado por Pearl (14,15). Se recomienda al lector referirse a la revisión presentada por Sauer y VanderWeele (16).

## INTERPRETACIÓN DE LOS RESULTADOS DEL MODELO

Los modelos logísticos son comúnmente utilizados en este tipo de estudios y los programas estadísticos arrojan una tabla con las variables incluidas en el modelo y el intercepto, su error estándar, y el resultado de una prueba de hipótesis (Tabla 2). Los coeficientes de la regresión logística se interpretan una vez se eleva a la base de los logaritmos naturales por el coeficiente, produciendo un Odds Ratio (OR).

TABLA 2. PRESENTACIÓN DE RESULTADOS DE MODELOS LOGÍSTICOS

VARIABLE	COEFICIENTE	ERROR ESTÁNDAR	P-VALOR
Intercepto	$\beta_0$	$SE_0$	$P_1$
Var 1	$\beta_1$	$SE_1$	$P_2$
Var 2	$\beta_3$	$SE_2$	$P_3$
Var 3	$\beta_2$	$SE_3$	$P_4$

$$OR_1 = e^{\beta_1}$$

$$OR_2 = e^{\beta_2}$$

$$OR_3 = e^{\beta_3}$$

Los intervalos de confianza se calculan teniendo en cuenta que la incertidumbre alrededor del estimador existe en el espacio logarítmico. Es decir, que primero se calculan los intervalos de confianza alrededor de cada  $\beta$  y después se transforma de la misma manera en que

se hizo con las medidas centrales. En el caso de  $\beta_1$  se puede escribir como:

$$IC95\% = \left[ e^{\beta_1 - z_{1-\alpha} \times SE_1}, e^{\beta_1 + z_{1-\alpha/2} \times SE_1} \right]$$

En donde  $\alpha$  es el nivel de confianza (usualmente 0,05, pero como está dividido en dos se convierte en cada extremo en 0,025), SE es el error estándar.

Una vez se tiene esta medida, se puede proceder a la interpretación, que procederá de acuerdo con la naturaleza de la variable. Por ejemplo, si la variable A en el modelo de arriba estuviese medida como dicotómica, la interpretación de un OR de 2 sería que, siendo todas las otras variables iguales entre dos poblaciones, aquella que esté expuesta tendrá unos "Odds" del doble de una población no expuesta. Si esta interpretación le parece complicada, otra forma de entender la interpretación de un OR es la siguiente: Supongamos que el sujeto A tiene 50.000 pesos en la billetera, y el sujeto B tiene 150.000 pesos en la billetera. Se puede decir que el sujeto B tiene 3 veces el dinero del sujeto A en su billetera (3 X 50.000 = 150.000), pero también se puede decir que el sujeto B tiene 2 veces más dinero que el sujeto A (2 X 50.000 + 50.000 = 150.000).

Los valores  $p$  reportados por los paquetes estadísticos son frecuentemente malinterpretados, y su análisis va más allá de determinar si son mayores o menores a 0,05. Para la interpretación se encuentran los siguientes problemas:

1. Se debe saber a qué prueba estadística corresponden los valores, y sus hipótesis nula y alterna.
2. Los umbrales pueden ser muy distintos de 0,05, particularmente en casos de múltiples comparaciones.
3. Si la medida de asociación se asume constante a través de distintos estudios, un mayor tamaño muestral reduce el valor  $p$ .
4. Su valor no dicta la significancia clínica de la asociación encontrada.

Imaginemos que se conduce un estudio con 50 pacientes, encontrando un OR de 3, con un valor  $p$  asociado

a la prueba de Chi-cuadrado de 0,04. Asumiendo que el OR de 3 corresponda a la realidad, si en un estudio subsecuente se muestrean 1000 pacientes, el valor  $p$  será mucho más pequeño, a pesar de que el OR no sea distinto. De forma similar, el intervalo de confianza se hará cada vez más angosto. Es aquí donde el juicio del experto clínico es más valioso, dando interpretación al tamaño del efecto observado, considerando la plausibilidad del mecanismo, su discordancia con el conocimiento clínico establecido, etc.

## CONFLICTO DE INTERESES

El autor no declara tener algún conflicto de interés con relación al tema desarrollado en este artículo o con la revista médica sanitas.

## FINANCIACIÓN

No se recurrió a ninguna convocatoria o fuente de financiación externa para la realización de este trabajo.

## REFERENCIAS

1. Cornfield J. A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix. *J Natl Cancer Inst.* 1 de junio de 1951;11(6):1269-75.
2. Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *J Natl Cancer Inst.* 1 de abril de 1959;22(4):719-48.
3. Miettinen O. Estimability and estimation in case-referent studies. *Am J Epidemiol.* febrero de 1976;103(2):226-35. <https://doi.org/10.1093/oxfordjournals.aje.a112220>
4. Snow J. On the mode of communication of cholera. [Internet]. 1ra ed. London, United Kingdom: John Churchill; 1855. Disponible en: <https://collections.nlm.nih.gov/ext/cholera/PDF/0050707.pdf>
5. Lazcano-Ponce E, Salazar-Martínez E, Hernández-Avila M. Estudios epidemiológicos de casos y controles. Fundamento teórico, variantes y aplicaciones. *Salud pública Méx.* 2001;43:135-50. <https://doi.org/10.1590/S0036-36342001000200009>
6. Gómez MG-, Danglot-Banck C, Alvarado SGH, Torre GG de la. El estudio de casos y controles: su diseño, análisis e interpretación, en investigación clínica. *Rev Mex Pediatr.* 2003;70(5):257-63.
7. Hengartner MP. Raising Awareness for the Replication Crisis in Clinical Psychology by Focusing on Inconsistencies in Psychotherapy Research: How Much Can We Rely on Published Findings from Efficacy Trials? *Front Psychol.* 2018;9:256. <https://doi.org/10.3389/fpsyg.2018.00256>
8. Thompson WD. Statistical analysis of case-control studies [Internet]. 1996 [citado 11 de septiembre de 2019]; Disponible en: <http://iris.paho.org/xmlui/handle/123456789/15455>
9. Elashoff JD, Lemeshow S. Sample Size Determination in Epidemiologic Studies. En: Ahrens W, Pigeot I, editores. *Handbook of Epidemiology.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 559-94. [https://doi.org/10.1007/3-540-26577-5\\_15](https://doi.org/10.1007/3-540-26577-5_15)
10. Hosmer DW, Lemeshow S, Sturdivant RX. Model-Building Strategies and Methods for Logistic Regression. En: *Applied Logistic Regression.* John Wiley & Sons, Ltd; 2013. p. 89-151. <https://doi.org/10.1002/9781118548387.ch4>
11. Mundry R, Nunn CL. Stepwise model fitting and statistical inference: turning noise into signal pollution. *Am Nat.* enero de 2009;173(1):119-23. <https://doi.org/10.1086/593303>
12. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology.* 2003;14(3):300-6. <https://doi.org/10.1097/01.EDE.0000042804.12056.6C>
13. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol.* 2013;177(4):292-8. <https://doi.org/10.1093/aje/kws412>
14. Hernán MA, Robins J. Causal inference. Boca Raton: Chapman & Hall / CRC, forthcoming; 2019.
15. Pearl J. Causal diagrams for empirical research. *Biometrika.* 1995;82(4):669-88. <https://doi.org/10.1093/biomet/82.4.669>
16. Sauer B, VanderWeele TJ. Use of Directed Acyclic Graphs [Internet]. Agency for Healthcare Research and Quality (US); 2013 [citado 19 de julio de 2019]. Disponible en: <https://www.ncbi.nlm.nih.gov/books/NBK126189/>