ARS MEDICA
revista de ciencias médicas

Informe Especial
https://doi.org/10.11565/arsmed.v50i2.2090

Cálculo del riesgo relativo utilizando regresión logística binaria mediante el método de duplicación de casos

Calculation of relative risk using ordinary logistic regression with the doubling-of-cases method

Jaime Cerda-Lorca[1,*], Luis Villarroel-del Pino[1]

Resumen

La investigación clínica y epidemiológica examina asociaciones entre diversas exposiciones y eventos de interés. En estudios con diseño prospectivo, el riesgo relativo (RR) es una medida comúnmente utilizada para expresar los resultados. No obstante, para controlar o ajustar variables confundentes, es necesario recurrir a modelos de regresión, siendo el modelo de regresión logística binaria el más empleado. Sin embargo, este modelo calcula *odds ratios* (OR), que bajo ciertas circunstancias pueden sobreestimar el RR, lo que puede llevar a una valoración inadecuada del tamaño del efecto de la exposición. Este artículo, de carácter docente, describe aspectos epidemiológicos y matemáticos del método de duplicación de casos *(doubling-of-cases method*), una alternativa simple y novedosa para estimar el RR utilizando un modelo de regresión logística binaria, evitando la sobreestimación del tamaño del efecto de la exposición.

Palabras clave: riesgo relativo; regresión logística binaria; método de duplicación de casos; sesgo de confusión; epidemiología; bioestadística.

Abstract

Clinical and epidemiological research examines associations between various exposures and outcomes. In prospective studies, relative risk (RR) is a commonly used effect measure to express the results. However, to control for or adjust confounding variables, regression models are necessary, with the binary logistic regression model being the most commonly used. Nevertheless, this model calculates odds ratios (OR), which under certain circumstances may overestimate the RR, potentially leading to an inaccurate assessment of the exposure effect size. This educational article describes the epidemiological and mathematical aspects of the doubling-of-cases method, a simple and innovative alternative for estimating RR using a binary logistic regression model, thereby avoiding the overestimation of the exposure effect size.

Keywords: relative risk; binary logistic regression; doubling-of-cases method; confounding bias; epidemiology; biostatistics.

Fecha de envío: 2024-08-16 - Fecha de aprobación: 2025-04-16

Introducción

La investigación clínica y epidemiológica explora la existencia de asociaciones entre diversas exposiciones y eventos de interés. Cuando el diseño de investigación es prospectivo, es decir, el punto de partida del estudio es la exposición, observándose en sujetos expuestos y no-expuestos hacia el futuro la ocurrencia de un evento de interés (Noordzij *et al*., 2009), una medida de efecto comúnmente utilizada para expresar los resultados es el riesgo relativo (RR), debido a su simpleza de cálculo e interpretación: se trata de un cociente entre la incidencia o riesgo absoluto del evento de interés entre los grupos expuesto y no-expuesto, o viceversa. Un problema frecuente consiste en que ante la necesidad de controlar o ajustar variables confundentes, se hace necesario el uso de modelos de regresión, siendo el más utilizado el modelo de regresión logística binaria, por su simpleza y amplia disponibilidad.

(1) Escuela de Salud Pública. Facultad de Medicina. Pontificia Universidad Católica de Chile. Santiago. Chile
** Autor de correspondencia: jcerdal@uc.cl*

ISSN: 0719-1855. Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile.

Sin embargo, este modelo tiene el inconveniente de calcular *odds ratios* (OR), los cuales bajo ciertas circunstancias sobreestiman el valor del RR, conduciendo a una inadecuada valoración del tamaño del efecto de la exposición (Tripepi *et al.,* 2008). El presente artículo, de carácter docente, describe aspectos epidemiológicos y matemáticos del método de duplicación de casos *(doubling-of-cases method)*, una alternativa simple y novedosa para estimar el valor del RR utilizando un modelo de regresión logística binaria, evitando la sobreestimación del tamaño del efecto de la exposición.

Un problema de tipo práctico y matemático

Cuando un diseño de investigación es de tipo observacional, en contraposición a los diseños de tipo experimental, es prácticamente de regla la existencia del sesgo de confusión, producido por una acción de una o más variables confundentes, definidas como variables asociadas tanto a la exposición como al evento de interés, sin ser parte de un mecanismo causal entre ambas. El fenómeno de confusión conduce al planteamiento de asociaciones espurias, siendo necesario su control o mitigación a través de diversas estrategias, tanto metodológicas como estadísticas: las primeras corresponden a restricción, pareamiento y aleatorización, mientras que las segundas corresponden a análisis estratificado (también llamado procedimiento de Mantel-Haenszel) y el uso de modelos de regresión, los cuales entregan medidas de efecto ajustadas por una o más variables confundentes, siendo esta última la estrategia más utilizada (van Stralen *et al*., 2010).

En estudios clínicos y epidemiológicos prospectivos (*i.e.* estudios de cohorte concurrente y no-concurrente, ensayos clínicos aleatorizados, ensayos de campo), los investigadores acostumbran a calcular en primer lugar RR crudos o no-ajustados, para luego proceder al cálculo de RR ajustados por variables confundentes a partir de modelos de regresión tales como el modelo log-binomial, modelo Poisson y modelo Poisson con errores estándar robustos, entre otros (Knol *et al*., 2012). Sin embargo, el uso de estos modelos es relativamente infrecuente, por razones de orden práctico y matemático: muchos investigadores no poseen los conocimientos para construir e interpretar los resultados que entregan estos modelos, los programas estadísticos de uso más frecuente a menudo no incluyen la opción de construir estos modelos y matemáticamente la construcción de estos modelos no siempre es posible: el modelo log-binomial puede no lograr convergencia (definida como el proceso mediante el cual el algoritmo de estimación alcanza una solución estable para los parámetros del modelo) y el modelo Poisson puede presentar sobredispersión (situación donde la variabilidad de los datos es mayor que la que el modelo Poisson predice), requiriendo alternativas de modelaje de mayor complejidad.

Ante esta realidad, para calcular RR ajustados en estudios prospectivos los investigadores suelen optar por el modelo de regresión logística binaria, de propiedades muy ventajosas: se enseña de rutina en cursos de estadística básica, se encuentra ampliamente disponible en programas estadísticos de uso frecuente y generalmente no ofrece restricciones de tipo matemático. Sin embargo, su uso conlleva una importante limitante, cual es el calcular OR en lugar de RR. Los OR son una medida de efecto más difícil de interpretar por parte de investigadores y usuarios y suelen sobreestimar el valor del RR cuando el número total de casos en un estudio supera el 10% de la muestra total (Cerda *et al*., 2013). La sobreestimación del RR puede tener consecuencias importantes, como es el conducir a investigadores y usuarios a tomar decisiones erróneas y sus consiguientes consecuencias.

Surge entonces la siguiente pregunta: ¿existe alguna manera de estimar RR ajustados, sin necesidad de renunciar al uso del modelo de regresión logística binaria y sus ventajas? La respuesta sorprendentemente es afirmativa y será detallada en los párrafos siguientes.

Una alternativa simple y novedosa: el método de duplicación de casos

La estimación de RR utilizando un modelo de regresión logística binaria se conoce como método de duplicación de casos *(doubling-of-cases method)*. Este método, simple y novedoso, fue propuesto originalmente por Miettinen (1982) y validado por Díaz-Quijano (2012) y Ning *et al.,* (2022).

En un estudio prospectivo los datos se ordenan en una tabla de contingencia (Tabla 1), a partir de la cual el RR se calcula según la fórmula y el OR según la fórmula . El método de duplicación de casos consiste en la construcción de un conjunto de datos expandidos, en que cada caso del conjunto original es duplicado y re-clasificado como un no-caso, conservando la información de sus covariables, de modo que los no-casos expuestos pasan de ser a ser y los no-casos no-expuestos pasan a ser a ser . En el conjunto de datos expandidos, el OR' se calcula según la fórmula , siendo su resultado idéntico al entregado por la fórmula del RR calculado en base al conjunto de datos originales (OR' = RR). El método de duplicación de casos, por consiguiente, aporta la siguiente novedad: en un estudio prospectivo, el RR calculado en base al conjunto de datos originales a partir de un modelo log-binomial (o alguna de sus alternativas) tiene un valor idéntico al OR' calculado en base al conjunto de datos expandidos a partir de un modelo de regresión logística binaria.

ISSN: 0719-1855. Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile.

Tabla 1: comparación entre el conjunto de datos originales y el conjunto de datos expandidos, obtenido mediante el método de duplicación de casos.

Conjunto de datos originales			
	Casos	**No-casos**	**Total**
Expuestos	a	b	a + b
No-expuestos	c	d	c + d
Conjunto de datos expandidos: método de duplicación de casos			
	Casos	**No-casos**	**Total**
Expuestos	a	a' + b	a + a' + b
No-expuestos	c	c' + d	c + c' + d

El riesgo relativo calculado en base al conjunto de datos original equivale a *RR = [(a/a+b) / (c/c+d)]*, mientras que el *odds ratio* calculado en base al conjunto de datos expandidos, obtenido mediante el método de duplicación de casos, equivale a *OR' = [(a/a'+b) / (c/c'+d)]* , siendo su resultado idéntico (RR = OR').

Una importante limitación del método de duplicación de casos

El valor del OR' calculado mediante el método de duplicación de casos es idéntico al valor del RR calculado en base al conjunto de datos originales a partir de un modelo log-binomial (o alguna de sus alternativas), sin embargo, su intervalo de confianza al 95% es diferente. Matemáticamente esto se explica por el hecho de que el conjunto de datos originales y expandidos poseen un tamaño diferente y por el hecho de que la fórmula de cálculo del RR y OR es diferente, específicamente en lo concerniente a su error estándar, valor crítico para la construcción de intervalos de confianza. El método de duplicación de casos aumenta el error estándar, por consiguiente, los intervalos de confianza al 95% son más amplios para el OR', en comparación al RR calculado a partir del modelo log-binomial (o algunas de sus alternativas). Este hecho tiene importantes consecuencias prácticas, pues la toma de decisiones no solamente considera la magnitud de la medida de efecto, sino también su precisión (es decir, la amplitud del intervalo de confianza). Asimismo, la valoración que se hace de la certeza de la evidencia, utilizando metodología GRADE, toma en consideración la imprecisión de la medida de efecto, dimensionada a través de su respectivo intervalo de confianza (Kirmayr *et al*., 2021; Quilodrán *et al*., 2021).

Lamentablemente el método de duplicación de casos no contempla una solución simple y novedosa para corregir el tamaño del intervalo de confianza del OR'. Al respecto, Ning y colaboradores (2022) sugieren el cálculo de intervalos de confianza utilizando errores estándar robustos *(robust sandwich-type standard errors)*, solución matemáticamente correcta, sin embargo, introduce una complejidad insoslayable al método de duplicación de casos, ya que esta opción no siempre está disponible en los programas estadísticos de uso frecuente y le resta simplicidad. Por su parte, Díaz-Quijano (2012) advierte que no se puede establecer un factor de corrección simple y práctico para este problema debido a que, en una regresión multivariable, el error estándar de cada predictor depende de su correlación con todas las demás variables incluidas en el modelo, de modo que investigadores y usuarios deben estar conscientes que el error de tipo II podría ser mayor. Asimismo, señala que esta estrategia debería emplearse únicamente cuando la regresión logística binaria es el único método disponible. Desde un punto de vista epidemiológico, es posible afirmar que el método de duplicación de casos ofrece una alternativa más conservadora que el cálculo de RR a partir de modelos log-binomial (o alguna de sus alternativas), situación que resulta menos problemática en comparación a su contrafactual: si el método de duplicación de casos entregase intervalos de confianza más estrechos (es decir, más precisos), se podría incurrir en error de tipo I, considerado de mayor gravedad por parte de la comunidad científica que el error de tipo II.

Ejemplo

Un estudio de cohorte concurrente evaluó la asociación entre la exposición a una sustancia tóxica (expuestos versus no-expuestos) y la ocurrencia de una determinada enfermedad (presente versus ausente). Sobre un total de 140 sujetos expuestos a la sustancia tóxica, 78 presentaron la enfermedad (riesgo absoluto = 55,7%), mientras que sobre un total de 140 sujetos no-expuestos a la sustancia tóxica, 45 presentaron la enfermedad (riesgo absoluto = 32,1%). El RR de enfermedad entre expuestos versus no-expuestos equivale a *RR* = (78÷140) ÷ (45÷140) = 1,733; por su parte, el OR de enfermedad entre expuestos versus no-expuestos equivale a *OR* = (78÷62) ÷ (45÷95) = 2,656. Se puede apreciar que el OR sobreestima el valor del RR ya que el total de casos en la muestra (n=123) supera el 10% de la muestra total (n=280).

En este estudio, la variable sexo corresponde a una variable confundente, siendo ajustada mediante un modelo de regresión logística binaria (Tabla 2), el cual entregó un *ORa* = 3,642 (IC95% 2,059-6,445). El mismo procedimiento se repitió utilizando un modelo

ISSN: 0719-1855. Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile.

log-binomial, el cual entregó un *RRa* = 2,000 (IC95% 1,498-2,671) (Tabla 3). La aplicación del método de duplicación de casos condujo a la creación de un conjunto de datos expandidos, utilizándose un modelo de regresión logística binaria para estimar el RR ajustado, el cual entregó un *ORa'* = 2,000 (IC95% 1,235-3,239). Nótese que este valor es idéntico al RR ajustado entregado por modelo log-binomial, sin embargo, su intervalo de confianza es más amplio (es decir, menos preciso). Lo mismo ocurre al comparar dicho intervalo de confianza al 95% con el entregado por los modelos Poisson y Poisson con errores estándar robustos (Tabla 3).

Tabla 2: ejemplo numérico de riesgo relativo (RR) y *odds ratio (OR)* calculados en base al conjunto de datos original, y OR' calculado en base al conjunto de datos expandidos (método de duplicación de casos).

Conjunto de datos originales (n=280)			
	Casos	**No-casos**	**Total**
Expuestos	78 (a)	62 (b)	140 (a + b)
No-expuestos	45 (c)	95 (d)	140 (c + d)
RR = [(a/a+b) / (c/c+d)]			
RR = [(78/140) / (45/140]			
RR = 2,0			
OR = [(a/b) / (c/d)]			
OR = [(78/62) / (45/95]			
OR = 2,7			
Conjunto de datos modificados: método de duplicación de casos (n=403)			
	Casos	No-casos	Total
Expuestos	78 (a)	78 (a') + 62 (b)	218 (a + a' + b)
No-expuestos	45 (b)	45 (b') + 95 (d)	185 (b + b' + d)
OR' = [(a/a'+b) / (c/c'+d)]			
OR' = [(78/78+62) / (45/45+95]			
OR' = 2,0			

Tabla 3: comparación del Riesgo Relativo (RR), *odds ratio (OR)* y sus respectivos intervalos de confianza al 95%, calculados en base a diferentes modelos de regresión.

Modelo de regresión	**Medida de efecto ajustada por la variable sexo, e intervalo de confianza al 95%.**
Modelo de regresión logística binaria	OR = 3,642 (2,059–6,445)
Modelo log-binomial	RR = 2,000 (1,498–2,671)
Modelo Poisson	RR = 2,000 (1,339–2,987)
Modelo Poisson con errores estándar robustos	RR = 2,000 (1,488–2,688)
Método de duplicación de casos: regresión logística binaria	OR' = 2,000 (1,235–3,239)

El riesgo relativo calculado en base a diversos modelos de regresión presenta un valor idéntico al *odds ratio* calculado en base al conjunto de datos expandidos, obtenido mediante el método de duplicación de casos. Sin embargo, su intervalo de confianza al 95% es más amplio (es decir, menos preciso).

Por último, cabe destacar que en estudios de tipo transversal (e.g. estudios de prevalencia), la razón de prevalencias y el *odds ratio* de prevalencias se calculan a partir de la misma fórmula que el RR y OR de un estudio prospectivo, por consiguiente, en estudios transversales también es factible aplicar el *doubling-of-cases method.*

Conclusión

En estudios prospectivos, el método de duplicación de casos representa una alternativa simple y novedosa para la estimación de RR a partir de un modelo de regresión logística binaria. El método es especialmente útil en circunstancias en que el OR sobreestima el valor del RR y se requiere el uso de modelo de regresión logística binaria, debido a razones prácticas y matemáticas. Sin embargo, su principal limitante es aumentar la amplitud del intervalo de confianza. En consecuencia, el método de duplicación de casos tiene menos potencia estadística que otros métodos como el modelo log-binomial o el modelo Poisson con errores estándar robustos. Por esta razón es recomendable utilizarlo cuando se tiene tamaños muestrales grandes, de modo de mitigar esta disminución de potencia.

ISSN: 0719-1855. Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile.

Reconocimientos

Contribuciones declaradas por los autores:

Jaime Cerda-Lorca: Conceptualización; Análisis Formal; Investigación; Escritura – Borrador Original; Escritura – Revisión y Edición. **Luis Villarroel-del Pino:** Escritura – Revisión y Edición

Conflicto de interés: los autores declaran no presentar conflictos de interés.

Fuentes de financiamiento: En la redacción del presente artículo los autores no recibieron aporte de fondos de ninguna institución, pública, privada, comercial ni sin fines de lucro.

Referencias

Cerda J, Vera C & Rada G. (2013). Odds ratio: aspectos teóricos y prácticos. *Revista médica de Chile* **141**(10), 1329-1335. https://dx.doi.org/10.4067/S0034-98872013001000014

Díaz-Quijano FA. (2012). A simple method for estimating relative risk using logistic regression. *BMC medical research methodology* **12**, 14. https://doi.org/10.1186/1471-2288-12-14

Kirmayr M, Quilodrán C, Valente B, Loezar C, Garegnani L & Franco J. (2021). The GRADE approach, Part 1: how to assess the certainty of the evidence. Metodología GRADE, parte 1: cómo evaluar la certeza de la evidencia. *Medwave* **21**(2), e8109. https://doi.org/10.5867/medwave.2021.02.8109

Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP & Groenwold RH. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne* **184**(8), 895–899. https://doi.org/10.1503/cmaj.101715

Miettinen O. (1982). Design options in epidemiologic research. An update. *Scandinavian Journal of Work, Environment & Health* **8** Suppl 1:7-14.

Ning Y, Lam A & Reilly M. (2022). Estimating risk ratio from any standard epidemiological design by doubling the cases. *BMC medical research methodology* **22**(1), 157. https://doi.org/10.1186/s12874-022-01636-3

Noordzij M. Dekker FW, Zoccali C & Jager KJ. (2009). Study designs in clinical research. *Nephron. Clinical practice* **113**(3), c218–c221. https://doi.org/10.1159/000235610

Quilodrán C, Kirmayr M, Valente B, Pérez-Bracchiglione J, Garegnani L & Franco J. (2021). The GRADE approach, Part 2: Evidence to decision frameworksoutlining decision-making in health. Metodología GRADE, parte 2: de la evidencia a la decisión esquematizando la toma de decisiones en salud. *Medwave* **21**(4), e8182. https://doi.org/10.5867/medwave.2021.04.8182

Tripepi G, Jager KJ, Dekker FW, & Zoccali C. (2008). Linear and logistic regression analysis. *Kidney international* **73**(7), 806–810. https://doi.org/10.1038/sj.ki.5002787

van Stralen KJ, Dekker FW, Zoccali C, & Jager J. (2010). Confounding. *Nephron. Clinical practice* **116**(2), c143–c147. https://doi.org/10.1159/000315883

ISSN: 0719-1855. Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile.