Colegio Mexicano de Inmunología Clínica y Alergia, A. C. CMICA
Revista Alergia México raM

Statistical methods for effect size analysis

Métodos estadísticos para el análisis del tamaño del efecto

Mario Enrique Rendón-Macías,[1] Irma Susana Zarco-Villavicencio,[2] Miguel Ángel Villasís-Keever[3]

Abstract

Informing in the studies about the effect size of an intervention or the impact of the factor(s) about an outcome, allows better decision-making for the application of the results in clinical practice. This article presents different methods to analyze the effect size, which can be through direct or indirect statistical methods. Within the direct methods, there's the difference in means between groups and the difference of absolute or relative frequencies. Among the indirect methods, Cohen's "d" family (which are based on standard deviation values), the "r and R^2" family, measures of association (e.g. OR, RR, HR), and impact measures (e.g. NNT) are shown. The decision to use any of these methods depends on the objectives of the study and the measuring scale that is used to assess the results, as well as the data distribution. In order to enhance the understanding of the methods described in this article, examples are included, and the need to include level of precision (e.g. confidence intervals) is highlighted, along with the clinical decision thresholds for a better interpretation.

Key words: Effect size; Mean difference, Cohen's "d"; Association measures; Clinical decision

Resumen

Informar en los estudios sobre el tamaño del efecto de una intervención o del impacto de factor(es) sobre un desenlace, permite tomar mejores decisiones para la aplicación de los resultados a la práctica clínica. En este artículo se presenta la manera de analizar el tamaño del efecto, lo cual puede ser mediante métodos estadísticos directos o indirectos. Dentro de los métodos directos, se encuentra la diferencia de promedios entre grupos y la diferencia de frecuencias absolutas o relativas. Dentro de los métodos indirectos se muestran los índices de la familia de "d" de Cohen (que se basan en valores de desviación estándar), la familia de "r y R^2", medidas de asociación (RM, RR, HR) e impacto (NNT). La decisión del uso de cualquiera de los métodos descritos, depende de los objetivos del estudio, la escala de medición usada para evaluar los resultados y la distribución de los datos. Para facilitar la comprensión, se incluyen ejemplos y se resalta la necesidad de incluir los diferentes estadísticos con su nivel de precisión (ej. intervalos de confianza), junto con los umbrales clínicos de decisión, a fin de mejorar su interpretación.

Palabras clave: Tamaño del efecto; Diferencia de promedios; d de Cohen; Medidas de asociación; Decisión clínica

[1]Universidad Panamericana, Facultad de Medicina, Departamento de Salud Pública, Ciudad de México, México
[2]Instituto de Investigación en Psicología Clínica y Social A. C., Ciudad de México, México
[3]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Coordinación de Investigación Médica, Ciudad de México, México

Correspondencia: Miguel Ángel Villasís-Keever.
miguel.villasis@gmail.com

Recibido: 20-junio-2021
Aceptado: 23-junio-2021
DOI: 10.29262/ram.v658i2.949

Introducción

La investigación clínica ha permitido mejorar las condiciones de salud de la humanidad, lo cual se puede observar porque cada vez tenemos más personas longevas o por mejoría en la calidad de vida de pacientes con enfermedad.

Lo anterior se ha logrado por la incorporación en la práctica clínica de hallazgos de la investigación, tanto de medidas preventivas como de intervenciones terapéuticas. Para llevar a cabo la incorporación de los resultados de estudios, se requieren de ciertos factores, tales como el diseño apropiado (por ejemplo, ensayos clínicos aleatorizados para evaluación de intervenciones terapéuticas), los criterios de selección de los participantes, el tamaño de muestra, la consistencia de los resultados en diferentes estudios (es decir, que la mayoría coincidan del posible beneficio), así como el tamaño del efecto.

El tamaño del efecto se refiere a la diferencia de los resultados en la(s) variable(s) de resultado (*outcomes*) que se obtiene(n) entre los grupos que se estudian. De manera tradicional, esta diferencia se evalúa mediante la prueba de significancia estadística de la hipótesis nula (SEHN), lo que en inglés esto se ha denominado *null hypotesis significance testing* (NHST).[1] Todo lo cual, se puede resumir como la interpretación del valor de p, después de aplicar pruebas estadísticas como chi cuadrada, prueba de t Student, análisis de varianza, etcétera.

Sin embargo, SEHN solo mide la probabilidad condicional a la veracidad de la hipótesis nula de los datos observados (o más extremos), resumidos con una prueba estadística. Esto se calcula con muestras hipotéticas, con igual número de individuos repetidos infinitamente y asumiendo tengan distribución gaussiana o normal. Esto mismo, se puede expresar como la evidencia contra la hipótesis nula;[1,2] de tal manera que si la evidencia es tan extrema, o con un nivel crítico (en general, con una $p < 0.05$), se rechaza la hipótesis de nulidad, aceptando la posibilidad de una relación o asociación entre los fenómenos de estudio.[2] Si bien, la inferencia que se adopta (con el valor de p), es la posibilidad de que haya una magnitud diferente entre grupos, esto no es real ya que, al aplicar los resultados de investigaciones en la práctica clínica habitual, probablemente no se observarán los mismos resultados de los estudios. Por lo tanto, tiene implicaciones para tomar decisiones en el ámbito clínico, social y económico.

Para evitar el problema de SEHN, en las últimas décadas ha surgido el interés en cómo demostrar la magnitud de la diferencia. Así surge el *"tamaño del efecto" (effect size)*.[1,2,3] Tal es el interés de esta información que, en la actualidad, diversas revistas científicas en el área de la salud solicitan mostrar el tamaño del efecto, como requisito para publicar los resultados de un estudio científico.[4] De ahí que, el objetivo del presente artículo es informar los índices más usados para determinar el tamaño del efecto. Para facilitar la comprensión, en cada índice expuesto se incluyen ejemplos, además se proporcionan ligas a programas estadísticos en línea de acceso libre (Anexo) para realizar los cálculos estadísticos pertinentes de los resultados mostrados en los ejemplos. Los lectores interesados en las fórmulas específicas, las pueden consultar en referencias bibliográficas.[2,3,5,6,7]

Índices para calcular el tamaño de efecto

Podemos clasificar a los índices de tamaños de efecto en dos grandes grupos: directos e indirectos, según la naturaleza de la medición de la variable de resultado (Figura 1).

I. Tamaño de efecto mediante valores directos

Estos índices comparan resultados entre dos o más grupos, de acuerdo con su escala de medición (cuantitativa o cualitativa) para que su interpretación sea clara y sencilla. El tamaño o magnitud del efecto se comprueba mostrando la lejanía de los promedios entre grupos, ajustado por su varianza, o bien, por la diferencia en las frecuencias absolutas o porcentuales.

Cuando no hay diferencia entre grupos (datos muy similares o se traslapan), la magnitud es igual a "cero" (es decir, el valor nulo). Pero, dado que los datos entre grupos pudieran tener variación —relacionado con el tamaño de muestra—, el tamaño de efecto deberá calcularse junto con su intervalo de confianza a 95 % (IC 95 %) para estimar su precisión.[8,9,10]

Por otro lado, aunque uno pudiera aceptar que un tamaño de efecto es estadísticamente significativo, también se debe tener en cuenta el umbral de decisión clínica.[8,11] Es decir, a fin de estar en posición de aplicar los resultados de un estudio en la práctica clínica, el intervalo de confianza del tamaño del efecto debería estar por arriba de ese umbral clínico establecido.[8]

Por lo anterior, cuando se presenta un tamaño de efecto en medidas con valor directo es necesario establecer los siguientes dos puntos:

- Si el tamaño del efecto mostrará una diferencia de promedios correspondiente a una medida cuantitativa, o bien, a una relativa (proporción o porcentaje)
- Cuál o cuáles son los valores umbrales de significancia clínica.

Esto último es determinado por la utilidad planteada de los resultados. Schultz *et al*.[12] sugieren cuatro posibles juicios de decisión para considerar puntajes umbrales clínicamente significativos:

- Cuando los individuos regresan a su funcionamiento normal, o bien, experimenten cambio significativo en los síntomas.
- Cuando el efecto represente cambios en la calidad de vida de los pacientes.
- Cuando los cambios sean importantes para la sociedad.
- Si el resultado impacta en la vida de los pacientes.

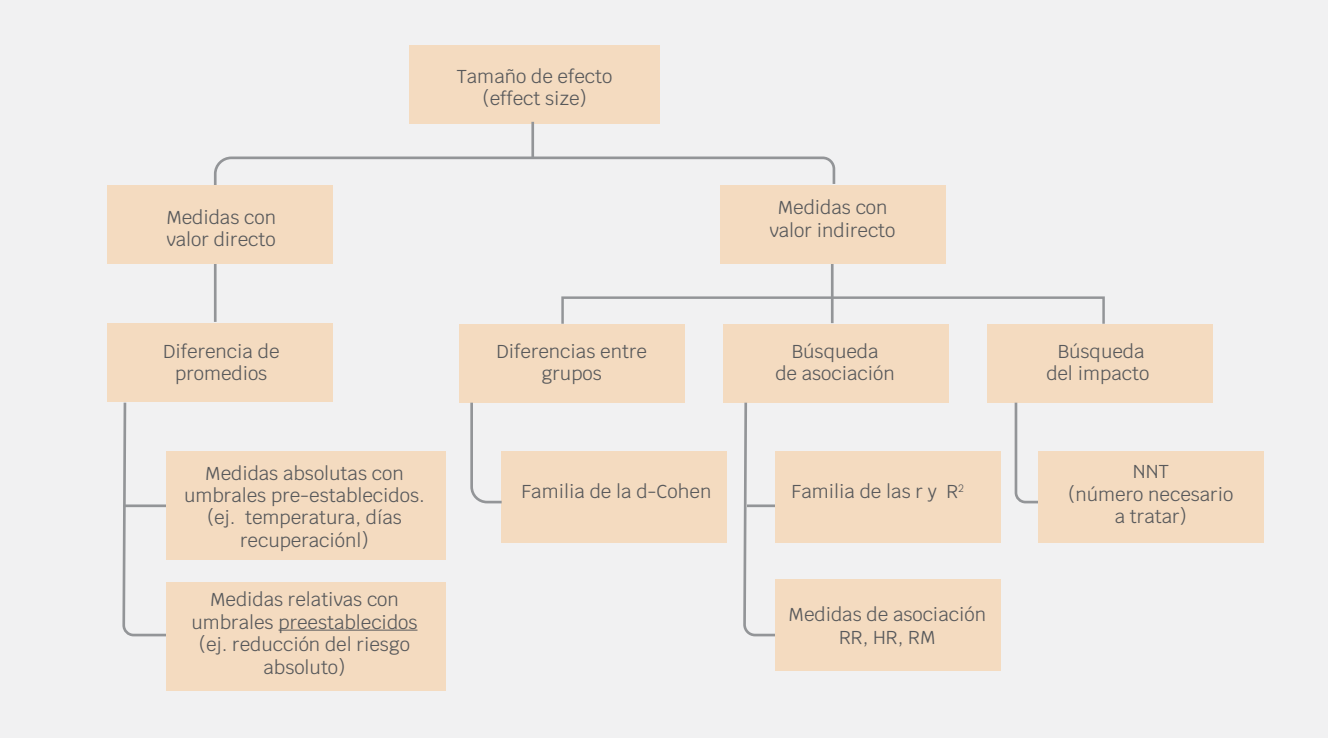

Figura 1. Tipos de estimadores para determinar el tamaño del efecto. RR = riesgo relativo, HR = hazard ratio, RM = razón de momios

Ejemplo e interpretación

A) Tamaño de efecto por diferencia de promedios

En un estudio se deseaba determinar el beneficio de usar probióticos de manera oral en pacientes con urticaria crónica. Se estudiaron a 30 individuos con dieta adicionada con probióticos y 30 con placebo. El efecto se midió con el número de ronchas presentes después de cinco días de tratamiento. El promedio del número de ronchas antes del estudio era 55 ± 4, tanto en el grupo experimental (probióticos) como en el control (placebo). Posterior a los cinco días de tratamiento, el promedio de lesiones en el grupo experimental fue de 15 ± 2 lesiones y 25 ± 3 para el control.

Para interpretar adecuadamente este resultado se necesita (Figura 2):

- Determinar si el objetivo del estudio fue una superioridad (mayor eficacia) del tratamiento experimental sobre el control, de equivalencia (el efecto es similar al grupo control), o bien, de no inferioridad (el efecto no es menor que lo observado en el grupo control).
- Establecer el umbral clínico para determinar la eficacia. Los autores desean que el tratamiento impacte de manera importante a los pacientes; así, determinaron que después del tratamiento, tener 10 lesiones menos que el grupo control sería clínicamente significativo.
- Calcular la diferencia encontrada (tamaño del efecto) y su precisión (IC 95 %). Utilizando un calculador en línea (MedCalc for Windows, versión 19.4, Ostend, Belgium [Anexo]), encontramos que la diferencia entre los promedios de los grupos fue de 10 lesiones (IC 95 % = 8-11).

Para contrastar las ventajas del tamaño de efecto con la prueba de SEHN, usamos t de Student para grupos independientes; el valor obtenido fue de 15.9 (58 gL), por lo que $p < 0.0001$. Con esta información uno puede concluir que la diferencia entre los dos grupos fue estadísticamente significativa. Sin embargo, al analizar el tamaño del efecto y de acuerdo con el IC 95 %, los datos obtenidos del valor inferior (8 lesiones menos), está por debajo del umbral de significancia clínica (10 lesiones), por lo tanto, aunque el tratamiento con probióticos parece ser benéfico, el tamaño del efecto no es contundente.[8] De esta forma, si se decide adoptar este tratamiento, entonces el médico debe anticipar que, aunque habrá mejoría, no todos los pacientes tendrán el mismo efecto terapéutico.

B) Tamaño de efecto por diferencia de proporciones

En este mismo estudio, los autores incrementaron el número de pacientes por grupo para evaluar la seguridad del uso de probióticos. Se analizaron 120 pacientes por grupo para comparar la frecuencia de diarrea (efecto secundario) entre el

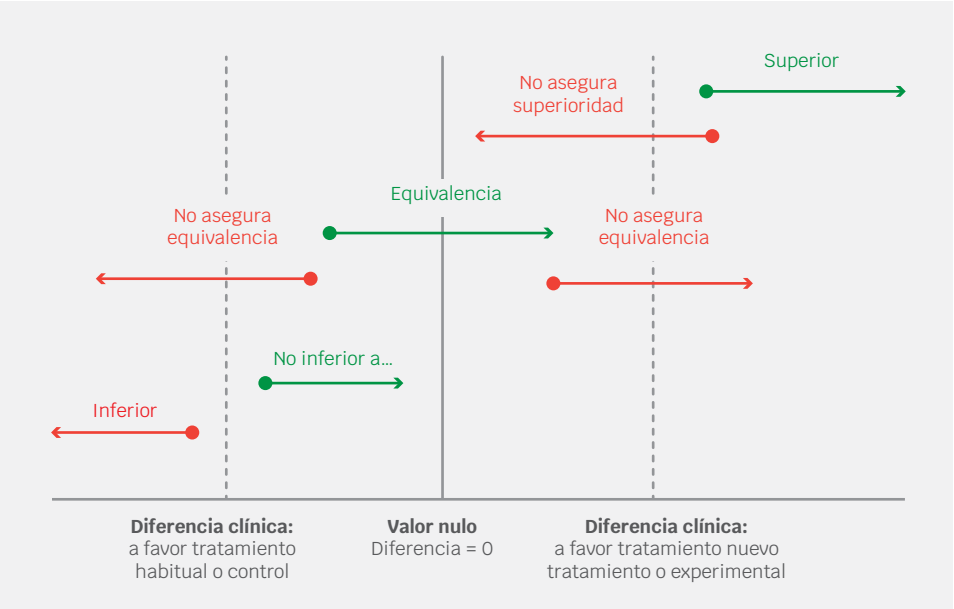

Figura 2. Análisis de resultados, basado en el tamaño del efecto (diferencia de promedios o porcentaje) para estudios de tratamiento según el objetivo: superioridad, equivalencia o no inferioridad.

grupo experimental y el control. Para comparar entre los dos tratamientos, se utilizó como índice a la *diferencia en riesgo absoluto* (RRA).[13,14] Para esta variable, los autores apostaban que no habría diferencia en la incidencia de diarrea entre los dos grupos, lo cual corresponde a una hipótesis de *equivalencia* (Figura 2). Como era poco probable que no hubiera casos de diarrea (diferencia igual a cero), consideraron que habría equivalencia si la diferencia entre las incidencias se mantenía en un intervalo de −5 % a +5 % (umbrales).

En los resultados se observó que en el grupo experimental, 18/120 pacientes (15 %) presentaron diarrea y 12/120 (10 %) del grupo control. De esta forma, la diferencia es de 5 % (20 % menos 15 %), es decir, hubo mayor porcentaje de diarrea en el grupo experimental (IRA o incremento en el riesgo absoluto).[15]

Con la prueba SHEN, mediante chi cuadrada el valor fue de 0.95 (1 gl), p = 0.32, con lo cual se establece que no hay diferencia estadísticamente significativa entre grupos, por lo que se pudiera concluir que los probióticos no incrementan la frecuencia de diarrea. Sin embargo, el IC 95 % de la diferencia (del tamaño del efecto) varía de −3 % (más diarrea en el grupo placebo) a 15 % (más diarrea en el grupo de probióticos). Por lo anterior, no es posible comprobar la hipótesis de equivalencia, porque es más probable la ocurrencia de episodios de diarrea en el grupo experimental (Figura 2).

II. Tamaño de efecto mediante valores indirectos

En cualquier investigación, la manera de evaluar o medir el fenómeno en estudio depende de la naturaleza o forma de medición de la(s) variable(s) de resultado. Cuando se desea evaluar fenómenos complejos (en contraste con datos precisos como la presión arterial, el peso corporal o los resultados de estudios de laboratorio), se pueden utilizar instrumentos que incluyen dominios o múltiples variables, en los cuales incorporan criterios, que en muchas ocasiones son subjetivos.[16] Este tipo de mediciones son comunes en la investigación médica (ej. escalas pronósticas), en psicología (ej. escalas de depresión), o en estudios económicos (ej. escalas de calidad de vida). Estos instrumentos tienen métricas donde los resultados se expresan en la dirección (a favor o en contra) de cómo ocurre un fenómeno, o también por los cambios que se presentan al otorgar una intervención. En estos escenarios, el tamaño de efecto se establece de *manera estandarizada* (ver más adelante), por lo que la evaluación del tamaño del efecto será indirecta.[3,9,16,17,18]

Existen múltiples estimadores indirectos para determinar cuán diferentes son los promedios estandarizados entre dos o más grupos; los más utilizados son los conocidos como familia de *"d"* de Cohen. En el Cuadro 1 se señalan los más usados y los requisitos para su empleo.[3,7,16,19]

Para usar este tipo de estadísticos se requieren dos o más grupos a comparar, que las variables tengan escala de medición cuantitativa, y que los datos tengan distribución normal. Al igual que con los métodos directos, entre más alejados los valores entre los grupos, se puede afirmar que la diferencia es significativa. Para definir qué tan grande es esa distancia, en lugar de usar los promedios de los grupos, se utilizan las desviaciones estándar (de ahí el término, *estandarizada*) para establecer la diferencia entre los grupos, lo que representa qué tan variable son los resultados obtenidos de cada grupo. Así, entre más grande es la deviación estándar, menos significativo son los resultados y viceversa.

Para interpretar el estadístico *"d"* de Cohen, puede ocurrir que sea igual a "0", lo que significa que no hay diferencia

Cuadro 1. Familia de estadísticos d Cohen

Nombre de estadístico	Utilidad
d_{pop} Cohen	Para comparación de dos grupos independientes, cuando la varianza poblacional es conocida y la varianza se calcula con una n de la población
d_s Cohen	Para comparación de dos grupos independientes cuando la varianza poblacional se desconoce y la varianza se calcula con n de la muestra (n-1)
g Hedges	Para comparación de dos grupos independientes pequeños, donde se debe corregir el sesgo asociado a muestras pequeñas (muy usado en metaanálisis)
ΔGlass	Para comparación de dos grupos independientes, cuando la manipulación experimental puede afectar la desviación estándar
d_z Cohen	Para grupos correlacionados o dependientes
η^2 o ωx^2	Cuando son más de dos grupos independientes

entre grupos; mientras que cuando la diferencia entre los promedios es menor a la varianza (o desviación estándar [DE]), el tamaño del efecto se considera como pequeño (Figura 3, gráfico superior). Por el contrario, el efecto será grande si la diferencia de los promedios es mayor a la distancia de su varianza (Figura 3, gráfico inferior). Para ser más práctico y valorar la magnitud (o contundencia) de las diferencias, basado en el valor de *"d"*, se ha propuesto que, cuando se obtiene un valor entre 0 y 0.19, se considera como no efecto o que la diferencia es intrascendente. Mientras que con valores de 0.2 a 0.49, la diferencia es pequeña; de 0.5 a 0.79, como moderada; de 0.8 a 1.29, grande; y ≥ 1.3, muy grande.[2,19,20]

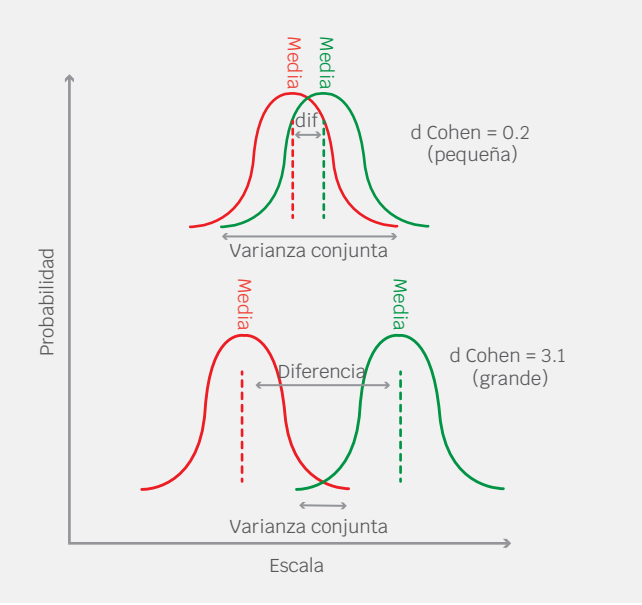

Figura 3. Diferencia de Cohen (d Cohen) que calcula la relación entre la diferencia del promedio de dos grupos con distribución normal (verde y roja), con relación a la varianza conjunta. Valores bajos es un tamaño de efecto pequeño, valores grandes implican un tamaño de efecto grande.

Ejemplo e interpretación

A) Tamaño de efecto indirecto por diferencia en desviación estándar

En un estudio (hipotético) para determinar el grado de alivio a la disnea, se utiliza una escala análoga visual de 10 cm, donde 10 es igual a sensación de ahogo extremo y 0 a una respiración tranquila. Hay dos grupos de 10 pacientes, el primero recibe el medicamento A y el segundo recibe placebo.

Después de brindar el tratamiento, el promedio de disnea en el grupo A fue de 4 ± 2 (DE) y en el placebo de 6 ± 3; estos datos, al ser analizados con la prueba t Student se obtiene un valor 1.75 (gl 18), p = 0.096. Sin embargo, debemos observar que la diferencia puntual entre los dos promedios es igual a 2 (resta de 6 – 4), es decir, el medicamento A redujo dos puntos la disnea, pero este dato no da una idea completa de los resultados. Entonces, al calcular el valor de d Cohen (al usar las ligas del Anexo) da una cifra de 0.79 (diferencia moderada), es decir, los promedios se separan en 79 %. A pesar de esta diferencia, los autores pudieron concluir que no es estadísticamente significativa, y rechazar la recomendación de este medicamento.

Si este mismo estudio se hubiera realizado con 50 pacientes en cada grupo, con los mismos resultados de promedio y DE, el valor de t Student hubiera sido 3.92, (98 gl), p = 0.0002. Entonces, los autores rechazarían la hipótesis nula y recomendarían el medicamento. Sin embargo, con el estadístico d de Cohen, a pesar del incremento de pacientes, el tamaño del efecto se mantiene sin cambios (0.79), ya que no se afecta por el tamaño de muestra.[2,7] Dado que el IC 95 % para el valor d tiene una amplia variación (0.21 a 1.36), se puede determinar que el tamaño de efecto es muy impreciso, ya que va de un efecto pequeño a uno muy grande.

Con lo anterior, podemos aseverar que el índice de Cohen ofrece más información que sólo la comparación de promedios, ya que analiza la variabilidad entre grupos, lo cual sin duda ayudará para identificar, en su caso, la conveniencia de usar una intervención en la práctica clínica habitual.

Otro beneficio que tiene utilizar valores estandarizados, es poder hacer la comparación de los resultados con distintos instrumentos. Para esto, ahora asumamos que, en el mismo estudio, se evaluó el nivel de esfuerzo para respirar de manera subjetiva en una escala de siete puntos (0, sin esfuerzo, a 7, máximo esfuerzo). Los autores deseaban también establecer si el medicamento mejoraba el esfuerzo respiratorio. En los 50 pacientes con el medicamento A, el promedio del esfuerzo fue de 3 ± 1.5, mientras que en el grupo control fue de 5 ± 1. Nuevamente, la diferencia puntual es de 2 unidades menos para el grupo experimental. Cuando se analiza con la prueba de t Student, el valor es de 7.8 (98 gl), p = 0.0001, por lo que los autores pueden rechazar la hipótesis nula y afirmar que el medicamento reduce el esfuerzo para respirar. Mientras que con el cálculo de "d" Cohen, la comparación entre grupos del esfuerzo respiratorio es de 1.56 (IC 95 % = 0.93-2.2), que se interpreta como un tamaño de efecto grande o muy grande.

Con esta última información y, a pesar que las unidades de medición no son similares, al comparar los resultados del tamaño de efecto indirecto de la disnea (IC 95 % = 0.21-1.36), el medicamento parece tener mayor impacto sobre la sensación de la fuerza requerida para respirar que con la sensación de disnea. Esto no pudo haber sido observado si solamente se hubiera tenido el análisis de la comparación de los promedios de grupos, ya que la diferencia fue de dos puntos para ambos.

Otra forma de comprender la diferencia mediante el análisis del efecto indirecto es con el coeficiente Cohen U3, el cual se le conoce como medida de no superposición. Compara el porcentaje de una población que supera la mitad superior de los casos, con un grupo de contraste (comparador). El valor va de 0.5 o 50 % a 1.0 o 100 %. El efecto es importante si es superior a 0.7 o 70 % y es mucho mejor si es mayor a 90 % (Figura 4). En el ejemplo que estamos empleando, el esfuerzo respiratorio con el valor inferior del IC 95 %, el resultado es de 0.93 o 93 % y para la disnea de 0.79 o 79 %. Con estos datos, la interpretación es que, con ambos instrumentos de medición, el medicamento A es superior al placebo, pero la magnitud del efecto parece ser mejor evaluada con el esfuerzo respiratorio.

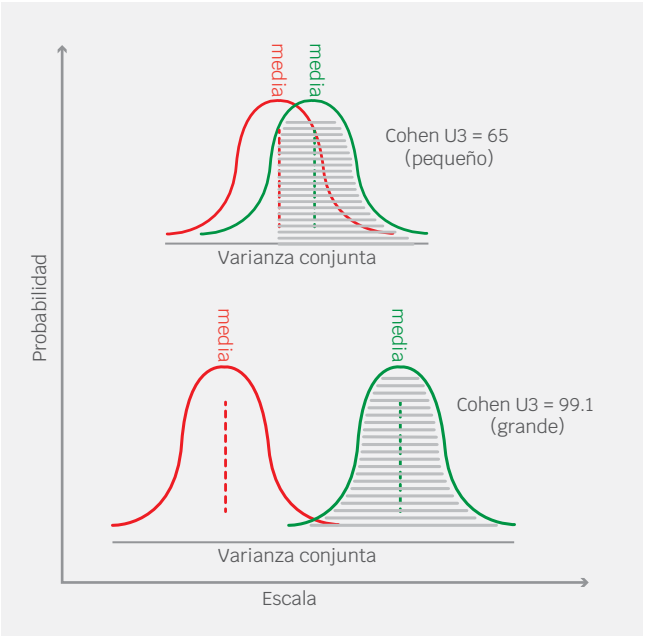

Figura 4. Cohen U3 o medida de no superposición, donde se toma el porcentaje de una población (verde) que supera la mitad superior de los casos de otra población (roja), equivale al área gris. Gráfico superior, efecto U3 pequeño; gráfico inferior, efecto U3 grande (50 a 100 %).

B) Porcentaje de solapamiento o coeficiente de solapamiento (OVL)

Este estadístico expresa el porcentaje de superposición de las distribuciones de dos grupos (Figura 5). Entre más grande es el valor (mayor solapamiento) , el tamaño del efecto es menor y viceversa. En nuestro ejemplo, el esfuerzo respiratorio tuvo un valor de 0.44 o 44 % y la evaluación de la disnea de 0.67 o 67.0 %.

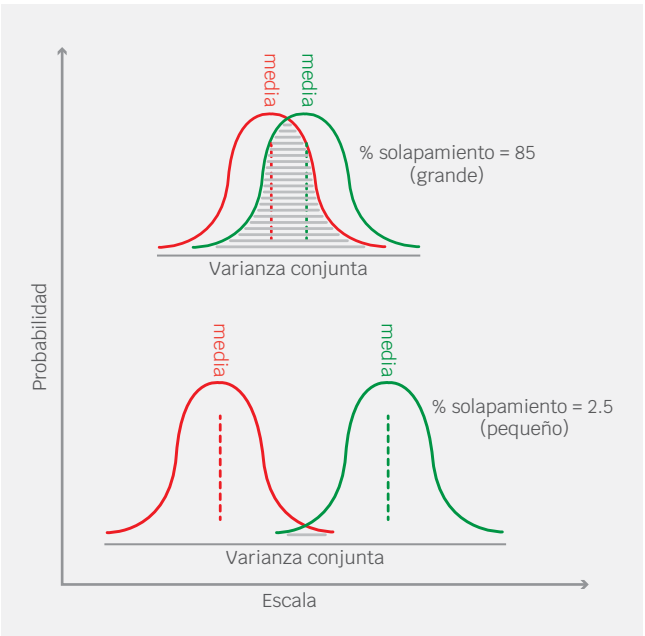

Figura 5. Porcentaje de solapamiento (OVL). La medida de superposición corresponde al área gris. El efecto es pequeño cuando el solapamiento es grande (gráfico superior). El efecto es grande cuando el solapamiento es pequeño (gráfico inferior).

C) Probabilidad de superioridad (P_{sup})

También conocido como tamaño de efecto en lenguaje común (CL), área bajo la curva del receptor operante (AUC), o A por su versión no paramétrica. En caso que la intervención en estudio tenga beneficio, entonces se considera como la probabilidad que una persona seleccionada al azar en el grupo experimental tenga una puntuación más alta, en comparación a la probabilidad de otra persona seleccionada al azar en el grupo control. En el caso de la evaluación de la percepción del esfuerzo respiratorio, este fue de 0.86 o 86 % y para la evaluación de la disnea de 0.71 o 71 %. Con estos datos, también se comprueba que el medicamento A es superior al placebo.

III. Tamaño del efecto indirecto para medidas de asociación[18,21]

En la medición de la fuerza de asociación se utilizan dos familias: la familia de las "r", así como las medidas de asociación para variables nominales. La familia de las "r" podemos considerarlas en dos grupos: *medidas de variabilidad asociada* y *medidas de variabilidad explicada.* Las primeras son los coeficientes de correlación (siendo las más conocidas, r de Pearson, rho de Spearman, Phi, Eta^2 y el coeficiente biserial). Estos coeficientes miden la variabilidad conjunta de dos variables. En general, van de +1.0 (una medida aumenta con igual fuerza como la otra) a -1.0 (una medida aumenta, con igual fuerza con la otra que disminuye), pasando por el "0", donde no hay efecto de una sobre la otra.

La interpretación puede ser diferente de acuerdo al contexto; en estudios de ciencias básicas, valores cercanos a +1.0 a -1.0 se requieren para hablar de que existe una fuerte asociación o que el tamaño de la asociación es significativo. En campos como la sociología o psicología, valores de "r" alrededor de +0.5 o -0.5 pueden ser importantes. Cohen propuso la siguiente escala: efectos pequeños valores de 0.1 a 0.29, medianos de 0.3 a 0.49, grandes de 0.5 a 0.69 y muy grandes ≥ 0.7.[20] Para determinar la amplitud del coeficiente se recomienda calcular su IC 95 %.

Las medidas de variabilidad explicadas evalúan la relación entre dos o más variables, considerando la dependencia de una a otra, asumiendo que el cambio de la primera dictará cambios en la otra. Los coeficientes más usados son el R^2 y los pseudo R^2, que van de 0 a 1.0. Los valores pueden ser totales, si se calculan tomando todas las variables estudiadas como un conjunto, o bien, ser emitidas en R^2 parciales, donde se analiza la asociación de cada variable de manera independiente. Los valores de 0.04 a 0.24 son considerados efectos pequeños, de 0.25 a 0.63 como medianos y ≥ 0.64 como grandes.[17]

Medidas de asociación clinico-epidemiológica.

Este tipo de indicadores de la magnitud del efecto fueron presentados en un artículo previo en esta revista.[15] Incluyen el riesgo relativo (RR), la razón de hazard (HR) o razón de tasas de incidencia y la razón de momios (RM). Se usan para evaluar asociación de factores de riesgo para el desarrollo de un evento desfavorable (o desenlace), como una enfermedad o una complicación. Se interpretan como cuántas veces es más probable un resultado, relacionado con una exposición determinada.

Valores > 1.0 se consideran asociados con el evento; mientras que los < 1.0 se consideran preventivos o protectores para el desarrollo del evento. Entre más alejados del 1.0, los valores serán más significativos o trascendentes; así, entre más cerca del cero (< 0.5), entonces podemos afirmar que este factor definitivamente es preventivo. Por el contrario, cuando los valores son > 3.0, se asocian fuertemente con el riesgo (o daño) para que se presente una enfermedad.[8,14,21,22,23,24] Sin embargo, la interpretación clínica depende del contexto de la asociación, dado que valores pequeños como 1.2 pueden ser importantes en campos de Salud Pública porque incrementan 20 % el riesgo a la aparición de algún resultado no deseado. Por ejemplo, este 20 % puede significar la aparición de alguna enfemedad en un número importante de personas, entre mayor sea el tamaño de la población.

IV. Tamaño del efecto indirecto en medidas de impacto

Un estimador usado en efectos de tratamiento es el número necesario para tratar (NNT).[14] Esta medida de impacto va de 1 a infinito, siendo valores < 5 indicativos de alta eficacia del tratamiento a considerar. Se calcula con la recíproca de la diferencia de riesgos (1/RRA).[15]

En el estudio de urticaria crónica que ya comentamos, asumamos que en el grupo con probióticos 25/30 (83.3 %) pacientes lograron eliminar sus lesiones contra 12/30 (40 %) en el grupo control. La prueba de chi cuadrada es de 10.15 (1 gl), p = 0.0014, lo cual se interpreta, que el uso de probióticos es mejor. Con estos mismos datos, el NNT es de 2.3, es decir, por cada dos pacientes tratados con probióticos, se logra que se cure uno más que lo observado en el grupo placebo. El IC 95 % es de 1.5 a 4.6, con lo cual el tamaño del efecto es muy significativo,[25] por lo que estaríamos en la posición de recomendar el uso de probióticos.

Conclusiones

Las medidas de tamaño de efecto buscan dar información sobre la magnitud, distancia o asociación encontrada entre grupos, de acuerdo al efecto de intervenciones o efectos de factores relacionados con desenlaces. Bien seleccionados permiten traducir los resultados en decisiones clínicas o epidemiológicas útiles, comparar resultados entre estudios, planear estudios posteriores y explicar fenómenos de forma más sencilla.

En esta revisión no agotamos todos los estimadores de tamaño de efecto sugeridos en la literatura, sino solo los más utilizados. Los mostrados se basan en estadística clásica o frecuentista, por lo que también se invita a conocer las propuestas para medir el tamaño de efecto mediante estadística bayesiana.[26,27]

Anexo
Fuentes para los cálculos

Social Science Statistics-Effect Size Calculators
https://www.socscistatistics.com/effectsize

Psychometrica- Computation of Effect Sizes
https://www.psychometrica.de/effect_size.html

GraphPad-t Test Calculator
https://www.graphpad.com/quickcalcs/ttest2/

Psychologist-Interpreting Cohen's d Effect Size
https://rpsychologist.com/cohend/

Referencias

1. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011;24(1):18-23. DOI: 10.1038/ajh.2010.205
2. Sullivan GM, Feinn R. Using effect size-or why the p value is not enough. J Grad Med Educ. 2012;4(3):279-282. DOI: 10.4300/JGME-D-12-00156.1
3. Kelley K, Preacher KJ. On effect size. Psychol Methods. 2012;17(2):137-152. DOI: 10.1037/a0028086
4. Alhija FN-A, Levy A. Effect size reporting practices in published articles. Educ Psychol Measurment. 2009;69(2):245-265. DOI: 10.1177/0013164408315266
5. Grissom RJ, Kim JJ. Effect. sizes for reseach: univariate and multivariate application. Segunda edición. Inlgaterra: Taylor & Francis Group; 2012.
6. Ventura-León J. Otras formas de entender la d de Cohen. Rev Evaluar. 2018;18(3):73-78. DOI: 10.35670/1667-4545.v18.n3.22305
7. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4:863. DOI: 10.3389/fpsyg.2013.00863
8. Martínez-Ezquerro JD, Riojas-Garza A, Rendón-Macías ME. Significancia clínica sobre significancia estadística. Cómo interpretar los intervalos de confianza a 95. Rev Alerg Mex. 2017;64(4):477-486. DOI: 10.29262/ram.v64i4.334
9. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007;82(4):591-605. DOI: 10.1111/j.1469-185X.2007.00027.x
10. Silva-Ayçaguer LC, Suárez-Gil P, Fernández-Somoano A. The null hypothesis significance test in health sciences research (1995-2006): statistical analysis and interpretation. BMC Med Res Methodol. 2010;10:44. Disponible en: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-10-44
11. O'Brien SF, Yi QL. How do I interpret a confidence interval? Transfusion. 2016;56(7):1680-1683. DOI: 10.1111/trf.13635
12. Schulz R, O'Brien A, Czaja S, Ory M, Norris R, Martire LM, et al. Dementia caregiver intervention research: in search of clinical significance. Gerontologist. 2002;42(5):589-602. DOI: 10.1093/geront/42.5.589
13. Schechtman E. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat--which of these should we use? Value Health. 2002;5(5):431-436. DOI: 10.1046/J.1524-4733.2002.55150.x
14. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry. 2006;59(11):990-996. DOI: 10.1016/j.biopsych.2005.09.014
15. Rendón-Macías ME, García H, Villasís-Keever M. Medidas de riesgo, asociación e impacto en los estudios de investigación clínica. Cómo interpretarlas para su aplicación en la atención médica. Rev Alerg Mex. 2021;68(1):65-75.
16. Peterson SJ, Foley S. Clinician's guide to understanding effect size, alpha level, power, and sample size. Nutr Clin Pract. 2021;36(3):598-605. DOI: 10.1002/ncp.10674

17. Pek J, Flora DB. Reporting effect sizes in original psychological research: a discussion and tutorial. Psychol Methods. 2018;23(2):208-225. DOI: 10.1037/met0000126
18. Maher JM, Markey JC, Ebert-May D. The other half of the story: effect size analysis in quantitative research. CBE Life Sci Educ. 2013;12(3):345-351. DOI: 10.1187/cbe.13-04-0082
19. Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. Front Psychol. 2012;3:111. DOI: 10.3389/fpsyg.2012.00111
20. Cohen J. A power primer. Psychol Bull. 1992;112(1):155-159. DOI: 10.1037//0033-2909.112.1.155
21. Ialongo C. Understanding the effect size and its measures. Biochem Med (Zagreb). 2016;26(2):150-163. DOI: 10.11613/BM.2016.015
22. Grzybowski A, Mianowany M. Statistics in ophthalmology revisited: the (effect) size matters. Acta Ophthalmol. 2018;96(7):e885-e888. DOI: 10.1111/aos.13756.
23. Lininger M, Riemann BL. Statistical primer for athletic trainers: using confidence intervals and effect sizes to evaluate clinical meaningfulness. J Athl Train. 2016;51(12):1045-1048. DOI: 10.4085/1062-6050-51.12.14
24. Calin-Jageman RJ. The new statistics for neuroscience majors: thinking in effect sizes. J Undergrad Neurosci Educ. 2018;16(2):E21-E25.
25. Bender R. Calculating confidence intervals for the number needed to treat. Control Clin Trials. 2001;22(2):102-110. DOI: 10.1016/s0197-2456(00)00134-3
26. Kelter R. Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Med Res Methodol. 2020;20(1):88. Disponible en: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-00968-2
27. Kelter R. Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Res Notes. 2020;13(1):452. DOI: 10.1186/s13104-020-05291-z

ORCID
Miguel Ángel Villasís-Keever, 0000-0002-8566-0811; Mario Enrique Rendón-Macías, 0000-0001-7310-6656;
Irma Susana Zarco-Villavicencio, 0000-0002-1881-0031