Colegio Mexicano de Inmunología Clínica A.C. | CMICA
Revista Alergia México
Metodología de la investigación

Clinical significance vs statistical significance. How to interpret the confidence interval at 95 %

Significancia clínica sobre significancia estadística. Cómo interpretar los intervalos de confianza a 95 %

José Darío Martínez-Ezquerro,[1] Alberto Riojas-Garza,[2] Mario Enrique Rendón-Macías[2,3]

Abstract

The validity of a study depends on its proper planning, execution and analysis. If these are sufficiently correct, the decision to apply the recommendations issued depends on the expected clinical effect. This effect may have random variations, hence the need to use statistical inference. For years the p-value has been used to determine this statistical significance and the confidence intervals to measure the magnitude of the effect. In this review we present a proposal of how to interpret the 95 % confidence intervals (CI 95 %) as estimators of the expected effect variability based on considering the threshold or value of clinical significance and the null value of the difference or rejection of statistical significance. Thus, an association or effect where the CI 95 % includes the null value (no effect or difference) is interpreted as inconclusive; one between the null value and the clinical threshold (without including them) as possibly inconsequential; one that does not include the null value but the clinical threshold as yet not conclusive and one beyond the clinical threshold as conclusive.

Keywords: P-value; Confidence interval 95 %; Clinical decision

Este artículo debe citarse como: Martínez-Ezquerro JD, Riojas-Garza A, Rendón-Macías MA. Significancia clínica sobre significancia estadística, como interpretar los intervalos de confianza a 95 %. Rev Med Alerg Mex. 2017;64(4):477-486

[1]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Hospital de Pediatría, Unidad de Investigación Médica en Genética Humana. Ciudad de México, México
[2]Universidad Panamericana, Escuela de Medicina. Ciudad de México, México
[3]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Hospital de Pediatría, Unidad de Investigación Médica en Epidemiología Clínica. Ciudad de México, México

Correspondencia: Mario Enrique Rendón-Macías.
drmariorendon@gmail.com

Recibido: 2017-11-02
Aceptado: 2017-11-20

Resumen

La validez de un estudio depende de su adecuada planeación, ejecución y análisis. Si estas son suficientemente correctas, la decisión de aplicar las recomendaciones emitidas depende del efecto clínico esperado. Este efecto puede tener variaciones aleatorias, de ahí la necesidad de usar la inferencia estadística. Durante años se ha usado el valor de p para determinar esta significancia estadística y los intervalos de confianza para medir la magnitud del efecto. En esta revisión se presenta una propuesta de cómo interpretar los intervalos de confianza a 95 % (IC 95 %) como estimadores de la variabilidad del efecto esperado, con base en considerar el umbral o valor de significancia clínica y el valor nulo de diferencia o rechazo de significancia estadística. Con ello, una asociación en la cual el IC 95 % incluye el valor nulo (no efecto o diferencia) es interpretado como no concluyente; uno entre el valor nulo y el umbral clínico (sin incluirlos) como posiblemente intrascendente; uno que no incluye al valor nulo, pero sí al umbral clínico como aún no contundente y uno más allá del umbral clínico como contundente.

Palabras clave: Valor de p; Intervalo de confianza a 95 %; Decisión clínica

Abreviaturas y siglas

DM, diferencia de medias
gl, grado de libertad
IC, intervalo de confianza
RM, razón de momios
TSSC, *Total Symptoms Severity Complex*

Introducción

La investigación médica tiene como propósito mejorar la atención de los pacientes mediante estudios diseñados de tal forma que permitan evaluar objetivamente el impacto de factores o intervenciones en la prevención o recuperación de la salud. Estos estudios se realizan con muestras lo más representativas posible de los pacientes que se beneficiarán a futuro de las intervenciones o recomendaciones emitidas. Debido a la necesidad de realizar investigaciones en muestras, la inferencia estadística es fundamental para predecir si el efecto encontrado sucederá igual en los no participantes. Esta inferencia se ha sustentado por años en la prueba de hipótesis con base en el valor de p y, más recientemente, en la estimación del efecto y su variabilidad con el cálculo de los intervalos de confianza a 95 %.[1,2] Sin embargo, su interpretación aún sigue siendo inadecuada, tanto por algunos lectores, editores y revisores de artículos científicos, como por los propios investigadores.[3]

En esta revisión retomamos el significado de las pruebas de hipótesis y el valor de p, así como la estimación de la precisión a través de los intervalos de confianza a 95 %. Además, proponemos una estrategia de interpretación de los resultados basada en la significancia clínica y estadística mediante umbrales de decisión. Para ello se presenta un ejemplo con cuatro distintos escenarios.

Prueba de hipótesis y valor de p

El valor de p es una medida propuesta por Ronald A. Fisher en 1925[4] como una guía numérica sobre la fuerza de la evidencia estadística en contra de la hipótesis nula.[5] En su prueba de significancia, actualmente conocida como prueba de hipótesis, se comienza estableciendo una hipótesis nula estadística. En esta, dos grupos en comparación son equivalentes y la diferencia entre ellos y su error estándar se utilizan para construir una relación crítica; el valor de dicha relación corresponde a un valor de probabilidad llamado "p", el cual denota la probabilidad

bajo la hipótesis nula de encontrar una diferencia al menos tan grande como el valor observado en esta diferencia.

Cuando el valor de p es menor a un límite preestablecido, llamado alfa y usualmente fijado en 0.05, se rechaza la hipótesis nula y se tiene que aceptar la hipótesis de "no igualdad". Por tanto, se dice que el resultado observado es "estadísticamente significativo".[6] Con este resultado se diría que si una diferencia en un efecto a favor de algún tratamiento, controlando todos los factores confusores, fuera una variación aleatoria, sería una muy rara (5 % o menos de las veces) de una distribución donde el promedio del efecto es cero. Por ello, se interpreta como una duda suficiente para rechazar esta posibilidad (hipótesis nula) y se opta por atribuir la explicación a un efecto a favor de uno de los tratamientos o factores de riesgo evaluados.

Es evidente que a menor valor de p, mayor evidencia estadística contra la hipótesis nula.

Este énfasis en $p < 0.05$ o $p \geq 0.05$ para definir si los resultados son o no estadísticamente significativos debería rechazarse dado que transmite poca información. Además, el valor de 0.05 es un valor de corte arbitrario que promueve un pensamiento perezoso[7] y prácticas de pirateo de valores de p (*p-hacking*),[8] al pretender ejercer de manera incorrecta la dicotomía "todo o nada" para interpretar y evaluar los resultados de estudios cuya intención es su aplicación práctica inmediata. En estos casos de aplicación práctica, las decisiones necesitan equilibrar costos y beneficios, así como considerar la magnitud de ambos.

Por lo tanto, el tamaño del efecto siempre es importante en experimentos de aplicación práctica y confiar únicamente en las pruebas de hipótesis nula es inapropiado, debiendo distinguirse la significancia estadística de la significancia clínica o práctica.[9] Esto no quiere decir que las pruebas de hipótesis y los valores de p no sean útiles en la interpretación de experimentos, si no que el análisis estadístico primario debe sustentarse en análisis más descriptivos e informativos.[10]

Precisión e intervalos de confianza a 95 % (IC 95 %)

Los intervalos de confianza (IC) representan el rango o intervalo de valores calculados mediante métodos estadísticos (media poblacional, mediana, variancia, probabilidad o cualquier otra cantidad desconocida) que teóricamente incluyen el parámetro verdadero y que tras la repetición de muestreos tienen una probabilidad fija de contener el parámetro. El nivel de confianza de 95 % significa que el intervalo de confianza abarca el valor verdadero en 95 de 100 estudios desarrollados.[11,12] Los IC son reportados con rangos o intervalos y estimadores puntuales. Los intervalos describen los valores inferiores y superiores (límites) de incertidumbre o márgenes de error. Su amplitud describe la magnitud del efecto o diferencia entre los grupos: cuando es grande (usualmente con tamaños de muestra pequeños) representa mayor incertidumbre y que los resultados no son claros; cuando es pequeña representa mayor certeza y claridad.

Los estimadores puntuales de una muestra poblacional son valores específicos que pueden ser la media de la muestra, la puntuación de la diferencia de grupos, el tamaño del efecto, la razón de momios (RM) o el riego relativo (RR), entre otros, y no representan el valor verdadero, sino el mejor estimador del valor verdadero del promedio de la muestra.[13]

A partir del ensayo de Rothman (1978) sobre la contraproducente y ocasionalmente engañosa postura dicotómica para clasificar los resultados de un estudio en una decisión de "todo o nada" mediante una prueba de significancia estadística,[10] ocurrió un giro en el estilo para reportar los resultados de investigación hacia los IC, con o sin el acompañamiento de los valores de p.[14]

Al elegir una medida que cuantifique el grado de asociación o efecto en los datos y luego calcular un intervalo de confianza, se resume la fuerza de asociación de los datos y se permite la variación aleatoria de forma simple y sin ambigüedad. La posición exacta del límite del intervalo no es relevante para una interpretación apropiada. El límite de un intervalo de confianza depende también del grado de confianza, que por lo general se selecciona arbitrariamente a 90, 95 o 99 %; aunque estos intervalos difieren en anchura, rara vez generarán interpretaciones distintas porque la localización precisa de los límites al intervalo tiene poca consecuencia práctica. Lo que rige la interpretación es la posición aproximada del intervalo en su escala de medida.[10]

Canalizar todo el interés en la localización precisa del límite de un intervalo de confianza sería el equivalente a "todo o nada" de las pruebas de hipó-

tesis para la significancia estadística y, como ocurre con el pirateo de los valores de p, resultaría en prácticas de pirateo de los intervalos de confianza (*CI-hacking*). Sin embargo, la localización del intervalo de confianza en zonas de decisión preestablecidas es muy orientadora para la interpretación clínica.

Para explicar mejor cómo interpretar los intervalos de confianza se presenta el siguiente ejercicio hipotético con representaciones gráficas para ayudar a demostrar su valor clínico e interpretación.

Ejercicio

Un equipo de investigadores realizó un ensayo clínico aleatorizado doble ciego con dos grupos paralelos de pacientes con rinitis alérgica estacional. El objetivo fue probar si el tratamiento A (nuevo fármaco) es más eficaz que el B (más recomendado al momento) para atenuar o eliminar los síntomas de la enfermedad. Para evaluar la severidad de los síntomas se utilizó la escala TSSC (*Total Symptoms Severity Complex*). Esta escala califica la severidad en un rango de 0 a 14 (nula a máxima gravedad, respectivamente). En el estudio se incluyeron pacientes con puntuaciones pretratamiento de 10 a 14. El resultado fue evaluado a los 14 días de tratamiento. Ningún participante abandonó o se perdió durante el estudio.

A fin de comparar las pruebas de hipótesis (valor de p) y los intervalos de confianza a 95 % se describirán tres estudios hipotéticos con 5, 40 y 400 o 4000 participantes por grupo. Asimismo, la respuesta medida por la TSSC será analizada resumiendo el efecto con una medición cuantitativa (diferencia de las medias de los tratamientos con sus desviaciones estándares) o una cualitativa (respuesta al tratamiento en frecuencias simples y porcentajes, si la puntuación postratamiento fue ≤ 2).

Escenario hipotético1

Estudio con una muestra de cinco pacientes por grupo y con la variable de resultado medida de forma cuantitativa. Al obtener el promedio de la puntuación de la TSSC en ambos grupos, la media del grupo con tratamiento A fue menor a la del B, con una diferencia de menos 1 (Cuadro 1a). Si se realiza una prueba de hipótesis para grupos independientes con t de Student se obtiene un valor de 0.87 (8 gl). De considerar una hipótesis nula unilateral (tratamiento A es igual o menor a B) el valor de p es 0.20, por arriba del umbral de 0.05, y, por tanto, no se rechaza la hipótesis nula; ambos tratamientos parecen ser igualmente eficaces. De considerar una hipótesis nula de dos colas (la media del efecto entre los grupos es igual y contra alguno es diferente), el valor de p es 0.40, aún menos significativa.

¿Qué sucede si aumentamos el número de participantes en ambos grupos? En el mismo Cuadro 1a se observa que en ambos grupos la variabilidad entre sus participantes (desviaciones estándares) es menor y la diferencia entre las medias de ambos grupos disminuye ligeramente (menos 0.7 a favor del grupo A). Con el incremento de sujetos, la significancia estadística mejora: para una prueba de una cola, el valor de t de Student es de 1.78, con 78 gl, lo que da un valor de $p = 0.03$, meritorio de rechazar la hipótesis nula (media del grupo $A \geq B$), aunque no para una prueba de dos colas, con $p = 0.07$ ($A = B$). Los autores pudieran reportar la primera opción al justificar la hipótesis con dirección (una cola) ante la premisa de buscar si el nuevo tratamiento es mejor.

Ahora bien, si el número de participantes se incrementa sustancialmente a 400 por grupo, la diferencia entre las medias es de solo menos 0.5 puntos para el tratamiento A respecto al B, sin embargo, el valor de t de Student es de 6.21, con 789 gl. Con ello, tanto una prueba de una cola (unidireccional) como de dos colas (bidireccional) dan una $p < 0.0001$; en ambas, la hipótesis nula sería rechazada.

Como se puede ver en el ejemplo, la significancia estadística depende del tamaño de la muestra, siempre y cuando las condiciones del estudio sean iguales. El supuesto es que el efecto promedio no cambia y lo que se controla es la variabilidad entre los sujetos. A medida que aumenta el número de participantes es factible que la mayoría tenga valores cercanos a ese promedio y con ello se disminuye la variancia. Sin embargo, como observamos, el efecto real en la reducción de los síntomas entre grupos es mínimo e, incluso, sin una posible implicación clínica. Entonces, ¿cuánto cambia la decisión terapéutica o la satisfacción del paciente en una valoración de gravedad entre 3.7 y 4.3 si ambas son cercanas a una puntuación de 4? En realidad, ambas muestran una severidad baja.

Escenario 2

Para este ejemplo, los resultados del estudio hipotético serán los del Cuadro 1b. La diferencia es más notoria entre los tratamientos si el estudio incluyó solo

Cuadro 1. Resultados postratamiento

Pacientes por grupo	Tratamiento A Media ± 1 DE	Tratamiento B Media ± 1 DE	Diferencia de medias	IC 95 % de la diferencia	p 1 cola 2 colas
a) Escenario 1					
5	3.8 ± 1.9	4.8 ± 1.7	-1.0	−3.6 a 1.6	0.20 0.40
40	3.8 ± 1.8	4.5 ± 1.7	-0.7	−1.48 a 0.08	0.03 0.07
400	3.75 ± 1.3	4.3 ± 1.2	-0.55	−0.36 a −0.7	< 0.0001 < 0.001
b) Escenario 2					
5	3 ± 2.5	7.0 ± 3.5	-4	−8.3 a 0.4	0.04 0.07
40	2.5 ± 1	8.5 ± 3	-6	−5 a −7	< 0.0001 < 0.0001
400	2.2 ± 0.5	8.3 ± 1.2	−6.1	−5.9 a −6.2	< 0.0001 < 0.0001

DE, desviación estándar; IC, intervalo de confianza al 95 %

cinco pacientes. El tratamiento A redujo cuatro puntos el promedio en comparación con el tratamiento B. Si aplicamos la t de Student dará una puntuación de 2.08, con 8 gl, y un valor de p unidireccional de 0.04 o una bidireccional de 0.07, con lo cual se rechaza la hipótesis nula solo para la hipótesis unilateral.

Al aumentar el tamaño de muestra a 40 individuos por grupo, la diferencia se incrementa a 6 puntos a favor del tratamiento A; el valor de la prueba de t de Student es de 12, con 78 gl, y la prueba unidireccional o bidireccional dará una $p < 0.0001$, con lo que se rechaza la hipótesis nula y se acepta un efecto a favor del tratamiento A como posible explicación de esta rara distribución. Este tamaño de muestra ya es suficiente para tomar una decisión de preferir el tratamiento A si no hay contraindicación. Realizar un estudio con más pacientes no sería ético y sí más costoso para llegar a la misma conclusión, sin embargo, para fines del ejercicio mostramos el efecto de incrementar a 400 pacientes en cada grupo. Como se muestra en el Cuadro 1b, no cambia la interpretación.

En este escenario, desde el ensayo con cinco pacientes había una diferencia más notable y, por lo tanto, al incrementar un poco el número de participantes se evidenció la significancia estadística.

En los ejemplos anteriores observamos una evaluación a través de una escala cuantitativa, sin embargo, en la clínica suele ser más importante determinar si un paciente tiene mejoría en la resolución de su enfermedad, objetivo de un tratamiento. Por ello, si consideramos que una puntuación ≤ 2 en la escala TSSC es equivalente a la resolución del problema, podemos analizar a los pacientes como curados (puntuación ≤ 2) y no curados (puntuación > 2).

Escenario 3

Con un desenlace dicotómico, en el Cuadro 2a se observa que con cinco pacientes por grupo no hay diferencia en el riesgo de no curarse si los sujetos reciben el tratamiento A o B (20 % de cada grupo, RR = 1) y con valor de p = 1.0 no se rechaza la hipótesis nula (uni ni bidireccional). Al incrementar el número de pacientes, la proporción sube aproximadamente 30 % en ambos grupos y nuevamente el riesgo es

cercano a 1 (RR = 0.86) y aún no estadísticamente significativo. Finalmente, con 4000 pacientes en cada grupo, el porcentaje de falla entre los grupos es poca (22.5 contra 25 %), aunque el riesgo relativo de no curarse es estadísticamente menor con el tratamiento A (RR = 0.9), tanto con una hipótesis unilateral (p = 0.004) como bilateral (p = 0.009). Nuevamente, la significancia estadística depende del tamaño de muestra (si las condiciones de los ensayos son iguales), aún con un efecto pequeño.

Escenario 4

El Cuadro 2b muestra un resultado donde el tratamiento A logra un control de los síntomas mucho mayor que el tratamiento B. Con cinco pacientes se observa una diferencia de 60 % más de curados (RR = 0.25), a pesar de lo cual la diferencia no es estadísticamente significativa (p = 0.13 bilateral y p = 0.06 unilateral) y, por lo tanto, no se concluye que el tratamiento A sea mejor. Si se aumenta a 40 sujetos por tratamiento, esta diferencia se hace altamente significativa desde el punto de vista estadístico y más aún si el tamaño de la muestra se incrementa a 400 individuos por grupo. Nuevamente, el segundo incremento de pacientes se hace innecesario por el costo, pero principalmente por implicaciones éticas.

Intervalos de confianza

Como se comentó, la prueba de hipótesis con el valor de p solo permite establecer si se puede o no rechazar una hipótesis nula sin dar información sobre la magnitud y dirección del efecto. Los intervalos de confianza permiten determinar la variabilidad del efecto en un rango estimado de posibles muestras, en la que una significancia estadística de 5 % de error en la estimación equivale a 95 % de probabilidades de obtener ese intervalo. La amplitud del intervalo depende principalmente del tamaño de la muestra y del control (calidad) en la medición; de esta forma, a medida que ambas aumentan es menor el tamaño del intervalo. Entre más pequeño el intervalo es más preciso el estimado.

En los escenarios expuestos se observa que el intervalo de confianza se redujo a medida que aumentaron los participantes (Figuras 1a y 1b). Asimismo, cuando el intervalo de confianza no incluye al estimado nulo (valor 0 en una diferencia o 1 en

Cuadro 2. Pacientes que mantuvieron puntuaciones arriba de dos TSSC

Pacientes por grupo	Tratamiento A		Tratamiento B		RR	IC 95 %	p
	n	%	n	%			
a) Escenario 3							
5	1	20	1	20	1	0.08 a 11.9	1.00
40	11	27.5	13	32.5	0.84	0.43 a 1.6	0.62
4000	900	22.5	1000	25	0.90	0.83 a 0.97	0.008
b) Escenario 4							
5	1	20	4	80	0.25	0.04-1.52	0.13
40	7	17.5	27	67.5	0.39	0.24-0.62	0.0001
400	71	17.7	269	67.2	0.40	0.34-0.46	< 0.001

RR, riesgo relativo; IC, intervalo de confianza a 95 %

una medida de asociación como el RR) hay suficiente confianza para decir que es estadísticamente diferente. Con cinco pacientes, el intervalo indica que la diferencia entre las puntuaciones con el tratamiento A comparadas con las del B puede ser desde menos 3.6 puntos a favor del tratamiento A, hasta menos 1.6 puntos para el tratamiento B y existe la posibilidad de una diferencia de 0 (Cuadro 1a); por tanto, no hay confianza en tomar partido sobre alguno. Cuando se aumentó el tamaño a 400 sujeto por grupo, la diferencia a favor del tratamiento A fue de menos 0.36 a 0.7. Hay confianza en 95 % de las ocasiones de aseverar que el tratamiento A reduce los síntomas más que el tratamiento B.

Intervalos de confianza y significancia clínica

Hasta ahora, tanto las pruebas de hipótesis como el IC 95 % han permitido determinar si una diferencia encontrada es estadísticamente significativa; es decir, no es una diferencia en gran parte atribuida a la variación aleatoria propia de una muestra pequeña.[11] Asimismo, han permitido mostrar que el incremento en el tamaño de la muestra a un nivel suficientemente grande logrará una significancia estadística, pero en las decisiones médicas las implicaciones clínicas son fundamentales. En cualquier investigación debe considerarse la estimación de un nivel de diferencia o asociación suficiente para justificar la realización

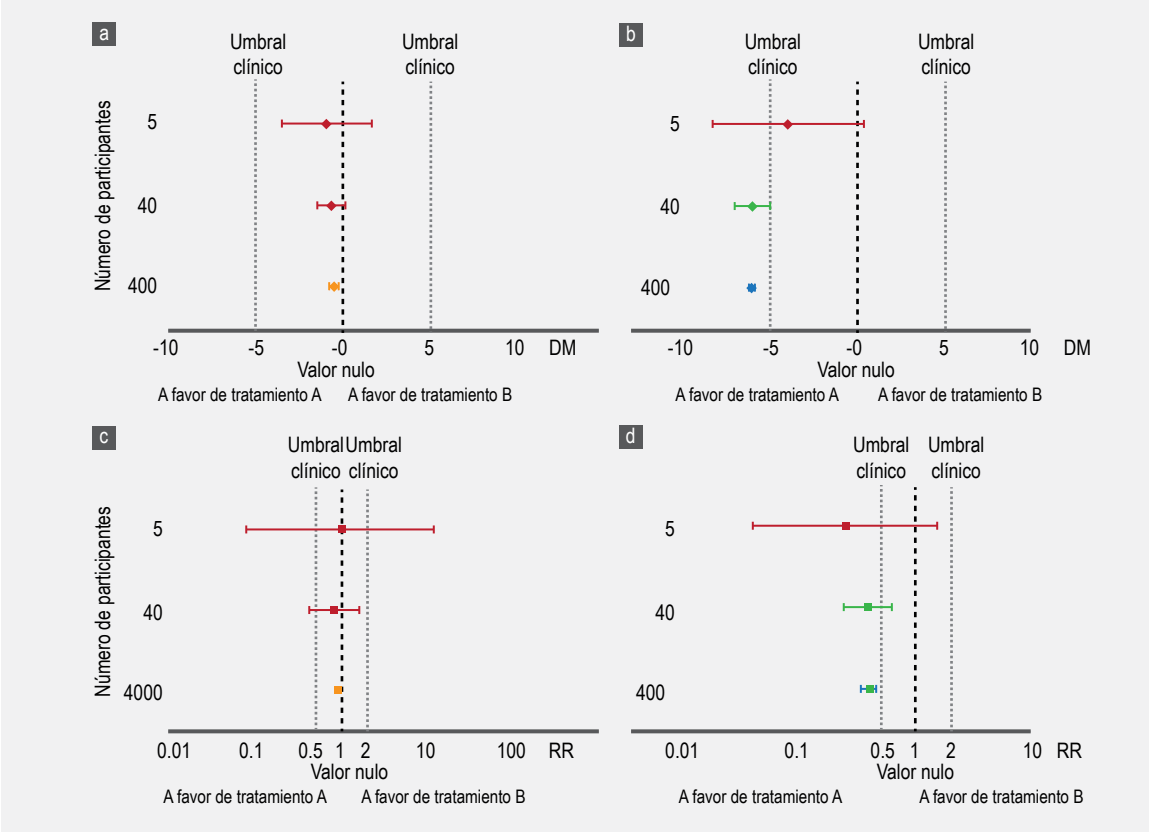

Figura 1. Localización de los intervalos de confianza 95 % (límite menor y mayor de cada línea horizontal) y del estimador (cuadrado o rombo) con relación al valor nulo [0 para A y B y 1 para C y D]). Los umbrales clínicos se establecieron para una diferencia de medias (DM) [A y B] de 5 unidades y para estimador de asociación (riesgo relativo o RR) [C y D] en 0.5 y 2 (a favor de A y a favor de B respectivamente). Los intervalos de confianza pueden interpretarse como diferencia o asociación no concluyente (rojo), diferencia o asociación intrascendente (naranja), diferencia o asociación no contundente (verde) y diferencia o asociación contundente (azul).

de acciones inmediatas. En nuestro estudio hipotético, si se desea concluir en recomendar o no el tratamiento A sobre el B se debe determinar a partir de cuál diferencia en la puntuación se tendrían implicaciones en los síntomas de los pacientes. En una escala de 0 a 14, quizá una diferencia de más de 5 puntos mostraría que con determinado tratamiento habría más pacientes con ausencia de síntomas o casos menos graves.

Si se establece este nivel de diferencia de 5 puntos como el clínicamente significativo para determinar una respuesta más eficaz, a este nivel se le denomina "umbral clínico", diferente del "umbral de significancia estadística o valor nulo".

Por otro lado, la decisión pudiera basarse en un estimado de asociación como el riesgo relativo (RR), razón de momios (RM), razón de hazard (HR). En estos casos, aunque los umbrales clínicos pudieran fijarse a criterio de quien realiza la investigación, en general se acepta que un factor es considerado como de riesgo o protección importante cuando incrementa o disminuye dos o más veces el riesgo de presentar algún desenlace; este nivel equivale a un RR (OR o HR) ≤ 0.5.

Con estos dos umbrales y el IC 95 % del estimador del efecto estudiado podemos establecer diferentes interpretaciones para un estudio.[1,2,15]

Interpretaciones

- *Diferencia o asociación no concluyente*. Sucede cuando el intervalo de confianza a 95 % del estimado del efecto incluye al valor nulo (independientemente de dónde esté situado el valor puntual del estimado). En todos los escenarios hipotéticos con cinco pacientes se dio este fenómeno, así como en los escenarios 2 y 3 con 40 pacientes. En los primeros dos escenarios, los intervalos de confianza a 95 % incluyeron el 0 y para las dos últimas el valor nulo de 1 (Figura 1; líneas horizontales de color rojo). En este escenario, el estimado encontrado podría estar a favor de cualquier tratamiento o no apoyar ninguno, de ahí que no se pueda obtener una conclusión. Esto no significa que haya o no un efecto real, solo que el estudio no aporta evidencias suficientes a favor o en contra de uno de los tratamientos.
- *Diferencia o asociación clínicamente intrascendente*. Sucede cuando el intervalo de confianza a 95 % se localiza entre el umbral estadístico y el clínico, sin tocar ninguno. Aquí hay significancia estadística (el intervalo no incluye al valor nulo), pero está por debajo del umbral clínico, es decir, el efecto es tan escaso como para tener un impacto en la decisión médica. Esta condición sucede en el escenario 2 con 400 pacientes, donde el umbral estimado de una diferencia clínicamente importante se fijó en 5, y en el escenario 3 con 4000 participantes, con un umbral clínico a un riesgo de 0.5 (a favor del tratamiento A) o arriba de 2 (a favor del tratamiento B) (Figura 1; líneas horizontales de color naranja). Desde el punto de vista estadístico se puede rechazar la hipótesis nula, pero aun cuando el estudio esté bien realizado, el impacto real de la recomendación o maniobra parece ser escaso. Hay que aclarar que los efectos pequeños pueden tener grandes repercusiones en algunos estudios epidemiológicos, dado que el número de beneficiarios o perjudicados (según el estudio) se incrementa en grandes poblaciones.
- *Diferencia o asociación clínicamente no contundente*. Sucede cuando el intervalo de confianza a 95 % está por arriba o distal del umbral estadístico (valor nulo), pero incluye el umbral clínico (predeterminado). Por ello, no hay duda de un efecto superior de un tratamiento sobre otro desde el punto de vista estadístico, pero pudiera no ser tan importante como se esperaba, es decir, aún no hay suficiente evidencia para atribuir un efecto clínico esperado. Esta condición se muestra en los escenarios 2 y 4 con 40 pacientes: en el escenario 2, el umbral clínico se estableció en una diferencia de 5 puntos entre los tratamientos; en el escenario 4, un riesgo relativo de 0.5 (a favor del tratamiento A) o arriba de 2 (a favor del tratamiento B) (Figura 1; Líneas horizontales color verde).
- *Diferencia o asociación clínicamente contundente*. Sucede cuando el intervalo de confianza a 95 % se encuentra por arriba o distal al umbral clínico preestablecido (en consecuencia, tampoco incluye el valor nulo). Esto significa que aun con variación en el estimado analizado (diferencia o asociación), este se encontrará en 95 % de las probabilidades en un valor de importancia clínica para la toma de decisiones inmediatas. En los escenarios 2 y 4, con ensayos de 400 pacientes por grupo, se observa esta condición (es-

cenario 2 con umbral clínico por arriba de cinco puntos de diferencia y escenario 4 con umbral clínico menor a un RR de 0.5) (Figura 1; líneas horizontales de color azul).

Otros aspectos por considerar

Con la clasificación anterior, el clínico puede valorar mejor los resultados de un estudio. Cuando el intervalo de confianza de un resultado se encuentra situado en el área de contundencia no solo hay confianza en que el efecto es altamente posible de suceder, sino la magnitud a la cual pudiera suceder. Por otro lado, si el intervalo está en el escenario de no contundencia, el médico debería ser prudente en la recomendación. Si bien hay evidencia sobre el efecto buscado asociado con el o los factores considerados, este no es de la magnitud esperada o aún se requiere incrementar el número de ensayos para tener una mayor precisión.

Cuando se presenta una condición de ausencia de posible efecto clínico, pero con una significancia estadística (condición de intrascendencia), lo recomendado es desechar la asociación o relación. En este caso, incrementar el tamaño de la muestra no mejorará el apoyo al fenómeno estudiado, solo incrementará la significancia estadística y generará mayor confusión en la interpretación.

Cuando la condición de diferencia o asociación no concluyente está presente, es recomendable analizar la amplitud del intervalo de confianza. Si esta es muy grande y el estimado puntual se ubica en la zona de contundencia, es necesario realizar más estudios para aclarar el escenario. Esto sucedió en los escenarios 2 y 4 con pocos pacientes. Cabe resaltar que la descalificación errónea de esta asociación solo por la ausencia de significancia estadística es muy frecuente. Contrario a esto, si el intervalo de confianza es muy corto (preciso) y el estimado puntual se localiza en el área de intrascendencia, no vale la pena continuar realizando más ensayos o incluyendo más pacientes dado que esto solo aumentará la significancia estadística, pero no la significancia clínica.

Por último, aún en un escenario en el que se obtenga un IC 95 % en el área de contundencia, la confirmación de la asociación o efecto real requiere el análisis profundo de los objetivos y supuestos bajo los cuales se realizó un estudio y asegurar la ausencia de sesgos importantes en la selección, ejecución e interpretación de los resultados de un estudio,[12,16] así como validaciones posteriores.

Conclusiones

La interpretación de asociaciones o diferencias en estudios clínicos debe basarse en la implicación clínica de estas. Para mejorar la interpretación se propone analizar la localización de los intervalos de confianza a 95 % de los estimadores con relación al umbral preestablecido de importancia clínica y el valor de la hipótesis nula. Con ello se pueden establecer asociaciones o diferencias contundentes, no contundentes, intrascendentes y no concluyentes, que facilitarán la toma de decisiones clínicas inmediatas.

Referencias

1. Argimon JM. El intervalo de confianza: algo más que un valor de significación estadística. Med Clin. 2002;118(10):382-384. DOI: http://dx.doi.org/10.1016/S0025-7753(02)72393-2
2. van-Rijn MHC, Bech-A, Bouyer J, van-Den-Brand JAJG. Statistical significance versus clinical relevance. Nephrol Dial Transplant. 2017;32 Suppl 2:ii6-ii12. DOI: http://dx.doi.org/10.1016/S0025-7753(02)72393-2
3. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-50. DOI: http://dx.doi.org/10.1007/s10654-016-0149-3
4. Fisher RA. Statistical methods for research workers. Escocia: Genesis Publishing; 1925.
5. Goodman S. A dirty dozen: Twelve p-value misconceptions. Semin Hematol. 2008;45(3):135-140. DOI: http://dx.doi.org/10.1053/j.seminhematol.2008.04.003
6. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998;51(4):355-360. DOI: http://dx.doi.org/10.1016/S0895-4356(97)00295-3
7. Gardner MJ, Altman DG. Confidence intervals rather than P values: Estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986;292(6522):746-750. DOI: http://dx.doi.org/10.1136/bmj.292.6522.746

8. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015;13(3):e1002106. DOI: http://dx.doi.org/10.1371/journal.pbio.1002106
9. Frick RW. The appropriate use of null hypothesis testing. Psychol Methods. 1996;1(4):379-390. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.6113&rep=rep1&type=pdf
10. Rothman J. A show of confidence. N Engl J Med. 1978;299(24):1362-1363. DOI: http://dx.doi.org/10.1056/NEJM197812142992410
11. Du-Prel JB, Hommel G, Röhrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009;106(19):335-339. DOI: http://dx.doi.org/10.3238/arztebl.2009.0335
12. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev. 2016;23(1):103-123. DOI: http://dx.doi.org/10.3758/s13423-015-0947-8
13. Page P. Beyond statistical significance: clinical interpretation of rehabilitation research literature. Int J Sports Phys Ther. 2014;9(5):726-736.
14. Savitz DA, Tolo KA, Poole C. Statistical significance testing in the American Journal of Epidemiology, 1970-1990. Am J Epidemiol. 1994;139(10):1047-1052. DOI: http://dx.doi.org/10.1093/oxfordjournals.aje.a116944
15. Manterola C, Pineda V, Grupo MINCIR. El valor de "p" y la "significación estadística". Aspectos generales y su valor en la práctica clínica. Rev Chil Cir. 2008;60(1):86-89. DOI: http://dx.doi.org/10.4067/S0718-40262008000100018
16. Clark ML. Los valores p y los intervalos de confianza: ¿en qué confiar? Rev Panam Salud Publica. 2004;15(5):293-296. Disponible en: https://scielosp.org/pdf/rpsp/v15n5/21999.pdf