ANÁLISIS BASADO EN LA EVIDENCIA

Jorge Mario Estrada

EJE 2

Analicemos la situación

Fuente: Shutterstock/370946246

AREANDINA
Fundación Universitaria del Área Andina
MIEMBRO DE LA RED ILUMNO

ÍNDICE

ÍNDICE

INTRODUCCIÓN

Los escenarios en los que la toma de decisiones se da en el proceso de atención de un paciente son diversos. Asimismo, la evidencia que se publica acerca de dichas decisiones sobrepasa la capacidad humana para su conocimiento. Por tal motivo, el profesional debe ser capaz de discriminar entre la buena y mala evidencia, esto hace necesario que conozcamos los errores que se pueden dar dentro del desarrollo de una investigación científica, así como los conceptos básicos para reconocer la validez de los estudios en diferentes ámbitos de aplicación. Decidir sobre la terapia de un paciente, introducir una nueva prueba diagnóstica y conocer el pronóstico de un paciente son situaciones que ameritan reconocer cuándo se está ante un resultado válido que da respuesta a las interrogantes propuestas.

Error aleatorio y sesgo en estudios clínicos

Los resultados de una investigación científica son propensos al error, es decir, se pueden obtener resultados que no reflejen el estado verdadero del fenómeno que se está investigando. La desviación del resultado puede ser atribuida a dos fuentes de error y su presencia en gran magnitud no permite tener conclusiones válidas.

Un estudio se desarrolla a partir del vacío de conocimiento en un área y con relación a un estado de un fenómeno desconocido por quien inicia el proceso de investigación. Puede afirmarse que el investigador siempre parte de la existencia de una verdad que desconoce, pero que, a través de un estudio clínico o epidemiológico, pretende aclarar (conocer) y ampliar su conocimiento del fenómeno para tomar decisiones.

En la búsqueda de aclarar esta verdad que subyace en una realidad, la forma en que se determinen los procedimientos y algunos de estos, *per se*, pueden introducir errores que lleven a subestimar o sobreestimar la verdad, esto se denomina en el ámbito de la metodología de la investigación científica sesgo o desvío del resultado.

De manera general, este desvío puede ser provocado por dos tipos de errores que se pueden cometer durante la planeación y el desarrollo de una investigación: el error aleatorio y el error sistemático o sesgo.

Sesgo

Desvío o falso resultado en un proceso de investigación científica.

Figura 1.
Fuente: Shutterstock/243096655

Error aleatorio

Este error es constante en toda investigación, debido a que generalmente se trabaja con una muestra y no con toda la población, es decir, con un subgrupo que representa dicha población y en el que también está representado el fenómeno sobre el cual se quiere tener conocimiento. Sin embargo, esta medición del fenómeno en una única muestra no deja de ser más que una aproximación al valor real del fenómeno. Esta aproximación es conocida en el ámbito de la estadística como una estimación puntual. En estas estimaciones, al ser generalmente calculadas en una muestra única, cabe siempre la posibilidad de que se dé un valor diferente si se selecciona una segunda muestra de la misma población. Ocurriría lo mismo si se tomara una tercera muestra, quedando claro que al obtener conclusiones sobre una estimación de una única muestra se tiene una alta relación con lo que se denomina incertidumbre. La incertidumbre se define mediante la siguiente pregunta: si tomase una muestra en la que se estimó un valor y luego tomase una segunda muestra para estimar el mismo valor, ¿los resultados serían diferentes? La pregunta que queda es: ¿cómo podría establecer que bajo condiciones de desarrollo de una investigación y solo teniendo la posibilidad de obtener una única muestra esta represente un valor acertado frente al valor real del fenómeno investigado?

El error aleatorio está directamente relacionado con trabajar con una porción de individuos (muestra) y no con el total de la población (universo) que permitiese conocer el valor real del fenómeno estudiado. Para efectos prácticos, el error aleatorio siempre está presente y puede ser causal de que las estimaciones sean alejadas de una verdad y, por ende, las conclusiones derivadas sean incorrectas. Sin embargo, la estadística antepone herramientas para controlar el problema de azar (error aleatorio) que se produce al trabajar con una única muestra de una población. Estas herramientas están encaminadas a la minimización del error aleatorio. Primero minimizándolo desde el diseño, es decir, ajustando un tamaño de muestra que sea lo suficientemente grande y permita controlar un grado de error máximo admisible por el investigador; segundo, usando técnicas de estadística de muestreo y estimación que cuantifiquen el error aleatorio presente en una medición y permitan reportarlo de manera tal que el lector pueda estar informado y lo integre a sus conclusiones.

Como un ejemplo de la presencia de este error se expone un caso:

Ejemplo

Suponga que se desea conocer el valor real del efecto (verdad subyacente que poca o ninguna vez se conoce) que tiene un nuevo antihipertensivo sobre la mortalidad por eventos cardiovasculares en una población de pacientes con hipertensión moderada a severa. Este efecto es expresado en la forma de diferencia de riesgo absoluto de muerte (existen otras medidas que serán revisadas más adelante) por eventos cardiovasculares. Este efecto puede tomar valor cero (0) indicando que no hubo diferencias en la mortalidad por estos

eventos entre quienes toman el antihipertensivo nuevo comparado contra algún medicamento convencional. Puede tomar valores de uno (1) indicando que el antihipertensivo nuevo reduce comparativamente con el antihipertensivo convencional el 100% de las muertes. Además, supongamos que este valor real del efecto es de 0,7. Posterior a esto, tenemos tres estudios diferentes con igual tamaño de muestra y probando la efectividad del nuevo antihipertensivo. Cada estudio obtuvo una muestra de 20 pacientes, 10 recibieron el nuevo fármaco y los restantes el medicamento convencional. En ambos grupos se midió la mortalidad. Los resultados reportados fueron:

Estudio 1 n = 20	Estudio 2 n = 20	Estudio 3 n = 20
DR = 0,46	DR = 0,65	DR = 0,73

Tabla 1.
Fuente: propia

Como se aprecia, los valores del efecto estimado puntualmente en cada estudio de hecho son diferentes. La población de donde fueron extraídas las muestras para estos estudios fue la misma, es decir, pacientes hipertensos susceptibles de beneficiarse de la nueva terapia.

Como se observa en el ejemplo anterior, por simple azar los resultados son diferentes en cada estudio. Si no se conoce el verdadero valor del efecto (normalmente es lo que intenta un estudio conocer), ¿cómo se puede saber que uno de estos estudios es una buena aproximación a ese valor verdadero, contando con la oportunidad de realizar el estudio una sola vez por costos? Esto deja entrever la presencia del error aleatorio en todo estudio realizado; sin embargo, de nuevo la estadística provee técnicas de estimación (serán revisadas más adelante) que se basan en la información de una única muestra, pero con la capacidad de calcular el nivel de error aleatorio o por azar que está relacionado con la medida puntual obtenida en la muestra, y permite hacer la inferencia sobre cuál sería la variabilidad observada si se hubiera realizado el mismo experimento, pero en muestras repetidas. Algunas técnicas nos permiten estimar qué valor de probabilidad está asociado a cometer un nivel de error cuando se trabaja con esta única muestra.

En conclusión, el error aleatorio está presente en todos los estudios. Puede ser identificado antes del estudio, controlado o minimizado, así como cuantificado para su asociación con el resultado puntual obtenido en la muestra. El azar y su control en los estudios clínicos y epidemiológicos tiene manejo con técnicas estadísticas.

Error sistemático o sesgo

Este error tiene un origen y un control totalmente distintos al aleatorio. Se origina cuando la implementación de procedimientos dentro de la investigación tiende a desviar sistemáticamente un resultado, alejándose del valor real del fenómeno estudiado.

Principalmente, se encuentra presente el sesgo en estudios dentro de la implementación de procedimientos de medición como tal, por ejemplo, cuando para determinar el efecto de un factor de riesgo ante una condición, esta es medida bajo protocolos o instrumentos no aptos o no validados para tal fin. Este es el caso de la determinación del uso de una prueba diagnóstica para su clasificación, la prueba *per se* puede presentar problemas con sus probabilidades de clasificación (sensibilidad y especificidad), es decir, en su capacidad para detectar la enfermedad o para descartarla. De tal manera, este error genera una mala clasificación entre los sujetos y los errores están relacionados directamente con el desvío en la estimación real del fenómeno estudiado. Las desviaciones de estos resultados son: mostrar una magnitud y fuerza más altas de las reales o, por el contrario, más bajas. En el peor de los casos, serían contrarias a las reales.

El sesgo principalmente se ve expresado cuando los procedimientos escogidos por el investigador hacen que las mediciones se desvíen de su valor real, siendo de esta manera aspectos metodológicos del diseño de la investigación los responsables del sesgo. Dichos aspectos metodológicos son controlados solo cuando se planean, no siendo posible su control posterior cuando se obtienen resultados. El problema del error sistemático es eminentemente metodológico, es decir, de proceder. No existe ningún control más que tomar o aplicar correctamente los procedimientos en una investigación.

Ejemplo

Continuando con el ejemplo anterior, la diferencia en los resultados de los tres estudios expuestos puede ser atribuida a que los grupos analizados (10 sujetos por grupo en cada estudio) hayan presentado discrepancias entre ellos por la presencia diferenciada de un factor pronóstico. Por ejemplo, uno de los grupos podría tener en promedio una edad 10 años mayor, comparado con el otro. Esta diferencia de edad puede tener como efecto que los pacientes respondan con menor efectividad a la intervención y, así, el efecto visto no sea real.

Como se puede evidenciar, el sesgo y el error aleatorio influyen sobre los resultados finales de cualquier investigación y son los responsables de la inexactitud y la imprecisión de una estimación. En términos prácticos, un paso fundamental en el análisis basado en la evidencia es tratar de identificar si los resultados observados en una investigación se vieron afectados por algunos de estos errores y, de esta manera, evaluar la validez de las estimaciones y su disponibilidad para ser aplicadas en la práctica diaria.

Como se especificó en el eje 1, cuatro escenarios prácticos permiten establecer la aplicación de métodos basados en la evidencia. Las evidencias (investigaciones clínicas) que se recuperan con búsquedas deben ser analizadas con una mirada crítica en su desarrollo metodológico y en los resultados reportados. Para cada uno de estos escenarios se desarrollan aspectos metodológicos y de interpretación de los resultados posibles que permiten establecer estándares mínimos a evaluar en la validez de los resultados en cada escenario, es decir, en cada tipo de estudio que responde a un tipo de pregunta clínica planteada.

Ensayos clínicos

Figura 2.
Fuente: Shutterstock/691186021

Cada estudio debe ser evaluado en cuanto a la validez de su resultado, lo cual puede ser evidente mediante la valoración de aspectos que pudieron haberse afectado por la presencia de error aleatorio y del sesgo.

La situación de duda o desconocimiento sobre la efectividad de los resultados clínicos de un tratamiento o intervención en el paciente está sobre la mesa. Solo puede ser abordada con evidencia de un estudio clínico experimental, es decir, un ensayo clínico en el que la intervención es comparada de manera controlada frente a una alternativa, comúnmente una terapia estándar o, en su defecto, con un placebo. En la literatura científica, habitualmente estos estudios son denominados artículos que "prueban la efectividad de una terapia" y son los que de manera mayoritaria aportan el máximo nivel de evidencia en la toma de decisiones, esto atribuido al carácter experimental del diseño en este tipo de estudio.

Como todo diseño de investigación, este está sujeto a errores aleatorios y sistemáticos que pueden detectarse bajo el no cumplimiento de aspectos metodológicos mínimos en su desarrollo, implementación y análisis. Se abordarán a continuación algunos de estos aspectos a tener en cuenta en un futuro análisis crítico de ensayos clínicos.

Validez de los resultados en un ensayo clínico

Uno de los principales aspectos a determinar en un ensayo clínico para observar una diferencia entre los grupos evaluados (intervención versus control) es que dichos grupos están conformados de manera similar. Se podría afirmar que deberían estar sin diferencias al inicio del ensayo, esto permitiría establecer fácilmente que ningún factor pronóstico con presencia mayoritaria en uno de los grupos pudiera favorecer el resultado final.

Una de las razones por las que esto sucede (diferencias en factores pronósticos entre grupos al inicio del estudio) es que actúa la preferencia del paciente o del médico en determinar si un paciente es asignado al grupo de tratamiento o control. Diferentes factores o variables pronósticos pueden estar relacionados con la preferencia a recibir una intervención u otra dentro de un estudio. La edad del paciente, la severidad de su enfermedad y la comorbilidad son algunos ejemplos de estos factores, a los que también se les puede atribuir su contribución al resultado visto en una variable de desenlace principal tomada como medida de efectividad de la terapia, es decir, pueden tener un efecto sobre la medida evaluada, por ende, si se observan diferencias entre dos grupos en un ensayo clínico, podrían no solo ser o solo ser atribuibles a estos factores pronósticos y no a la intervención o terapia en evaluación.

Aleatorización

Para dar solución al sesgo que se produce por tener grupos desbalanceados frente a factores pronósticos conocidos, el ensayo debe implementar dentro de su diseño la aleatorización, importante técnica que consiste en la asignación aleatoria (probabilística) que da iguales posibilidades de ser asignado a un grupo de tratamiento como al grupo de control. El poder de la aleatorización es que los grupos de tratamiento y control tienen más probabilidades de ser equilibrados con respecto a los factores pronósticos conocidos y desconocidos. Aunque la aleatorización es una técnica poderosa, no siempre crea grupos con pronóstico similar. Los investigadores pueden cometer errores que comprometen la aleatorización o la misma puede fallar debido a la mala suerte. Como punto suplementario, los investigadores que están reclutando pacientes en un ensayo pueden consciente o inconscientemente distorsionar el equilibrio entre los grupos, por tal motivo, esta aleatorización debe ser también encubierta al investigador. Generalmente, se utilizan centros externos que asignan mediante diferentes estrategias a qué brazo del estudio pertenece el paciente.

Otra causa del desequilibrio que se puede dar al inicio de un ensayo clínico es la baja cantidad de pacientes, de tal manera que el bajo número no permite que los mismos se asignen de manera homogénea. Finalmente, se puede verificar cómo funcionó la aleatorización cuando se despliegan resultados para ambos grupos en términos de los factores pronósticos tomados en cuenta en el estudio. Las diferencias deben ser mínimas o nulas. Cuando esta aleatorización no funciona, el resultado no está perdido.

Existen técnicas estadísticas que ajustan los resultados por dichas diferencias observadas y deben ser aplicadas por los investigadores.

Enmascaramiento

Si la aleatorización funciona, los grupos de tratamiento y control comienzan con un pronóstico similar. La aleatorización, sin embargo, no ofrece garantías de que los dos grupos seguirán siendo equilibrados en los factores pronósticos medidos al inicio. El enmascaramiento es la óptima estrategia para mantener el balance del pronóstico.

En la medida que sea posible, se deben enmascarar los siguientes actores en el desarrollo del ensayo clínico: pacientes, investigadores, recolectores de información, personal que mide el o los desenlaces principales y analista de datos, cada uno con efecto distinto. La aplicación del enmascaramiento garantizará mayor validez del resultado.

Pérdida de seguimiento

La realización de un ensayo clínico suele ser demorada y difícil de completar correctamente. Si se controla adecuadamente, menos del 80 % de los pacientes, los resultados pueden verse sesgados.

Idealmente, al final de un ensayo se debe conocer en cada paciente el estado de la variable clínica de efectividad seleccionada. Cuanto mayor sea el número de pacientes al cual se les desconoce su estado final (pacientes perdidos en el seguimiento), mayor afectación en la validez tendrá el estudio. La razón es que los pacientes con peores pronósticos pueden ser retenidos, mientras que quienes tengan resultados adversos o su resultado es bueno, pueden desaparecer.

Tiempo de estudio

Los estudios deben permitir tiempo suficiente para que los resultados que se están midiendo se manifiesten. El lector debe usar su juicio clínico para decidir si esto era cierto para el estudio que está evaluando y si la duración del seguimiento era apropiada para los resultados que le interesan.

Intención a tratar

Los investigadores también pueden afectar la aleatorización si se omiten de los análisis pacientes que no reciben el tratamiento asignado o, peor aún, cuentan los eventos que ocurren en pacientes no adherentes que fueron asignados al tratamiento versus el grupo de control. Dichos análisis sesgan los resultados si las razones de la no adherencia están relacionadas con el pronóstico.

Los investigadores previenen este sesgo cuando siguen el principio de intención de tratar y analizan a todos los pacientes del grupo al que fueron asignados al azar, independientemente de su adherencia.

Resultados en un ensayo clínico

Tras evaluar en primera instancia los aspectos metodológicos más importantes que garantizan en gran medida el control de sesgos en los ensayos clínicos, pasamos a evaluar los resultados que se reportan en estos estudios, los cuales son variados de acuerdo con el tipo de variable (cuantitativa o categórica) que fue medida como desenlace. Ante todo, debe tenerse claro que las medidas utilizadas frecuentemente son de origen estadístico y están basadas en la comparación de, al menos, dos grupos (tratamiento y experimental). Para efectos de sencillez en los elementos básicos de las medidas, se utilizarán únicamente ejemplos de un ensayo con dos brazos: un grupo que recibió el tratamiento nuevo o experimental y otro que recibió un tratamiento control, estándar o, como mínimo, placebo.

Las variables que pueden ser seleccionadas para medir la efectividad de una intervención pueden clasificarse de dos formas. Las llamadas cuantitativas hacen referencia a datos de tipo numérico. Generalmente, aluden a biomarcadores o bioclínicas como la tensión arterial, la concentración de un metabolito en sangre y la carga viral, todas dan como resultado de medírselas en el sujeto un valor numérico.

La otra forma de variable es la categórica. Asigna al sujeto un atributo entre unos posibles, por ejemplo: la curación (sí, no), la mortalidad (muerto/vivo), la aparición de infarto o no, la infección o no, etc. Se pueden colocar más de dos categorías entre sus posibilidades; sin embargo, para dar sencillez didáctica, se trabajará con variables categóricas con únicamente dos posibles resultados, también llamadas dicotómicas.

Variable cuantitativa

La medida más utilizada para estimar la asociación entre un tratamiento y un desenlace es utilizar un estimador que resuma el resultado de dicha variable cuantitativa en cada grupo en evaluación (intervención vs. control) y compare la diferencia entre los resultados. Estimadores como la media (o promedio) son frecuentemente utilizados, de esta manera, se procede a calcular el promedio entre el grupo de intervención y el promedio en el grupo de control, luego, se restan ambos promedios y se obtiene la medida denominada diferencia de medias.

Diferencia de medias

Estimador de las diferencias de dos medias en dos poblaciones independientes, clasificado junto con otros estimadores como tamaño del efecto.

Si dicha diferencia es igual a cero (0), esto se interpreta como que no hay diferencia alguna entre los resultados de un grupo experimental comparado con el grupo control. Por el contrario, si dicha diferencia absoluta tiene un valor distinto a cero, se apreciará que existe una diferencia en el resultado clínico evaluado atribuible a la intervención o al control.

Ejemplo

Supongamos que se evalúa un nuevo régimen profiláctico de anticoagulante (única dosis/día vs. una dosis/día fraccionada) en pacientes con riesgo de trombosis venosa profunda. Para evaluar su efectividad, se tomó como variable de desenlace la TP en segundos. Al finalizar el ensayo, se estimó que, por término medio, el grupo de una dosis diaria obtuvo un TP de 25 segundos, mientras que el grupo de una dosis/día fraccionada obtuvo un TP promedio de 35 segundos. La diferencia de medias sería:

$$m_1 = 25$$

$$m_2 = 35$$

$$Diff = 35\text{-}25 = 10$$

Por tanto, la diferencia obtenida corresponde a que, en promedio, el régimen de una dosis fraccionada alargó 10 segundos más el TP por encima del grupo de una única dosis, mostrando un efecto mayor sobre la anticoagulación.

Variable categórica

Cuando la variable de desenlace se mide de una forma dicotómica, es decir, con la posibilidad de dos atributos de tipo infectado/no infectado, curado/no curado, vivo/muerto, los resultados del ensayo clínico se pueden representar en una tabla de contingencia de 2x2.

Figura 3.
Tabla de 2x2 para representar los resultados de un ensayo clínico y las medidas para resumir la efectividad de un tratamiento
Fuente: propia

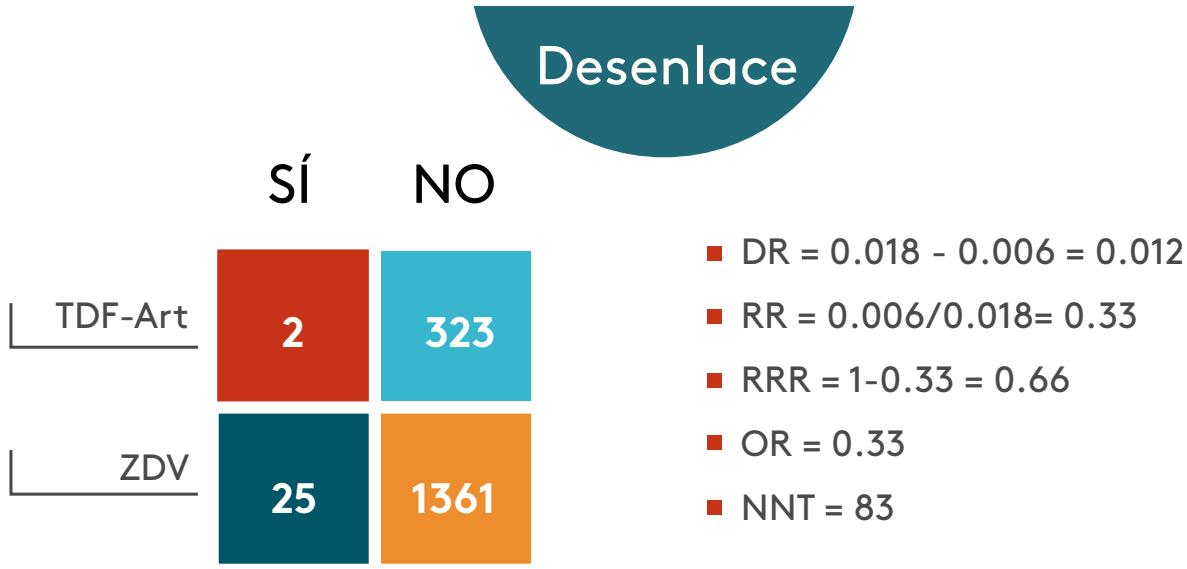


Figura 4. Resultados de un ensayo clínico comparando Tenofovir (TDF) con base en ART con Zidovudina únicamente (ZDV) para prevención de VIH perinatal
Fuente: Fowler et ál. (2016)

La primera medida a ser calculada es el riesgo. Se denomina riesgo a la posibilidad de desarrollar el evento o desenlace dentro de un grupo al que se fue asignado (experimental o control). También es llamado tasa del evento y se le asigna el nombre de riesgo de base al calculado entre el grupo de control. Es importante aclarar que cuando se denomina riesgo no obligatoriamente implica peligro, pues en algunos ensayos el desenlace principal no necesariamente es negativo (mortalidad, complicaciones), también puede ser positivo (curación, mejoramiento en severidad de la enfermedad). Entonces, no se hablaría de riesgo, sino de proporción de sujetos que hacen el evento. De manera general, se maneja el término riesgo. Se debe tener en cuenta siempre el desenlace.

Riesgo (proporción de evento) grupo experimental = a/a+b; valores posibles [0,1]

Riesgo (proporción de evento) grupo control = c/c+d; valores posibles [0,1]

Las siguientes medidas se desprenden del uso de los riesgos en ambos grupos.

Una primera medida intuitiva, como se mostró en la diferencia de medias, es obtener la diferencia entre los riesgos de ambos grupos, esta es denominada diferencia de riesgos (DR) o reducción de riesgo absoluta.

DR = (a/a+b) - (c/c+d); DR con valores posibles en el intervalo cerrado [0,1]

La interpretación de la DR es la menos intuitiva, debido a que está evaluando de manera absoluta la diferencia de riesgo generada a causa de la intervención. Clínicamente, no podemos valorar de una manera fácil cuánto riesgo es suficiente o cuánto riesgo es alto o tan alto como para alarmarse. Valores de DR como en el ejemplo de la

tabla 3, el riesgo de infección perinatal de VIH para los grupos control (ZDV) y experimental (TDF-Art) fueron de 0,018 (1,8 %) y 0,006 (0,6 %), respectivamente, mostrando una diferencia de riesgo de 0,012 (1,2 %), para lo cual es claro que cuantitativamente hubo una diferencia de 1,2% en el riesgo de infección entre los dos antirretrovirales, siendo este resultado a favor de la terapia antirretroviral con TDF. Sin embargo, valorar si dicha diferencia tiene una magnitud importante se hace difícil, pues no tenemos la percepción clara cuantificada del riesgo medido. Quizá para el paciente no se ha valorado de manera importante esta diferencia como para someterse a una nueva terapia.

Otra forma de medida de resultado, pero con un carácter relativo, a diferencia de la anterior, es comparar los grupos mediante un cociente. Es medida es llamada riesgo relativo (RR) o razón de riesgos.

RR = a/a+b / c/c+d; RR con valores posibles en intervalo abierto (0,)

La interpretación del RR resulta más fácil y con mejor percepción sobre el resultado, esto tanto para el clínico como para los pacientes. Al ser un cociente entre el riesgo (proporción del evento) del grupo experimental (numerador) y el riesgo (proporción del evento) del grupo control (denominador), muestra cuántas veces más riesgo de padecer un desenlace se tiene perteneciendo a un grupo comparado con el otro. Tomando de nuevo del ensayo de Fowler et ál. (2016), se estimaron tasas de infección de VIH perinatal que dan como resultado un RR de 0,33 (0,006/0,018), lo que significa que el TDF-Art redujo en una tercera parte el riesgo de infección entre quienes lo recibieron comparado con quienes recibieron ZDV.

Las dos anteriores medidas son algunas de las formas comúnmente vistas en las publicaciones de ensayos clínicos; sin embargo, existen otras medidas más informativas, incluso para transmitir los resultados de estos estudios a los mismos pacientes en quienes serán implementados las intervenciones evaluadas. Este es el caso de la reducción del riesgo relativo (RRR).

RRR = DR / riesgo del grupo control = (c/c+d - a/a+b)/(c/c+d)

Operando con las fracciones se encuentra que:

RRR = 1- RR; RRR con valores posibles en el intervalo cerrado [0,1]

La RRR hace referencia a la proporción de riesgo de base que es removida por el tratamiento en prueba. Por ejemplo, con el ensayo de transmisión perinatal de VIH, la RRR es 0,66, es decir, se estima que el TDF reduce en un 66 % el riesgo de transmisión de perinatal de VIH comparado con ZDV o, lo que es mejor, reduce en dos terceras partes el riesgo de infección perinatal que se da con ZDV.

Otra medida similar al RR para evaluar la relación entre un tratamiento y un desenlace es la razón de posibilidades (en inglés *odds ratio*) (OR). Se estiman las *odds* (posibilidades) en cada grupo (experimental y control), se calcula el cociente de las *odds* y esto da la OR. El resultado es equivalente al RR, su expresión de cálculo es:

OR = (a/b)/(c/d) = ad/cb; OR con valores posibles en el intervalo abierto (0,)

También se puede expresar el impacto del tratamiento por el número de pacientes que se necesitaría tratar para prevenir un evento o desenlace. Esto es llamado el número necesario para tratar (NNT).

NNT = 1/RD; NNT con valores posibles en el intervalo abierto (0,)

Si se dice que un RD = 1,2 %, entonces, si se tratan 100 pacientes con TDF-Art, se evita un evento de transmisión de infección perinatal. Ahora bien, ¿cuántos pacientes se requiere tratar para evitar una infección perinatal de VIH?, la respuesta sería NNT = 1/0,012 = 83 pacientes.

Daño o etiología

Figura 5.
Fuente: Shutterstock/146441699

En este nuevo escenario práctico, las preguntas clínicas están dirigidas a conocer el daño producido por una exposición o el papel etiológico que juega una exposición frente a una condición que, generalmente, es una enfermedad o evento negativo.

Ahora bien, los estudios que permiten dar respuesta a este tipo de pregunta bien podrían estar en el marco de un diseño de investigación tipo ensayo clínico. Aunque los investigadores lo realizan para determinar si los agentes terapéuticos son beneficiosos, también deben buscar efectos dañinos y, a veces, pueden hacerse descubrimientos sobre los efectos negativos de la intervención en sus resultados primarios. Este diseño tiene el poder de proveer la evidencia más contundente frente a una exposición, ya que incluye metodológicamente la técnica de aleatorización, la cual permitirá tener grupos balanceados, en términos de factores pronósticos conocidos y desconocidos, y controlará el efecto de estos sobre cualquier desenlace evaluado.

De otra manera, existen varias razones por las que un ensayo clínico no puede implementarse para evaluar el daño producido por una exposición. Se mencionan las tres más importantes que tienen las justificaciones más fuertes. En primer lugar, se considera que no es ético aleatorizar pacientes a exposiciones que de manera anticipada (bajo hipótesis de investigación) se espera que traigan efectos nocivos. En segundo lugar, a menudo hay preocupación por los efectos adversos raros y graves que pueden llegar a ser evidentes solo después de que decenas de miles de pacientes han consumido un medicamento durante años. Incluso ensayos clínicos muy grandes no logran detectar una asociación entre una exposición (tratamiento) y algún evento adverso. En tercer lugar, los ensayos clínicos a menudo no reportan adecuadamente la información sobre el daño.

Según lo indicado hasta el momento, el clínico o practicante interesado en contestar una pregunta clínica orientada sobre el efecto dañino de una exposición debe dirigir la búsqueda de evidencia hacia los estudios observacionales, como son los estudios de cohorte y los de casos y controles.

En un estudio de cohorte el investigador identifica a grupos expuestos y no expuestos de pacientes, cada uno de una cohorte, y luego los sigue en el tiempo, controlando la ocurrencia del resultado previsto. El diseño de la cohorte es similar a un ensayo clínico, pero sin aleatorización. La determinación de si un paciente recibió la exposición de interés resulta de la preferencia del paciente o de la casualidad.

Los estudios de casos y controles también evalúan las asociaciones entre las exposiciones y la enfermedad. Los resultados raros o los que tardan mucho tiempo en desarrollarse pueden amenazar la viabilidad de los estudios de cohorte. El estudio de casos y controles proporciona un diseño alternativo que se basa en la identificación inicial de los casos, es decir, en pacientes que ya han desarrollado la enfermedad y en la selección de controles, personas que no tienen la enfermedad de interés. Utilizando diseños de casos y controles, los investigadores evalúan la frecuencia relativa de la exposición previa a ser caso o control.

Teniendo claro lo anterior, al momento de encontrar un estudio de este tipo para tomar decisiones sobre el paciente, se debe tener, al igual que en cualquier investigación clínico-epidemiológica, la posibilidad de la presencia del error aleatorio y del sesgo. Por tanto, profundizaremos en aspectos metodológicos que tienen relación con la validez de los resultados (sesgos) y con la forma de reportar los resultados de estos estudios y la valoración de la precisión de las estimaciones (error aleatorio).

Validez

Argumentos necesarios para calificar un resultado como certero o verdadero de una realidad.

Validez de los resultados en estudios observacionales (cohorte y casos y controles)

Estudios de cohorte

Comparabilidad entre expuestos - no expuestos y su ajuste

Figura 6.
Fuente: Shutterstock/566945398

El investigador puede desarrollar un estudio de cohorte retrospectiva o prospectivamente. En el primer caso, la exposición y el desenlace o daño evaluado ocurrieron antes del inicio del estudio, por tanto, el investigador debe retroceder para recolectar datos sobre el desenlace y la exposición, para lo cual debe tenerse en cuenta la calidad de las fuentes de donde serán tomados los datos. En un estudio prospectivo, el investigador inicia conformando los grupos de exposición entre individuos libres del daño. Para ambos casos, se debe evaluar qué diferencias existen en factores pronósticos, probablemente también relacionados con la aparición de la enfermedad o el daño. Este desbalance es ocasionado porque los grupos son conformados por el estado natural de cada participante, es decir, el investigador no interviene en asignar dicha exposición (como sí ocurre en los ensayos clínicos), de tal modo que los grupos son conformados de forma no aleatoria y bajo esta situación pueden ser diferentes.

Los estudios de cohorte de exposiciones potencialmente dañinas darán resultados sesgados si el grupo expuesto al agente nocivo y el grupo no expuesto comienzan con diferentes características basales que les dan un pronóstico diferente. Este desbalance no puede ser evitado desde el diseño del estudio (como sí ocurre en ensayos clínicos por la aleatorización), dado que es propio del diseño de observacionales, pero, desde una forma analítica (métodos estadísticos), se puede controlar el efecto de dicho desbalance en los factores pronósticos identificados.

Debe tenerse en cuenta que, para realizar dicho ajuste por diferencias entre los grupos, las variables pronósticas deben ser medidas de manera exacta, es decir, debe ponerse especial atención en los métodos de medición que el investigador utiliza para valorar su presencia. También aplica esta condición para la medición de la exposición y el mismo desenlace o daño evaluado, ya que sesgos de información relacionados a estas variables generan estimados sesgados de la relación exposición daño/enfermedad.

Medición del desenlace en los grupos

No se desea que un grupo sea estudiado más exhaustivamente que el otro porque esto podría conducir a reportar una mayor incidencia de daño/enfermedad en el grupo más estudiado. Condiciones especiales dentro de la cohorte pueden hacer que los investigadores busquen más exhaustivamente el daño/enfermedad entre ellos y hagan ver que hay un mayor riesgo de padecer el daño/enfermedad entre un grupo comparado con el otro, mostrando al final una asociación exposición-enfermedad falsa o aumentada en magnitud.

Desenlace

Medida que se toma como la variable efecto o dependiente en la comparación de dos grupos independientes o dependientes.

Seguimiento de los individuos

Al igual que sucede en los ensayos clínicos, la pérdida de seguimiento puede introducir sesgo porque los pacientes que se pierden pueden tener resultados diferentes frente a los pacientes que están disponibles para la evaluación. Se espera que el seguimiento sea lo suficientemente largo para que los eventos se expresen y lo adecuadamente completo sobre todos los individuos participantes.

Estudios de casos y controles

Al igual que los estudios de cohorte, los estudios de casos y controles son susceptibles a variables de confusión no medidas, particularmente cuando la exposición varía con el tiempo.

Medición de la exposición en los grupos

En los estudios de casos y controles la determinación de la exposición es un problema clave. Si los casos tienen un mejor recuerdo para reportar la exposición que los pacientes control, el resultado será una asociación espuria. También existen problemas en relación con los métodos con los que se valora y clasifica el grado de exposición entre casos y controles. Esto puede generar una mala clasificación y que las estimaciones de la asociación sean sesgadas.

En el caso de estudios observacionales también incluyen otros diseños como el corte transversal y la serie de casos. En el primero, a partir de una muestra aleatoria de una población, se obtienen mediciones en un momento dado del tiempo sobre la exposición y enfermedad, por tanto, el problema principal de este diseño radica en que no hay una clara dirección de la temporalidad entre exposición y daño, siendo difícil establecer el papel causal de la exposición, ya que no se puede determinar qué ocurrió primero: la exposición o la enfermedad.

En el diseño de serie de casos, el problema radica en no proporcionar un grupo de comparación, por lo que es imposible determinar si el resultado observado probablemente habría ocurrido en ausencia de la exposición.

Resultados en estudios de cohorte y de casos y controles

Los resultados de los estudios de cohorte y de casos y controles también se pueden presentar en una tabla de 2x2. Las medidas para expresar la relación exposición-enfermedad utilizadas para el análisis son RR y OR.

Figura 7. Tabla 2x2 para resultados de un estudio de cohorte (parte A) para estudio de casos y controles (parte B) y OR en resultados de un estudio de caso y controles entre asociación de infección oral por VPH y cáncer orofaríngeo (parte C)
Fuente: D'Souza et ál. (2007)

Diagnóstico

En este nuevo escenario o situación práctica, la pregunta PICO planteada hace referencia al ejercicio práctico del diagnóstico, es decir, la situación en que se ve involucrado el proceso de llegar a un diagnóstico en un paciente. El proceso diagnóstico proviene de un fenómeno probabilístico dirigido así: el practicante, con base en una entrevista, que incluye características del paciente, signos y síntomas, además de la epidemiología local de las enfermedades, establece un listado de posibles diagnósticos con valores de probabilidad para cada uno (probabilidad pretest). Posterior a esto, establece la realización de pruebas diagnósticas que intentan dar información adicional, lo que está relacionado con valorar de nuevo la probabilidad del diagnóstico ante esta nueva información (resultado de un test), la cual es llamada probabilidad postest y, finalmente, tomar una decisión.

En la práctica general, muchos test diagnósticos son objeto de evaluación en su desempeño diagnóstico, es decir, la capacidad que tienen para clasificar correctamente a los sujetos. Esto hace necesario que la toma de decisiones frente a las pruebas enviadas para un diagnóstico esté fundamentada en el análisis crítico de la literatura que soporta la utilidad de la prueba, allí es cuando surge desde los estudios la evidencia con la que se cuenta. Esta evidencia debe ser evaluada frente a la ocurrencia del sesgo y del error aleatorio, lo cual permita aplicarla de manera correcta y con evidencia científica.

Clásicamente, estos estudios de evaluación de pruebas diagnósticas se desarrollan bajo un diseño de investigación de corte transversal o de casos y controles, y algunos bajo un diseño de cohorte; sin embargo, para simplificar el desarrollo teórico se tendrán en cuenta los que son implementados como estudios descriptivos de corte transversal. Las medidas que resumen los resultados de estos estudios son iguales, independientemente del diseño utilizado por el investigador.

El estudio descriptivo de corte transversal utilizado en la evaluación de pruebas diagnósticas consiste en la selección de una muestra representativa de una población con el cumplimiento de criterios mínimos para sospecha de un diagnóstico (espectro de signos y síntomas). En esta muestra de sujetos se aplica el test o prueba nueva o en evaluación y se corrobora en ellos mismos, mediante una prueba *gold* estándar o de referencia, la presencia de manera real de la enfermedad o condición de diagnóstico en cuestión. Los datos que el investigador recolecta en este tipo de estudios dependen de los resultados arrojados por la prueba en evaluación. De esta manera, si el resultado dado por el test es un valor cuantitativo, por ejemplo, diluciones de una serología, nivel de hemoglobina y concentración de un biomarcador, los datos serán de tipo numérico. Si, por el contrario, los resultados son de tipo positivo/negativo, reactivo/no reactivo, se dará un dato de tipo cualitativo dicotómico. Para facilidad del lector y según lo que frecuentemente se ve en la literatura, se tomará en cuenta para el análisis de estos tipos de estudios aquellos que entregan datos de tipo cualitativo dicotómicos.

Validez de los resultados en estudios sobre pruebas diagnósticas

Espectro de la enfermedad

Los participantes en un estudio son seleccionados o muestreados de una población objetivo de personas que presentan signos y síntomas probablemente asociados a una enfermedad que se está investigando. Idealmente, esta muestra debe reflejar la población objetivo en todas las formas importantes, de modo que la frecuencia y las posibilidades de diagnóstico sean un reflejo de lo que sucede en la población. Cuanto más representativa sea la muestra, más exacta serán las probabilidades de enfermedad resultantes. Para verificar si esta representación de la población blanco está en la muestra, existen cuatro formas. En primer lugar, encontrar la definición que los investigadores dan al problema clínico que se presenta, porque esto determina la población objetivo de la cual los pacientes del estudio van a ser extraídos.

En segundo lugar, examinar el contexto del que los pacientes son reclutados. Los pacientes con el mismo problema clínico podrían presentarse en cualquiera de los diferentes entornos clínicos, un primer nivel de atención, urgencias o clínicas de referencia. La elección de dónde buscar atención puede involucrar varios factores, incluyendo la enfermedad, la disponibilidad de los diversos entornos, los hábitos de referencia de su médico o preferencias del paciente. Los pacientes en niveles más complejos de atención involucran mayor severidad, permitiendo esto que la posibilidad de presentación de la enfermedad sea más alta que en otros niveles. En tercer lugar, los métodos de los investigadores para identificar a los pacientes en cada sitio y cómo pudieron evitar algunos de los pacientes. Idealmente, deben reclutarse una muestra de todos los pacientes que buscan atención en los sitios de estudio para el problema clínico durante un periodo especificado. Si los pacientes no se incluyen consecutivamente, la inclusión de pacientes con diferentes trastornos subyacentes reduce la representatividad de la muestra y, de esta manera, reduce la confianza en la validez de los resultados. En cuarto lugar, examinar el espectro de gravedad y características clínicas de los pacientes en la muestra del estudio. ¿Se encuentran todas las variaciones importantes de este problema clínico en la muestra?

Evaluación diagnóstica definitiva

Los artículos sobre pruebas diagnósticas proveerán evidencia válida solo si los investigadores llegan a los diagnósticos correctos para los pacientes del estudio. Esto se logra verificando diferentes aspectos dentro del estudio.

La evaluación diagnóstica debe ser capaz de detectar todas las posibles causas del problema clínico en consideración. Debe haber exhaustividad en la búsqueda del diagnóstico, si no, las probabilidades de enfermedad estimadas serán sesgadas.

Deben estar descritos suficientemente todos los procedimientos diagnósticos llevados a cabo en el paciente, así como las reglas de decisión frente a llegar al diagnóstico definitivo, esto tiene que ver con la estandarización de los procedimientos realizados, de manera tal que los haga reproducibles en un contexto fuera de la investigación. Al no haber dicha estandarización, se puede provocar inexactitud en las estimaciones de las probabilidades de diagnóstico.

Por último, el estándar de referencia utilizado para dar clasificación final a cada sujeto en el estudio debe tener características que lo hagan reproducible. Tanto la prueba que se está evaluando como la estándar de referencia deberían llevarse a cabo en todos los pacientes del estudio. Por ejemplo, si la prueba bajo investigación resulta negativa, esto puede haber llevado a que el investigador decida no realizar la prueba de oro en pacientes con este resultado, lo cual generará una situación denominada en el diseño de estos estudios, como verificación parcial de la enfermedad. Esto debe estar informado dentro del artículo, además de cómo fueron tratados los análisis de datos en estos casos. Algo importante en este punto es que tanto la evaluación por estándar de referencia como por prueba deben haberse evaluado de manera enmascarada entre sí, garantizando la independencia de los resultados por la prueba en evaluación y por la prueba de oro.

Resultados en estudio de evaluación de pruebas diagnósticas

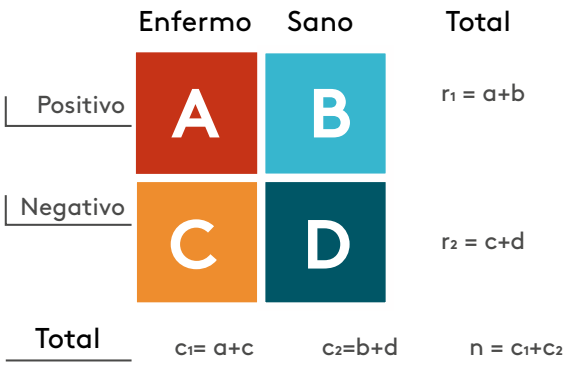
Partiendo de que los datos obtenidos en un estudio de evaluación de pruebas diagnósticas y los valores del test en evaluación tienen la forma de positivo/negativo, además de que la condición establecida por la prueba de referencia también constituye un resultado dicotómico (enfermo/sano), estos datos pueden organizarse en una tabla de 2x2.

Pruebas diagnósticas

Resultados de test (cuantitativos-cualitativos) objetivos o subjetivos, que clasifican a un individuo en términos de ausencia o presencia de un resultado.

La tabla de ejemplo muestra los resultados de un estudio de corte transversal realizado por Miller et ál. (2008), en el que se evalúa el desempeño diagnóstico de una tomografía de multidetección angiográfica utilizada para la identificación de la enfermedad coronaria (obstrucción).

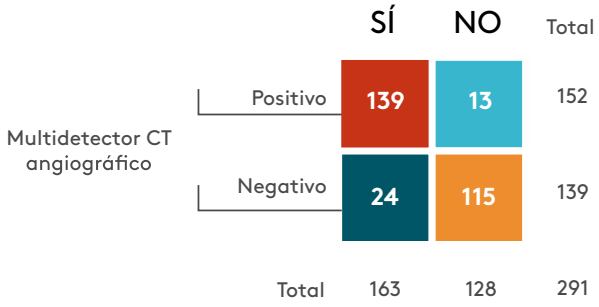
Gold estándar



- Prevalencia de la enfermedad o probabilidad de enfermar: $p = c_1/n$
- Sensibilidad (Se) = a/c_1 Falsos negativos (FN) = c/c_1
- Especificidad (Sp) = d/c_2 Falsos positivos (FP) = b/c_2
- Valor predictivo positivo (VPP) = a/r_1
- Valor predictivo negativo (VPN) = d/r_2
- Razón de probabilidad positiva (RP+) = Se/(1-Sp)
- Razón de probabilidad negativa (RP-) = (1- Se)/Sp

Figura 8. Tabla de 2x2 para estudios de evaluaciones de pruebas diagnósticas y sus medidas de desempeño
Fuente: propia

Enfermedad coronaria aguda



- Prevalencia de la enfermedad o probabilidad de enfermar p = 0.56
- Se = 0.85 FN = 0.15
- Sp = 0.9 FP = 0.10
- VPP = 0.91
- VPN = 0.82
- RP+ = 8.5
- RP- = 0.16

Figura 9. Resultados de un estudio evaluando el desempeño diagnóstico del multidetector CT angiográfico comparado con la angiografía convencional para el diagnóstico de enfermedad coronaria aguda
Fuente: Miller et ál. (2008)

Las medidas utilizadas están dadas en términos de las probabilidades de clasificación correcta de los sujetos, calculadas a partir de los resultados arrojados por el test en evaluación. La probabilidad de ser clasificado como positivo, dado que se está enfermo, es denominada sensibilidad y es estimada a través de la proporción de positivos entre los enfermos. Da cuenta de la capacidad del test para detectar a los sujetos en estudio como enfermos. La probabilidad de ser clasificado como negativo, dado que se está sano, es denominada especificidad y es estimada a través de la proporción de negativos entre los sanos. Lógicamente, asociadas a estas probabilidades de clasificación correcta también están las probabilidades de clasificación incorrecta: la proporción de falsos positivos y falsos negativos. Los valores posibles para estas medidas están en el intervalo cerrado [0,1]. Para lo cual, sensibilidad y especificidad de una prueba con valores de 100 % sería lo ideal.

Las probabilidades de clasificación correctas e incorrectas son medidas intrínsecas del desempeño de una prueba que, finalmente, solo son de utilidad e informativas para la investigación. En la praxis médica del diagnóstico hay otras medidas que suelen ser más informativas frente a la toma de decisiones para confirmar un diagnóstico sobre un individuo que consulta con una sintomatología y le es enviada una prueba diagnóstica, la cual a vuelta de consulta trae un resultado positivo o negativo. A partir de esta información, al médico o practicante le es de más utilidad conocer la probabilidad de que el paciente tenga o no la enfermedad sospechada, con base en que trae un resultado de un test. Esta es la utilidad de calcular los valores predictivos, los cuales son: predictivo positivo, que refleja la probabilidad de tener la enfermedad sospechada cuando se trae un resultado positivo, y el predictivo negativo, que se refiere a la probabilidad de estar sano cuando el resultado de la prueba fue negativo. Estos valores predictivos se estiman a través de la proporción de enfermos entre los sujetos con resultados positivos y la proporción de sanos entre sujetos que tienen un resultado negativo.

Un inconveniente ampliamente descrito por la teoría de probabilidades y que puede demostrarse empíricamente es la fuerte asociación o dependencia entre los valores predictivos y la prevalencia o probabilidad de enfermar, lo cual dificulta el uso de estos en la práctica, principalmente. Lo anterior se atribuye a que la prevalencia de la enfermedad varía dependiendo de las características de la población y el sitio donde se realiza el estudio. A mayor prevalencia, el valor predictivo positivo de la prueba aumenta. Si la prevalencia es baja, el valor predictivo negativo aumenta. Dada esta relación, y que la prevalencia varía según el sitio y las características de la población base de donde se extrae la muestra, los valores predictivos pueden cambiar fácilmente de acuerdo con el contexto de desarrollo del estudio, situación que no sucede con las probabilidades de clasificación correctas e incorrectas.

Para solventar dicho traspié, se plantean dos nuevas medidas que eliminan esta dependencia entre los valores predictivos y la prevalencia de la enfermedad. Las razones de probabilidad (*likelihood ratio*) permiten evaluar el desempeño diagnóstico de la prueba bajo la aplicación práctica de su resultado en escenarios clínicos, independientemente de la prevalencia. La razón de probabilidad positiva es el cociente entre la probabilidad de tener un resultado positivo entre enfermos y la probabilidad de tener un resultado positivo entre sanos, por tanto, su interpretación es: cuántas veces es más probable un

resultado positivo entre enfermos comparados con los sanos. En referencia a la razón de probabilidades negativa, que corresponde al cociente entre la probabilidad de tener un resultado negativo entre enfermos y la probabilidad de tener un resultado negativo entre sanos, de la misma forma su interpretación sería: cuántas veces es más probable ver un resultado negativo entre enfermos comprados con los sanos.

Precisión de los resultados en los estudios clínico-epidemiológicos

Este apartado del eje hace referencia a la cuantificación del error aleatorio presente en los resultados de un estudio. Debido a que las formas de expresar dicho error son transversales a las diferentes formas de análisis de los resultados, se explicarán de manera general, siendo su diferencia particular solo en la aplicación de fórmulas explicitas de acuerdo con la medida a la cual se le pretende calcular el nivel de error estadístico asociado a su estimación. Sin embargo, no es objeto de este documento presentar una exhaustiva formulación que se aleja del objetivo principal del análisis basado en la evidencia.

La idea de evaluar el grado de error aleatorio en los resultados de un estudio es determinar la precisión de los resultados obtenidos. Se parte de que el resultado puntual obtenido con los datos de una muestra es solo una aproximación al valor verdadero en la población de interés. Como se mencionó, esta medida puntual está asociada a un error por variabilidad muestral (si se repite el estudio con una muestra diferente de la misma población, el resultado puntal será diferente). Técnicas estadísticas de inferencia permiten calcular el grado de error asociado a esta variabilidad muestral, permitiendo obtener un posible rango de valores, en el que podría estimarse se encuentra el valor verdadero en la población, teniendo como insumos únicamente los datos de la muestra. Las dos técnicas principalmente utilizadas son los intervalos de confianza y los contrastes de hipótesis, que son aplicables a todos los tipos de análisis revisados hasta el momento (RR, OR, DR, Se, Sp, VPP, VPN, etc.).

Intervalos de confianza

Los intervalos de confianza son una técnica inferencial que permite estimar por medio de un rango posible de valores en los cuales, con cierto nivel de seguridad o confianza, se espera que esté el valor verdadero del parámetro poblacional. Para realizar esto, los intervalos de confianza integran dentro de su formulación una cantidad denominada "error estándar", la cual corresponde a la variabilidad que se da en una estimación si se calculara en muestras diferentes.

Intervalos de confianza

Técnica de inferencia estadística para generalizar resultados a una población partiendo de una muestra aleatoria de la misma.

Este error estándar es la clave para construir el intervalo, pues supone la cantidad de incertidumbre que se genera haciendo los cálculos en diferentes muestras. Asociándolo al valor puntual obtenido en la muestra del estudio, se puede hablar de que el valor aproximado se encuentra en la medida si se hubiera evaluado toda la población.

La forma genérica para cualquier intervalo de confianza que se le construya a una medida tiene la siguiente forma:

Estimador muestral coeficiente de confiabilidad x error estándar

Los estimadores muestrales corresponden a las medidas estudiadas anteriormente, estos son: RR, OR, RD, Se, Sp, etc., los cuales pueden ser calculados con las expresiones matemáticas mostradas en secciones anteriores. En relación al coeficiente de confiabilidad, este corresponde a un valor constante preestablecido para un nivel de confianza. Si el investigador diseñó el estudio con un nivel de confianza de 95 %, el coeficiente a utilizarse será 1,96, pero sí se determinó un nivel de confianza de 90 %, dicho valor cambiará a 1,64.

En cuanto al error estándar, es diferente para cada estimador muestral, es decir, hay una expresión matemática según la medida a ser utilizada; sin embargo, no es objeto de este curso profundizar en dichos aspectos estadísticos, debido a que regularmente estos intervalos de confianza son reportados como requisitos para valorar la precisión de los resultados en la publicación de los estudios clínico-epidemiológicos.

Amplitud del intervalo de confianza

Un aspecto importante de la interpretación de los intervalos de confianza es la amplitud. El intervalo expresa el rango de posibles valores que una medida toma en la población, entre más estrecho sea, mayor precisión expresa sobre el valor verdadero, permitiendo al lector del estudio conocer el tamaño del efecto estudiado en toda población de interés a la cual se pretende generalizar el resultado.

Retomando el ejemplo del ensayo clínico comparando los antirretrovirales TDF vs. ZDV (Fowler et ál., 2016), se estimó con los datos del estudio de manera puntual que la diferencia de riesgos para TDF vs. ZDV era de 1,2 %, es decir, los pacientes que recibieron TDF disminuyen su riesgo absoluto de infección perinatal por VIH en 1,2 puntos porcentuales, comparados con los que recibieron ZDV. A esta diferencia los autores reportaron un intervalo de confianza al 95 % de 0,4 % a 2,1 %. El intervalo permite fijar que el valor del efecto en la reducción absoluta de riesgo de infección por VIH perinatal en la población general de interés oscila entre 0,4 % y 2,1 % con un 95 % de seguridad. Como se puede observar, este intervalo no incluye el valor de cero, lo que dentro del 95 % de posibilidades dice que esta reducción siempre será observada si se repitiese este estudio múltiples veces, indicando que un efecto significativo, desde el punto de vista estadístico, se da en la población.

Lectura recomendada

Intervalos de confianza

Roberto Candia y Gianella Caiozzi

Determinantes importantes de la amplitud de los intervalos de confianza son el tamaño de la muestra y el número de eventos/desenlaces que ocurren en el estudio, cuando el evento es dicotómico (muerte/vivo, infectado/no infectado, curado/no curado). Si bien los estimados con tamaños de muestra grandes proveen intervalos de confianza estrechos, también es cierto y demostrable empíricamente que tamaños de muestra pequeños donde la tasa del desenlace (número de eventos de interés por grupo) es alta, pueden proveer intervalos de confianza igual o más precisos que con los tamaños de muestra grandes.

Contrastes de hipótesis

Para cada tratamiento existe un verdadero efecto subyacente que cualquier experimento individual solo puede estimar. Los investigadores usan métodos estadísticos para avanzar en la comprensión de este verdadero efecto. Un acercamiento desde lo estadístico es comenzar con lo que se llama hipótesis nula y se trata de rechazar dicha hipótesis.

La hipótesis nula es un modelo propuesto por el investigador que se da bajo la premisa de que los datos obtenidos en una muestra siguen ese modelo. Generalmente, la afirmación subyacente en el modelo es que no hay diferencia entre los tratamientos que se comparan.

En este contexto, el procedimiento estadístico (contraste de hipótesis) implementado permite, utilizando los datos obtenidos en la muestra puntual, determinar si son consistentes con la hipótesis nula (modelo propuesto). Si los resultados muestran que los datos no son compatibles con el modelo propuesto, habría entonces evidencias en contra de la hipótesis nula y quedaría solo la opción de rechazarla. Por el contrario, si los datos muestran una compatibilidad con la hipótesis nula planteada, sería el caso en que los resultados muestran evidencia a favor de la hipótesis nula y, por ende, no podría ser rechazada, no por lo menos con los datos obtenidos en la muestra.

El procedimiento de contraste de hipótesis lleva a únicamente a la toma de dos posibles decisiones: se rechaza o no la hipótesis nula. Como consecuencia de este rechazo o no, dos errores se pueden cometer: el primer error es el rechazo de una hipótesis nula, cuando esta en realidad era verdadera. En palabras más comunes, sería concluir que un tratamiento es efectivo o que la asociación vista entre una exposición-enfermedad existe cuando en realidad no era así.

Este error es denominado error tipo I o alfa. El segundo error está asociado a la decisión de no rechazar la hipótesis nula cuando esta en realidad era falsa o lo que es lo mismo, afirmar que un tratamiento no era efectivo cuando en realidad sí lo era. Este error es denominado error tipo II o beta.

	Rechazo	No rechazo
Hipótesis nula verdadera	Error tipo I	Decisión correcta
Hipótesis nula falsa	Decisión correcta	Error tipo II

Tabla 2. Errores aleatorios tipo I y II asociados con un contraste de hipótesis
Fuente: propia

Dentro del contraste, estos errores no ocurren al mismo tiempo. En la medida en que el error tipo I disminuye, como consecuencia, aumenta la certeza (confianza) sobre el rechazo y el error tipo II se incrementa. De forma contraria, si reduzco el error tipo II, aumenta la probabilidad de un contraste con poder para detectar diferencias observadas.

Instrucción

Lo invito a que realice el control de lectura sobre intervalos de confianza.

Valores de p

Para el control del error aleatorio alfa dentro de un contraste de hipótesis se le ha establecido una frontera máxima permitida, lo cual establece que si los datos sobrepasan dicha frontera el error aleatorio asociado a una decisión sea mínimo y, por tanto, aumente la plausibilidad de una conclusión sobre la hipótesis nula. Este nivel alfa máximo permitido es arbitrario y colocado antes de llevar a cabo el estudio o experimento. El valor tomado es objeto de discusión; sin embargo, en estudios clínico-epidemiológicos convencionalmente se ha establecido en 5 % o 0,05. De tal modo que, cuando es llevado a cabo el contraste, se produce un valor de probabilidad denominado "valor de p", el cual es comparado contra el nivel alfa preestablecido y permite tomar una decisión frente a la hipótesis nula planteada. Si este valor de p del contraste sobrepasa esta frontera del nivel de error alfa, el resultado obtenido con dicha muestra es incompatible con el modelo propuesto de la hipótesis nula, siendo su observación tan extrema o más extrema que la vista en menos del 5 % de las veces si se repitiera el estudio. Esta situación da evidencia empírica para rechazar la hipótesis nula y establecer que se está ante un resultado del estudio estadísticamente significativo.

Finalmente, el contraste de hipótesis presenta serias limitaciones en el análisis de resultados desde el punto de vista práctico. La primera hace referencia al valor umbral de error alfa que se antepone en un contraste. Este valor es arbitrario, no existe ninguna fórmula o justificación práctica que permita decidir, por ejemplo, entre un 5 % o 10 % o, porque no, entre un 1 % y 5 %. Dicha discusión permanece abierta entre los expertos.

En segundo lugar, el contraste es un método que reduce la conclusión de si un efecto visto es real o no a únicamente dos radicales posibilidades: existe o no efectividad, existe o no asociación, es o no un factor pronóstico. No deja opciones en un continuo para que se den puntos intermedios también válidos, como, por ejemplo, "es casi eficaz", "sugiere una asociación" o "muy poco probable que sea efectivo".

En tercer lugar, hace más de 15 años el Comité Internacional de Editores de Revistas Biomédicas lanzó su recomendación para la publicación de resultados de estudios, dando como sugerencia el uso más frecuente de intervalos de confianza que de valores p, debido a sus claras limitaciones.

Son múltiples las perspectivas de error que pueden estar involucradas en los resultados arrojados por una investigación científica en el ámbito de las ciencias de la salud; sin embargo, uno de los elementos necesarios para garantizar una evaluación adecuada de la evidencia es el reconocimiento del error sistemático, atribuido a lo que se llama en epidemiologia sesgo y que desvía el resultado de la verdad, para lo cual se entregaron elementos que deben ser evaluados en cada tipo de estudio, de acuerdo con el escenario clínico (terapia, diagnóstico, pronóstico, etc.). También está el error aleatorio que, si bien es un tema estadístico, es necesario conocer las fuentes de este posible error y las formas de controlarlo y expresarlo en los resultados para que quien lea esté informado y tome adecuadas decisiones.

Lectura recomendada

El valor de "p" y la "significación estadística". Aspectos generales y su valor en la práctica clínica

Carlos Manterola, Viviana Pineda y Grupo Mincir

Instrucción

Para finalizar, lo invito a revisar la animación que hemos preparado sobre escenarios clínicos.

BIBLIOGRAFÍA

D'Souza, G. et ál. (2007). Case-control study of human papillomavirus and oropharyngeal cancer. *The New England Journal of Medicine*, *356*(19), 1944-1956.

Fowler, M. et ál. (2016). Benefits and risks of antiretroviral therapy for perinatal HIV prevention. *The New England Journal of Medicine, 375*(18), 1726-1737.

Miller, J. M. et ál. (2008). Diagnostic performance of coronary angiography by 64-Row CT. *The New England Journal of Medicine, 359*(22), 2324-2336.