ISSN 0034-7000
REV. ARG. CARDIOL., 57, Nº 3, 96-100

Editorial

Significación estadística y decisión médica

Cuando en un trabajo de investigación se comprueba que una intervención terapéutica modifica favorablemente y con significación estadística un índice hemodinámico, un conjunto de variables clínicas, o la expectativa de vida de una determinada patología, puede concluirse que esta propuesta debe ser sistemáticamente empleada en la práctica clínica.

La situación inversa suele también ser tomada como verdad absoluta: si no hay diferencias estadísticamente significativas entre la terapéutica ensayada y el placebo, no hay justificación científica y aun ética para su empleo.

Este razonamiento se ha visto fortalecido por una serie de trabajos multicéntricos que, al menos por ahora, son estadísticamente cuestionables. Estos *ensayos clínicos controlados* han modificado profundamente algunas estrategias terapéuticas y sus conclusiones son en principio universalmente aceptadas.

Algunas características de estos trabajos son las siguientes:

1. Son desde el punto de vista metodológico sencilllos, ya que no requieren aparatología especial o estudios complicados. Pretenden aclarar el impacto de una propuesta terapéutica sobre un único aspecto o "punto final" como son las condiciones de vida o más frecuentemente la mortalidad.
2. Con este diseño metodológico los centros participantes no requieren alta complejidad. Se incluye así un gran número de hospitales, habitualmente de varios países, y no infrecuentemente de diferentes continentes, convirtiéndose en investigaciones internacionales. De este modo el número de pacientes incorporado es elevado y lo es en un corto período, aspecto este último esencial para la investigación.
3. Son "controlados" con placebo, randomizados y generalmente a doble ciego, requisito este último importante cuando el punto final no es mortalidad.
4. El sostén económico es brindado por una empresa farmacéutica que debe demostrar en forma incuestionable que su producto de investigación básica es clínicamente útil.
5. Algunos ejemplos de estos ensayos son los siguientes:
   a) Cirugía de revascularización y tratamiento médico en la enfermedad coronaria.
   b) Tratamiento trombolítico endovenoso en la fase aguda del infarto.
   c) Antiagregantes en los diferentes cuadros clínicos de la cardiopatía isquémica.
   d) Drogas vasodilatadoras e inotrópicas en la insuficiencia cardíaca.
   e) Betabloqueantes en el período postinfarto agudo.
6. Aunque con diferente metodología, las conclusiones presentadas en términos de significación estadística pretenden simplemente demostrar que las diferencias halladas no son producto del azar.

Determinantes de la significación estadística

Habituados a aceptar el valor "P" como expresión del impacto de la intervención ensayada sobre la incidencia o magnitud del evento estudiado, el internista se ha encontrado con una nueva modalidad en la que son actualmente presentadas las conclusiones. Los resultados son definidos ahora como el cambio inducido por el tratamiento en la incidencia del evento o punto final, pero expresando ese cambio en relación con el ocurrido en el grupo control. El tratamiento trombolítico en el infarto es un buen ejemplo para explicar este punto. La estreptoquinasa administrada en las tres primeras horas reduce la mortalidad en la etapa aguda un 26% (*modificación relativa del riesgo*) con respecto a la ocurrida en el grupo control (de 12% a 9,2%).[1] Este hallazgo puede también ser expresado como *riesgo relativo*, que en el caso anterior será de 0,74 (la mortalidad del grupo tratado es el 74% de la del grupo control) (Fig. 1).

Por supuesto que a este parámetro hay que agregarle otro que exprese que la diferencia no es producto del azar. El *intervalo de confianza*

define este aspecto. Indica el rango en el cual puede ubicarse la modificación relativa de la incidencia del evento (o riesgo relativo) con una probabilidad definida de no ser consecuencia del azar. Si volvemos al ejemplo de la estreptoquinasa, con una reducción de la mortalidad del 26%, los límites del intervalo de confianza del 95% hallados en el trabajo antes citado fueron de 0,37 a 0,13. En otras palabras, entre esos límites hay sólo un 5% (*nivel de significación*) de probabilidad de haber hallado un resultado (en este caso reducción de la mortalidad por la estreptoquinasa) falsamente positivo (error tipo 1). Cuando el límite del intervalo de confianza alcanza la unidad (riesgo relativo 1, modificación del evento 0%), la diferencia hallada puede ser consecuencia del azar (diferencia no significativa) y no por la intervención ensayada (Fig. 2).

De este modo la diferencia significativa se alcanza no sólo con un desplazamiento amplio en términos relativos de la media de la incidencia del punto final, sino que debe agregarse a ello un intervalo de confianza estrecho. Cuanto menor es la diferencia hallada, más pequeño debe ser el intervalo de confianza (Fig. 2).

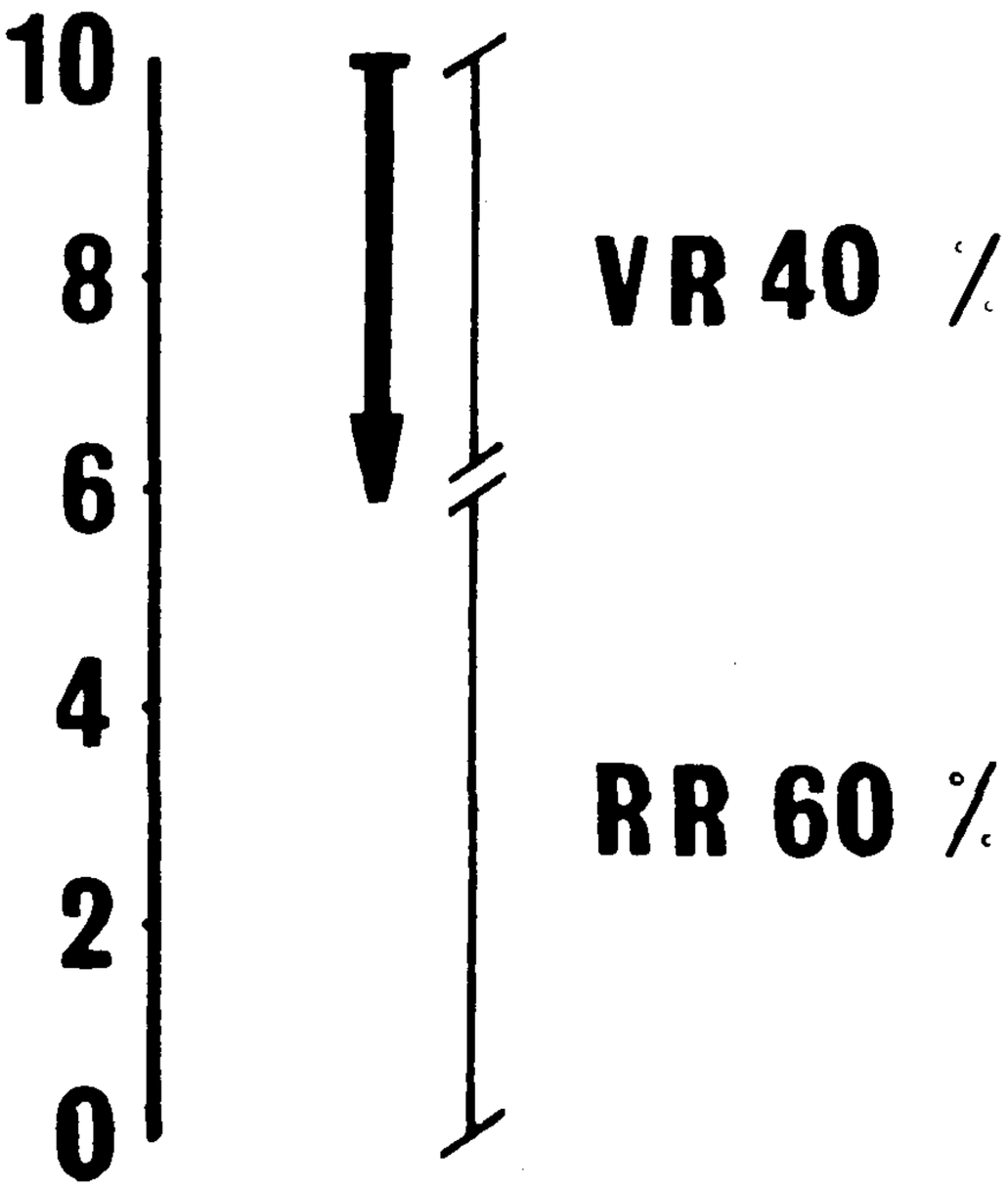

Fig. 1. Expresión gráfica del resultado de una hipotética intervención que reduce la incidencia de un evento. Este hallazgo puede ser expresado como variación relativa (VR) del 40% o riesgo relativo (RR), en este caso del 60%.

Veamos ahora qué factores intervienen en la modificación relativa del evento (o riesgo relativo) y cuáles en el intervalo de confianza. El primero es fácil aceptar que estará condicionado por la eficacia de la intervención (por ejemplo capacidad de la estreptoquinasa para reducir la mortalidad).

El intervalo de confianza depende del número de pacientes que fueron incluidos en el estudio. Cuanto mayor es la población estudiada (*tamaño de la muestra*) menor será el intervalo de confianza.

Ahora bien, la posibilidad de detectar la variación de un evento, definido por su media y el intervalo de confianza correspondiente, depende de la incidencia de dicho evento en la población en estudio. Por ejemplo, si una estrategia terapéutica reduce la mortalidad el 26% con un intervalo de confianza entre 0,37 y 0,13, según el ejemplo anteriormente comentado, la probabilidad de detectar ese efecto será mayor

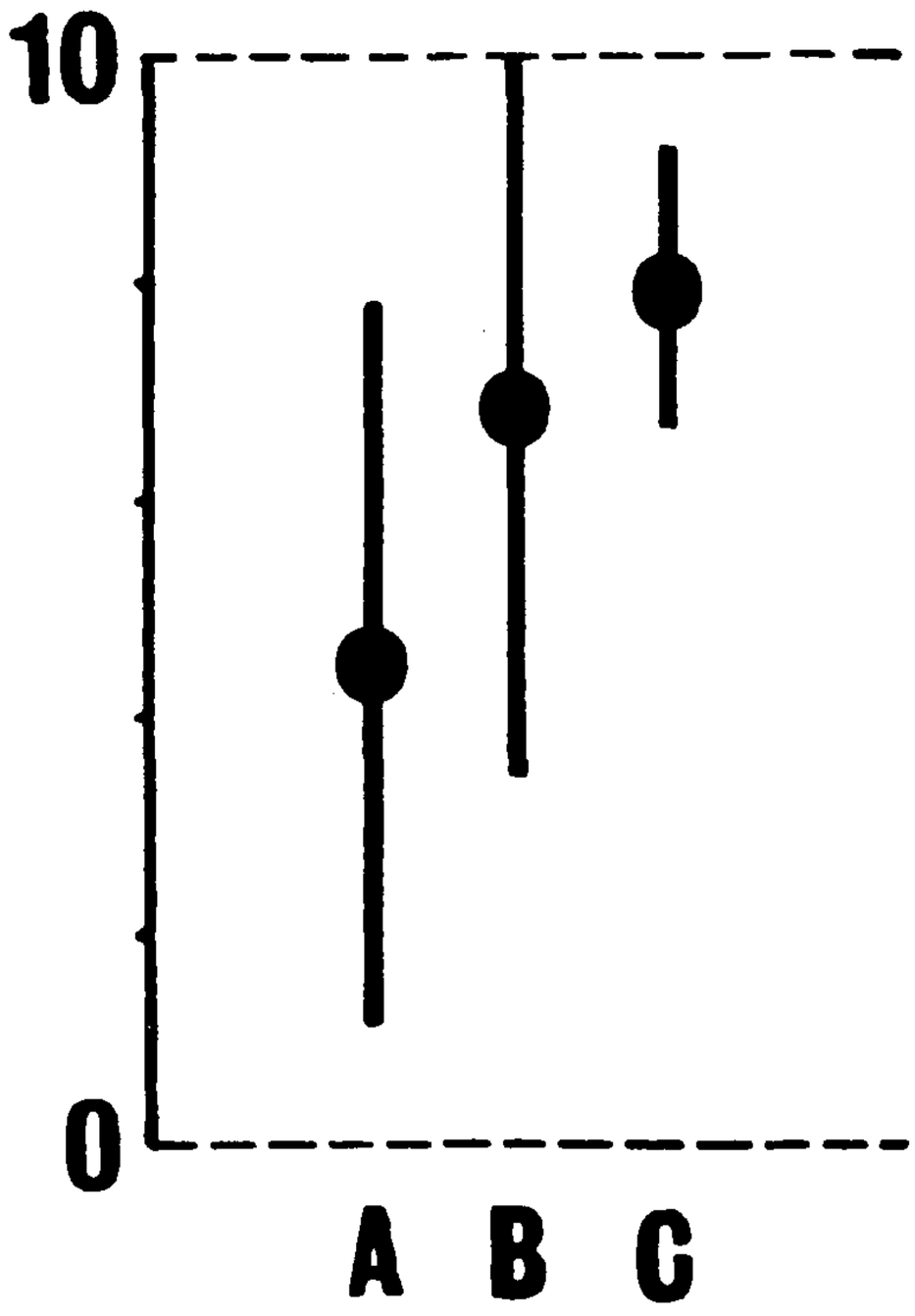

Fig. 2. El gráfico ilustra los resultados de tres intervenciones, A, B y C, con su valor medio y los correspondientes intervalos de confianza. A y B tienen igual intervalo de confianza. La intervención A logra significación estadística debido a que su valor medio se encuentra ampliamente desplazado. C también adquiere significación estadística por un intervalo de confianza muy estrecho, este último debido a las características o número de la población estudiada.

si la mortalidad de la población en estudio (grupo control) es del 10% que si es del 5%. Ello se debe a que esa probabilidad depende de la diferencia en números absolutos entre los pacientes fallecidos en el grupo control y en el que recibió la droga. También en este caso esta dificultad puede superarse adaptando el tamaño de la muestra a la mortalidad esperable en la población en estudio.

De este modo el número de pacientes a incluir en el ensayo debe calcularse a partir de la modificación del evento que se estima ha de inducir la intervención y de la incidencia del mismo esperable en la población control. Si la muestra es inadecuada, es posible no detectar una variación significativa a pesar de que la intervención es realmente eficaz (resultado falsamente negativo o error tipo 2). La *potencia* del estudio define la posibilidad de cometer ese error. Si por ejemplo es del 90%, indica que hay sólo un 10% de probabilidad de tener un resultado falsamente negativo.

De este modo es claro que un efecto leve (por ejemplo de sólo un 10%, riesgo relativo 0,90) puede "alcanzar significación estadística" si la mortalidad de la población en estudio es alta o si el número de pacientes es elevado.

Por supuesto que todas las consideraciones anteriores son también válidas cuando los resultados son expresados por el valor "P". La ventaja de definir los hallazgos en estos términos es la facilidad con que los mismos pueden trasladarse a la práctica clínica. Si la mortalidad es del 10% y es descendida al 7% (reducción del 30%), 3 de cada 100 pacientes tratados serán "salvados", pero si es del 40% con igual tasa de reducción, 12 de cada 100 serán beneficiados por la medida. Con el valor de P este análisis no es factible. Por ejemplo, si en una investigación se demuestra que un procedimiento terapéutico reduce la mortalidad con un nivel de significación de P menor de 0,05, es imposible trasladar este hallazgo a una situación clínica definida.[2]

La modificación en más en la incidencia de un evento puede también ser representada de esta forma. Por ejemplo, si se investiga la acción de la digital en enfermos con insuficiencia cardíaca, el punto final puede ser la frecuencia de pacientes en clase funcional I. El efecto beneficioso de la droga queda definido si la digital aumenta el porcentaje de enfermos en esta clase funcional (si se eleva en un 30%, el riesgo relativo será de 1,3). El intervalo de confianza convertirá esa diferencia en significativa toda vez que no se alcance la unidad.

Por otro lado, si se trata del estudio de una variable discreta, la misma queda definida por la media y el desvío estándar. La diferencia entre grupo control y el tratado puede también ser expresada como la modificación en términos relativos de una media con respecto a la otra. En este caso el intervalo de confianza depende del número de casos y de la dispersión (desvío estándar) de los datos de cada población. El intervalo de confianza se abre si el número de enfermos estudiados es escaso y el desvío estándar es amplio.

La decisión médica

El "punto final" del accionar médico debe ser necesariamente el alivio de los síntomas y la prolongación de la expectativade vida. Modificar favorablemente la fracción de eyección, el volumen minuto o un registro electrocardiográfico sólo adquiere significación clínica si es expresión de aquel objetivo. El problema es que no conocemos con precisión la relación entre esas variables subclínicas y las de real significación clínica, y menos aún en qué forma la variación de unas predice la modificación de las otras. Por ejemplo, la relación entre fracción de eyección y mortalidad ha sido confirmada en diferentes trabajos, pero no conocemos ni por aproximación la relación exacta entre variación de fracción de eyección y variación de mortalidad.

Claro está que este tipo de dificultad ha sido obviado por los ensayos clínicos controlados. Con ellos se demuestra no sólo que una intervención modifica con significación estadística la evolución clínica, sino que permite calcular en términos absolutos el número de pacientes "mejorados o salvados" por una intervención.

Pero todas las intervenciones, aun las aparentemente más inocuas, pueden tener efectos colaterales. El análisis de estos últimos y su peso, tanto por el número como por su significación clínica, es minuciosamente llevado a cabo en los estudios multicéntricos. La ecuación riesgo/beneficio, tan en boga en la literatura médica, analiza la relación entre las ventajas y los efectos indeseables del ensayo terapéutico. Si la comparación resulta favorable, suele concluirse en la conveniencia de su implementación. El cardiólogo clínico se ve entonces impulsado a emplear sistemáticamente el tratamiento en cuestión. Pero veamos algunos aspectos que cuestionan este razonamiento.

Si el beneficio de la intervención en términos absolutos es escaso, por ejemplo menos de 2

por cada 100 pacientes tratados, aunque la relación riesgo/beneficio pueda justificarlo, la sistematización del ensayo es cuestionable. Un solo paciente con mala evolución complica los resultados estadísticos en tal proporción que muchos enfermos deberán ser incluidos para balancear la ecuación.

El costo económico, que en ciertos casos puede adquirir una significación especial, no suele ser incluido en el análisis del estudio multicéntrico. Este enfoque adquiere una preponderancia particular en nuestro país pero también en naciones desarrolladas, y motiva discusiones frecuentes en la literatura médica, como ha sido últimamente comunicado.[3]

Pero el error que adquiere mayor significación por su frecuencia es el que resulta de trasladar el hallazgo de una intervención a una población que difiere de aquella en la cual fue comprobado. Por ello la comparación de poblaciones debe realizarse con sumo cuidado. Por ejemplo: dos poblaciones con infarto de miocardio pueden ser aparentemente coincidentes en el tiempo de evolución de la necrosis, la localización, grados de falla de bomba, etc., pero si la mortalidad en un grupo es mayor que en el otro, este aspecto pesa mucho más que todos los anteriores. La diferencia no sólo es cuantitativa sino cualitativa. Por ejemplo, podría ocurrir que en una serie con una mortalidad del 12% la participación de la disfunción contráctil sea preponderante y que en una del 6% sea prácticamente nula, interviniendo en este último caso otras etiologías, como la ruptura de estructuras cardíacas o los disturbios del ritmo en los cuales la estreptoquinasa no tiene efecto o resulta aun contraproducente. De manera que una reducción del 26% hallada en una determinada población podría tener una significación menor o nula cuando se aplica a otra de bajo riesgo, no sólo como consecuencia del cálculo matemático, sino porque el beneficio de la medida en términos relativos también sería menor.

De este modo, una intervención que ha demostrado variar significativamente en forma relativa un evento puede tener escasa relevancia clínica cuando se analiza en términos absolutos o simplemente ofrecer una relación riesgo/beneficio que se halla *borderline*.

Discutamos ahora una situación opuesta. ¿Es posible sostener una decisión médica fundamentada sólo en una tendencia sin el respaldo que brinda la significación estadística? En nuestra opinión, en este caso el desafío es aún mayor.

Según fue discutido, la falta de significación puede ser sólo la consecuencia del número reducido de pacientes que participaron en el estudio, es decir de un resultado falsamente negativo. ¿Cómo descubrir entonces si una propuesta es válida debido a que la falta de significación es sólo consecuencia de ello? ¿Cómo tomar posición aquí y ahora en cuestiones no estadísticamente confirmadas? Tal vez la coincidencia de resultados de diferentes investigaciones resulte útil en este sentido.

Si hay una mayoría de estudios que comprueban que un tratamiento determinado disminuye la mortalidad en el infarto de miocardio o en la insuficiencia cardíaca, la no significación estadística tal vez esté condicionada por el número reducido de pacientes estudiados.

El *metanálisis* es una técnica estadística que permite recopilar los resultados de diferentes trabajos que, habiendo explorado el mismo tratamiento, no tuvieron significación estadística por el escaso número de pacientes incluidos. Cuando se recopilaron varios estudios con estreptoquinasa de resultados no concluyentes, pudo comprobarse que en la mayoría de ellos la mortalidad descendía, y cuando con las técnicas del metanálisis se calculó la media y el intervalo de confianza de la totalidad de esas investigaciones, se demostró que se reducía lo suficiente y con un intervalo de confianza tan estrecho que alcanzaba significación estadística.[4] Este hallazgo fue el primer indicio sugestivo acerca del beneficio del tratamiento trombolítico en el infarto. Luego, tal vez como consecuencia de ello, llegaron los estudios multicéntricos que confirmaron aquella presunción.

Claro que para un metanálisis se requiere recopilar "todos" los estudios con la intención de no cometer un sesgo en la selección de los mismos. Además debe disponerse de la información detallada del trabajo. Por otro lado, es imposible disponer de un metanálisis en todas las cuestiones de decisión clínica.

El cardiólogo clínico puede sospechar que una tendencia sistemática hallada en diferentes estudios es sugestiva de haberse cometido un error tipo 2 (falso negativo). La decisión médica puede entonces apoyarse en este criterio hasta que un ensayo clínico randomizado confirme su impresión. Este análisis es independiente de aquellas conductas "heroicas" que, justificadas por la gravedad extrema del cuadro, han sido implementadas con el objeto de revertir una situación clínicamente terminal. Pero este razonamiento no tiene justificación científica.

El problema de los subgrupos

Cuando un estudio diseñado para demostrar el efecto de una intervención ha fallado en confirmar esa hipótesis en la población total, ninguna conclusión a nivel de subgrupo es válida, ya que el análisis en este caso es retrospectivo. Una nueva investigación se impone, limitada ahora a esa subpoblación.

Pero, de haberse confirmado que la terapéutica modifica el punto final en toda la población, ¿qué valor tiene el análisis retrospectivo tendiente a buscar subgrupos donde el efecto resulta de mayor significación o por el contrario aquellos en que el mismo carezca de valor?

Si una vez más ejemplificamos con el infarto de miocardio, es posible conformar subgrupos según el tiempo de evolución, la localización y los resultados de combinar ambas variables. Cuando los autores del ISIS II[5] demostraron la falacia del análisis a nivel de subgrupos, al hallar que los pacientes de un determinado signo astrológico no se beneficiaban con el trombolítico, pareció evidente que la única información válida era aquella que afectaba a la totalidad de la población.

Si esta conclusión fuera cierta, la estreptoquinasa no tendría mayor eficacia en el subgrupo con menor tiempo de evolución. Sin embargo en la mayoría de los trabajos se ha demostrado que, a menor tiempo de evolución, mayor es la eficacia en cuanto a reducción de la mortalidad.

Presentadas así las cosas, la ecuación no parece tener solución, ya que no resulta sencillo aceptar el valor de los hallazgos en un caso y no en el otro. Nuevamente aquí el problema puede ser encarado relevando toda la información comunicada para aplicar la técnica del metanálisis. Claro que esta información puede no estar disponible. Además el número de subgrupos es infinito, y si se quiere, cada paciente puede ser considerado como un subgrupo. También aquí el cardiólogo clínico puede ensayar el siguiente razonamiento: si diferentes estudios han sido coincidentes en que en ese subgrupo la medida terapéutica en discusión modifica favorablemente la incidencia de un evento en igual sentido y en forma sistemática, es posible que esto no resulte consecuencia del azar. Seguramente ha de ser difícil que nuevas investigaciones sean coincidentes en demostrar la relación entre el horóscopo y la respuesta al tratamiento trombolítico.

Una conclusión obvia

El ensayo clínico controlado ha significado un progreso notable en la investigación al explorar aspectos esenciales para el médico asistencial. Sus resultados respaldados estadísticamente demuestran que la investigación ha alterado el curso evolutivo de una enfermedad. Tal vez confirmen o permitan dilucidar una hipótesis fisiopatológica. La decisión para implementar una terapéutica ha de apoyarse necesariamente en estas investigaciones.

Pero de alguna manera la toma de decisiones en la práctica cotidiana también seguirá siendo artesanal. El traslado en forma sistemática de un resultado con validez estadística a la indicación médica, o su rechazo en caso contrario, puede conducir a conductas erróneas que en definitiva priven al paciente de la mejor opción terapéutica y al médico de ejercer con plenitud una profesión que debe entenderse, aun en nuestros días, más como un arte que como una ciencia exacta.

Arturo Cagide, Raúl Oliveri
Servicio de Cardiología, Hospital Italiano
de Buenos Aires

BIBLIOGRAFIA

1. Gruppo Italiano per lo studio della steptochinasi nell' infarto miocardico (GISSI): Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Lancet, 397, 1986.
2. Evans S, Mills P, Dawson J: The end of the P value? Br Heart J *60*: 177, 1988.
3. Goldman L, Sia B, Cook F, Tutherford J, Weinstein M: Costs and effectiveness of routine therapy with long-term beta-adrenergic antagonist after acute myocardial infarction. N Engl J Med *319*: 152, 1988.
4. Yusuf S, Collins R, Peto R et al: Intravenous and intracoronary fibrinolitic therapy in acute myocardial infarction: oveview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. Eur Heart J *6*: 556, 1985.
5. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group: Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17.187 cases of suspected acute myocardial infarction. Lancet, 349, 1988.