bit
Boletín de Información Farmacoterapéutica de Navarra
VOLUMEN 16, Nº 1 ENERO 2008

15!
1993-2008
QUINCE AÑOS
DE INFORMACIÓN
INDEPENDIENTE

La interpretación de los resultados de un ensayo clínico aleatorizado

IDOIA GAMINDE
SERVICIO DE DOCENCIA, INVESTIGACIÓN Y DESARROLLO SANITARIO. DEPARTAMENTO DE SALUD. GOBIERNO DE NAVARRA
JUAN ERVITI
SERVICIO DE PRESTACIONES FARMACÉUTICAS. SNS-O

En este artículo se revisan algunos de los conceptos claves que nos ayudan a interpretar los estudios que son capaces de demostrar la eficacia de los tratamientos farmacológicos: el papel de los sesgos, el papel del azar, la relevancia clínica, la validez externa de los ensayos, las variables intermedias y compuestas, el análisis de las variables secundarias, y de los subgrupos, los ensayos de no inferioridad, la influencia de los estudios individuales en el metanálisis y, finalmente, la opinión del paciente

La interpretación de los resultados de un ensayo clínico aleatorizado

Para interpretar correctamente un ensayo clínico aleatorizado (ECA) lo primero que hay que hacer es evaluar la robustez de los resultados: ¿se debe el beneficio del tratamiento a la forma en la que se ha realizado el estudio?

Los sesgos

Sesgo es un término que se refiere a cualquier error sistemático debido al diseño, ejecución o interpretación del estudio. Los errores en la asignación aleatoria y en el enmarascamiento (ciego) son los sesgos más habituales.

La asignación aleatoria consiste en distribuir cada participante a uno de los grupos de tratamiento por azar. Así, se pretende que los grupos incluidos en el ensayo sean semejantes en todas las características relevantes menos en una: la intervención que cada uno recibe[1]. La ausencia de aleatorización (o errores en el proceso) puede dar lugar a grupos de estudio no comparables. De hecho, en muchas revisiones sistemáticas se excluyen los ensayos no-aleatorizados debido a los sesgos que pueden aparecer al no aleatorizar. Es muy importante que la secuencia de aleatorización de los pacientes a cada grupo permanezca oculta para los pacientes y evaluadores.

El enmascaramiento es el procedimiento por el que se asegura que los sujetos participantes en un ECA, los observadores, o ambos, no conocen el tratamiento asignado u otra característica que pudiera sesgar los resultados[1].

Los estudios con enmarascamiento incorrecto pueden sobreestimar el efecto en un 41% y los estudios en los que no está claro, en un 30%[2].

El papel del azar

Una vez evaluados los posibles sesgos relacionados con la forma de aleatorización y el enmascaramiento, nos plantearemos si los datos obtenidos son reales o se deben al azar.

La aleatorización y el enmascaramiento son aspectos clave del ECA

La comparación de la eficacia de dos intervenciones se puede realizar a través de la comprobación de hipótesis (utilizando pruebas de significación estadística), o a través de técnicas de estimación [utilizando intervalos de confianza (IC)].

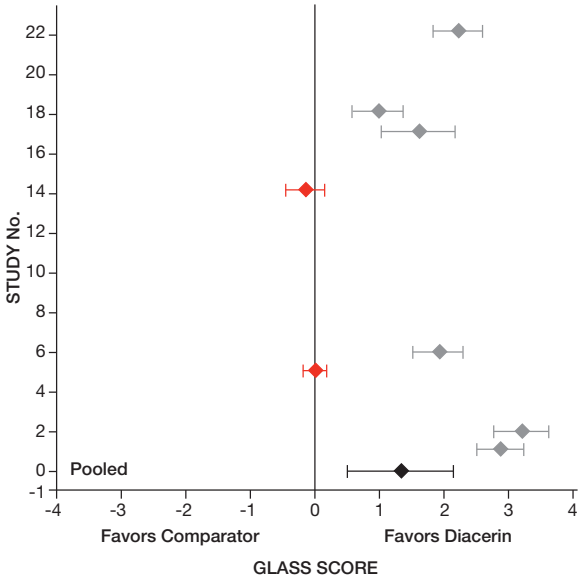
Las pruebas de significación estadística calculan la probabilidad de que los resultados observados entre los grupos del ECA puedan ser debidos al azar, en el supuesto de que ambas intervenciones fueran igual de eficaces (hipótesis nula cierta). Esta probabilidad es el grado de significación estadística y se representa con la letra "p". Generalmente se adopta el valor p=0,05 como punto de corte por debajo del cuál se considera que se puede rechazar la hipótesis de igualdad entre ambas intervenciones (con un 95% de confianza) y concluir que el resultado es "estadísticamente significativo"[3]. Solo nos permite rechazar o no la hipótesis nula de "no diferencia" entre los dos grupos, pero nada dice de la magnitud de la diferencia o del sentido de esta.

Para evaluar el papel del azar son más útiles los intervalos de confianza. Los IC dan una idea de la magnitud o relevancia del efecto observado. Permiten conocer entre qué límites es probable que se encuentre la verdadera diferencia. Habitualmente se trabaja con IC del 95%, lo que significa que, en un 95% de las veces, un IC correctamente construido debería de contener el verdadero valor de la variable de interés. Por ejemplo, si se encuentra que la diferencia observada entre dos tratamientos es del 22% (p<0,05; IC 95% = 17 a 27%), esto significa que podemos tener un 95% de confianza de que la diferencia real entre los dos tratamientos se encuentra entre el 17% y el 27%.

Una característica útil de los IC es que permiten decir si se ha alcanzado o no la significación estadística, al igual que en una prueba de hipótesis. Cuando la medida de efecto es la diferencia entre las intervenciones, si el IC incluye el valor 0 se concluye que el resultado no es estadísticamente significativo, que no se ha podido demostrar diferencia en el efecto del tratamiento entre el grupo de tratamiento y el control. Cuando, en vez de una diferencia absoluta nos interesa una media relativa como el riesgo relativo, si el IC incluye el valor 1 se concluye que el resultado no es estadísticamente significativo[3].

Por ejemplo, en la figura 1 vemos distintos IC de la diferencia entre la diacereína y el placebo en el alivio del dolor. Dos de los estudios cruzan la línea de "no diferencia" (el intervalo de confianza incluye el valor 0), mientras que otros

Figura 1. Metanálisis de la eficacia de la diacereína frente a placebo en el dolor.



Fuente: [1] Rintelen B, Neumann K, Leeb B. A meta-analysis of controlled clinical studies with diacerein in the treatment of osteoarthritis. Arch Intern Med. 2006;166(17):1899-906.

seis estudios muestran la superioridad de la diacereína frente al placebo. También podemos observar la diferente amplitud de los IC, que reflejan la precisión de las estimaciones, los intervalos más estrechos son más precisos que los intervalos grandes.

Otra de las ventajas de los IC sobre la comprobación de hipótesis tradicional es que dan información adicional. Los límites superior e inferior del IC nos informan sobre cuán pequeño o grande puede ser el verdadero efecto. Si el IC es estrecho, podemos tener confianza en que cualquier efecto fuera de este rango ha sido descartado por el estudio. Esta situación se presenta cuando el tamaño de muestra del estudio es muy grande y la estimación del verdadero efecto es muy precisa. Es decir, que el estudio tiene "poder" suficiente para detectar un efecto. Pero, si el estudio es pequeño, el IC es muy amplio, capturando un rango muy diverso de tamaños de efecto. De esta manera los estimadores del tamaño de efecto pueden ser bastante imprecisos. Es un estudio con poco "poder" y nos da menos información.

Posibles errores en la interpretación de los resultados.

Los IC, al igual que los valores "p", nos ayudan a interpretar los hallazgos de la investigación a la luz de los efectos del azar. En la interpretación, sin embargo, hay algunos escollos.

Nos podemos equivocar al ver efectos que no son reales. El IC nos muestra que la diferencia es "estadísticamente significativa". Por tanto, concluiremos que los dos tratamientos son diferentes. Sin embargo, sólo porque es poco probable observar una diferencia grande sólo por azar, esto no significa que sea cierto. Por definición, uno de cada 20 resultados significativos podría ser falso (un 5%) y las diferencias encontradas serían fruto del azar. Por eso, el azar puede equivocarnos al hacernos creer que existen diferencias entre los grupos cuando en realidad no las hay. A esto se le denomina error tipo I. La probabilidad de cometer un error de este tipo se conoce como "α" y se suele expresar como el *grado de significación p.* Un valor de p=0,05 hace referencia a que existe un α=0,05.

Otro posible error que podemos cometer es concluir de un resultado no significativo que no hay efecto, cuando en realidad sí existe. Esto es el error tipo II. El hecho de asimilar la "no significación" con ausencia de efecto es un malentendido frecuente y perjudicial. Un IC no significativo simplemente nos dice que la diferencia observada es consistente con que no hay verdadera diferencia entre los dos grupos. Sin embargo, no podemos rechazar esta posibilidad. Sólo porque no hemos encontrado un efecto de tratamiento significativo no quiere decir que este no exista. La probabilidad de cometer un error de este tipo suele denotarse por ß y su complementario, 1-ß, es lo que se conoce como poder estadístico o potencia estadística. Expresa la probabilidad de detectar una diferencia estadísticamente cuando realmente existe esa diferencia.

A la hora del diseño, por tanto, debe establecerse la magnitud mínima de la diferencia o asociación que se considere de relevancia clínica, así como el poder estadístico que se desea para el estudio y, de acuerdo con ello, calcular el tamaño de la muestra necesaria.

Significación estadística y relevancia clínica

La significación estadística a veces se interpreta de manera incorrecta al asociarla con un resultado importante. Las pruebas de significación sólo se

Aun en ECAs bien diseñados, se producirá un falso positivo cada 20 resultados, simplemente por azar

Es necesario determinar la relevancia clínica de un ensayo estadísticamente significativo

preguntan si los datos que produce un estudio se pueden deber al azar o no. Rechazar la equivalencia entre dos intervenciones no significa necesariamente que aceptamos que hay una diferencia importante entre ellas. Un estudio grande puede encontrar que una diferencia pequeña es estadísticamente significativa. Es, por lo tanto, diferente evaluar la relevancia clínica de esta diferencia. En la evaluación de la importancia de resultados estadísticamente significativos, es el tamaño del efecto (no el tamaño de la significación) lo que es relevante.

Para mostrar la relevancia clínica de los resultados de un ECA la forma recomendada de presentar los resultados debe incluir la reducción relativa del riesgo (RRR), la reducción absoluta del riesgo (RAR) y el número necesario de pacientes a tratar para reducir un evento (NNT). Consideremos para su cálculo este ejemplo: mueren el 15% de pacientes en el grupo de intervención y un 20% en el grupo control. El riesgo relativo, que es el cociente entre los expuestos al nuevo tratamiento o actividad preventiva y los no expuestos, es en este caso (0,15/0,20=0,75). El riesgo de muerte de los pacientes que reciben el nuevo tratamiento relativo al de los pacientes del grupo control fue de 0,75. La RRR es el complemento del RR, es decir, (1-0,75)* 100 = 25%. El nuevo tratamiento reduce el riesgo de muerte en un 25% relativo al que ha ocurrido en el grupo control. La reducción absoluta del riesgo (RAR) sería: 0,20-0,15= 0,05 (5%). Podríamos decir, por tanto, que de cada 100 personas tratadas con el nuevo tratamiento podemos evitar 5 casos de muerte. La siguiente pregunta sería: si de cada 100 personas tratadas con el nuevo tratamiento podemos evitar 5 casos de muerte, ¿cuántos tendríamos que tratar para evitar un solo caso de muerte? En otras palabras ¿cuál es el NNT? Su cálculo requiere una simple regla de tres que se resuelve dividiendo 1/RAR. En este caso 1/0,05 = 20. Por tanto, la respuesta es que necesitamos tratar a 20 pacientes con el nuevo tratamiento para evitar un caso de muerte[4].

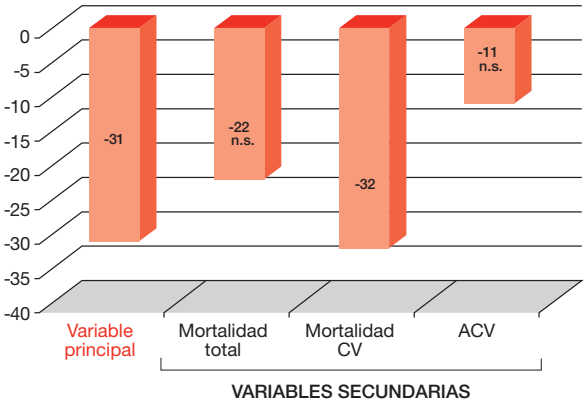
Este modo de presentar los resultados cuantifica el esfuerzo a realizar para conseguir la reducción de un episodio desfavorable. El presentar los resultados sólo como reducción porcentual del riesgo relativo (RRR), aunque es técnicamente correcto, tiende a magnificar el efecto de la intervención al describir del mismo modo situaciones muy dispares. Cambios pequeños en el riesgo basal absoluto de un hecho clínico infrecuente conducen a grandes cambios en el número de pacientes que necesitamos tratar con la intención de prevenir uno. Por tanto, es necesario determinar la relevancia clínica de un ensayo que presente resultados estadísticamente significativos. Cuanto más reducido es el NNT, la magnitud del efecto del tratamiento es mayor. Si no se encontrase eficacia en el tratamiento, la reducción absoluta del riesgo sería cero y el NNT sería infinito. Como sucede en las estimaciones de otros parámetros, se debe expresar el NNT con intervalos de confianza para estimar la incertidumbre que dicho parámetro presenta[4].

Recursos para calcular la relevancia clínica de los resultados

- Calculadora *on line* que calcula la relevancia clínica de los resultados de ECA junto con su correspondientes IC: http://www.healthcare.ubc.ca/calc/clinsig.html
- Listado de NNTs [http://www.jr2.ox.ac.uk/bandolier/band50/b50-8.html].

Un ejemplo de ello es la distinta visión que podemos obtener sobre la eficacia de las estatinas en la prevención primaria de episodios cardiovasculares según centremos nuestra atención en los datos de riesgo relativo, riesgo absoluto o únicamente valoremos el hecho de que se encuentran diferencias significativas entre los grupos en estudio. De los ensayos realizados, el estudio WOSCOP[5] fue el que encontró mayores diferencias en la variable principal (infarto no fatal + mortalidad coronaria). Si nos fijamos en el porcentaje de reducción de riesgo relativo en las distintas variables, aparentemente la eficacia de la intervención es notable (figura 2).

Figura 2. Reducción del riesgo relativo con pravastatina frente a placebo (estudio WOSCOP)



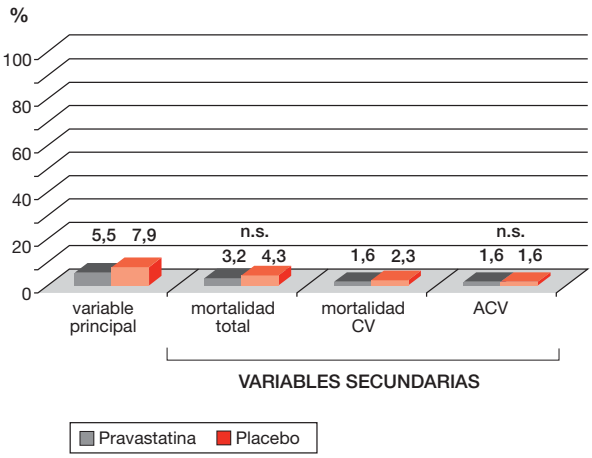
Sin embargo, si consideramos la reducción de riesgo absoluto, veremos que la magnitud del efecto de la intervención es bastante pobre, al margen de que las diferencias fueran o no estadísticamente significativas (figura 3). En el grupo placebo se produjo alguno de los episodios incluidos en la variable principal en el 7,9% de los pacientes. Por otro lado, en el grupo intervención un 5,5% de los pacientes desarrollaron algún episodio. Es decir, un 2,4% de los pacientes obtuvieron algún beneficio, mientras que el 97,6% restante no se benefició del tratamiento tras 5 años tomando el fármaco (figura 4). Si lo expresamos en términos de NNT, se necesitaría tratar a unos 42 pacientes durante 5 años para evitar uno de los episodios incluidos en la variable principal.

En el resto de ensayos clínicos publicados en prevención primaria cardiovascular los resultados todavía han sido peores en cuanto a la magnitud del efecto del medicamento, al margen de que las diferencias fueran estadísticamente significativas o no.

Hay muchos otros casos en los que la magnitud del efecto es muy pequeña (de relevancia clínica discutible) a pesar de que los ensayos mostraron que había diferencias estadísticamente significativas.

Un ejemplo es la eficacia de los anticolinérgicos en la incontinencia urinaria. En un ensayo frente a placebo, la solifenacina demostró reducir el número de micciones diarias de una media de 11 a 10. Aun sin tener en cuenta los efectos adversos del medicamento... ¿podría considerarse clínicamente relevante la reducción de una micción al día respecto a un total de 11 micciones diarias?[6]

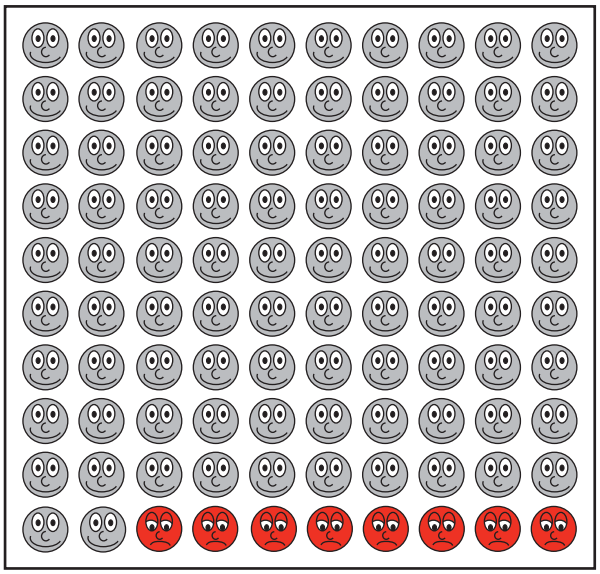
Figura 3. Incidencia de distintos episodios cardiovasculares en los grupos pravastatina y placebo (estudio WOSCOP)



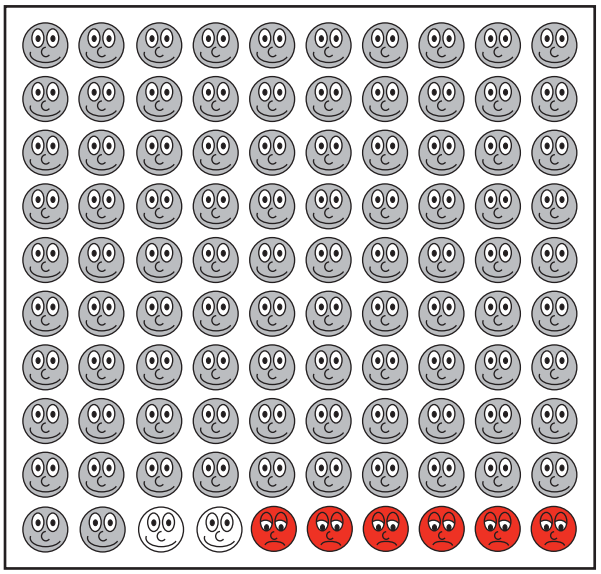
Mención aparte merece el caso de los estudios que miden los efectos de los fármacos mediante el uso de escalas. En estas ocasiones se evalúa la diferencia de la puntuación obtenida en el grupo tratado respecto al control. Si lo comparamos con variables importantes como la incidencia de infartos, por ejemplo, la relevancia clínica de una mejora o empeoramiento de determinada magnitud en una escala es más difícilmente objetivable.

Figura 4. Representación visual de los resultados de pravastatina vs placebo tras 5 años de tratamiento

Resultados del grupo placebo tras 5 años de seguimiento



Resultados del grupo pravastatina tras 5 años de seguimiento



No experimenta infarto ni muerte de origen coronario

Evitan un infarto o muerte de origen coronario

Experimenta infarto o muerte de origen coronario

Hay que reflexionar sobre las características de los pacientes para poder juzgar la validez externa de los datos

En este sentido, un metanálisis que valora los efectos de rivastigmina, donepezilo, galantamina y memantina sobre la demencia vascular observa una mejora de los pacientes tratados con estos fármacos de entre 1 y 2 puntos sobre una escala de 70 puntos (ADAS-cog). Las diferencias fueron estadísticamente significativas pero los propios autores se cuestionan la relevancia clínica de este hallazgo[7].

Validez externa. Extrapolar más allá del ensayo clínico

Otra cuestión que hay que tener en cuenta es que los resultados de un determinado estudio se refieren a los pacientes que han participado en ese estudio. Incluso si un efecto se considera como probablemente real y clínicamente relevante, hay otra pregunta a la que tenemos que responder: ¿pueden aplicarse los resultados a otros grupos de pacientes y a un paciente concreto? Ni los IC, ni los valores p nos pueden ayudar a interpretarlo. La evaluación de la validez externa se realiza sobre las características de los pacientes (criterios de inclusión y exclusión, proporción de cumplidores, etc.) y el nivel asistencial en el que se ha desarrollado el ensayo.

Otras cuestiones metodológicas relevantes

Las medidas de variables intermedias

No es infrecuente que se empleen variables intermedias *(surrogate endpoints)* en lugar de resultados finales con verdadera importancia clínica. Una variable intermedia es una medida de laboratorio o fisiológica utilizada como sustituta de una variable de resultados que mide cómo se siente el paciente, cómo funciona, o cuánto vive. A menos que se conozca con certeza que las variables intermedias están relacionadas con resultados clínicos relevantes, lo más prudente es interpretarlas con mucha cautela.

Ejemplos de variables intermedias y variables a las que sustituyen

- La tensión arterial como una medida intermedia del ictus.
- El grado de aterosclerosis en una angiografía coronaria como una medida del infarto de miocardio o muerte coronaria.
- El retorno de la circulación espontánea en el paro cardiaco como medida del retorno a una función neurológica o de supervivencia.
- La densidad ósea como medida del riesgo de fracturas.
- El perfil lipídico como medida del riesgo cardiovascular.
- La hemoglobina glucosilada (HbA1c) como medida de control del paciente diabético.
- El PSA como marcador del cáncer de próstata.

Lo ideal sería que las decisiones se tomaran sobre variables como la calidad de vida relacionada con la salud, morbilidad (infarto o ictus) o incluso mortalidad. Cuando utilizamos estas variables intermedias para realizar inferencias sobre determinados beneficios esperados, estamos asumiendo que hay una verdadera relación entre la medida intermedia y la de resultados, que hay una conexión clara entre el cambio en los valores de la intermedia y el resultado clínicamente relevante. Incluso, cuando las variables intermedias dan información fiable sobre los resultados clínicamente relevantes para los pacientes, el efecto de la medida intermedia tiene que ser grande, robusto y de suficiente duración como para poder realizar inferencias con credibilidad[8].

Por ejemplo, la terapia hormonal sustitutiva se empleó durante muchos años en mujeres postmenopáusicas con la creencia de que podría aportar beneficios de tipo cardiovascular por el hecho de que mejora el perfil lipídico. Sin embargo, cuando se realizó el ensayo *Women´s Health Initiative*[9] se observó que la incidencia de ictus y enfermedad coronaria era superior en las mujeres que tomaban la terapia hormonal, a pesar de esa mejora del perfil lipídico. La variable intermedia (perfil lipídico) no fue un buen predictor de los resultados en la variable importante (ictus o enfermedad coronaria).

Las variables compuestas

Las variables compuestas *(composite endpoints)* son aquellas en las se juntan dos o más variables y se consideran una medida única de resultados. Habitualmente se justifican bajo el supuesto de que el efecto de cada uno de los componentes es similar y que los pacientes atribuirán la misma importancia a cada uno de los componentes. Sin embargo, esto no es siempre así. Para interpretar correctamente las variables compuestas Montori y colaboradores[10] sugieren al clínico realizarse las siguientes preguntas:

- ¿Las variables individuales que componen la variable combinada son de igual importancia para los pacientes?
- ¿Se registró un número similar de episodios en las variables más y menos importantes?
- ¿Es posible que las variables individuales tengan reducción del riesgo similar?

- ¿La relevancia clínica de las variables individuales es similar?
- Entre las variables individuales… ¿las estimaciones puntuales de las reducciones del riesgo son similares y los IC lo suficientemente estrechos?

La respuesta a estas cuestiones determinará si es necesario examinar las variables individuales de manera separada.

Un ejemplo de variable combinada inadecuada es la utilizada en el ensayo MIRACL[11], en el que se compara la eficacia de atorvastatina frente a placebo en el síndrome coronario agudo. La variable propuesta es la suma de las siguientes: muerte + infarto no fatal + parada cardiaca con resucitación + isquemia miocárdica recurrente sintomática objetivada que requirió rehospitalización de emergencia.

Respondiendo a las preguntas anteriormente expuestas, podríamos realizar los siguientes comentarios.

No se puede decir que todos los componentes tengan la misma importancia para los pacientes. No es comparable la "muerte" o un "infarto" a sufrir una "rehospitalización por isquemia", por ejemplo.

La frecuencia de aparición de los distintos fenómenos que se incluyeron en la misma variable fue muy distinta. La mayor incidencia observada corresponde a la rehospitalización (7,3%), en contraposición a la parada cardiaca (0,5%).

En cuanto a la reducción de riesgo observado en cada una de las variables individuales, en las variables más sólidas (muerte e infarto) no se encontraron diferencias estadísticamente significativas entre los grupos en estudio [RR = 0,92 (0,75-1,13)]. Sin embargo, la rehospitalización contó con una reducción de riesgo mayor [RR = 0,74 (0,57-0,95)], lo que contribuyó a que la variable final registrase diferencias en el límite de la significación estadística [RR = 0,84 (0,70-1,00)].

Al margen de estas puntualizaciones, la variable combinada elegida no es sencilla sino más bien "extraña" y no parece que la lógica nos haga pensar en algo así para valorar la eficacia de una intervención sobre la disminución de episodios coronarios.

Por otro lado, además de razonar sobre si la variable elegida es adecuada para la enfermedad que se estudia, siempre es importante fijarnos en cuáles son los medicamentos evaluados para poder opinar sobre la variable combinada. Puede darse el caso de que esté compuesta por variables importantes todas ellas, pero la inclusión o exclusión de alguna de ellas de la variable combinada haga que se favorezca o perjudique a alguno de los grupos en estudio.

Los resultados de variables combinadas deben tomarse con cautela

Análisis de variables secundarias

La variable principal es aquella que permite responder al objetivo del estudio. El tamaño muestral estará calculado de manera que se incluya un número suficiente de individuos que permita obtener una información fiable sobre los resultados en dicha variable principal. Por otro lado, en todos los ensayos suelen describirse otras variables, que pueden ser interesantes, pero que no suelen contar con un número de casos suficiente como para poder establecer conclusiones sólidas. Se trata de las variables secundarias. La proliferación de variables secundarias que se recogen en cada ensayo clínico hará que alguna de ellas sea estadísticamente significativa, solo por azar.

Un ejemplo claro es el estudio ELITE, en el que se comparó la eficacia de un ARA II (losartán) frente a un IECA (captoprilo) en pacientes con insuficiencia cardiaca. La variable principal combinada consistía en muerte y/o ingreso hospitalario por insuficiencia cardiaca. No se registraron dife-

Tabla 1. Resultados principales del ensayo ELITE

Variable principal	Losartán (n=352)	Captoprilo (n = 370)	Reducción de riesgo (IC)	p
Muerte y/o ingreso por insuficiencia cardiaca	33 (9,4%)	49 (13,2%)	0,32 (-0,04 to 0,55)	0,075
Otras variables				
Mortalidad total (variable secundaria)	17 (4,8%)	32 (8,7%)	**0,46 (0,05-0,69)**	**0,035**

Tabla 2. Resultados principales del ensayo ELITE II

Variable principal	Losartán (n=1.578)	Captoprilo (n = 1.574)	Hazards ratio (IC)	p
Mortalidad total (variable principal)	280 (17,7%)	250 (15,9%)	**1,13 (0,95-1,35)**	**0,16**
Otras variables				
Muerte súbita (variable secundaria)	130 (8,2%)	101 (6,4%)	1,30 (1,00-1,69)	

Los resultados de variables secundarias y post hoc solo sirven para generar hipótesis, NUNCA para tomar decisiones clínicas

rencias significativas entre ambos grupos. Sin embargo, se observó una reducción estadísticamente significativa de la mortalidad total en el caso del losartán respecto al captoprilo. El hecho de que se detectasen diferencias estadísticamente significativas en esta variable secundaria hizo pensar en una superioridad del losartán frente al captoprilo en la reducción de la mortalidad.

Por ello, se diseñó otro ensayo clínico (ELITE II), en el que la variable principal fue mortalidad total y se reprodujeron las características del ensayo anterior. Sin embargo, en este estudio no se encontraron diferencias significativas en la mortalidad, con lo que se rechazó la hipótesis generada con el estudio ELITE. Por otro lado, se observó una reducción de la muerte súbita (variable secundaria) de un 30% en el grupo captoprilo respecto al losartán. Sería tan incorrecto considerar que el losartán ofrece alguna ventaja en reducción de la mortalidad frente al captoprilo como decir que éste tiene mejores resultados que el losartán en prevenir la muerte súbita[12,13] (tablas 1 y 2).

En un ensayo, sólo podemos obtener información fiable de los resultados de la variable principal. Los datos de las variables secundarias sirven para generar hipótesis que tendrán que ser evaluadas en posteriores ensayos correctamente diseñados.

Análisis de subgrupos

Si un problema de salud puede variar en función de diferentes características, puede resultar práctico planificar la estimación del parámetro en los distintos subgrupos de interés. Si se desea realizar un análisis de subgrupos, deberá tenerse en cuenta en el cálculo del tamaño muestral y en el método de selección de los sujetos. De no hacerlo así, se perderá la precisión en la estimación del parámetro en cada subgrupo en relación con la obtenida cuando se analiza el total de la muestra, ya que el número total de sujetos será claramente inferior. En el momento de interpretar los resultados de un análisis de subgrupos hay que considerar tres aspectos importantes:

- La definición de los subgrupos y el análisis de los resultados en los mismos debe estar planificado de forma previa al desarrollo del ensayo. De no ser así, deberíamos ser cautos a la hora de interpretar los resultados.
- Es importante valorar la relevancia clínica de las diferencias encontradas entre los distintos subgrupos.
- El tratamiento estadístico debe ser el correcto. Hay veces que los autores simplemente publican las diferencias obtenidas en la variable principal en cada uno de los subgrupos y luego lo comparan entre dichos subgrupos. Este no es un abordaje adecuado. El hecho de que el resultado sea estadísticamente significativo en un subgrupo y no en el resto de subgrupos no significa que haya una diferencia real entre subgrupos. Las diferencias encontradas pueden estar condicionadas por el distinto tamaño muestral en cada subgrupo u otros motivos. Debe realizarse un test de interacción entre subgrupos para poder concluir con un mínimo de seguridad que existe alguna diferencia entre los distintos subgrupos[14].

Análisis *post hoc*

El análisis *post hoc* se trata de la realización de un subgrupo que no estaba previamente definido en el protocolo del estudio. Normalmente se seleccionan aquellos pacientes en los que la intervención ha sido más eficaz y se les agrupa intentando buscar alguna característica común.

Un ejemplo puede ser el estudio TROPOS[15] en el que se evaluó la eficacia del ranelato de estroncio frente a placebo en la prevención de fracturas de cadera. Los resultados no consiguieron demostrar eficacia alguna del medicamento [RR = 0,85 (0,61-1,19)]. Se realizó un análisis *post hoc* y se observó que las mujeres de más riesgo (edad media = 80 años, con fracturas previas en el 60% de los casos y con densidad ósea < - 3,5 DE) tenían resultados en el límite de la significación estadística [RR = 0,64 (0,41-1,00)].

Los resultados de los análisis *post hoc* no son válidos para demostrar o refutar la hipótesis del estudio y solo sirven para generar otras hipótesis que tendrán que ser demostradas adecuadamente. En ningún caso debemos condicionar nuestra práctica clínica por los datos obtenidos en un análisis *post hoc*. Este tipo de estrategias de análisis sirven para generar hipótesis que tendrán que ser comprobadas en ensayos adecuadamente diseñados. Otra posible utilidad es la de generar alertas de seguridad en base a datos de farmacovigilancia.

Ensayos de "no inferioridad"

Últimamente están proliferando los ensayos de equivalencia terapéutica y de "no inferioridad". Los primeros tratan de demostrar que dos intervenciones son similares desde un punto de vista clínico. Para ello se define **de forma arbitraria** la magnitud de la máxima diferencia clínica que el investigador considera que se puede aceptar para

considerar equivalentes los tratamientos. A este concepto se le denomina “delta (∂)”. Así, en los estudios de equivalencia se pretende demostrar que los efectos del fármaco en estudio se encuentran dentro del rango “ ±∂” cuando se compara con el control. Los estudios de “no inferioridad” se fijan en un solo lado del rango del intervalo, de modo que persiguen comprobar que el fármaco en estudio no está por debajo del valor -∂ respecto al control.

Los ensayos de equivalencia se han utilizado ampliamente para valorar los nuevos medicamentos, pero han perdido fuerza a favor de los ensayos de no inferioridad. Este tipo de ensayos son aceptados por las autoridades reguladoras para aprobar nuevos medicamentos o nuevas indicaciones. El uso de estos ensayos implica la actitud de no intentar probar ninguna ventaja del medicamento en cuestión sobre el comparador. Un ejemplo es el estudio COMPASS[16] en el que el trombolítico saruplasa era equivalente a la estreptoquinasa en el postinfarto a pesar de tener un 50% más de mortalidad. Hay quien opina que los ensayos de no inferioridad no son éticos ya que exponen al paciente a experimentos clínicos sin ninguna seguridad de que el medicamento examinado no es peor que el tratamiento estándar y sin realmente analizar cuánto es mejor[17].

Metanálisis. Influencia de la calidad de los estudios individuales

Los metanálisis van a dar una información agregada de los resultados de distintos ensayos individuales. Por ello, es importante tener en consideración todas las cuestiones mencionadas hasta aquí para cada uno de los estudios incluidos. La distinta calidad de los mismos puede condicionar los resultados finales del metanálisis de forma importante.

Un ejemplo gráfico es un metanálisis recientemente publicado sobre la eficacia de la diacereína frente a placebo en el tratamiento de la artrosis[18]. Para ello se incluyeron diecinueve ensayos que valoraban los resultados de este fármaco en el dolor y en la movilidad. La conclusión de los autores es que la diacereína produce unos efectos beneficiosos en ambos parámetros, si bien de pequeña magnitud. Sin embargo, dos de los ensayos eran muy superiores en calidad al resto y en ambos no había diferencias significativas entre la diacereina y el placebo. Si comparamos estos dos ensayos con el resto, observamos notables diferencias en la puntuación media de la escala de JADAD[19] (4,5 frente a 2,7), en la duración media de los estudios (24 frente a 2,8 meses) y en el número medio de pacientes incluidos en cada ensayo (404 frente a 119) (figura 1).

Si se hubieran incluido en el metanálisis únicamente los estudios de calidad las conclusiones hubieran sido opuestas a las enunciadas por los autores.

El paciente debería tener información clara en términos de riesgos absolutos para poder elegir más libremente

Y el paciente... ¿qué opina sobre el tratamiento que le proponemos?

Hoy en día se acepta que el paciente debería implicarse en la decisión terapéutica. Su opinión es particularmente importante en los casos en los que el tratamiento ofrece beneficios marginales, como ocurre en algunas patologías crónicas. Sin embargo, la medicina basada en la aplicación de guías de práctica clínica normalmente no contempla la opinión del enfermo.

Los pacientes no suelen ser adecuadamente informados sobre el tratamiento que se les prescribe. Las guías basan habitualmente sus recomendaciones en ensayos clínicos a gran escala que presentan los resultados de forma compleja. Este hecho, sumado al escaso tiempo que dispone el médico en su consulta, hace que la comunicación entre el facultativo y el paciente sea más difícil. Sin embargo, hay estudios que muestran que, cuando se le explica al paciente la eficacia del tratamiento en clave de reducción de riesgo absoluto, la mayor parte de ellos declinan tomar el tratamiento recomendado en las guías.

La opinión del paciente bien informado puede fomentar un enfoque más crítico de la investigación clínica y de la medicina basada en las guías de práctica clínica[20].

Conclusiones

Debe prestarse especial atención a la asignación aleatoria de los pacientes a los grupos en estudio y al enmascaramiento para evitar los principales sesgos de diseño de los ECA.

Independientemente de que el ensayo esté bien diseñado, en uno de cada veinte hallazgos se dará un resultado estadísticamente significativo que en realidad no lo es, simplemente por azar.

La significación estadística no significa necesariamente que el hallazgo sea clínicamente relevante. Es el tamaño del efecto el que determina la importancia, no la presencia de significación estadística.

Es preferible utilizar IC que valores "p". Ambos expresan la significación estadística pero el IC nos da información adicional como el tamaño del efecto y la precisión del resultado (amplitud del intervalo).

Es importante fijarse en las características de los pacientes incluidos en el ensayo para evaluar la validez externa de los resultados obtenidos.

Los resultados obtenidos en variables combinadas hay que analizarlos con criterio y cautela.

Los resultados obtenidos de variables secundarias y análisis post hoc solo sirven para generar hipótesis que tendrán que ser comprobadas en ensayos adecuadamente diseñados. No debemos trasladar a la práctica clínica la información procedente de estas variables.

Los ensayos de "no inferioridad" pretenden probar que el fármaco en estudio no es peor que el control, asumiendo la irrelevancia clínica de una cierta variabilidad en los resultados de ambos fármacos, que es establecida de forma arbitraria por los investigadores.

Los datos de un solo ensayo clínico no pueden justificar el cambio de la práctica clínica. Se necesita más evidencia que apoye o rechace los hallazgos del ensayo.

Debería trasladarse al paciente de forma objetiva y entendible la información de los resultados de los ensayos clínicos, de modo que pueda decidir sobre su tratamiento de forma más libre.

Revistas Secundarias / Journal Clubs

Realizan un resumen con comentario crítico de los ECAs, ayudan y facilitan la lectura crítica

ACP: http://www.acpjc.org/

Evidence Based Medicine [http://ebm.bmjjournals.com/]

Evidence Based Medicine en castellano [http://ebm.isciii.es/sumarios.asp]

Evidecence Based Mental Health [http://ebmh.bmjjournals.com/]

Evidecence Based Mental Health en castellano [http://ebmh.isciii.es/]

Evidence Based Nursing [http://ebn.bmjjournals.com/]

Evidencias en Pediatría [http://www.aepap.org/EvidPediatr/index.htm]

Recursos en la red sobre medicina basada en la evidencia en la red

http://www.fisterra.com

http://www.infodoctor.org/rafabravo/

http://www.redcaspe.org/

BIBLIOGRAFÍA

1. Laporte J. Principios Básicos de Investigación Clínica. Barcelona: AstraZeneca 2001

2. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273(5):408-12.

3. Argimon J. El intervalo de confianza: algo más que un valor de significación estadística. Medicina Clínica. 2002;118(10):382-4.

4 Pita S, López de Ullibarri I. Número necesario de pacientes a tratar para reducir un evento CAD ATEN PRIMARIA 1998 [cited; 96-8]. Available from: http://www.fisterra.com/mbe/investiga/5nnt/5nnt.asp

5. Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, Macfarlane PW, et al. Prevention of Coronary Heart Disease with Pravastatin in Men with Hypercholesterolemia. N Engl J Med 1995;333(20):1301-8.

6. Chapple CR, Rechberger T, Al-Shukri S, Meffan P, Everaert K, Huang M, et al. Randomized, double-blind placebo- and tolterodine-controlled trial of the once-daily antimuscarinic agent solifenacin in patients with symptomatic overactive bladder. BJU International 2004;93(3):303-10.

7. Kavirajan H and Schneider LS. Efficacy and adverse effects of cholinesterase inhibitors and memantine in vascular dementia: a meta-analysis of randomised controlled trials. Lancet Neurol 2007;6:782-92.

8. Freemantle N, Calvert M. Composite and surrogate outcomes in randomised controlled trials. BMJ 2007;334(7597):756-7.

9. Women"s Health Intitiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. JAMA 2002;288(3):321-333.

10. Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse JW, Pacheco-Huergo V, Bryant D, et al. Validity of composite end points in clinical trials. BMJ 2005;330(7491):594-6.

11. Schwartz GG, Olsson AG, Ezekowitz MD, Ganz P, Oliver MF, Waters D, et al. Effects of Atorvastatin on Early Recurrent Ischemic Events in Acute Coronary Syndromes: The MIRACL Study: A Randomized Controlled Trial. JAMA 2001;285(13):1711-8.

12. Pitt B, Poole-Wilson PA, Segal R, Martinez FA, Dickstein K, Camm AJ, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomised trial - the Losartan Heart Failure Survival Study ELITE II. Lancet 2000;355: 1582-7

13. Pitt B, Segal R, Martinez FA, Meurers G, Cowley AJ, Thomas I, et al. Randomised trial of losartan versus captopril in patients over 65 with heart failure (Evaluation of Losartan in the Elderly Study, ELITE). Lancet 1997;349:747-52.

14. Fletcher J. Subgroup analyses: how to avoid being misled. BMJ 2007;335:96-7.

15. Reginster JY, Seeman E, De Vernejoul MC, Adami S, Compston J, Phenekos C, et al. Strontium Ranelate Reduces the Risk of Nonvertebral Fractures in Postmenopausal Women with Osteoporosis: Treatment of Peripheral Osteoporosis (TROPOS) Study. J Clin Endocrinol Metab 2005;90(5):2816-22

16. Tebbe U, Michels R, Adgey J, Boland J, Caspi A, Charbonnier B, et al. Randomized, double-blind study comparing saruplase with streptokinase therapy in acute myocardial infarction: the COMPASS Equivalence Trial. Comparison Trial of Saruplase and Streptokinase (COMPASS) Investigators. J Am Coll Cardiol 1998;31(3): 494-6

17. Garattini S, Bertele V. Non-inferiority trials are unethical because they disregard patients" interests. Lancet 2007;370:1875-7.

18. Rintelen B, Neumann K, Leeb B. A meta-analysis of controlled clinical studies with diacerein in the treatment of osteoarthritis. Arch Intern Med 2006;166(17):1899-906.

19. Jadad A, Moore R, Carroll D, Jenkinson C, Reynolds D, Gavaghan D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clin Trials 1996;17:1-12.

20. Penston J. Patient´s preferences shed light on the murky world of guideline-based medicine. Journal of Evaluation in Clinical Practice 2007;13:154-9.