Sociedad Iberoamericana de Información Científica

Colección Trabajos Distinguidos Serie Medicina Farmacéutica

Volumen 1, Número 3, Abril 2004

Sociedad Argentina de Medicina Farmacéutica

Sociedad Brasileira de Medicina Farmacéutica

Expertos Invitados

EL NUMERO DE PACIENTES NECESARIO A TRATAR ES UNA IMPORTANTE MEDIDA DE RESULTADO EN ENSAYOS CLINICOS

Columnista Experto de SIIC

Dr. Marcos Gastón Duffau Toro

Profesor Titular de Pediatría. Dirección de la asignatura Investigación para Médicos en el Programa de Especialización en Pediatria. Preside la Comisión de Evaluación Académica de la Facultad de Medicina de la Universidad de Chile.

En los ensayos clínicos donde se comparan dos formas de manejo de los integrantes de la investigación, se pueden describir varias maneras de expresar los resultados. Desde luego, si éstos son dicotómicos y se ordenan los valores numéricos en una tabla de 2 x 2, podríamos comparar la distribución de los valores obtenidos con la distribución teórica bajo la hipótesis nula de que no hay diferencias de los grupos en comparación: chi cuadrado, entonces. Ello nos dará en forma general la información de cuán probable es que los resultados que se han volcado en la tabla tengan una distribución atribuible o explicable por el azar. Si no se cumplen los requisitos que exige esta prueba de hipótesis, se pueden analizar los datos empleando la prueba de Fisher-Irwin de probabilidades exactas. Igualmente, se puede manejar como la comparación de dos proporciones por una aproximación a «Z» así como analizarlos de acuerdo con sus intervalos de confianza respectivos. Como sea, se obtendrá "p", es decir el error alfa. Hasta aquí no se tendrá más información. Sin embargo, para cada grupo tenemos a disposición el total de casos y la parte que tuvo el resultado en estudio. De tal manera, veremos que la información logra ir bastante más lejos que lo indicado hasta este instante. Si consideramos los dos grupos sugeridos, sometidos a tratamiento experimental y control, esperando un resultado determinado, tendríamos:

	Resultado esperado		
	Sí	No	
Grupo experimental	a	b	a+b
Grupo control	c	d	c+d
Total	a+c	b+d	a+b+c+d

Asignando valores para uso posterior, supongamos que:
a = 12; b = 104; c = 28; d = 101.

El riesgo corresponde a la probabilidad de ocurrencia de un suceso (generalmente no deseado). De este modo, si del grupo experimental se espera que tenga con menor frecuencia un resultado adverso («sí», en la tabla), es decir que el tratamiento experimental tenga efecto benéfico, la presentación de la información puede ser como a continuación:

Riesgo absoluto (RA) en el grupo control = c/(c+d) = 28/129 = 0.217
Riesgo absoluto (RA) en el grupo experimental = a/(a+b) = 12/116 = 0.103
Riesgo relativo (RR) = (RA del grupo experimental)/(RA del grupo control)
= 0.103 / 0.217 = 0.48
IC 95% del RR, 0.25 – 0.89 (por Statcalc de EpiInfo)

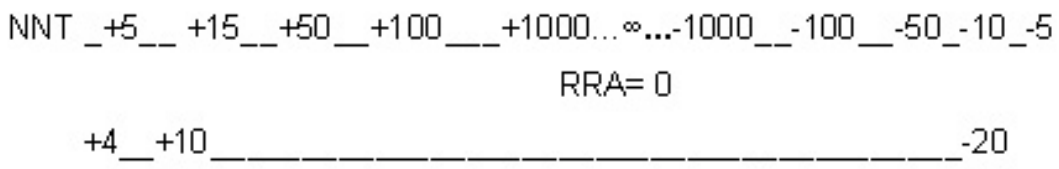
El riesgo relativo (RR) señala cuánto más probable o menos probable es el efecto o resultado en el grupo experimental que en el grupo control. Al valor obtenido se puede agregar su intervalo de confianza, generalmente del 95%, que nos indicará las cifras entre las cuales se encontraría, con un 95% de probabilidad, el RR del universo del que proceden los casos estudiados. Para que el RR encontrado en la investigación pueda considerarse de interés, su IC no debe incluir el valor «1», puesto que esa cifra indica riesgos absolutos iguales. La llamada «reducción del riesgo absoluto» (RRA) es simplemente la disminución del riesgo observada del grupo control al grupo experimental y corresponde entonces a su diferencia

RA del grupo control - RA del grupo experimental = RRA
0.217 – 0.103 = 0.114
21.7% - 10.3% = 11.4%
IC 95% de la RRA, 0.0239 – 0.2041 (2.39% - 20.41%) (por MedCalc)

Otra expresión de interés es la «reducción relativa del riesgo» (RRR) (no se equivoque, esta no es «reducción del riesgo relativo») que es la proporción que representa la reducción del riesgo absoluto (RRA) respecto del riesgo basal, que sería el del grupo control. Entonces

(RRA/RA del grupo control) x 100 = RRR

Si tenemos el valor de la reducción del riesgo absoluto y ésta es de, como se estableció en el ejemplo, 0.114 o 11.4%, podemos interpretarla en el sentido que de cada 100 sujetos tratados con el esquema experimental 11.4 de ellos obtendrán el beneficio (adicional) que representa el nuevo tratamiento. Ahora, 1 de ellos obtendrá el beneficio adicional por cada 100/11.4 = 8.77 tratados (la cifra se redondea al entero superior, de modo que quedaría en 9 casos). Esto representa el llamado «número necesario a tratar» (NNT) para obtener 1 caso adicional beneficiado. Habitualmente se le expresa como el valor inverso de la reducción del riesgo absoluto. Para el ejemplo planteado, en que el tratamiento experimental reduce el riesgo de un suceso indeseado, 9 pacientes deben ser tratados en el grupo experimental para evitar un episodio adicional no deseado. Conviene recordar que la expresión de un resultado sólo como RR plantea un problema, ya que por tratarse de una relación entre dos riesgos absolutos, un mismo valor del RR puede representar riesgos absolutos totalmente diferentes. Así, un RR de 2 puede ser el resultado de 0.40 / 0.20 o 0.05 / 0.025 , etc. En ambos casos la RRR es 50% , sin embargo NNT en la primera situación es de 5 casos y en el segundo es 40 casos, lo que con toda seguridad afectará las decisiones clínicas. El intervalo de confianza del NNT no parece ser un asunto definitivamente resuelto, pero se le puede estimar utilizando los valores inversos de los extremos del intervalo de confianza de la RRA : 1 / 0.0239 o 100 / 2.39 = 41.8 ,es decir 42 casos a 1 / 0.2041 o 100 / 20.41 = 4.89, es decir 5 casos. Entonces, para la situación propuesta, NNT es 9 pacientes con un IC del 95% de 5 a 42. Cuando el IC de la RRA pasa por cero, es decir cuando los riesgos absolutos no se pueden considerar significativamente diferentes, el IC 95% del NNT presentará un problema de interpretación. Supongamos una RRA de 10% (0.10) con IC 95% de -5% a +25%. En tal caso NNT sería 10 con extremos del IC de –20 a + 4. Lo que inmediatamente llamaría la atención es que el valor del NNT , 10, no parece estar contenido en su intervalo de confianza. Esto no sería tal si vemos lo siguiente:



Una lectura de este IC 95% sería, por ejemplo, que se requiere tratar 10 pacientes para obtener un caso adicional con el beneficio deseado variando desde 4 pacientes para obtener tal beneficio en uno adicional hasta 20 para reducir en un caso los favorecidos, comparado con el tratamiento de contraste.

Puesto que los casos de un grupo son heterogéneos en su nivel de riesgo, se podría hacer un intento (un tanto difícil) de caracterizar el paciente individual en cuanto a su probabilidad específica (expresada como una fracción) de un evento respecto al valor medio con el que se obtuvo el NNT del grupo. Una vez estimada esta fracción para un caso determinado, se divide NNT por ella y se obtendrá el NNT para ése tipo de paciente.

Cuando el tratamiento experimental aumenta la probabilidad de un suceso favorable los índices se modifican un tanto y así tendremos el aumento relativo del beneficio (ARB) en que se comparará el incremento de frecuencia de sucesos favorables con la frecuencia basal dada por el grupo control. Así, si en éste último la frecuencia es 30% y en el grupo experimental 40%, el incremento de 10% se expresará en relación al 30% del grupo control y obtendremos

(10% / 30%) x 100 = 33.33%

Este valor es en definitiva el ARB. El aumento absoluto del beneficio (AAB) se expresa como la diferencia simple entre los dos grupos respecto a la frecuencia de sucesos beneficiosos considerados. En el ejemplo, AAB sería de 10%. Su valor inverso corresponde al NNT, que aquí sería 100/10 = 10 es decir, 10 casos deben recibir el tratamiento experimental para obtener la ocurrencia de un suceso favorable adicional, comparado con el grupo control.

Si el tratamiento experimental causa cierto daño, es decir cuando aumenta la probabilidad de un suceso no deseado, comparado con el tratamiento control, nuevamente la situación cambia algo en el análisis de los resultados. Así, el aumento relativo del riesgo (ARR) se calcularía del mismo modo que el ARB, sólo que en vez de referirse al aumento del beneficio lo hace respecto al aumento del riesgo. Nuevamente tenemos aquí la opción de medir el cambio del riesgo absoluto, en este caso aumento del mismo (ARA), lo que corresponde a la diferencia de riesgo en ambos grupos de tratamiento.

Como en los casos anteriores, el valor inverso de ARA proporcionará el número de pacientes con tratamiento experimental necesarios para generar un caso adicional de daño comparado con la situación del grupo control (NND).

El valor de NNT puede ser obtenido también en los casos en que se dispone sólo del riesgo relativo (RR) o de la razón de disparidad (OR):
A partir del RR.
Si el RR es menor que 1:

NNT = 1 / (1 – RR) x Fc
Fc= Frecuencia esperada en los controles.

Con los datos de la tabla vemos que el grupo experimental tiene menor riesgo que los controles 0.103 vs. 0.217, lo que indicaría que el tratamiento nuevo es «protector» y el RR es inferior a 1, es decir, 0,4746. Entonces,

NNT = 1 / (1 – 0.4746) x 0.217 = 8.77

y se redondea a 9, cifra que coincide con la encontrada previamente. Si el RR es mayor que 1:

NNT = 1 / (RR – 1) x Fc.

El mismo ejemplo anterior Ud. podría verlo desde el punto de vista inverso, es decir que el riesgo en los controles es mayor que en el grupo experimental y el RR sería 0.217 / 0.103 = 2.10.

Empleando la fórmula propuesta para el caso en que RR es mayor que 1:

NNT = 1 / (2.1068 – 1) x 0.103 = 8,77

que se redondea a 9.
Como se puede ver, el referente es el grupo experimental, que aquí pasaría a ser «control» respecto del cual se informará del resultado del otro grupo. Lo que cambia, de una a otra forma de cálculo, es la lectura del resultado. En el primer caso «se requiere tratar a 9 pacientes con el tratamiento experimental ("protector") para evitar el riesgo en un paciente adicional respecto del grupo (tratamiento) control". En el segundo caso "se requiere tratar a 9 pacientes para obtener un caso adicional padeciendo el riesgo asociado al tratamiento control". Cuando se dispone de OR, se puede obtener NNT de la siguiente manera: Si OR es menor que 1:

NNT = 1 - [p0 (1 – OR)] / (1 - p0) p0 (1 - OR)

Si OR es mayor que 1 :

NNT = 1 + [p0 (OR - 1)] / (1 - p0) p0 (OR - 1)
p0 = Frecuencia esperada en el paciente (exposición en los controles).

En un ejemplo:

		Casos	Controles	
Exposición	(+)	100	150	250
	(-)	20	80	100
	Total	120	230	350

p0 = Exposición en controles = 150/230 = 0.65
OR= 2.67
NNT = 1 + [0.65 (2.67 – 1)] / (1 – 0.65) 0.65 (2.67 – 1) = 5.48

cifra que se aproxima a 6. Se lee así: Por cada 6 expuestos, se agrega un "caso" adicional.

Umbral de NNT. Si bien en muchas oportunidades, conociendo el resultado de NNT se podrá intuitivamente considerar que es una cifra de interés, en otros casos se requerirá establecer un umbral, que si se trata de obtener un beneficio por ejemplo, permitirá decir que si NNT está por debajo de él es claramente conveniente, pero si está por encima muy probablemente ya no lo es. Dos formas de estimación que se han propuesto, trabajan con el aspecto económico una de ellas y con el aspecto clínico la otra.

NNT calculado por programas de computación. Mencionaremos que variados programas ofrecen esta función pero, como siempre, si el usuario no está familiarizado con las bases bioestadísticas que subyacen, es muy probable que pueda equivocarse en la introducción de la información o en la interpretación de los resultados.

BIBLIOGRAFÍA

1. -Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence based medicine. How to practice and teach EBM.London: Churchill Livingstone, 1997: 136-141 , 168-170.
2. -Altman DG.: Confidence intervals for the number needed to treat. BMJ , 1998; 317: 1309- 1312.
3. -Altman DG, Andersen PK: Calculating the number needed to treat for trials where the outcome is time to an event. BMJ,1999;319:1492-1495.
4. -Smeeth L., Haines A., Ebrahim S.: Numbers needed to treat derived from meta- analyses. BMJ, 1999; 318:1548 – 1551.

DISEÑO ESTADÍSTICO DE ESTUDIOS CLÍNICOS EN FASE II Y SUS APLICACIONES EN CÁNCER DE MAMA: ACTUALIZACIÓN DE UNA INVESTIGACIÓN SOBRE LAS MODALIDADES DE EVALUACIÓN Y COMUNICACIÓN DE LA TOXICIDAD EN EL MISMO CONTEXTO CLÍNICO

Columnista Experto de SIIC
Dr. Ottaiano A, Di Maio M, De Maio E y Perrone F

Médico especialista en Oncología

"Estudio de modalidades de valoración y comunicación de toxicidad en estudios prospectivos no comparativos de quimioterapia en cáncer de mama".

Resumen

Recientemente realizamos una investigación [1] acerca de la forma de valorar y registrar la toxicidad en trabajos clínicos prospectivos no comparativos en cáncer de mama publicados entre 1995 y 1999 en siete revistas distinguidas de la especiali dad (*Annals of Oncology*; *Breast Cancer Research and Treatment*; *British Journal of Cancer*; *Cancer*; *Clinical Cancer Research*; *European Journal of Cancer* y *Journal of Clinical Oncology*). Los artículos incluid os se seleccionaron en forma manual; dos investigadores independientes completaron un formulario de registro (*study report form*,SRF). En estos formularios se captaron dos clases de información: datos sobre las características del estudio incluido (revista, año de publicación, año de inicio del trabajo, organización del estudio, apoyo de un patrocinador, número de ramas de la investigación, indicación de la fase del trabajo, diseño estadístico y puntos primarios de evaluación) y datos de las varia bles asociadas con el registro de toxicidad (escala empleada para codificar la toxicidad, indicación de exámenes planificados y su momento de realización, uso de tablas descriptivas de toxicidad, tipo de mediciones y frecuencia con la que se realizó hemo grama). La idoneidad de las modalidades de valoración de toxicidad se estableció en función de cuáles fueron las evaluaciones indicadas por los autores y en qué momento fueron planificadas. Las modalidades pudieron ser adecuadas cuando se publicaron los detalles de ambos interrogantes e inadecuadas cuando uno o ambos parámetros estaban ausentes. Las revistas se clasificaron en dos subgrupos según el factor de impacto: impacto muy alto, representado por el *Journal of Clinical Oncology*, que siempr e tuvo un puntaje cercano o superior a 7, y las de impacto alto en la que se incluyeron las seis publicaciones restantes, cuyo factor de impacto estuvo siempre por encima de dos pero por debajo de 4. Los datos se cruzaron en tablas de eventos fortuitos con cinco variables relacionadas con el contexto (número de instituciones participantes; patrocinador; presencia de un diseño estadístico identificable y presencia de un rótulo explícito de fase II). Finalmente evaluamos las asociaciones entre las modali dades de valoración y registro de la toxicidad y de los factores relacionados con el contexto por la prueba de chi cuadrado.

Entre los 122 estudios seleccionados y revisados encontramos que la escala de la OMS fue la utilizada más frecuentemente (45.9%) para evaluar toxicidad, seguida por las escalas CTC (35.2%). Las modalidades de valoración de toxicidad se comunicaron en forma inadecuada o no se comunicaron en más del 20% de los estudios. La toxicidad fue una variable primaria de evaluación en el 45.9% de los estudios y se resumió predominantemente por paciente (69.7%). Se identificaron tres patrones de frecuencia de solicitud de recuento de blancos: semanal (la modalidad más común); una vez al finalizar cada ciclo (la menos habitual) y más de una vez por semana. En el 21.3% de los trabajos no hubo información en relación con este parámetro. En la mayoría de los art ículos, la toxicidad y su gravedad fueron comunicadas en forma completa (82.8% y 68.9%, respectivamente). En forma llamativa notamos que un factor de alto impacto se asoció significativamente con un uso más frecuente de las escalas CTC (p= 0.001) y con m ayor frecuencia de hemogramas (p= 0.002). En los trabajos que refirieron los resultados de investigaciones multicéntricas más frecuentemente se adoptaron mediciones por paciente para comunicar la toxicidad (p= 0.006). La indicación explícita de la fase d e estudio se correlacionó con

el uso más frecuente de tablas para comunicar la toxicidad (p= 0.0006). Asimismo, la presencia de un patrocinador se correlacionó con un incremento relevante del uso de escalas CTC (p= 0.0006). El diseño estadístico identifi cable también se asoció en forma significativa con el uso de escalas CTC (p= 0.006) y con la aplicación de tablas para referir la toxicidad (p = 0.05). En forma similar, los estudios de inicio más reciente (1993-1997 *versus* 1986 a 1992) se asociar on con mayor uso de escalas CTC (p = 0.03) y de tablas para mostrar la toxicidad (p = 0.05). No hubo correlación significativa de las modalidades de valoración de la toxicidad y de su registro según el año de publicación y las variables principales de an álisis del estudio. Dada la amplia diversidad de modalidades de registro y comunicación de la toxicidad observada, en nuestra opinión los estándares actuales deberían ser revisados y compaginados para mejorar la confiabilidad de cada dato.

Aspectos metodológicos ocultos en estudios publicados de fase II de tratamiento de cáncer de mama

Durante el análisis de los datos previos otro defecto importante fue la falta de un diseño estadístico formal el cual sólo pudo identificarse en un tercio de los trabajos seleccionados (34.4%), de manera que la mayoría carecía de un plan estadístico de e studio y una estimación *a priori* del tamaño de la muestra. Por otro lado, observamos que un diseño estadístico se asoció con el uso más frecuente de las escalas CTC y de tablas de toxicidad. Por este motivo, en el trabajo actual prestamos mayor at ención a la aplicación de diseños en fase II en el contexto clínico del estudio previo y también analizamos datos de 23 trabajos de terapia hormonal que habían sido eliminados del ensayo anterior por el bajo índice de toxicidad. De hecho, esta actualizac ión se basó en 145 ensayos de tratamiento de cáncer de mama publicados en las mismas revistas entre 1995 y 1999. En la misma revisamos la magnitud de las estrategias estadísticas aplicadas en estudios en fase II de cáncer de mama. Recientemente se ha pu blicado un artículo extenso al respecto.[2].

Aplicación de los diseños de fase II a la investigación clínica en cáncer de mama: actualización del "Estudio de modalidades de valoración y comunicación de toxicidad en estudios prospectivos no comparativos de quimioterapia en cáncer de mama"

La investigación de drogas antineoplásicas se realiza con estudios en fase I, luego en fase II y, finalmente, con ensayos clínicos prospectivos en fase III. Las investigaciones en fase II tienen por objetivo evaluar si existe evidencia de acción antitumo ral que justifique estudios futuros con la droga experimental; así se reduce la probabilidad de planificar investigaciones prolongadas, costosas y no éticas con terapias ineficaces. La metodología de los estudios en fase II intenta minimizar el número de pacientes tratados con terapias posiblemente inútiles, reducir el riesgo de concluir erróneamente que el nuevo fármaco es ineficaz o de rechazar en forma equivocada un fármaco potencialmente útil. El diseño estadístico de los estudios en fase II puede a gruparse según las principales características: principales parámetros o criterios de valoración, número de tratamientos, tipo de estructura de la inferencia, cantidad de estadios, número de drogas (tabla 1).

TABLA 1. **Características principales del diseño estadístico en estudios en fase II.**

Punto de evaluación
Respuesta
Respuesta + índice de progresión
Tiempo hasta el evento (progresión o muerte)
Respuesta más toxicidad
Número de brazos
Uno
Múltiples (aleatorizado)
Estructura de inferencia
Prueba de la hipótesis
Estimación
Bayesiana
Teoría de selección
Número de estadios
Un-estadio (no es posible la interrupción precoz)
Múltiples estadios (la interrupción prematura se basa en inactividad, toxicidad o ambos)
Número de drogas
Un único agente
Combinación de drogas

También se tuvieron en cuenta otras variables relacionadas con la planificación de la fase de los estudios, pertinentes para esta actualización: presencia de un estudio en fase I, tipo de tratamiento experimental, cantidad de drogas (agentes en forma ais lada o en combinación), número de pacientes enrolados. La aleatorización no se consideró *per se* un diseño estadístico identificable. Se consideró que los trabajos no estuvieron diseñados cuando no se reconoció un método en la planificación del tam año de la muestra. Los artículos rotulados como de fase II pero planificados con métodos que son típicos en los trabajos en fase III se consideraron con diseño, a pesar de lo inapropiado que pudiese ser el plan estadístico. Se registraron otras variables relacionadas con el éxito del estudio como la duración y los resultados. Estos últimos se definieron como negativos cuando estuvieron explícitamente comunicados en esta forma o cuando fueron ambiguos pero estuvieron seguidos por la convicción categórica de que la droga en cuestión no era apta para estudios futuros. Aquellos artículos en los cuales se concluyó que el tratamiento debería ser posteriormente evaluado en estudios en fase III se consideraron positivos. La duración de la investigación se defi nió como el tiempo transcurrido desde el inicio del trabajo hasta su publicación, utilizando al año como medición de ambas; usualmente no se dispuso de información más precisa. Las correlaciones entre la presencia o no de un diseño estadístico identifica ble y de otras variables se establecieron con la prueba de chi cuadrado. Los valores de $P \leq 0.05$ se consideraron significativos. Las variables de contexto clínicamente significativas en el análisis univariado se incorporaron posteriormente en el model o de regresión logística de multivariado. Las asociaciones se comunicaron como *odds ratios* (OR) con intervalo de confianza de 95% (IC 95%). Se aplicó la prueba de orden de suma (*rank-sum*) de Mann-Whitney para comparar la duración de los est udios con diseño estadístico o sin él.

Las características generales y metodológicas de los estudios se muestran en la tabla 2 y 3.

TABLA 2. **Características generales de los 145 estudios seleccionados.**

	N (%)
Revista	
Ann Oncology	19 (13.1)
Cancer	13 (9.0)
Breast Cancer Res Treat	21 (14.5)
Br J Cancer	15 (10.3)
Clin Cancer Res	10 (6.9)
Eur J Cancer	18 (12.4)
J Clin Oncol	49 (33.8)
Año de publicación	
1995	29 (20.0)
1996	31 (21.4)
1997	25 (17.2)
1998	25 (17.2)
1999	35 (24.1)
Año de inicio del estudio	
Sin referencia	50 (34.5)
1986-1992	44 (30.3)
1993-1997	51 (35.2)
Centros participantes	
Unico	71 (49.0)
Múltiples	74 (51.0)
Apoyo de un patrocinador	
Ninguno o sin referencia	77 (53.1)
Parcial o total	68 (46.9)
Tratamiento experimental	
Quimioterapia	123 (84.8)
Terapia endócrina	18 (12.4)
Otras	4 (2.8)
Número de drogas	
Unico agente	62 (42.8)
Combinación	83 (57.2)

TABLA 3. **Características metodológicas de los 145 estudios seleccionados.**

	N (%)
Estudio previo en fase I	
Sin referencia	55 (37.9)
Con Referencia	90 (62.1)
Tipo de estudio	
Unico brazo	134 (92.4)
Aleatorizado	11 (7.6)
Punto primario de evaluación	
Respuesta (± toxicidad)	129 (89.0)
Sólo toxicidad	10 (6.9)
Otros	6 (4.1)
Número de pacientes enrolados (por cuartilos)	
≤ 26	36 (24.8)
27-38	36 (24.8)
39-50	36 (24.8)
≥ 51	37 (25.6)
Indicación de la fase de estudio	
No explícita	24 (16.6)
Fase II explícita (en el título o el texto)	121 (83.4)
Diseño estadístico del estudio	
No identificable	94 (64.8)
Identificable	51 (35.2)
Resultados del estudio	
Negativo	29 (20.0)
Positivo	116 (80.0)

En 50 (34.5%) no se dispuso de información sobre la fecha de inicio del trabajo. En los restantes 95, la duración promedio (tiempo entre el inicio y la publicación del estudio) fue de 4.5 años (DE 2.2). La mitad tuvo una organización multicéntrica. No se informó patrocinador en el 53.1% de los casos. Más de la mitad de las investigaciones (57.2%) evaluó una combinación de drogas y no un único agente. En el 37.9% de los artículos no hubo referencia de un estudio previo en fase I. Como era de esperar, la respuesta tumoral fue el punto primario de análisis en el 89% de los trabajos, en forma aislada o en simultáneo con el registro de toxicidad; este última fue el único parámetro de evolución en el 6.9% de las investigaciones. En número promedio de pacient es enrolados fue de 39 (rango intercuartilo: 26 a 51). En 24 (19.3%) de los ensayos no hubo indicación explícita de la fase del estudio; no se identificó un diseño estadístico en 94 (64.8%) trabajos. Entre estos estaban los 24 estudios en los cuales no h abía indicación explícita de la fase de investigación. La referencia de un estudio previo en fase I, el inicio del trabajo en años más recientes, el tratamiento experimental con un único fármaco, la organización multicéntrica y el apoyo de un patrocinado r se asociaron significativamente con la presencia de un diseño estadístico específico en el análisis de variables únicas (tabla 4). El tratamiento con un único agente (OR 2.35; IC 95%: 1.01-5.51) y la organización multicéntrica (OR 3.24; IC 95%: 1.47-7. 15) fueron factores predictivos independientes de la presencia de un plan estadístico en el modelo de regresión logística de múltiples variables (tabla 4). Tal como se muestra en la tabla 5, los trabajos con planificación estadística más frecuentemente s e publicaron en revistas de alto impacto y tuvieron, en forma global, menor duración: transcurrió alrededor de un año menos entre el momento de inicio y de publicación en comparación con aquellos sin planificación estadística (3.9 *versus* 4.9 años). No se encontró asociación entre el diseño estadístico del estudio y los resultados finales en general.

TABLA 4. **Análisis de variables únicas y múltiples de la asociación entre el diseño estadístico del estudio y las variables de contexto.**

Variables de contexto	No. (%) con diseño estadístico identificable (*)	P (chi-cuadrado)	OR (IC 95%)
Año de inicio del estudio		0.04	
1986-1992	10 (22.7)		1
Sin referencia	17 (34.0)		1.81 (0.67-4.89)
1993-1997	24 (47.1)		2.60 (0.98-6.91)
Centros participantes		0.0005	
Unico	15 (21.1)		1
Múltiples	36 (48.6)		3.24 (1.47-7.15)
Apoyo patrocinador		0.03	
Ninguno (sin referencia)	21 (27.3)		1
Parcial o total	30 (44.1)		1.04 (0.44-2.43)
Número de drogas		0.004	
Combinaciones	21 (25.3)		1
Unico agente	30 (48.4)		2.35 (1.01-5.51)
Estudio previo en fase I		0.02	
Sin referencia	13 (23.6)		1
Con Referencia	38 (42.2)		1.90 (0.83-4.32)

(*) Porcentajes marginales de las tablas de contingencia con inclusión de los 145 artículos: OR = *odds ratio*. IC 95% = intervalo de confianza de 95%

TABLA 5. **Asociación entre el diseño estadístico y las variables de valoración.**

Variables de valoración	Diseño estadístico identificable		P
	No (n=94)	Sí (n=51)	
Factor de impacto, n (%)			
Alto	69 (73.4)	27 (52.9)	0.01(*)
Muy alto	25 (26.6)	24 (47.1)	
Resultado del estudio, n (%)			
Negativo	17 (18.1)	12 (23.5)	0.43(*)
Positivo	77 (81.9)	39 (76.5)	
	No (n=61)	Si (n=34)	
Duración en años, media (DE)	4.9 (2.4)	3.9 (1.7)	0.01(**)

(*) Prueba de chi-cuadrado. (**) Prueba Mann-Whitney

Discusión, críticas y debates sobre los datos actualizados

En esta actualización sólo 51 de los 145 trabajos (35.2%) seleccionados para el análisis tuvieron un diseño explícito; la organización multicéntrica y el tratamiento con un único fármaco experimental fueron las variables que más se relacionaron con la pr esencia de un diseño estadístico. Los artículos con un plan estadístico formal fueron de menor duración y se publicaron con un factor de alto impacto. El índice de estudios sin un diseño formal es "demasiado" elevado (64.8%). En más de 20 artículos no se hizo referencia en ninguna parte a un diseño estadístico o en fase II. Si bien los mismos se comunicaron como estudios prospectivos es difícil saber si fueron realmente prospectivos o simplemente fueron la recolección retrospectiva de datos. Por ende, s us resultados son cuestionables porque las dos formas de valorar resultados (prospectiva *versus*

retrospectiva) pueden dar lugar a información muy distinta.

El hallazgo de que el diseño estadístico fuese más frecuente en estudios con un único fármaco en comparación con investigaciones de dos drogas en forma simultánea debe considerarse con mucha atención. De hecho, el objetivo de los estudios en fase II de combinación de drogas no es simplemente el de mostrar eficacia sino también revelar que la actividad alcanza un nivel suficiente de interés que justifica la realización de estudios más amplios en fase III. La falta de modelo estadístico complica l a interpretación de los resultados aún más que en los primeros estudios en fase II cuyo objetivo es mostrar, al menos, algo de actividad incluso cuando sea muy baja.

Sin embargo, el índice de artículos con comunicación de un diseño estadístico es mayor que el recientemente encontrado por Mariani y Marubini[3] quienes mostraron que en sólo el 19.7% de los 308 estudios en fase II de cáncer publicados du rante 1997 se identificaba un modelo estadístico. No obstante, estos investigadores prestaron atención a todos las revistas disponibles a través de Medline mientras que nuestra búsqueda se limitó a unas pocas revistas de muy buena calidad (por ejemplo, a quellas con un factor de impacto constantemente superior a 2 publicadas durante 1994-1999); es por ello que nuestros datos también deben considerarse negativos.

El índice bajo de estudios con planificación estadística puede tener varias explicaciones. En primer lugar, las asociaciones entre un plan estadístico y la organización multicéntrica y un inicio más reciente indican que la difusión de la cultura d e la metodología es cada vez mayor, particularmente en aquellas situaciones en las que ciertos aspectos metodológicos y estadísticos específicos se tienen en cuenta durante la planificación de la investigación. Sin embargo, en una enfermedad frecuente co mo lo es el cáncer de mama, es posible alcanzar el tamaño de la muestra requerida para la mayoría de los estudios en fase II en muchas unidades clínicas y esto favorece el inicio de ensayos en fase II sin planificación. Además, la forma usual de resumir los datos provenientes de estudios en fase II sobre una determinada droga es una forma elemental de comunicar el índice de respuesta o de toxicidad, en el mejor de los casos con intervalos de confianza; desafortunadamente, los datos rara vez son interpre tados y presentados acorde con el plan estadístico del estudio. Aún así, este tipo de interpretación requeriría una homogeneidad sustancial en los métodos para la planificación estadística que no podemos analizar en esta revisión por el escaso número de artículos con diseño encontrados.

Otro problema es la interpretación errónea del papel de los trabajos en fase II en investigación clínica. En forma ideal, deberían realizarse uno o unos pocos estudios en fase II para cada nueva droga o combinación de fármacos, inmediatamente desp ués de la investigación en fase I y, en caso de resultados positivos, antes de trabajos en fase III. La mayoría de los planes estadísticos incluyen aspectos éticos y operativos coherentes con este contexto. Lamentablemente, muchos de los artículos que re visamos no reúnen este paradigma fundamental. En aproximadamente el 40% de los estudios no se menciona un análisis previo en fase I. Algunos trabajos que abordan drogas no nuevas podrían leerse en forma optimista como estudios confirmatorios en fase II, pero más bien parecen tener la apariencia de un diseño hecho a medida acorde con la práctica clínica común. Por último, muchos artículos tienden a dar mensajes definitivos acerca de la utilidad clínica de la droga, a pesar de que deberían ser dados por estudios en fase III. Las futuras investigaciones de seguimiento podrían evaluar cuántos de estos estudios en fase II con hallazgos positivos culminan realmente en estudios en fase III.

Los ensayos en fase II aleatorizados[4] son particularmente proclives a este tipo de error, especialmente cuando se incluye un estándar o un brazo control como base de comparación. El pasaje desde el abordaje de selección (que es en sí pr eliminar) al abordaje de la evaluación de la hipótesis podría asociarse con un riesgo inaceptablemente alto de resultados falsos positivos. [5] Tal como se estableció en forma reciente, a menos que el estudio de seguimiento en fase III esté gar antizado por algún mecanismo externo -regulaciones gubernamentales para la aprobación de un nuevo fármaco; el diseño de selección puede ser más dañino que beneficioso por la propensión a ser usado en forma incorrecta. [6]

No encontramos diferencia en el número de pacientes enrolados en trabajos según la presencia o

no de un plan estadístico. Por supuesto, en los trabajos que carecen de plan no pudimos verificar *a posteriori* si el número de pacientes tratados era el adecuado. La selección de revistas de alto impacto podría nuevamente ser una posible explicación. Es factible que tales revistas acepten estudios bien planificados o sólo aquellos no planificados con un tamaño razonable de muestra (ni demasiado alto ni demasiado bajo). Sin embargo, esto no significa que se produzca información de la misma cantidad y calidad, independientemente del diseño estadístico ya que la interpretación de la mayoría de los estudios sin planificación sólo se deja al criteri o de sus Autores, frecuentemente no relacionado con los objetivos propuestos y la literatura de contexto. Si bien en el grupo de artículos que revisamos no hubo diferencia entre el índice de resultados negativos entre los ensayos con y sin planificación , los Autores usualmente tendieron a hacer hincapié en los hallazgos positivos y a minimizar los negativos. Sin el control adecuado de los hallazgos falsos positivos y falsos negativos, muchos trabajos con un bajo índice de respuesta son presentados como "bien tolerados". Es preocupante por ejemplo que la distribución de los índices de respuesta oscilara entre el 32% y el 94% en 16 trabajos limitados a la quimioterapia de primera línea en enfermedad en estadio IV que concluyen con un mensaje "positivo".

Los resultados de la revisión llaman la atención

Nuestra revisión demostró que sólo una minoría de los estudios en fase II en cáncer de mama publicados entre 1995 y 1999 en revistas de alta calidad tiene un buen diseño estadístico, fenómeno que se observó particularmente en aquellos con organización m ulticéntrica. La falta de un diseño formal aparentemente no indujo diferencias sustanciales en el número de pacientes enrolados y en el índice de resultados "positivos". Sin embargo, se asoció con un tiempo mayor desde el inicio hasta la publicación y co n un factor de impacto menor. Un número bastante grande de los trabajos seleccionados fue cuestionable por el hecho de que no parecieron ser verdaderamente prospectivos. En forma global, parece requerirse mayor aplicación de una metodología estadística e n la planificación de estudios en fase II en cáncer de mama con la finalidad de aumentar la confiabilidad de los resultados y de reducir el número de publicaciones innecesarias y a veces cuestionables.

BIBLIOGRAFÍA

1. Perrone F, De Maio E, Maione P, et al. Survey of modalities of toxicity assessment and reporting in noncomparative prospective studies of chemotherapy in breast cancer. J Clin Oncol 2002; 20: 52-7.
2. Perrone F, Di Maio M, De Maio E, et al: Statistical design in phase II clinical trials and its application in breast cancer. Lancet Oncol 2003; 4: 305-311.
3. Mariani L, Marubini E. Content and quality of currently published phase II cancer trials J Clin Oncol 2000; 18: 429-436.
4. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treatment Rep 1985; 69: 1375-1381.
5. Sylvester RJ. A bayesian approach to the design of phase II clinical trials. Biometrics 1988; 44: 823-836.
6. Liu PY, LeBlanc M, Desai M. False positive rates of randomized phase II designs. Control Clin Trials 1999: 20: 343-352.