

## ARTÍCULO DOCENTE

# Lectura crítica de la literatura biomédica (I): cómo valorar el riesgo de sesgo de un estudio



Jaime Latour-Pérez<sup>a,\*</sup> y Álvaro Rodríguez-Lescure<sup>b</sup>

<sup>a</sup> Servicio de Medicina Intensiva, Hospital General Universitario de Elche, Elche, España

<sup>b</sup> Servicio de Oncología Médica, Hospital General Universitario de Elche, Elche, España

Recibido el 29 de julio de 2014; aceptado el 3 de octubre de 2014

Disponible en Internet el 13 de diciembre de 2014

### PALABRAS CLAVE

Lectura crítica;  
Ensayos  
aleatorizados;  
Calidad de la  
evidencia;  
Guías de práctica  
clínica

### KEYWORDS

Critical appraisal;  
Randomized clinical  
trials;  
Quality of the  
evidence;  
Clinical practice  
guidelines

**Resumen** El ejercicio de una práctica clínica basada en la evidencia exige unas competencias básicas para valorar e interpretar la literatura biomédica. En este artículo se revisan los puntos críticos para determinar la validez de los resultados de un ensayo clínico, circunscribiéndonos al ensayo ACOSOG-Z0011.

© 2014 SESPM. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

**Critical appraisal of the biomedical literature (I): How to evaluate the risk of bias in a study**

**Abstract** Clinicians interested in evidence-based clinical practice need some basic competencies in critical appraisal of the medical literature. In this article we review the key points that determine the validity of the results of a specific clinical trial: the ACOSOG-Z0011 study.

© 2014 SESPM. Published by Elsevier España, S.L.U. All rights reserved.

## Introducción

En 1994, D.G. Altman publicó un famoso editorial en el que denunciaba el «escándalo de la investigación científica»<sup>1</sup>. En él señalaba que «gran parte de la investigación está

gravemente viciada por el uso de diseños inapropiados, muestras pequeñas (...), métodos incorrectos de análisis e interpretación defectuosa».

El crecimiento exponencial de la literatura biomédica ha contribuido a empeorar la situación en los últimos 20 años<sup>2,3</sup>, y hoy día cualquier lector avezado es consciente de que la publicación de un estudio en una revista de impacto no constituye en modo alguno una garantía de credibilidad.

Afortunadamente, en los últimos años se han desarrollado aproximaciones metodológicas racionales<sup>4,5</sup> para

\* Autor para correspondencia.

Correo electrónico: [jlatour@coma.es](mailto:jlatour@coma.es) (J. Latour-Pérez).

establecer el grado de evidencia y la fuerza de las recomendaciones incluidas en las guías de práctica clínica. Aunque inicialmente fueron diseñadas para su uso por paneles de expertos, estas metodologías ponen al alcance del clínico sin conocimientos estadísticos avanzados, herramientas sencillas para valorar críticamente la credibilidad de los resultados de los estudios y su aplicabilidad a sus enfermos individuales.

El objetivo de esta serie de artículos es mostrar la aplicación de la metodología *Grading of Recommendations Assessment, Development and Evaluation* (GRADE) a la valoración de la calidad de la evidencia y de su aplicación a un problema clínico concreto: la decisión de realizar vaciamiento ganglionar en mujeres con cáncer de mama y biopsia del ganglio centinela positiva.

## Escenario

Mujer de 53 años con adenocarcinoma ductal infiltrante de mama de 2,5 cm de diámetro, mal diferenciado (grado 3 de Scarff-Bloom-Richardson) y con invasión linfovascular, subsidiaria de cirugía conservadora. La biopsia selectiva del ganglio centinela evidencia 2 ganglios afectados de los 2 biopsiados, uno de ellos con un foco de 3,5 mm de diámetro y el otro con 2 focos de 1 y 2,2 mm, respectivamente. El estudio fenotípico correspondía a un perfil luminal B, con 70% de RE, 80% de RP, HER2neu negativo y ki67 del 30%.

Hasta hace unos años el tratamiento estándar en estos casos habría sido la lumpectomía asociada a vaciamiento axilar. Sin embargo, desde la publicación del estudio ACOSOG-Z0011<sup>6</sup> este abordaje está cuestionado, y algunas guías desaconsejan el vaciamiento ganglionar. Por otro lado, la enferma comenta que conoce un caso de cáncer de mama con vaciamiento axilar que ha quedado con un edema linfático importante, y teme que le pueda ocurrir lo mismo.

En este punto nos planteamos 2 preguntas: 1) ¿es segura la lumpectomía asociada a radioterapia (sin disección ganglionar) en estos casos?; y 2) ¿debemos recomendar a nuestra enferma vaciamiento ganglionar?

La primera es una pregunta de efectividad de una intervención, y se conoce bien que los estudios que en principio proporcionan una evidencia de mayor calidad para contestarla son los ensayos aleatorizados. La segunda, en cambio, plantea un problema de aplicabilidad, cuya resolución exige tomar en cuenta otros aspectos, además de la calidad de la evidencia.

Decidimos diferir la decisión hasta analizar pormenorizadamente el artículo ACOSOG-Z0011.

## Estudio ACOSOG-Z0011

Se trata de un ensayo aleatorizado multicéntrico paralelo, abierto, realizado en mujeres con cáncer de mama en estadio T1 o T2, susceptibles de tumorectomía, sin adenopatías palpables, con 1-2 metástasis en la biopsia del ganglio centinela. Las pacientes fueron asignadas aleatoriamente a disección ganglionar o a biopsia sin disección ganglionar. Todas las pacientes recibieron radioterapia. El tratamiento sistémico coadyuvante se dejó a criterio del clínico responsable. El desenlace clínico primario, registrado en [clinicaltrials.gov](http://clinicaltrials.gov) (NCT00003855) fue supervivencia

global y morbilidad relacionada con la disección ganglionar. El diseño inicial preveía reclutar 1.900 pacientes (950 por brazo), sin embargo el estudio se interrumpió con solo 891 pacientes debido a una mortalidad inferior a la prevista. Tras ajustar para el riesgo basal, el tratamiento conservador se asoció con una supervivencia no inferior al tratamiento con vaciamiento axilar (hazard ratio [HR] de mortalidad 0,87, IC 90%: entre 0,62 y 1,23).

## Interpretación de los resultados del estudio ACOSOG-Z0011

La hipótesis del estudio es que el tratamiento conservador (sin vaciamiento ganglionar) reduce las complicaciones quirúrgicas del tratamiento estándar sin comprometer sensiblemente la supervivencia global. En consecuencia, el estudio ACOSOG-Z0011 se plantea como un ensayo de «no inferioridad». En estos estudios solamente se pretende demostrar que el nuevo tratamiento no es peor que el antiguo, y se propone una diferencia aceptable que se considera como no relevante, en este caso un riesgo relativo de 1,30. En otras palabras, si la mortalidad con el tratamiento conservador supera en más de un 30% la mortalidad con el tratamiento estándar, entonces no podremos afirmar la no-inferioridad del tratamiento conservador. Si, por el contrario, demostramos la no inferioridad del tratamiento conservador, deberíamos evitar el vaciamiento ganglionar, con la consiguiente reducción de la morbilidad quirúrgica.

Los contrastes de hipótesis habituales (de 2 colas) tienen la siguiente estructura:

$$H_0: RR = 1; H' : RR \neq 1$$

Donde  $H_0$  = hipótesis nula;  $H'$  = hipótesis alternativa; RR = riesgo relativo (de mortalidad) del nuevo tratamiento frente a placebo o tratamiento estándar.

En este contexto el intervalo de confianza al 95% del RR nos informa tanto de su precisión como de la significación estadística: si el IC 95% no incluye la unidad, concluiremos que los resultados son estadísticamente significativos (con un nivel de significación del 5%) y aceptaremos la hipótesis alternativa (los 2 tratamientos tienen efectos diferentes).

A diferencia de los contrastes de hipótesis habituales (de 2 colas), los estudios de no inferioridad se plantean como un contraste de una cola:

$$H_0: RR > x; H' : RR \leq x$$

donde  $x$  es el umbral de aceptación (=1,30 en el estudio ACOSOG-AZ0011).

En este caso, el parámetro de interés es el intervalo de confianza de una cola, y el área de aceptación de  $H_0$  se sitúa entre 1,30 e infinito. Si el límite superior del intervalo de confianza es menor de 1,30 rechazaremos la hipótesis nula y concluiremos que el tratamiento conservador es no-inferior al tratamiento convencional.

Tal vez, el límite de 1,30 para definir la no-inferioridad pueda parecer a primera vista excesivo. Para aclarar este aspecto puede ser útil considerar el riesgo basal. Por ejemplo, si la mortalidad del grupo convencional es del 40%, un RR de 1,30 se corresponde con un aumento absoluto del

riesgo del 12% ( $[0,40 \times 1,30] - 0,40$ ); en cambio, si el riesgo basal es del 5%, el aumento del riesgo sería de solo un 1,5% ( $[0,05 \times 1,30] - 0,05$ ). En el estudio ACOSOG-Z0011 la mortalidad a los 5 años en el grupo convencional fue de 8,2%, por lo tanto el umbral de no inferioridad equivaldría a un aumento absoluto del riesgo del 2,46% ( $[0,082 \times 1,3] - 0,082$ ).

En el estudio de Giuliano et al. la magnitud del efecto se mide mediante el HR, pero a efectos prácticos podemos considerar el HR como una aproximación al RR. Tras ajustar para las diferencias pronósticas basales, el HR fue de 0,87 (IC 90% entre 0,62 y 1,23). Dado que el límite superior del IC 90% (de una cola) está por debajo del umbral prefijado (1,30) los autores rechazan la hipótesis nula y concluyen que la supervivencia global con el tratamiento conservador es no-inferior a la obtenida con vaciamiento ganglionar.

Antes de dar por bueno este resultado se decide examinar detenidamente el apartado de «Métodos», para explorar la validez del estudio.

## Riesgo de sesgo en el estudio ACOSOG-Z0011

El Manual de la Colaboración Cochrane propone un instrumento para valorar la validez de un ensayo aleatorizado<sup>7</sup>. Este instrumento examina 6 dominios que se han asociado empíricamente con la sobreestimación del efecto:

### Asignación mediante una lista aleatoria

La asignación aleatoria de los individuos a uno u otro de los brazos del estudio, si se realiza de forma apropiada, garantiza que los 2 grupos sean comparables en todas las variables pronósticas (conocidas o no) y, por tanto, que las diferencias observadas en el desenlace son atribuibles a la única variable en la que son diferentes (el grupo asignado). Este es el punto que confiere a los resultados de los ensayos aleatorizados una alta calidad de evidencia.

El estudio de Giuliano et al. no especifica el método para generar la secuencia de aleatorización<sup>6</sup>, sin embargo en un artículo previo del grupo<sup>8</sup> afirman que la asignación se hizo mediante ordenador, lo que hace suponer que se utilizó una lista aleatoria (o pseudoaleatoria).

### Ocultación de la secuencia de aleatorización

Para que los 2 grupos sean comparables no basta con que la asignación se haga a partir de una lista aleatoria. Es preciso, además, que el investigador desconozca el grupo al que va a ser asignado el siguiente paciente incluido en el estudio. El objetivo de la ocultación de la secuencia de aleatorización es conseguir que todos los pacientes tengan las mismas posibilidades de ser incluidos en el estudio, independientemente del grupo al que caigan asignados. Esto se consigue mediante procedimientos como el uso de sobres cerrados opacos, o el recurso a una central de aleatorización que obliga al investigador a registrar al paciente antes de conocer el tratamiento que se le va a aplicar.

En un estudio clásico, Schulz demostró que los estudios en los que no se clarificaba el método de ocultación de la secuencia de aleatorización sobreestimaban el efecto de la intervención<sup>9</sup>. El artículo de Giuliano<sup>6</sup> no especifica el método de ocultación de las secuencias de aleatorización,

sin embargo el referido estudio previo<sup>8</sup> sugiere que la asignación la realizó una central de aleatorización, en principio incompatible con la posibilidad de incluir preferencialmente en uno de los grupos a los pacientes de mejor pronóstico.

La ocultación de la secuencia de aleatorización (*concealed randomization*) no debe ser confundida con el enmascaramiento («ciego»). La ocultación de la secuencia de aleatorización se produce antes de la aleatorización y previene el sesgo de selección. El enmascaramiento (simple, doble, triple ciego) se produce después de la aleatorización, y su efecto consiste en conseguir que los 2 brazos se traten de idéntica forma y los resultados se midan de la misma manera (evitando el sesgo de medición).

### Enmascaramiento de pacientes, sanitarios y evaluadores

Por lo general, el impacto de la ausencia del enmascaramiento es menor que el de la ocultación de la secuencia de aleatorización, y depende en gran medida del tipo de desenlace estudiado: es mayor en los estudios que evalúan desenlaces subjetivos (como el dolor o la calidad de vida), y menos importante cuando se evalúan desenlaces «duros» como la mortalidad.

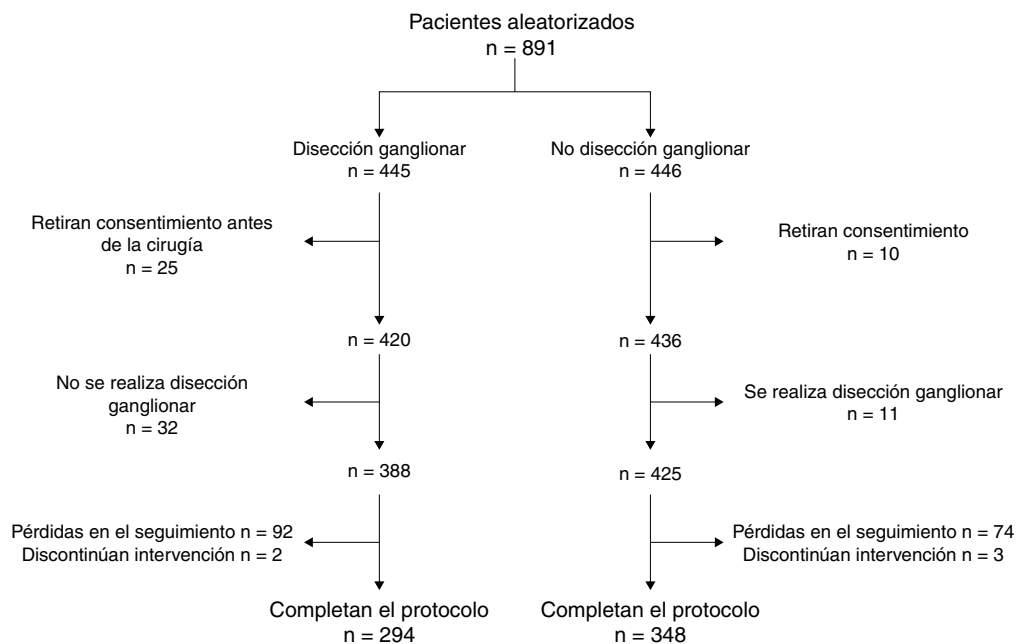
El estudio de Giuliano es un estudio «abierto» en el que tanto los pacientes como los investigadores conocen el brazo al que ha sido asignado el paciente. Teóricamente ello habría podido favorecer una mayor utilización de tratamiento adyuvante en el brazo de pacientes tratados sin disección ganglionar (cointervención), sin embargo la utilización de tratamiento adyuvante (96% frente a 97%) y de radioterapia (88,9% frente a 89,6%) fueron similares en ambos grupos. Por otro lado, el conocimiento del grupo asignado al enfermo podría haber contribuido a un sobrediagnóstico de linfedema en los pacientes tratados con vaciamiento axilar.

### Manejo de las pérdidas

Las pérdidas en el seguimiento, cuando se producen de forma aleatoria, no suponen una mayor dificultad. El problema se produce cuando las pérdidas de los 2 brazos tienen características pronósticas distintas. Ello puede hacer que 2 brazos inicialmente homogéneos se vayan haciendo diferentes conforme progresa el estudio, de modo que el efecto final observado no se deba al tratamiento evaluado, sino a las pérdidas en el seguimiento (sesgo de atrición).

Las guías CONSORT<sup>10</sup> proponen que los ensayos paralelos publicados incluyan un diagrama de flujo que presente de forma transparente el número de pacientes considerados para su posible inclusión, el número de pacientes aleatorizados, el número que recibió realmente cada uno de los tratamientos, el número de pérdidas en el seguimiento, etc. El artículo de Giuliano incluye una figura con un diagrama de flujo que no se ajusta a las recomendaciones CONSORT, por lo que resulta difícil identificar el número real de pérdidas en el seguimiento. En la figura 1 se intenta aclarar, en la medida de lo posible, estas pérdidas.

¿Cómo se deben analizar las pérdidas? El llamado análisis por protocolo incluye solo los pacientes que han completado al estudio, excluyendo las violaciones del protocolo (por



**Figura 1** Diagrama de flujo.

Reconstruido a partir de los datos de los estudios de Giuliano et al.<sup>6</sup> y Giuliano et al.<sup>11</sup>.

ejemplo pacientes que recibieron el tratamiento sin cumplir criterios de inclusión), *crossover* (por ejemplo pacientes asignados al grupo control que reciben el tratamiento experimental), pérdidas en el seguimiento, etc. En la medida que se excluye una parte de los pacientes, este análisis no preserva el efecto de la aleatorización. Por ejemplo, supongamos que los pacientes que rechazan el vaciamiento ganglionar tienen una enfermedad menos avanzada (de mejor pronóstico) que los pacientes que la aceptan. Si se excluyen estos pacientes del grupo tratado con vaciamiento axilar pero se mantienen en el grupo tratado de forma conservadora, entonces el análisis estaría favoreciendo al grupo conservador.

El análisis por intención de tratar intenta minimizar este sesgo analizando todos los pacientes en el grupo al que fueron asignados, independientemente de que finalmente no recibieran el tratamiento asignado o se perdieran sus datos de seguimiento.

El posible impacto de las pérdidas se puede valorar mediante el análisis en la peor de las situaciones posibles (*worst case analysis*). Para ello, se recalcula el efecto del tratamiento asumiendo que todas las pérdidas en uno de los brazos y ninguna de las pérdidas en el brazo alternativo son eventos (tabla 1). En el estudio de Giuliano et al. el número de pérdidas (el 19,4% de los pacientes incluidos en el estudio) es muy superior a la mortalidad observada (10,0%). Por lo tanto, el riesgo relativo es muy sensible al porcentaje de fallecimientos en cada uno de los brazos.

### Comunicación selectiva de resultados

Frecuentemente los estudios no comunican los resultados de todos los desenlaces predefinidos. Más concretamente, es más probable que se presenten los análisis con diferencias significativas que aquellos que no presentan diferencias

significativas. Esto contribuye a sobreestimar el beneficio de la intervención<sup>12</sup>.

El estudio de Giuliano et al. estaba prerregistrado en [clinicaltrials.gov](https://clinicaltrials.gov) (referencia NCT00003855). Aunque en dicho registro no se especifican con claridad los desenlaces primarios y secundarios, en el artículo están descritos todos los *endpoints* esperables.

### Otros dominios: grupos desequilibrados

Algunos ensayos presentan características específicas que cuestionan la credibilidad de los resultados. Una de ellas es el desequilibrio entre los grupos. En la tabla 1 del artículo de Giuliano et al. se puede ver que el grupo tratado con vaciamiento ganglionar tenía una mayor prevalencia de tumores en estadio T2 (32,1% frente a 29,4%) y de invasión linfovascular (40,6% frente a 35,2%) que el grupo tratado con solo biopsia.

El hecho de que estas diferencias sean o no significativas es irrelevante: de hecho, en los estudios pequeños (en los que son más frecuentes los grupos desequilibrados) diferencias importantes pueden no alcanzar significación estadística debido a falta de potencia (error de tipo II), mientras que en estudios grandes, diferencias irrelevantes puedan ser estadísticamente significativas. De hecho, las normas CONSORT desaconsejan el uso de la «p» para decidir sobre la existencia de diferencias en las características basales de los grupos.

La relevancia de estos desequilibrios basales en el estudio de Giuliano se pone de manifiesto en las diferencias entre las medidas de efecto crudas y las ajustadas para las covariables, que muestran de forma sistemática un HR más próximo a 1 cuando se toman en cuenta las covariables de los tratamientos.

**Tabla 1** Análisis por intención de tratar y análisis en la peor de las situaciones posibles en el estudio ACOSOG-Z0011

	Disección ganglionar	Tratamiento conservador
Pacientes (denominador)	420	436
Supervivientes %	91,8	92,5
Supervivientes n	$0,918 \times 420 \sim 386$	$0,925 \times 436 \sim 403$
Fallecidos	$420 - 386 = 34$	$436 - 403 = 33$
Mortalidad (análisis por intención de tratar)	$34/420 = 0,081 = 8,1\%$	$33/436 = 0,076 = 7,6\%$
Pérdidas en el seguimiento	92	74
Fallecidos entre los perdidos (peor de las situaciones posibles)	$0,00 \times 92 = 0$	$0,10 \times 74 \sim 7$
Mortalidad (análisis en la peor de las situaciones posible) 10% frente a 0% de las pérdidas = fallecidos	$(34 + 0)/420 = 0,081 = 8,1\%$	$(33 + 7)/436 = 0,092 = 9,2\%$
Riesgo relativo (análisis por intención de tratar)	$(33/436)/(34/420) = 0,935$	
Riesgo relativo (análisis en la peor de las situaciones posibles)	$((33 + 7)/436)/((34 + 0)/420) = 1,13$ (IC 90%: 0,787 a 1,633)	

## Conclusiones

En conclusión, una vez leído detenidamente el estudio podemos considerar que el ensayo ACOSOG-Z0011 presenta debilidades metodológicas importantes (en especial las pérdidas en el seguimiento y el desbalance en las variables pronósticas basales) que le restan credibilidad.

Por otro lado, somos conscientes de que la calidad de la evidencia no se puede valorar con un único estudio, y decidimos valorar el cuerpo de evidencia disponible. Este punto se abordará en el segundo artículo de esta serie.

## Responsabilidades éticas

**Protección de personas y animales.** Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

**Confidencialidad de los datos.** Los autores declaran que en este artículo no aparecen datos de pacientes.

**Derecho a la privacidad y consentimiento informado.** Los autores declaran que en este artículo no aparecen datos de pacientes.

## Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

## Bibliografía

- Altman DG. The scandal of poor medical research. *BMJ*. 1994;308:283-4.
- Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.

- Smith R. Medical research—still a scandal. 2014. [consultado 07 Noviembre 2014]. Disponible en: <http://blogs.bmj.com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/>
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924-6.
- CASPe[consultado 28 Feb 2014]. Disponible en: <http://www.redcaspe.org/drupal/>
- Giuliano AE, Hunt KK, Ballman KV, Beitsch PD, Whitworth PW, Blumencranz PW, et al. Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: A randomized clinical trial. *JAMA*. 2011;305:569-75.
- Higgins J, Green S. *Cochrane handbook for systematic reviews of interventions* version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011.
- Lucci A, McCall LM, Beitsch PD, Whitworth PW, Reintgen DS, Blumencranz PW, et al. Surgical complications associated with sentinel lymph node dissection (SLND) plus axillary lymph node dissection compared with SLND alone in the American College of Surgeons Oncology Group Trial Z0011. *J Clin Oncol*. 2007;25:3657-63.
- 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;273:408-12.
- Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *PLoS Med*. 2010;7:e1000251.
- Giuliano AE, McCall L, Beitsch P, Whitworth PW, Blumencranz P, Leitch AM, et al. Locoregional recurrence after sentinel lymph node dissection with or without axillary dissection in patients with sentinel lymph node metastases: The American College of Surgeons Oncology Group Z0011 randomized trial. *Ann Surg*. 2010;252:426-32.
- Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*. 2004;291:2457-65.