ARS MEDICA
revista de ciencias médicas

Artículo de práctica clínica
DOI: http://dx.doi.org/10.11565/arsmed.v39i1.9

Detención Precoz de Estudios Clínicos Randomizados: ¿Beneficio para los pacientes o para el investigador?: Sobrestimación de resultados, Reglas de Detención y Comités de Monitorización de Datos.

Roberto Candia Balboa [1,2.]
Gabriel Rada[1,3,4.]

Resumen - Los estudios clínicos randomizados detenidos precozmente por beneficio son cada vez más frecuentes en la literatura médica. La detención precoz se fundamenta en la necesidad ética de ofrecer rápidamente una intervención beneficiosa a toda la población en riesgo, especialmente a los pacientes que son parte de grupo control del estudio. A pesar que la detención se fundamenta en análisis que demuestran beneficio estadísticamente significativo, se sabe que los estudios truncados precozmente tienen riesgo de detectar diferencias que realmente no existen. Este fenómeno se produce porque habitualmente estos estudios son de pequeño tamaño muestral y están sustentados en análisis estadísticos repetitivos, realizados a medida que se reclutan los pacientes. El lector debe ser capaz de detectar estos estudios e identificar las herramientas para disminuir el riesgo de error y sesgo. El objetivo de la presente revisión es dar a conocer las herramientas disponibles para disminuir el riesgo, las estrategias estadísticas para sostener una adecuada detención temprana de un estudio, los comités de monitorización externa de los datos, la evidencia que sustentan estas afirmaciones y las iniciativas que se están llevando a cabo para cuantificar el riesgo y eventualmente ofrecer una solución.

Palabras clave: *Detención precoz por beneficio, estudios truncados por beneficio, O'Brien Fleming, Lan y De Mets, Comités de monitorización de datos, STOPIT-2.*

Abstract - Early stopping by benefit of randomized clinical studies are increasingly frequent in the medical literature. The early detection is based on the ethical need to quickly provide a beneficial intervention to the entire population at risk, particularly to patients who are part of study control group. Although the arrest was based on analyzes that show statistically significant benefit, it is known that early studies truncated have the risk to detect differences that do not exist. This phenomenon occurs because usually these studies have small sample size and are supported by repetitive statistical analyzes performed as patients are recruited. The reader should be able to detect these studies and identify tools to reduce the risk of error and bias. The aim of this review is to present the tools available to reduce risk, the statistical strategies to sustain an adequate early detection of an external monitoring data study committee, the evidence supporting these statements and initiatives that are being carried out to quantify the risk and eventually offer a solution.

***Keywords*:** *Early detection by benefit, stopped studies by benefit, O'Brien Fleming, Lan and De Mets, data monitoring committees, STOPIT-2.*

Fecha de envío: 01 de Junio de 2012 - Fecha de aceptación: 25 de Septiembre de 2012

Introducción

Es práctica habitual que los investigadores a cargo de ensayos clínicos randomizados (ECR) decidan realizar análisis interinos (evaluar los resultados antes de completar el tamaño muestral programado al inicio del estudio), con el objetivo de interrumpir el estudio precozmente si se detectan diferencias estadísticamente significativas. Esto motiva que los resultados sean obtenidos con un *n* pequeño, lo que aumenta el riesgo de *error* por azar, sobreestimando el efecto e incluso haciendo parecer efectivas intervenciones que no lo son (Grant *et al.*, 2005; Candia B *et al.*, 2006). En la actualidad no existe consenso entre los distintos grupos de expertos si es mejor completar siempre los estudios de acuerdo al tamaño muestral calculado al inicio, o

1) Unidad de Medicina Basada en Evidencia, Facultad de Medicina, Pontificia Universidad Católica de Chile. 2) Departamento de Gastroenterología, Facultad de Medicina, Pontificia Universidad Católica de Chile 3) Departamento de Medicina Interna, Facultad de Medicina, Pontificia Universidad Católica de Chile 4) Unidad docente asociada Hospital Dr. Sótero del Rio, Pontificia Universidad Católica de Chile.
*Autor de correspondencia: roberto.candia@gmail.com

ISSN: 0719-1855

interrumpir el estudio precozmente (por beneficio o por daño) en el momento que aparece una diferencia estadísticamente significativa, asumiendo el riesgo de *error*, muchas veces fundamentándose en razones éticas (Montori *et al.*, 2005; Candia B *et al.*, 2006; Bassler *et al.*, 2008; Peppercorn *et al.*, 2008).

El objetivo de éste artículo es dar a conocer al investigador algunas de las técnicas estadísticas creadas para disminuir el riesgo de error al interrumpir precozmente un ECR, las características que debieran tener los comités externos que vigilan la realización de tales análisis interinos y las ventajas y desventajas de la interrupción precoz por beneficio.

La encrucijada ética de una detención precoz

Cuando un investigador planifica un estudio, una de sus primeras tareas es calcular el número de pacientes necesarios para demostrar el beneficio de la terapia a investigar (cálculo de tamaño muestral). Este cálculo en algunas ocasiones puede ser impreciso, ya que el efecto real de la intervención puede ser mayor o menor al esperado. Esto nos pone frente a dos situaciones (Peppercorn *et al.*, 2008):

• Si el investigador detecta evidencia suficiente que la terapia en estudio es efectiva antes de completar el ECR, éste podría ser interrumpido para beneficiar a todos los pacientes (incluyendo los que están en el grupo control recibiendo placebo, por ejemplo).
• Al contrario, si el investigador detecta que la nueva terapia es dañina, el ECR debe ser interrumpido para evitar dañar a los que están en el grupo intervención recibiendo la nueva terapia.

Estas afirmaciones engloban a grosso modo la encrucijada ética que genera la interrupción precoz de un ECR:

• Por un lado está la responsabilidad ética con los pacientes reclutados para el estudio, a los cuales se les debe ofrecer la mejor terapia en el momento que se demuestre beneficio.
• Por otro lado está la responsabilidad ética con el resto de la comunidad, ya que una interrupción precoz genera resultados menos creíbles (un n pequeño tiene un mayor riesgo de error tipo I) (Grant *et al*., 2005; Montori *et al*., 2005; Candia B *et al*., 2006).

Hace más de 30 años que existe consenso entre los estadísticos en relación a éste último punto: el realizar múltiples análisis interinos con tamaños muestrales pequeños habitualmente sobrevalora las diferencias en los resultados (Betensky, 1998; Sankoh, 1999; Schulz & Grimes, 2005). Por lo tanto, si la interrupción es por beneficio, éste habitualmente está erróneamente sobrevalorado. A mayor número de análisis interinos, es más probable encontrar una *"diferencia estadística significativa"* espuria: si se analizan repetidamente los resultados de un estudio a medida que se reclutan pacientes, sólo por azar, en alguno de los cálculos se puede encontrar una diferencia estadística significativa, la que no necesariamente es real (*Figura 1*).

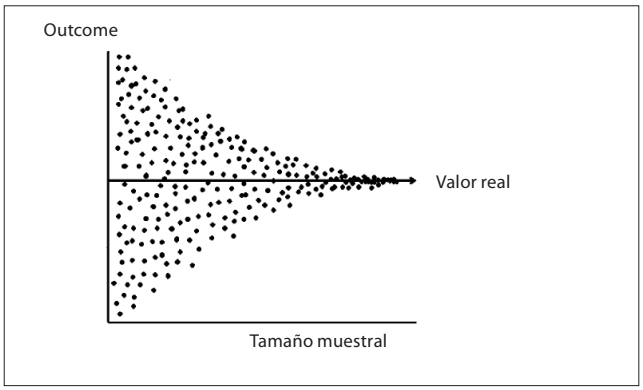

Figura 1.- Esquema que representa el comportamiento de los datos al realizar múltiples repeticiones de un estudio con distinto tamaño muestral. Cuando el estudio se repite con tamaños muestrales pequeños, la variación del outcome es azarosa. A mayor tamaño muestral, la variación del outcome del estudio se acerca con mayor precisión al valor real (el que se obtiene al aplicar la intervención a toda la población). Un estudio detenido precozmente genera datos imprecisos, ya que es el reflejo de múltiples análisis interinos realizados con un *n* pequeño (Candia B *et al.*, 2006).

Este fenómeno es análogo al que se produce al repetir un experimento en múltiples ocasiones: si la hipótesis nula es real (es decir, no existe diferencia entre los grupos), el realizar múltiples veces un experimento aumenta la probabilidad de encontrar diferencias que no existen en alguna de esas repeticiones, sólo por azar (error tipo I).

Por todo lo expresado anteriormente queda claro por qué esta encrucijada ética no está resuelta. Para algunos investigadores la interrupción precoz siempre llevará a una interpretación equivocada de los datos, por lo tanto, nunca se debiera aceptar ésta conducta, mientras que otros aún debaten activamente en qué situación ésta acción podría ser apropiada. Esta discusión parte de la posibilidad cierta que en algunos casos efectivamente una intervención puede tener un beneficio mayor al presupuestado inicialmente. En pocas palabras, tenemos por un lado la intención de evitar la inadecuada estimación del efecto, pero por otro queremos evitar que los estudios se prolonguen cuando ya sabemos que existe beneficio. En un intento de resolver este punto es que se han creado una serie de test estadísticos o reglas de detención, las que describiremos a continuación.

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

Técnicas o Reglas de Detención precoz por beneficio: que saber de ellas

Estas técnicas estadísticas fueron creadas como una forma de "corregir" el riesgo de error tipo I inherente al menor tamaño muestral y al alto número de análisis interinos que implica una detención precoz. Se fundamentan en el hecho que el nivel de significancia estadística necesario para interrumpir tempranamente un estudio por beneficio debe ser más exigente que el clásico p<0,05 exigido al final del estudio, donde el reclutamiento y el seguimiento de pacientes se ha completado. Así, exigiendo un valor *p* menor, con menor frecuencia se interrumpirá un estudio, por ende, es más difícil cometer un error tipo I y sobrestimar diferencias. Las técnicas más simples corresponden a Pocock (creada en 1977), Peto-Haybittle y O'Brien-Fleming (creada en 1979) (O'Brien & Fleming, 1979; Betensky, 1998; Schulz & Grimes, 2005), y todas ellas exigen que el número de análisis interinos sea programado a priori y a intervalos constantes (por ejemplo, si se realizaran 3 análisis, 2 interinos y el final, cada uno debe realizarse al completar un tercio del reclutamiento de pacientes). A continuación describiremos las características generales de cada una de éstas técnicas.

La regla de Pocock fue la primera en ser utilizada. Es una técnica estadística que calcula un valor *p* más exigente mientras mayor es el número de análisis interinos programados. A diferencia de otras técnicas, Pocock establece un valor *p* que es constante, es decir, es el mismo tanto en el primer como en el último análisis de datos. Esta técnica tiene 2 problemas:

• El primero está en el valor *p* necesario para demostrar diferencia estadística significativa al final del estudio, el que es mucho menor a 0,05. Por lo tanto, en ocasiones un valor clásicamente significativo puede no serlo al aplicar Pocock, por el sólo hecho de realizar análisis interinos. Por ejemplo, con sólo 2 análisis interinos, al aplicar la técnica de Pocock, el valor p necesario para demostrar diferencia estadística significativa es $p < ó = 0,029$ (ver tabla 1). Así, el hallazgo de un valor $p=0,03$ al final del estudio no permitiría establecer diferencia estadística significativa.
• El segundo es su baja exigencia para detener tempranamente un estudio. Por ejemplo, si se programan 2 análisis interinos (3 en total si agregamos el análisis final), basta con demostrar un $p < 0,022$ para detener el estudio con el primer análisis (Pocock, 2006).

Por estas razones, en la actualidad es una técnica poco utilizada.

La técnica de Peto-Haybittle establece un nivel de significancia alto para demostrar diferencia estadística en los análisis interinos, habitualmente un $p < 0,001$, el que es constante independiente del número de análisis interinos, y prácticamente no castiga el nivel de significancia estadística en el análisis final (a diferencia de Pocock). El defecto está en que el valor *p* no cambia si se programan 2 o 100 análisis interinos, lo que no se correlaciona con la realidad, ya que a mayor número de análisis, mayor es el riesgo de error tipo I, y mayor es el riesgo de sobreestimar los resultados (ver tabla 1).

La técnica de O'Brien-Fleming es simple y no posee los defectos antes mencionados. Para explicarla recurriremos al concepto de "gasto del valor *p*"[1]* (Muñoz N & Bangdiwala, 2000; Schulz & Grimes, 2005):

• Sabemos que sólo por azar a menor tamaño muestral, menor número de eventos observados y mayor riesgo de encontrar diferencias donde realmente no existen (es decir, mayor riesgo de error tipo I o error alfa) (Candia B *et al.*, 2006).
• Igualmente, a mayor número de análisis interinos mayor es la probabilidad que sólo por azar encontremos un resultado estadísticamente significativo irreal (también es un error tipo I)
• La técnica o regla de detención temprana de O'Brien-Fleming busca traducir éste riesgo de error a través del concepto de "gasto del valor *p*".
• Este "gasto" se refiere a la forma como se consume valor *p* a medida que se realizan análisis interinos, o, en otras palabras, la forma como se gasta la probabilidad tolerada de error tipo I para el estudio. En términos más simples: cuando un investigador planifica un estudio asume un valor inicial de error alfa tolerable, el que corresponde al estándar para definir diferencia estadística significativa: habitualmente 5%[2]*. Ésta probabilidad de 5% se "gasta" a medida que se realizan análisis interinos.
• A mayor número de análisis interinos programados, mayor es el gasto de valor *p*.
• Así, si se programa un análisis de resultados (sólo al completar el tamaño muestral programado al inicio del estudio) no se produce un "gasto" de la probabilidad de error alfa, por lo tanto puedo definir la detección de diferencias estadísticamente significativas con un valor $p = ó < a 0,05$. Se asume que al final del estudio el investigador posee la información "completa", por lo tanto puede utilizar $p < 0,05$ como límite de significancia estadística.
• Al realizar 2 o más análisis interinos el investigador "gasta" valor *p*. Es decir, dado que "gasté" error alfa al realizar más de

1) Estrictamente hablando el concepto de "gasto de p" o más bien "función de gasto" del valor p fue introducido por Lan y De Mets, para establecer reglas de detención en forma continua, sin la necesidad de programar los Análisis Interinos a priori, sin embargo, lo vamos a utilizar acá con una finalidad didáctica.*

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

un análisis de resultados, necesito obtener un valor *p* mucho menor a 0,05 en cada uno de esos análisis para demostrar diferencia.

• Lo anterior también se puede leer de la siguiente forma: dado que los análisis interinos son obtenidos con tamaños muestrales menores al programado al inicio del estudio, la información disponible con esos datos es menor, por lo que requiero de un límite de significancia mayor (es decir, un valor *p* mucho menor) para demostrar diferencia estadística.

• Mientras mayor es el número de análisis interinos programados, más temprano es el primer análisis en relación al reclutamiento de pacientes. O'Brien-Fleming ofrece un "gasto" progresivo del valor *p*: es mayor la exigencia mientras más temprano es el análisis interino. Así, con el primer análisis se debe obtener un valor *p* muy pequeño para demostrar diferencia estadística significativa, esta exigencia va disminuyendo a medida que aumenta el tamaño muestral (análisis interinos más tardíos).

• Con todo lo anterior se entiende que interrumpir tempranamente un estudio es difícil, improbable, ya que necesito encontrar un valor *p* muy pequeño para demostrar diferencia estadística, y es más pequeño mientras más temprano es el análisis interino.

• A diferencia de Peto, ésta técnica castiga el valor *p* final necesario para demostrar diferencia estadística significativa, pero el castigo es menor que el propuesto por Pocock *(ver tabla 1)*.

Con posterioridad se creó una técnica estadística para sustentar análisis interinos no programados y permitir de esa forma una mayor flexibilidad para la detención temprana de un ECR. La más utilizada es la propuesta por Lan y De Mets (DeMets & Lan, 1994; Betensky, 1998). Esta técnica utiliza una *función de gasto del error alfa*. Esta función es una fórmula matemática que define el valor *p* necesario para demostrar diferencia estadística a medida que se reclutan pacientes. De esta forma se pueden realizar análisis interinos no programados, asumiendo que el gasto del valor *p* se va a comportar de acuerdo a lo definido por esta función (Lan & DeMets, 1983; DeMets & Lan, 1994; Schulz & Grimes, 2005). Esta "función" puede ser aplicada al concepto teórico entregado por O'Brien-Fleming y es una de las más utilizadas en la actualidad.

Para comprender con mayor facilidad estas diferencias, en la Tabla 1 se exponen 3 estudios hipotéticos con 1, 2, 3 y 4 análisis interinos, el "gasto" del valor *p* asociado a cada situación y a cada técnica estadística antes descrita (Schulz & Grimes, 2005; Pocock, 2006).

Tabla 1.- Estudios hipotéticos, en donde se programan 1, 2, 3 y 4 análisis interinos más el análisis final. Se detallan las reglas de detención de Pocock, Peto y O'Brien Fleming en las distintas situaciones (Schulz & Grimes, 2005; Pocock, 2006).

En suma, todas las técnicas justifican la detención temprana de un ECR castigando el valor *p* necesario para la interrupción: mientras mayor es el número de análisis interinos programados mayor es el nivel de significancia estadística exigido para detener el estudio. Pocock establece reglas de detención donde el valor *p* es castigado en forma simétrica a lo largo de todo el análisis, incluido el análisis final. Peto establece un valor *p* altamente exigente uniforme para todos los análisis interinos, sin castigar en forma substancial el valor *p* del análisis final. O'Brien-Fleming castiga el valor *p* dependiendo no sólo de la cantidad de análisis interinos programados, también según si el análisis interino es más cercano al inicio del estudio (menor n, mayor castigo), además de castigar también el valor *p* en el análisis final.

Ahora, si bien estas "reglas" son ampliamente usadas, no existe consenso en su real utilidad, ya que a pesar de existir un menor riesgo de detención temprana dada la mayor exigencia estadística, el riesgo teórico de encontrar un resultado altamente significativo espurio dado sólo por azar, en contexto de un número bajo de eventos observados, nunca va a ser soslayado (Candia B *et al.*, 2006). Por lo tanto, estas técnicas sólo disminuyen la probabilidad de error sin aplacarlo totalmente.

Además no es claro que se deban utilizar como elemento único para decidir la interrupción de un ECR, ya que cada una de éstas técnicas fueron diseñadas para analizar sólo un aspecto de una terapia, no permitiendo evaluar en plenitud los efectos de la intervención, por ejemplo, la tolerancia a la droga, efectos adversos, adherencia a terapia, etc., todos elementos importantes al momento de analizar los resultados de un estudio (Sankoh, 1999; Sydes & Parmar, 2008). Por ésta razón se han creado iniciativas cuyo objetivo es evaluar el riesgo de error asociado a la detención precoz, el *n* mínimo necesario para disminuir este riesgo y determinar el momento más adecuado para realizar un primer análisis interino.

2) Esto significa que si yo repito el mismo estudio 100 veces, en 5 de esas repeticiones puedo encontrar diferencia donde realmente no existe, por lo tanto tolero un riesgo de error tipo I de 5%, lo que es igual a decir p<ó=0.05.*

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

Tabla 1

Número de Análisis Interinos	Análisis Interinos	Pocock	Peto- Haybittle	O'Brien-Fleming
1	1	0,029	0,001	0,005
	2 (reclutamiento completo)	0,029	0,0498	0,048
2	1	0,022	0,001	0,0005
	2	0,022	0,001	0,014
	3 (reclutamiento completo)	0,022	0,0495	0,045
3	1	0,018	0,001	0,0001
	2	0,018	0,001	0,004
	3	0,018	0,001	0,019
	4 (reclutamiento completo)	0,018	0,0492	0,043
4	1	0,016	0,001	0,00001
	2	0,016	0,001	0,0013
	3	0,016	0,001	0,008
	4	0,016	0,001	0,023
	5 (reclutamiento completo)	0,016	0,0489	0,041

Revisiones sistemáticas de estudios truncados por beneficio: ¿Error, Sesgo o solución al problema?

Como vimos en los párrafos anteriores, la detención precoz por beneficio tiene riesgo de error y sobrestimación de los resultados. La razón técnica ya ha sido expuesta y en términos de evidencia lo avala la iniciativa STOPIT-1 (Montori *et al.*, 2005).

Esta iniciativa correspondió a una revisión sistemática que buscó todos los estudios detenidos precozmente publicados por beneficio, independiente del tópico. Esta revisión muestra que los estudios truncados por beneficio son cada vez más frecuentes, habitualmente son publicados en revistas de alto impacto, a pesar que con frecuencia sobrestiman el beneficio, mostrando efectos de tratamiento de gran magnitud, muchas veces poco plausibles. Este fenómeno tiene una relación inversa con el número de eventos observados que definen el outcome: mientras menor es el número de eventos mayor es la sobrestimación del efectos, es decir, mayor es el error. Muchos de estos estudios fueron posteriormente rebatidos por estudios de mayor tamaño muestral, lo que se conoce como "regresión a la verdad" (Montori *et al.*, 2005; Candia B *et al.*, 2006).

Una forma de solucionar este fenómeno es a través de la realización de revisiones sistemáticas. Desde un punto de vista teórico, una revisión sistemática de tamaño razonable puede estimar un efecto con una precisión muy adecuada, a pesar de incluir estudios truncados por beneficio. Lo anterior ocurre porque los estudios detenidos precozmente tienen pocos eventos que definen el outcome, por lo tanto su peso al ser mezclados con otros estudios de mayor tamaño muestral, con más eventos y por tanto con resultados más representativos y de mayor precisión, permiten finalmente soslayar el error y corregir el hallazgo inicial (Hughes *et al.*, 1992). Esta afirmación fue evaluada mediante estudios de simulación computacional en donde se generaron meta-análisis sólo a partir de estudios ficticios truncados por beneficio, con sobrestimación de los resultados. Al mezclar estudios que han seguido reglas de detención adecuadas (O'Brien Fleming) los autores observaron, a través de este modelo computacional, que el riesgo de sobrestimar el efecto es mínimo. En la realidad habitual es poco probable que una revisión sistemática incluya sólo estudios detenidos precozmente, por lo tanto la mezcla se produce con estudios de mayor tamaño muestral, lo que minimizaría la sobrestimación del efecto (Goodman, 2007). Esta afirmación es discutida activamente por otros autores. Una revisión posterior evalúo el efecto de la inclusión de estudios truncados por beneficio en revisiones sistemáticas reales (Bassler *et al.*, 2007). Hasta el año 2007 los autores detectaron 96 revisiones sistemáticas que habían incluido al menos un estudio detenido precozmente por beneficio, de ellas el 46% incluyo 2 o más. En el 71% los autores de las revisiones no mencionaron la presencia de estudios con esta característica y, más aún, sólo el 2% la tuvo en consideración desde el punto de vista de la evaluación de la calidad metodológica. En el 17% de estas revisiones los

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

estudios detenidos precozmente por beneficio contribuyeron a más del 40% el "peso" global del meta-análisis. Por lo anterior se puede asumir que en las revisiones sistemáticas que incluyen estudios truncados por beneficio existe riesgo de sobrestimar los resultados, lo que se opone a los hallazgos generados por modelos computacionales.

Todos estos datos apoyan el concepto que los estudios truncados por beneficio pueden generar resultados erróneos, principalmente por sobrestimación del efecto de las terapias, sin embargo, el último punto abre una nueva arista en el problema: si bien la sobrestimación es azarosa, por lo tanto es un "error" desde el punto de vista metodológico, también es cierto que la dirección del error parece ser siempre la misma (sobreestimar más que minimizar los efectos de las terapias), además, dado que estos estudio se publican en revistas de alto impacto, pueden producir una especie de "congelamiento" en las publicaciones posteriores relacionadas al mismo tópico, especialmente aquellas con resultados cuyo efecto es menor o negativo, induciendo al sesgo de publicación. Este aspecto afecta directamente la metodología de revisiones sistemáticas que incluyen estudios detenidos por beneficio.

Con lo anterior podemos convenir que el error inicial se podría transformar en sesgo cuando los datos publicados inhiben la comunicación de nuevos estudios con resultados negativos (o menos "espectaculares"), induciendo un potencial sesgo de publicación y finalmente generando una sobrestimación genuina del efecto incluso en revisiones sistemáticas, dado el potencial alto peso de los datos truncados en los metanálisis correspondientes (Bassler *et al.*, 2007). Este aspecto persiste en activa discusión.

Iniciativa STOPIT-2 y Métodos Bayesianos

En la actualidad se está llevando a cabo un estudio metodológico internacional llamado STOPIT-2 (Bassler *et al.*, 2008; Briel *et al.*, 2009), el que pretende zanjar la disyuntiva antes expuesta. Para ello los autores se han planteado 3 objetivos globales:

• Cuantificar la magnitud del error generado por estudios truncados por beneficio mediante la comparación de sus resultados con los obtenidos de fuentes más confiables.
• Determinar los factores que predicen el error en estos estudios detenidos precozmente.

• Identificar si otras metodologías estadísticas solucionan el riesgo de error en la detención temprana.

Para resolver el primer punto los autores pretenden comparar los resultados de estudios truncados por beneficio con los obtenidos de revisiones sistemáticas del mismo tópico. Las revisiones sistemáticas, al tener un tamaño muestral mayor, nos proporcionan una estimación del efecto más confiable. Si mediante esta comparación se confirman y cuantifican estas diferencias los investigadores de STOPIT-2 pretenden identificar los factores que independientemente permiten predecir esta sobrestimación, con el objetivo de tomar medidas para solucionar el problema. Esto se realizará mediante análisis multivariados que evaluarán las reglas de detención utilizadas, la calidad metodológica de los estudios truncados y el número de eventos observados al momento de la detención (Silva & Benavides, 2001; Bassler *et al.*, 2008).

El último punto se refiere a la evaluación de métodos estadísticos Bayesianos como una alternativa para solucionar el problema. Los métodos estadísticos Bayesianos fueron propuestos por Thomas Bayes en el siglo XVII, basándose en una teoría estadística diametralmente distinta al método Frecuentista, que es el modelo que se utiliza casi de regla en la actualidad para evaluar hipótesis en medicina. Las razones de éste fenómeno se escapan de los objetivos de este artículo, pero están relacionados a la mayor simpleza de los métodos Frecuentistas. A continuación se describirán las diferencias entre ambos métodos:

• Los modelos Frecuentistas parten del supuesto que el parámetro poblacional a inferir es la constante y los estimadores puntuales medidos en las muestras (en los estudios) son aleatorios. Es por esta razón que cuando expresamos los resultados de un estudio se asume que tal valor es sólo una "estimación" del valor real, por lo tanto no es exacto y tiene riesgo de error (error alfa, error beta). Exceptuando el cálculo del tamaño muestral, este método no requiere de un estimador puntual inicial de la variable en estudio para la realización de los cálculos estadísticos posteriores, toda esa información se obtiene del estudio. Habitualmente las diferencias se estiman a través del cálculo de un "intervalo de confianza".
• El modelo Bayesiano asume que el parámetro poblacional a inferir con un estudio es aleatorio, mientras que los datos medidos en las muestras (en los estudios) son las constantes, es decir, todo lo contrario al modelo Frecuentista. Bajo este supuesto, lo único *real* son los datos obtenidos de los estudios y el parámetro poblacional se infiere a partir de la suma de un conocimiento a priori (otorgado por estudios previos o sólo por la experiencia) más el obtenido a través del estudio en desarrollo. Así, el valor poblacional del parámetro en estudio "varía" según el grado de información que manejemos: mientras en más ocasiones

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

se realiza un experimento, más información se obtiene. Finalmente todos estos datos son integrados, entregándonos una estimación del valor que buscamos, que es el que teóricamente se obtendría al aplicar la intervención a toda la población (Silva & Muñoz, 2000; Silva & Benavides, 2001; Grant *et al*., 2005). El modelo Bayesiano no posee el problema del cálculo del tamaño muestral: cada evento observado es información nueva que puede ser integrada y que su conjunto me permitirán deducir el parámetro poblacional que estoy buscando.

Considerando lo anterior, si los datos obtenidos en un ECR son analizados desde el punto de vista estadístico con un modelo Bayesiano, el análisis podría ser continuo: todo estudio podría ser interrumpido en el momento que los datos disponibles aplicados al teorema me otorguen un estimador del parámetro poblacional estable, que no variará a pesar de un mayor reclutamiento de pacientes y a un mayor número de eventos observados. Esta variación del efecto se cuantifica, en términos bayesianos, a través de un "Intervalo de Credibilidad", el que permite determinar la existencia de diferencias estadísticamente significativas.

Si bien todo lo anterior parece apoyar el uso de métodos Bayesianos para argumentar una interrupción temprana, esto no está probado por evidencia categórica, por ésta razón uno de los objetivos del STOPIT-2 es evaluar este punto (Silva & Benavides, 2001; Briel *et al.*, 2009).

Comités de Monitorización de Datos: Qué son y su utilidad en la detención temprana de un ECR.

Otra forma de disminuir el riesgo de error en la detención temprana es evitar tomar decisiones utilizando cómo único parámetro una regla detención estadística, dado los defectos antes mencionados de éstas técnicas. Bajo este concepto es que se han creado los Comités de Monitorización de Datos (CMD). Un CMD es un grupo multidisciplinario de profesionales cuya función es revisar periódicamente la eficacia y seguridad de los datos generados en un ECR y de esa forma evaluar la continuidad y/o eventual modificación del protocolo de un ECR.

Los cambios en el protocolo del estudio se refieren a la detención del reclutamiento de pacientes, lo que conlleva a la interrupción precoz del estudio; o al contrario, aumentar el tamaño muestral si la evidencia no es suficiente con los datos obtenidos al completar el *n* inicialmente estimado. También pueden sugerir cambiar el tiempo de seguimiento de los pacientes planificado por protocolo y de ésta forma informar y publicar los datos antes o después del tiempo inicialmente protocolizado (Grant *et al.*, 2005). La ventaja de contar con un CMD en un ECR se refiere a su capacidad de integrar toda la información disponible, tanto en términos de beneficio y efectos adversos de la terapia tanto con los datos obtenidos del ECR en desarrollo, como la evidencia proporcionada por otros estudios. Esta última muchas veces no está disponible al momento de planificar el protocolo inicial, pero puede aparecer a lo largo de su desarrollo y debe ser considerada al momento de tomar una decisión de interrupción precoz. Como se observa, la regla de detención es *sólo una* de las herramientas disponibles para decidir la modificación de un protocolo.

En la actualidad su utilidad práctica ha sido avalada por la FDA, la que recomienda su utilización, sin embargo, su estructura y funciones específicas son aún temas de discusión. El año 2005 el grupo de estudio DAMOCLES (Data Monitoring Committees: Lessons, Ethics and Statistics) propuso una serie de recomendaciones que se pueden resumir en los siguientes puntos (Sydes *et al.*, 2004; Grant *et al.*, 2005):

• Todo ECR debiera tener un CMD dentro de su protocolo, y su estructura y funciones deben ser establecidas a priori, idealmente antes de comenzar el reclutamiento de pacientes.
• El CMD es una entidad que cumple un rol asesor, ya que las decisiones que se toman son sugerencias para los investigadores a cargo del ECR, siendo finalmente éstos últimos los responsables del desarrollo del protocolo.
• El número de miembros del CMD es variable según las funciones asignadas. Este panel de expertos sugiere un número de 3 a 8 miembros, idealmente un número impar, en el caso que se requiera tomar una decisión vía votación.
• Para evitar conflictos de interés, los miembros deben ser externos al estudio, y así conservar su independencia al momento de tomar decisiones relacionadas al protocolo del ECR.
• En relación a los miembros del CMD se sugiere que debe existir al menos un presidente o líder, con experiencia en la toma de decisiones en términos de medicina basada en evidencia; profesionales clínicos relacionados y actualizados en el tema en estudio; un estadístico que conozca y/o esté familiarizado con temas clínicos; y un experto en temas éticos.
• Las decisiones tomadas por el CMD pueden ser equivocadas. Por ésta razón se debe insistir en una serie de puntos que minimizan el riesgo de error: la profesionalización de sus miembros, claridad en sus funciones y atribuciones, revisión minuciosa de la evidencia disponible y en constante aparición, resolución de diferencias de opinión por discusión activa de sus miembros.

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

Ahora, si bien los CMD permiten tener una apreciación más completa al momento de decidir una interrupción de un estudio, sus decisiones no están exentas de sesgo, como lo esboza una revisión sistemática de reciente publicación (Tharmanathan *et al.*, 2008), por lo tanto, si bien su presencia ayuda en la toma de decisiones, no garantiza que los resultados de un ECR detenidos precozmente sean confiables.

Conclusiones

El conocer como evaluar si un estudio tiene riesgo de error y/o sesgo es una herramienta necesaria en la actualidad, dada la gran cantidad de literatura que se publica día a día. La evaluación de la validez interna de un estudio es una herramienta crítica en la pesquisa de sesgo, sin embargo, el sesgo no es la única fuente de resultados alejados de la "verdad". El cálculo de tamaño muestral y la detención precoz de ensayos clínicos randomizados son elementos relevantes en el riesgo de error por azar en los resultados de un estudio, por lo tanto son puntos que también deben ser evaluados críticamente al momento de aplicar evidencia científica en nuestra práctica clínica. El conocer las reglas de detención precoz, la forma como se aplican y los CMD son las herramientas necesarias para evaluar este punto.

Nuestro país requiere de profesionales altamente capacitados y formados en la comprensión y revisión crítica de la evidencia clínica, ya que son ellos en conjunto con médicos subespecialistas expertos en cada patología los encargados de definir programas de salud, normas y protocolos frente a problemas clínicos específicos, basados en evidencia y costo-efectividad. Esto es crítico al momento de decidir aplicar a nivel masivo nuevas intervenciones avaladas por estudios idealmente con bajo riesgo de error por azar y de alta calidad metodológica y, en el futuro próximo, estos profesionales serán los encargados del desarrollo de evidencia clínica de alta calidad en nuestro país.

Referencias:

Bassler D, Ferreira-Gonzalez I, Briel M, Cook DJ, Devereaux PJ, Heels-Ansdell D, Kirpalani H, Meade MO, Montori VM, Rozenberg A, Schunemann HJ & Guyatt GH. (2007). Systematic reviewers neglect bias that results from trials stopped early for benefit. *Journal of clinical epidemiology* **60,** 869-873.

Bassler D, Montori VM, Briel M, Glasziou P & Guyatt G. (2008). Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of clinical epidemiology* **61,** 241-246.

Betensky RA. (1998). Construction of a continuous stopping boundary from an alpha spending function. *Biometrics* **54,** 1061-1071.

Briel M, Lane M, Montori VM, Bassler D, Glasziou P, Malaga G, Akl EA, Ferreira-Gonzalez I, Alonso-Coello P, Urrutia G, Kunz R, Culebro CR, da Silva SA, Flynn DN, Elamin MB, Strahm B, Murad MH, Djulbegovic B, Adhikari NK, Mills EJ, Gwadry-Sridhar F, Kirpalani H, Soares HP, Abu Elnour NO, You JJ, Karanicolas PJ, Bucher HC, Lampropulos JF, Nordmann AJ, Burns KE, Mulla SM, Raatz H, Sood A, Kaur J, Bankhead CR, Mullan RJ, Nerenberg KA, Vandvik PO, Coto-Yglesias F, Schunemann H, Tuche F, Chrispim PP, Cook DJ, Lutz K, Ribic CM, Vale N, Erwin PJ, Perera R, Zhou Q, Heels-Ansdell D, Ramsay T, Walter SD & Guyatt GH. (2009). Stopping randomized trials early for benefit: a protocol of the Study Of Trial Policy Of Interim Truncation-2 (STOPIT-2). *Trials* **10,** 49.

Candia B R, Letelier S LM & Rada G G. (2006). Estudios randomizados interrumpidos precozmente por beneficio: ¿Muy buenos o muy malos? *Revista médica de Chile* **134,** 1470-1475.

DeMets DL & Lan KK. (1994). Interim analysis: the alpha spending function approach. *Statistics in medicine* **13,** 1341-1352; discussion 1353-1346.

Goodman SN. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of internal medicine* **146,** 882-887.

Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, Elbourne DR, McLeer SK, Parmar MK, Pocock SJ, Spiegelhalter DJ, Sydes MR, Walker AE, Wallace SA & group Ds. (2005). Issues in data monitoring and interim analysis of trials. *Health technology assessment* **9,** 1-238, iii-iv.

Hughes MD, Freedman LS & Pocock SJ. (1992). The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* **48,** 41-53.

Lan G & DeMets DL. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659-663.

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, Bucher HC, Schunemann HJ, Meade MO, Cook DJ, Erwin PJ, Sood A, Sood R, Lo B, Thompson CA, Zhou Q, Mills E & Guyatt GH. (2005). Randomized trials stopped early for benefit: a systematic review. *Jama* **294,** 2203-2209.

Muñoz N SR & Bangdiwala SI. (2000). Análisis interino en ensayos clínicos: una guía metodológica. *Revista médica de Chile* **128,** 935-941.

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl

O'Brien PC & Fleming TR. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549-556.

Peppercorn J, Buss WG, Fost N & Godley PA. (2008). The dilemma of data-safety monitoring: provision of significant new data to research participants. *Lancet* **371,** 527-529.

Pocock SJ. (2006). Current controversies in data monitoring for clinical trials. *Clinical trials* **3,** 513-521.

Sankoh AJ. (1999). Interim Analyses: An Update of an FDA Reviewer's Experience and Perspective*. *Drug Information Journal* **33,** 165-176.

Schulz KF & Grimes DA. (2005). Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet* **365,** 1657-1661.

Silva LC & Benavides A. (2001). El enfoque bayesiano: otra manera de inferir. *Gaceta Sanitaria* **15,** 341-346.

Silva LC & Muñoz A. (2000). Debate sobre métodos frecuentistas vs bayesianos. *Gaceta Sanitaria* **14,** 482-494.

Sydes MR & Parmar MK. (2008). Interim monitoring of efficacy data is important and appropriate. *Journal of clinical epidemiology* **61,** 203-204.

Sydes MR, Spiegelhalter DJ, Altman DG, Babiker AB, Parmar MK & Group D. (2004). Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clinical trials* **1,** 60-79.

Tharmanathan P, Calvert M, Hampton J & Freemantle N. (2008). The use of interim data and Data Monitoring Committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. *BMC medical research methodology* **8,** 12.

ISSN: 0719-1855 © Dirección de Extensión y Educación Continua, Escuela de Medicina, Pontificia Universidad Católica de Chile. http://arsmedica.uc.cl