Alternativas al diseño de los ensayos clínicos

Pedro Muñoz Cacho
Técnico de Salud Pública. Unidad Docente de Medicina Familiar y Comunitaria de Cantabria

El primer «ensayo clínico» (EC) de la historia se ejecutó en el año 602 a. C. En realidad fue un estudio cuasi experimental: no hubo aleatorización, no fue ciego, con escaso número muestral (n = 4) en uno de los grupos, no se aportó información del número de pacientes del grupo de comparación, se llevó a cabo en niños o adolescentes y se planteó inicialmente como de no inferioridad aunque acabó mostrando superioridad. Sus principales sesgos potenciales fueron: corto período de observación (10 días), grupos no comparables en cuanto a la raza y medida del efecto principal muy subjetiva. Como fortaleza contaba con un período de seguimiento muy prolongado, en el que se confirman los hallazgos iniciales. Se publicó 76 años después de realizado, en lengua hebrea, en formato de resumen. Se trata del Libro de Daniel (versos 12 a 15), perteneciente a la Biblia.

La historia de los EC tiene muchos hitos interesantes; entre ellos, la aportación (en los años cuarenta) del epidemiólogo británico Austin Bradford Hill, que sentó las bases metodológicas de los EC tal como los entendemos en la actualidad. Desde entonces los partidarios y detractores han mantenido sus diferencias irreconciliables. Sin embargo, la supremacía de los EC se inicia en el año 1970, cuando la Food and Drug Administration (FDA) estableció que para la aprobación de nuevos medicamentos era requisito avalar los resultados con un EC. Desde ese momento la industria farmacéutica empezó a sobrepasar a los propios gobiernos y entidades académicas como primer patrocinador de EC. Esto ocurrió en los años noventa y continúa en la actualidad[1].

Algunos respetados metodólogos afirmaban que si al revisar un artículo no se encontraba la palabra «aleatorizar», era mejor pasar al siguiente. Otra frase célebre de uno de los más afamados defensores de los EC fue: «Hay que aleatorizar hasta que duela». Durante las últimas décadas la supremacía de los EC ha sido indiscutible, y en todas las clasificaciones para graduar la evidencia aparecen en la parte más alta como diseño individual. No obstante, en estos aproximadamente 80 años de vida de EC, no todo han sido luces. Por citar algunos de los más sonados desatinos, el de no identificar como eficaz el *bypass* aortocoronario con injerto. En ese EC, en la mayor parte de los pacientes con angina estable, no se detectó una mejoría en la supervivencia tras la intervención en comparación con los que recibieron tratamiento médico. El único resultado positivo fue la reducción en la incidencia y gravedad de la angina. Hasta Eugene Braunwald (el del famoso libro de cardiología) se rindió a la idea de que, efectivamente, la intervención solo valía para aliviar los síntomas, pero no prolongaba la supervivencia. Un editorial de *The New England Journal of Medicine* lo acredita[2]. Ahora sabemos la explicación: los pacientes estaban «demasiado sanos», los cirujanos eran «demasiado inexpertos», la mortalidad operatoria era demasiado alta y el análisis estadístico era sospechoso. Algunos eminentes cirujanos argumentaron que los EC no eran apropiados en cirugía. Entre ellos, René Favaloro, un cirujano argentino que trabajaba en el Cleveland Clinic y que estandarizó la técnica del *bypass* empleando la vena safena. Este cirujano argumentaba, refiriéndose a los EC: «... tomados de forma exclusiva pueden ser peligrosos»[3]. En esos años (1977), yo estudiaba Medicina y recuerdo muy claramente este mensaje: el *bypass* quita los síntomas, pero no mejora ni la mortalidad ni la incidencia de nuevos infartos. Es decir, que tampoco los EC nos han resuelto totalmente el problema, tampoco podemos confiar ciegamente en ellos. Algunas limitaciones de los EC son: validez externa limitada, generalizaciones a otros grupos de pacientes excluidos del estudio, insuficiente seguimiento o escaso número muestral para evaluar la duración del efecto o identificar efectos adversos raros pero graves. No obstante, la más importante de las limitaciones tiene que ver con los costes progresivamente más elevados. Los EC están a precio de oro: por ejemplo, uno en fase 3 puede costar 30 millones de dólares o más y uno con 14 000 pacientes en 300 centros puede alcanzar los 300 millones[4].

Otro aspecto que se debe considerar es que los resultados y las conclusiones de los EC comparados con los estudios observacionales, cuando se han analizado en profundidad, son esencialmente los mismos. Un excelente ejemplo es el artículo de Benson y Hartz[5]. Después de analizar 19 tratamientos sobre muy diferentes enfermedades, hallaron que solo en dos los efectos del tratamiento de los estudios observacionales se situaban fuera del intervalo de confianza de la medida combinada de los EC. La conclusión fue: «Encontramos poca evidencia de que la estimación del efecto en los estudios observacionales publicados después de 1984 sean consistentemente mayores o cualitativamente diferentes de los obtenidos en los ensayos aleatorizados y controlados». En la figura 1 se repro-

duce uno de los tratamientos analizados. Se presentan 24 estudios. Muy pocos estudios demuestran individualmente un beneficio significativo de la laparoscopia; sin embargo, el metanálisis sí lo demuestra. Además, la magnitud del efecto es la misma tanto en los estudios obsevacionales como en los EC.

Por lo comentado anteriormente, los sistemas para graduar la evidencia se han visto cuestionados[6]. Los instrumentos para graduar la evidencia son una parte importante de la medicina basada en la evidencia, y es necesario valorar la calidad de los artículos publicados y la elaboración de las guías de práctica clínica. Sin embargo, también hay imperfecciones que pueden conducir a tomar decisiones erróneas. Algunos de los problemas detectados en los actuales sistemas para graduar la evidencia se muestran en la tabla 1. Algunos merecen un comentario, como, por ejemplo, el número 4. Con el instrumento GRADE, para que un estudio que no sea EC pueda aumentar su evidencia ha de tener un tamaño del efecto significativamente grande, es decir, que paradójicamente el sistema premia estudios que pueden ser sesgados con respecto a estudios no-EC bien ejecutados pero con resultados que pueden estar en línea de lo observado con EC bien realizados. Con respecto al número 6, en España se ha podido comprobar que, incluso para expertos en el manejo del sistema GRADE, el método no era claro ni fácil de aplicar para 13 de 19 participantes (68,4 %)[7]. Finalmente, respecto al número 7, a los EC automáticamente se les asigna el mayor grado de evidencia, de forma que el resto de los artículos que utilizan otros diseños quedan por debajo en el nivel de evidencia. Aunque hay algunos instrumentos de graduación de la evidencia que permiten aumentar o disminuir la calidad de la evidencia en función de la calidad del estudio, no permiten que un EC deficiente sea calificado peor que un buen estudio no-EC.

Figura 1. *Odds ratio* para la infección tras laparoscopia comparada con laparotomía en pacientes apendicectomizados. Se presentan los resultados de 8 estudios observacionales y 16 ensayos clínicos. Tomada de Benson y Hartz[5]

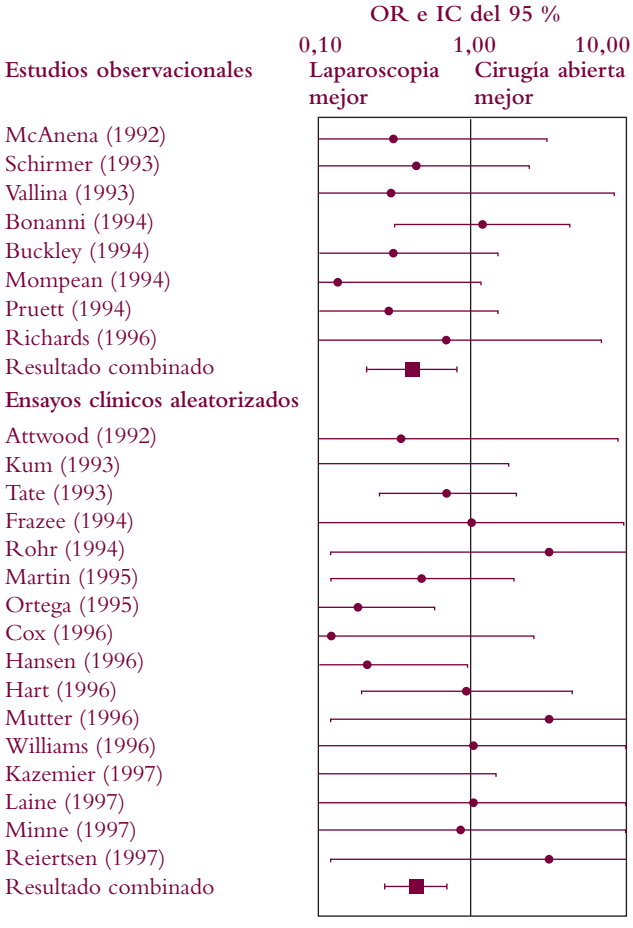

Tabla 1. Aspectos que son motivo de preocupación con respecto a los sistemas de graduación de la evidencia (traducida de Irving et al.[6])

1. Falta de información acerca de la validez y repetibilidad
2. Pobre validez concurrente
3. Puede no dar cuenta de la validez externa
4. Puede no ser inherentemente lógico
5. Proclive a la subjetividad
6. Sistemas complejos con instrucciones inadecuadas
7. Pueden estar sesgados hacia los ensayos clínicos aleatorizados y controlados
8. Pueden no abordar adecuadamente la variedad de estudios observacionales

Todos los diseños son capaces de tener sesgos, y hallazgos erróneos pueden provenir de cualquier tipo de diseño. Son conocidos los puntos fuertes y débiles de cada tipo de estudio; por tanto, lo recomendable es esforzarse en el diseño en aquellos aspectos en los que habitualmente son débiles[8].

Algunas soluciones que se han planteado para minimizar la falta de validez, sobre todo externa, de los EC clásicos es el empleo de EC pragmáticos, en los cuales se mantiene la aleatorización pero los criterios de selección son menos restrictivos, con lo cual se pretende que los pacientes sean similares a los que se atienden en las condiciones habituales de práctica clínica. Este debería ser el diseño de elección; sin embargo, no siempre se puede llevar a cabo, por motivos éticos, dificultades en su ejecución o retraso importante en el conocimiento de sus resultados. En ocasiones la diferencia entre EC pragmático y explicativo puede ser tan sutil que existe una herramienta para ayudar a los investigadores a determinar a qué tipo pertenece su estudio. Esta herramienta se llama PRagmatic Explanatory Continuum Indicator Summary (PRECIS), y actualmente está disponible la versión PRECIS-2[9,10].

Un enfoque relacionado es el diseño con aleatorización múltiple de cohortes. En este diseño se identifica una cohorte de pacientes, se pide su consentimiento para su posible inclusión en estudios posteriores y desde ese momento se registran las variables de resultado. De esta cohorte inicial se selecciona aleatoriamente

una cohorte a la que se administra el tratamiento que se pretende evaluar, comparando sus resultados con el resto de la cohorte no seleccionada que siguió recibiendo la asistencia estándar[4,11].

Otra variante son los EC basados en registros específicos de enfermedades, intervenciones o la propia historia clínica electrónica. Un ejemplo es el registro cardiovascular SWEDEHEART, iniciado en 2009 en Suecia. Para que la historia clínica electrónica pueda emplearse como fuente de información para los EC, se deben cumplir algunos requisitos; entre otros, estar sujetos a una auditoría formal para garantizar la validez de la información, unificar la codificación de los datos, etc.[4]. Este puede ser uno de los grandes retos de la atención primaria nacional.

Finalmente, otros diseños pueden darnos información de la aplicación de los tratamientos e intervenciones en condiciones de la práctica habitual. Entre los diseños observacionales destacan los estudios de cohortes, para los que se han propuesto unas listas-guía para facilitar su diseño, análisis e interpretación[12]. Por último, para la evaluación de la *mobile health* (mHealth), es decir, la tecnología basada en dispositivos móviles, tampoco se adapta de forma razonable el diseño de los EC clásicos. En el caso de la diabetes, la diversidad y cantidad de dispositivos móviles que emplean y los que se incorporarán en el futuro hacen que el tema resulte especialmente importante. Algunas propuestas son utilizar la metodología *continuous evaluation of evolving behavioral intervention technologies* (CEEBIT)[13,14]. También se ha consensuado una lista de comprobación de 16 ítems para mejorar la calidad de la evidencia relacionada con la mHealth[15].

Como resumen, se puede afirmar que ninguna investigación sobra. Toda la evidencia disponible se nos hace poca en muchas ocasiones. A veces deberíamos incorporar evidencia de la investigación básica, de la que frecuentemente no tenemos información de sus avances y nos ayudaría a explicar aparentes incongruencias. Sin olvidar si estamos ante lo que he dado en llamar (por mi cuenta y riesgo) el «caso análogo al paracaídas», que como sé que sois lectores curiosos lo vais a disfrutar. Se trata del artículo de Hayes et al.[16]. Doy por supuesto que el artículo que da origen a este lo conocéis, pero por si acaso: el artículo que plantea la cuestión en *BMJ* en el año 2003 es uno de Smith y Pell[17]. Aunque todo es importante, a veces descubrir que algo no es un «caso análogo al paracaídas» puede ser un gran avance si además podemos demostrarlo.

BIBLIOGRAFÍA

1. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard--lessons from the history of RCTs. N Engl J Med 2016;374:2175-81.
2. Braunwald E. Coronary-artery surgery at the crossroads. N Engl J Med 1977;297:661-3.
3. Favaloro RG. Critical analysis of coronary artery bypass graft surgery: a 30-year journey. J Am Coll Cardiol 1998;31(4 Suppl B):1B-63B.
4. De la Torre Hernández JM, Edelman ER. From nonclinical research to clinical trials and patient-registries: challenges and opportunities in biomedical research. Rev Esp Cardiol (Engl Ed) 2017;70:1121-33.
5. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878-86.
6. Irving M, Eramudugolla R, Cherbuin N, Anstey KJ. A critical review of grading systems: implications for public health policy. Eval Health Prof 2016;40:244-62.
7. Calderón C, Rotaeche R, Etxebarria A, Marzo M, Rico R, Barandiaran M. Gaining insight into the Clinical Practice Guideline development processes: qualitative study in a workshop to implement the GRADE proposal in Spain. BMC Health Serv Res 2006;6:138.
8. Frieden TR. Evidence for health decision making-beyond randomized, controlled trials. N Engl J Med 2017;377:465-75.
9. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009;62:464-75.
10. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ 2015;350:h2147.
11. Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the «cohort multiple randomised controlled trial» design. BMJ 2010;340:c1066.
12. Soto Álvarez J. Estudios observacionales para evaluar la efectividad clínica de los medicamentos. Uso de listas-guía para su diseño, análisis e interpretación. Aten Primaria 2005;35:156-62.
13. Modzelewski KL, Stockman MC, Steenkamp DW. Rethinking the endpoints of mHealth intervention research in diabetes care. J Diabetes Sci Technol 2018;12:389-92.
14. Mohr DC, Cheung K, Schueller SM, Hendricks Brown C, Duan N. Continuous evaluation of evolving behavioral intervention technologies. Am J Prev Med 2013;45:517-23.
15. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, et al.; WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016;352:i1174.
16. Hayes MJ, Kaestner V, Mailankody S, Prasad V. Most medical practices are not parachutes: a citation analysis of practices felt by biomedical authors to be analogous to parachutes. CMAJ Open 2008;6:1.e31-8.
17. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 2003;327:1459-61.