Limitaciones de los ensayos controlados aleatorios

Angus Deaton, Nancy Cartwright

9 de noviembre de 2016

The limitations of randomised controlled trials, VOX EU, CEPR

https://voxeu.org/article/limitations-randomised-controlled-trials

Traducción: Enrique A. Bour

En los últimos años, el uso de ensayos controlados aleatorios se ha extendido desde la evaluación del mercado laboral y los programas de bienestar a otras áreas de la economía y a otras ciencias sociales, quizás de manera más destacada en la economía del desarrollo y la salud. En esta columna se argumenta que parte de la popularidad de esos ensayos se basa en malentendidos sobre lo que son capaces de lograr, y se advierte contra las simples extrapolaciones de los ensayos a otros contextos.

Los ensayos controlados aleatorios (ECA) se han utilizado esporádicamente en la investigación económica desde los experimentos del impuesto sobre la renta negativo realizados entre 1968 y 1980 (véase Wise y Hausman 1985), y desde entonces se han utilizado regularmente para evaluar el mercado laboral y los programas de bienestar (Manski y Garfinkel 1992, Gueron y Rolston 2013). En los últimos años, se han difundido ampliamente en economía (y en otras ciencias sociales), tal vez de manera más destacada en la economía del desarrollo y la salud. La "revolución de la credibilidad" en econometría (Angrist y Pischke 2010) supuestamente libera la investigación empírica de los supuestos teóricos y estadísticos inverosímiles y arbitrarios, y los ensayos clínicos aleatorios se consideran el más "creíble" y "riguroso" de los métodos creíbles; de hecho, los diseños creíbles no relacionados con ensayos clínicos aleatorios suelen seguir un patrón lo más parecido posible a los ensayos clínicos aleatorios. Imbens (2010) escribe: "Los experimentos aleatorios ocupan un lugar especial en la jerarquía de la evidencia, a saber, en la parte superior".

En medicina, Pocock y Elbourne (2000) sostienen que sólo los ensayos clínicos aleatorios "pueden proporcionar una estimación fiable e imparcial de los efectos del tratamiento", y que sin esas estimaciones "ven peligros considerables para la investigación clínica e incluso para el bienestar de los pacientes". El vínculo entre sesgo y riesgo para los pacientes se considera obvio, sin que se intente demostrar que el diseño experimental de un ECA minimiza efectivamente el daño esperado para los pacientes. El Banco Mundial ha realizado muchos ECA relacionados con el desarrollo y hace afirmaciones que van mucho más allá de la imparcialidad. En su manual de aplicación se afirma que

"podemos estar muy seguros de que nuestro impacto medio estimado" (dado como la diferencia en las medias entre el grupo de tratamiento y el de control) "constituye el verdadero impacto del programa, ya que al construirlo hemos eliminado todos los factores observados e inobservados que de otro modo podrían explicar de forma plausible esas diferencias en los resultados" (Gertler y otros, 2011). Evidencia de alta calidad en efecto; la verdad es seguramente lo último en credibilidad.

¿Para qué sirven los ensayos controlados aleatorios?

En un documento reciente, sostenemos que parte de la popularidad de los ECA, tanto entre el público como entre algunos profesionales, se basa en malentendidos sobre lo que son capaces de lograr (Deaton y Cartwright 2016). Los ECA bien realizados podrían proporcionar estimaciones imparciales del efecto medio del tratamiento (EMT) en la población en estudio, siempre que no se introduzcan diferencias relevantes entre el tratamiento y el control después de la aleatorización, lo cual sirve para disminuir el cegamiento de sujetos, investigadores, recolectores de datos y analistas. La imparcialidad establece que, si repitiéramos el ensayo muchas veces, estaríamos en lo cierto en promedio. Sin embargo, casi nunca nos encontramos en esa situación, y con un solo ensayo (como es prácticamente siempre el caso) la imparcialidad no impide que nuestra única estimación esté muy lejos de la verdad. Si, como se cree a menudo, la aleatoriedad garantizara que los grupos de tratamiento y de control son idénticos, excepto por el tratamiento, entonces sí que tendríamos una estimación precisa - de hecho exacta - del EMT. Pero la aleatorización no hace nada de eso, ni siquiera en la línea de base; en cualquier ECA dado, nada asegura que otros factores causales estén equilibrados entre los grupos en el punto de la aleatorización. Los investigadores suelen comprobar el equilibrio en covariables observables, pero a menos que el dispositivo de aleatorización sea defectuoso, o que las personas rompan sistemáticamente su asignación, la hipótesis nula que subyace a la prueba es verdadera por construcción, de modo que la prueba no es informativa y no debe llevarse a cabo.

Angus Stewart Deaton (n. 1945, Edimburgo)
Nobel Prize, 2015
Angus Deaton and John Muellbauer, An Almost Ideal Demand System, The American economic review, 1980

Nancy Cartwright (n. 1944, Pensilvania)
Nancy Cartwright and Angus Deaton, *Understanding and misunderstanding randomized controlled trials, NBER, 2016*

Por supuesto, sabemos que el EMT de un ECA es sólo una estimación, no la verdad infalible, y como otras estimaciones, tiene un error estándar. Si se calcula adecuadamente, el error estándar del EMT estimado puede dar una indicación de la importancia de otros factores. Como entendió Fisher desde los primeros ensayos agrícolas, la

aleatorización, si bien no garantiza el equilibrio de los factores omitidos, nos da un método para evaluar su importancia. Sin embargo, incluso aquí hay trampas. Los estadísticos *t* para los EMT estimados de los ECAs no siguen en general la distribución *t*. Como ha sido recientemente documentado por Young (2016), una gran fracción de estudios publicados han hecho inferencias espurias debido a este *problema de Fisher-Behrens*, o por el fracaso de tratar apropiadamente los tests de hipótesis múltiples. Aunque la mayor parte de la literatura publicada es problemática, estas cuestiones pueden abordarse mediante mejoras en la técnica. Sin embargo, no es así en los casos en que los efectos de los tratamientos individuales están sesgados, como en los experimentos de atención de la salud, en los que uno o dos individuos pueden representar una gran parte del gasto (esto ocurrió en el Experimento de Salud Rand); o en microfinanciación, en los que unos pocos sujetos ganan dinero y la mayoría no (en los que la distribución *t* otra vez se quiebra). Una vez más, es probable que las inferencias sean erróneas, pero aquí no hay ninguna solución clara. Cuando hay efectos de tratamiento individuales atípicos, la estimación depende de si los valores atípicos son asignados a los tratamientos o a los controles, lo que causa reducciones masivas del tamaño efectivo de la muestra. La reducción de los valores atípicos solucionaría el problema estadístico, pero sólo a costa de destruir el problema económico; por ejemplo, en la asistencia sanitaria, son precisamente los pocos valores atípicos los que hacen o deshacen un programa. En vista de estas dificultades, sospechamos que una gran fracción de los resultados publicados de los ECA en economía del desarrollo y la salud no son fiables.

La "credibilidad" de los ensayos clínicos aleatorios proviene de su capacidad de obtener respuestas sin el uso de información previa potencialmente polémica sobre la estructura, como la especificación de otros factores causales o el detalle de los mecanismos a través de los cuales operan. Un público lego escéptico suele no estar dispuesto a aceptar conocimientos económicos previos e incluso dentro de la profesión hay diferencias sobre los supuestos o controles adecuados. Sin embargo, como siempre ocurre, la única vía para la precisión es a través de la información previa y el control de los factores que probablemente sean importantes, de la misma manera que en un experimento de laboratorio (no aleatorio) en física, biología o incluso economía, los científicos buscan una medición precisa mediante el control de los factores de confusión conocidos. La ciencia acumulativa ocurre cuando los nuevos resultados se construyen sobre los antiguos - o los socavan - y los ECA, con su negativa a utilizar la ciencia previa, lo hacen muy difícil. Y cualquier ECA puede ser cuestionado ex post examinando la diferencia entre tratamientos y controles tal como se asignaron realmente, y mostrando que factores importantes discutibles estaban distribuidos de forma desigual; la información previa se excluye por azar, pero reaparece en la interpretación de los resultados.

Un ECA bien realizado puede dar una estimación creíble de un EMT en una población específica, a saber, la "población en estudio" de la que se seleccionaron los tratamientos y controles. A veces esto es suficiente; si estamos haciendo una evaluación de un programa post hoc, si estamos comprobando una hipótesis que se supone que es generalmente cierta, si queremos demostrar que el tratamiento puede funcionar en algún lugar, o si la población de estudio es una muestra extraída al azar de la población de

interés cuyo EMT estamos tratando de medir. Sin embargo, la población en estudio no suele ser la población que nos interesa, especialmente si los sujetos deben ofrecerse voluntariamente para participar en el experimento y tienen sus propias razones para participar o no. Un famoso ejemplo temprano proviene de Ashenfelter (1981), que descubrió que las personas que se ofrecen como voluntarias para un programa de capacitación tienden a haber visto una disminución reciente de sus salarios; de manera similar, las personas que toman una droga pueden ser las que han fracasado en otras formas de terapia. De hecho, muchas de las diferencias de resultados entre los estudios experimentales y no experimentales pueden atribuirse no a diferencias de metodología, sino a diferencias en las poblaciones a las que se aplican.

El problema del "transporte"

En términos más generales, demostrar que un tratamiento funciona en una situación es una prueba extremadamente débil de que funcionará de la misma manera en otros lugares; este es el problema del "transporte": ¿qué se necesita para permitirnos utilizar los resultados en nuevos contextos, ya sean contextos de políticas o en el desarrollo de la teoría? Sólo puede abordarse utilizando conocimientos y comprensión previos, es decir, interpretando el ECA dentro de alguna estructura, la estructura por la que, de manera algo paradójica, el ECA obtiene su credibilidad al negarse a utilizar. Si queremos pasar de un ECA a una política, tenemos que construir un puente entre el ECA y la política. No importa cuán riguroso o cuidadoso sea el ECA, si el puente se construye mediante un símil de que el contexto de la política es de alguna manera similar al contexto experimental, el rigor en el ensayo no hace nada para apoyar una política; en cualquier cadena de pruebas, es el eslabón más débil el que determina la fuerza general del alegato, no el más fuerte. La utilización de los resultados de un ensayo controlado aleatorio no puede ser simplemente una cuestión de simple extrapolación del experimento a otro contexto. Los efectos causales dependen de los contextos en los que se derivan, y a menudo dependen de factores que pueden ser constantes dentro del contexto experimental pero diferentes en otros lugares. Incluso la dirección de causalidad puede depender del contexto. Tenemos más posibilidades de transportar los resultados si reconocemos la cuestión al diseñar el experimento - que en sí mismo requiere el compromiso de algún tipo de estructura - y tratamos de investigar los efectos de los factores que probablemente varíen en otro lugar. Sin una estructura, sin comprender por qué funcionan los efectos, no sólo no podemos transportar, sino que no podemos empezar a hacer economía del bienestar; el hecho de que una intervención funcione, y de que el investigador piense que la intervención hace que la gente se sienta mejor, no es garantía de que lo haga realmente. Sin saber por qué suceden las cosas y por qué la gente hace las cosas, corremos el riesgo de una inútil teorización causal casual ("cuento de hadas"), y hemos renunciado a una de las tareas centrales de la economía.

Referencias

Angrist, J and J-S Pischke (2010), "The credibility revolution in empirical economics: how better design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24(2), 3-30.

Ashenfelter, O (1978), "Estimating the effect of training programs on earnings," *Review of Economics and Statistics*, 60(1), 47–57.

Deaton, A and N Cartwright (2016), "Understanding and misunderstanding randomized controlled trials".

Garfinkel, I and C F Manski (1992), *Evaluating welfare and training programs*, Cambridge, MA: Harvard.

Gertler, P J, S Martínez, P Premand, L B Rawlings, and C M J Vermeersch (2011), *Impact evaluation in practice*, Washington, DC: The World Bank.

Gueron, J M and H Rolston (2013), *Fighting for reliable evidence*, New York: Russell Sage.

Imbens, G (2010), "Better late than nothing," *Journal of Economic Literature*, 48(2), 399-423.

Pocock, S J and D R Elbourne (2000), "Randomized trials or observational tribulations?" *New England Journal of Medicine*, 342, 1907-9.

Wise, D A and J A Hausman (1985), *Social Experimentation*, Chicago, IL: Chicago University Press for NBER.

Young, A (2016), "Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results," London School of Economics, Working Paper, Feb.