unicef
Office of Research-Innocenti

Síntesis metodológicas

Sinopsis de la evaluación de impacto n.º 8

Diseño y métodos cuasiexperimentales

Howard White y Shagun Sabarwal

RMIT UNIVERSITY
3ie International Initiative for Impact Evaluation
BetterEvaluation
unicef Office of Research-Innocenti

CENTRO DE INVESTIGACIONES INNOCENTI DE UNICEF

El Centro de Investigaciones Innocenti es la oficina de UNICEF especializada en investigación científica. El objetivo principal del Centro de Investigaciones es mejorar la comprensión internacional de una serie de cuestiones relacionadas con los derechos de la infancia, a fin de facilitar la plena aplicación de la Convención sobre los Derechos del Niño en todo el mundo. El centro tiene el cometido de establecer un marco integral de investigación y conocimiento dentro de la organización para brindar apoyo a los programas y políticas mundiales de UNICEF, y trabaja con los asociados para formular políticas con base empírica en favor de la infancia. Las publicaciones elaboradas por el centro contribuyen al debate global sobre la infancia y los derechos del niño e incluyen una amplia gama de opiniones.

Las opiniones expresadas corresponden a los autores o editores y se publican para estimular un mayor diálogo sobre métodos de análisis de impacto. Esta publicación no refleja necesariamente las políticas o perspectivas de UNICEF.

SINTESIS METODOLOGICAS DEL CENTRO DE INVESTIGACIONES

Las síntesis metodológicas del Centro de Investigaciones de UNICEF pretenden compartir prácticas de investigación, métodos, diseños y recomendaciones de reconocidos investigadores y analistas. La audiencia a la que van dirigidas es principalmente el personal de UNICEF que lleve a cabo, encargue o interprete los resultados de investigación y análisis para la toma de decisiones sobre programas, políticas y actividades de sensibilización.

Esta síntesis metodológica ha seguido un proceso de revisión por pares interna.

El texto no ha sido editado de acuerdo con los estándares de publicación oficiales y UNICEF declina toda responsabilidad por posibles errores.

Para consultas o descargas, pueden encontrar estas síntesis metodológicas en http://www.unicef-irc.org/KM/IE/

Recomendamos la siguiente cita para cualquier referencia al presente documento:

White, H., & S. Sabarwal (2014). Diseño y métodos cuasiexperimentales, *Síntesis metodológicas: evaluación de impacto n.º 8*, Centro de Investigaciones de UNICEF, Florencia.

Agradecimientos: Varios autores han proporcionado orientación en la preparación de esta síntesis. El autor y el Centro de Investigaciones de UNICEF desean agradecer a todos aquellos que han participado en la preparación de la presente publicación, especialmente a:

Por su contribución: Greet Peersman
Por su revisión: Nikola Balvin, Sarah Hague, Debra Jackson

Centro de Investigaciones Innocenti de UNICEF
Piazza SS. Annunziata, 12
50122 Florencia (Italia)
Tel.: (+39) 055 20 330
Fax: (+39) 055 2033 220
florence@unicef.org
www.unicef-irc.org

1. DISEÑO Y MÉTODOS CUASIEXPERIMENTALES: BREVE DESCRIPCIÓN

Al igual que los diseños experimentales, los diseños de investigación cuasiexperimentales contrastan hipótesis causales. Tanto en los diseños experimentales (ensayos controlados aleatorios) como en los cuasiexperimentales, el programa o política se considera como una «intervención» en la que se comprueba en qué medida un tratamiento —incluidos los elementos del programa o la política evaluados— logra sus objetivos, de acuerdo a las mediciones de un conjunto preestablecido de indicadores (véase la Síntesis n.º 7 (Ensayos controlados aleatorios)). No obstante, un diseño cuasiexperimental carece, por definición, de distribución aleatoria. La asignación a las condiciones (tratamiento versus ningún tratamiento o comparación) se lleva a cabo por autoselección (los participantes eligen el tratamiento), por la selección efectuada por los administradores (por ejemplo, funcionarios, profesores, autoridades, etc.) o por ambas vías[1].

Los diseños cuasiexperimentales identifican un grupo de comparación lo más parecido posible al grupo de tratamiento en cuanto a las características del estudio de base (previas a la intervención). El grupo de comparación capta los resultados que se habrían obtenido si el programa o la política no se hubieran aplicado (es decir, el contrafáctico). Por consiguiente, se puede establecer si el programa o la política han causado alguna diferencia entre los resultados del grupo de tratamiento y los del grupo de comparación.

Existen diferentes técnicas para crear un grupo de comparación válido, por ejemplo, el diseño de regresión discontinua y el emparejamiento por puntuación de la propensión, tratados más adelante, lo que reduce el riesgo de sesgo. El sesgo que puede resultar preocupante en este caso es el sesgo de «selección» —la posibilidad de que quienes son idóneos o que deciden participar en la intervención sean sistemáticamente diferentes de los que no pueden o no quieren participar—. Por tanto, las diferencias observadas entre los indicadores de interés de los dos grupos pueden deberse —en su totalidad o en parte— a un emparejamiento imperfecto en lugar de a la intervención.

Hay también métodos basados en la regresión no experimentales, como la estimación de variables instrumentales y los modelos de selección de la muestra (también conocidos como modelos de Heckman). Estas técnicas de regresión tienen en cuenta el sesgo de selección, mientras que los modelos de regresión simple, como el de mínimos cuadrados ordinarios, por lo general no lo hacen. También puede haber experimentos naturales basados en la aplicación de un programa o política que pueda considerarse equivalente a la distribución aleatoria o al análisis de series temporales interrumpidas, que analiza los cambios en las tendencias de los resultados antes y después de una intervención. Estos enfoques se usan muy poco y no se tratan en esta síntesis.

Los métodos de análisis de datos utilizados en los diseños cuasiexperimentales pueden ser ex post por diferencia única o de diferencia doble (también conocido como de diferencia en diferencias o DID).

[1] Shadish, William R., *et al.*, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, 2002, pág. 14.

Puntos principales

1. Al igual que los diseños experimentales, los diseños de investigación cuasiexperimentales contrastan hipótesis causales.
2. Un diseño cuasiexperimental carece, por definición, de distribución aleatoria.
3. Los diseños cuasiexperimentales identifican un grupo de comparación lo más parecido posible al grupo de tratamiento en cuanto a las características de referencia (previas a la intervención).
4. Existen diferentes técnicas para crear un grupo de comparación válido, como el diseño de regresión discontinua y el emparejamiento por puntuación de la propensión.

2. ¿CUÁNDO PROCEDE EMPLEAR MÉTODOS CUASIEXPERIMENTALES?

Los métodos cuasiexperimentales que implican la creación de un grupo de comparación se utilizan más a menudo cuando no es posible asignar de manera aleatoria los individuos o grupos a los grupos de tratamiento y los grupos de control. Este es siempre el caso para los diseños de evaluación de impacto ex post. También puede ser necesario utilizar diseños cuasiexperimentales para las evaluaciones de impacto ex ante, por ejemplo, en los casos en que los obstáculos éticos, políticos o logísticos, como la necesidad de una implantación geográfica gradual, descarten la aleatorización.

Los métodos cuasiexperimentales pueden utilizarse retrospectivamente, es decir, después de la intervención (en el periodo de tiempo t +1, en la tabla 1). En algunos casos, especialmente para las intervenciones que abarcan una duración más larga, pueden hacerse las estimaciones preliminares del impacto a mitad del periodo (tiempo t, en la tabla 1). No obstante, es muy recomendable que la planificación de la evaluación comience en todo caso antes de la intervención. Esto es especialmente importante, ya que los datos de referencia se deben recoger antes de exponer a los destinatarios a las actividades del programa o política (tiempo t -1, en la tabla 1).

Tabla 1. Calendario de la intervención y recopilación de datos para las evaluaciones de impacto con una muestra de gran tamaño

Anterior a la intervención	Intervención	Posterior a la intervención
t -1 Datos de referencia	t (Encuesta de término medio)	t +1 Datos finales

t = un periodo de tiempo específico

3. MÉTODOS CUASIEXPERIMENTALES PARA LA CREACIÓN DE GRUPOS DE COMPARACIÓN

Emparejamiento por puntuación de la propensión

¿Qué es el emparejamiento?

Los métodos de emparejamiento se basan en las características observadas a fin de crear un grupo de comparación mediante el empleo de técnicas estadísticas. Existen diferentes tipos de técnicas de emparejamiento, incluidos el emparejamiento crítico, las comparaciones emparejadas y la distribución secuencial, algunos de los cuales se tratan en la Síntesis n.º 6 (Sinopsis: Estrategias de atribución causal). Esta sección se centra en las técnicas de emparejamiento por puntuación de la propensión.

Un emparejamiento perfecto requeriría que cada individuo del grupo de tratamiento coincida con un individuo del grupo de comparación idéntico en todas las características observables pertinentes como la edad, la educación, la religión, la ocupación, la riqueza, la actitud frente al riesgo, etc. Obviamente, sería imposible. Encontrar una buena coincidencia para cada participante del programa por lo general implica estimar lo más fielmente posible las variables o factores determinantes que explican la decisión del individuo de participar en el programa. Si la lista de estas características observables es muy grande, entonces resulta difícil emparejarlas directamente. En esos casos, es más conveniente utilizar el emparejamiento por puntuación de la propensión.

¿Qué es el emparejamiento por puntuación de la propensión?

En el emparejamiento por puntuación de la propensión, no se empareja al individuo en función de cada una de las características observables, sino de su puntuación de la propensión, es decir, la probabilidad de que la persona participe en la intervención (probabilidad de participación prevista) dadas sus características observables. Por tanto, el emparejamiento por puntuación de la propensión coteja los individuos u hogares en tratamiento con otros semejantes, y posteriormente calcula la diferencia media en los indicadores de interés. En otras palabras, el emparejamiento por puntuación de la propensión asegura que las características medias de los grupos de tratamiento y de comparación sean similares, y esto se considera suficiente para obtener una estimación imparcial del impacto.

Cómo aplicar el emparejamiento por puntuación de la propensión

El emparejamiento por puntuación de la propensión consta de las cinco etapas siguientes:

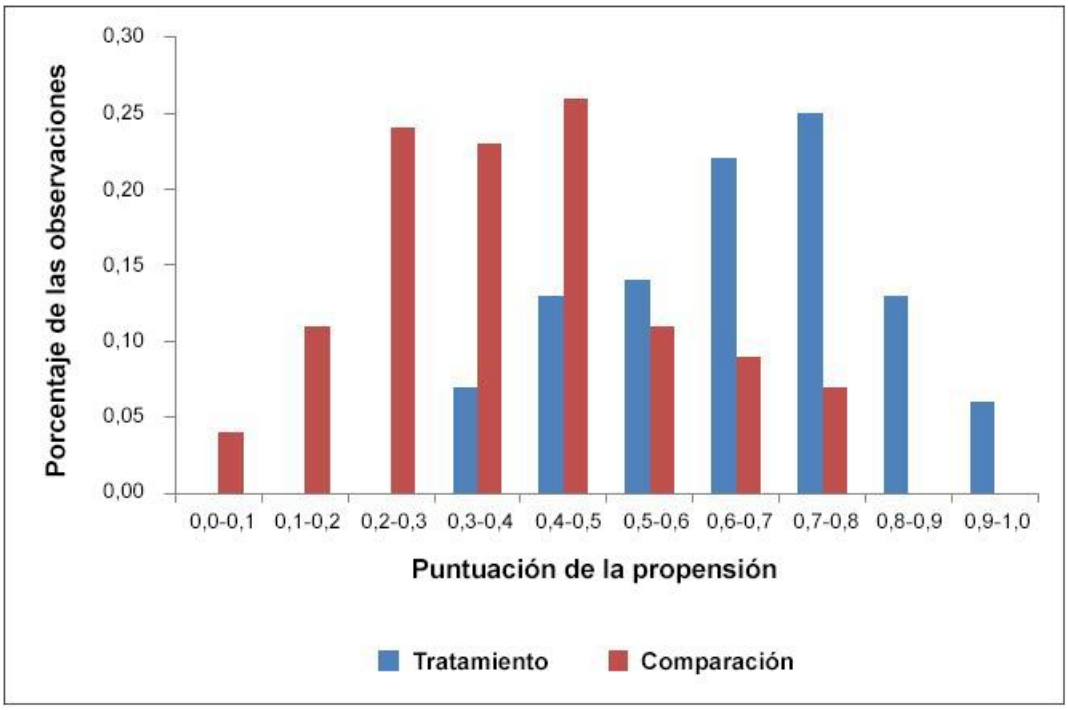
1. **Garantizar la representatividad**: asegurarse de que existe una encuesta con una muestra representativa de participantes y no participantes idóneos para la intervención. Para calcular las puntuaciones de la propensión, es preferible emplear los datos de referencia. Esta técnica, sin embargo, también puede utilizarse con datos finales: las variables que se emparejan deben ser variables no afectadas por la intervención.
2. **Calcular las puntuaciones de la propensión**: las puntuaciones de la propensión se construyen usando la «ecuación de participación», que es una regresión logit o probit en la que la participación en el programa es la variable dependiente (participa en el programa = 1, no participa en el programa = 0). Las características que se juzga que afectan la participación deben considerarse cuidadosamente y ser lo más exhaustivas posible, pero deben excluirse las que puedan haber sido afectadas por la intervención. Por esta razón, para calcular las puntuaciones de la propensión conviene usar datos de referencia, si se dispone de ellos.

3. **Seleccionar un algoritmo de emparejamiento**: cada miembro del grupo de tratamiento se empareja entonces con uno o más miembros del grupo de comparación. Hay diferentes maneras de hacerlo, como emparejar a cada participante con su «vecino más cercano» no participante. La media de los cinco vecinos más cercanos es la más utilizada habitualmente.

   Un individuo del grupo de comparación puede emparejarse con varias personas diferentes del grupo de tratamiento.

   Para que el emparejamiento sea válido, es indispensable comparar los «valores observados» de los participantes y los no participantes con la misma serie de características. Las observaciones efectuadas en el grupo de comparación cuya puntuación de la propensión sea inferior al menor valor observado en el grupo de tratamiento se descartan. Del mismo modo, también se descartan las observaciones efectuadas en el grupo de tratamiento cuya puntuación de la propensión sea superior al mayor valor observado en el grupo de comparación. Lo que queda se conoce como la «región de soporte común»; véase un ejemplo en el gráfico 1.

Gráfico 1. Ejemplo de una distribución de puntuaciones de la propensión en que la región de soporte común va de 0,31 a 0,80



Fuente: Datos elaborados por los autores solo a efectos ilustrativos.

El gráfico 1 muestra una distribución típica de puntuaciones de la propensión. La distribución del grupo de tratamiento está a la derecha de la del grupo de comparación, es decir, los individuos del grupo de tratamiento tienden a tener puntuaciones de la propensión más altas que los del grupo de comparación. Ningún miembro del grupo de tratamiento tiene una puntuación de la propensión menor que 0,3, y ningún miembro del grupo de comparación tiene una puntuación de la propensión mayor que 0,8. Así, al establecer la región de soporte común, se omite el 39% de las observaciones del grupo de comparación, con una puntuación de la propensión de 0 a 0,3, así como el 19% de las observaciones del grupo de tratamiento, con una puntuación de la propensión de 0,8 a 1. (En la práctica, se emplearía un valor de corte más preciso que el que muestra la clasificación cualitativa de los datos).

La tabla 2 muestra el emparejamiento de las variables seleccionadas de un análisis efectuado mediante emparejamiento por puntuación de la propensión para un estudio de impacto del acceso a agua potable en Nepal[2]. La columna «antes de emparejar» compara las características medias de los hogares con acceso a agua potable en el grupo de tratamiento con las de todos los hogares sin acceso a agua potable del grupo de comparación. Estos dos grupos de hogares son muy diferentes: es más probable que los que tienen acceso a agua potable se encuentren en entornos urbanos y tengan un nivel de educación y una situación económica mejores que aquellos sin acceso a agua potable. No obstante, cualquier diferencia relativa a la diarrea infantil en ambos grupos no se puede atribuir sencillamente al acceso al agua potable, puesto que existen muchas otras diferencias que pueden explicar por qué la incidencia de la diarrea infantil varía según los grupos.

Tabla 2. Características observables antes y después del emparejamiento (porcentaje del grupo que muestra la característica)

Variable	Antes de emparejar		Después de emparejar	
	Tratamiento (%)	Comparación (%)	Tratamiento (%)	Comparación (%)
Residente rural	29	78	33	38
Quintil más rico	46	2	39	36
Educación superior del cabeza de familia	21	4	17	17

Fuente: Bose, Ron, «The impact of Water Supply and Sanitation interventions on child health: evidence from DHS surveys», ponencia, Conferencia Bianual sobre la Evaluación de Impacto, Colombo, Sri Lanka, 22 al 23 de abril de 2009.

Después del emparejamiento se reducen sustancialmente las diferencias entre los dos grupos. El establecimiento de la región de soporte común descarta aquellos hogares sin acceso a agua potable que son muy diferentes a los que disponen de acceso a agua potable, de modo que los hogares emparejados del grupo de comparación son más urbanos y tienen un nivel de educación y una situación económica mejores que los hogares sin acceso a agua potable en su conjunto. Del mismo modo, también se han descartado de la evaluación los miembros menos similares del grupo de tratamiento.

4. **Verificar el equilibrio**: las características de los grupos de tratamiento y de comparación se cotejan para comprobar su equilibrio. Idealmente, no habrá diferencias significativas en las características observables medias entre los dos grupos. Ahora que los grupos de tratamiento y de comparación son similares en cuanto a sus características observables, la varianza en la incidencia de la diarrea infantil entre los grupos de tratamiento y de comparación puede atribuirse a diferencias tales como el acceso a agua potable.
5. **Estimar los efectos del programa e interpretar los resultados**: finalmente, la estimación del impacto, ya sea mediante diferencia única o doble, se efectúa, en primer lugar, calculando la diferencia entre el indicador del individuo de tratamiento y el valor medio de los individuos de comparación emparejados y, en segundo lugar, promediando todas estas diferencias.

[2] Bose, Ron, «The impact of Water Supply and Sanitation interventions on child health: evidence from DHS surveys», ponencia, Conferencia Bianual sobre la Evaluación de Impacto, Colombo, Sri Lanka, 22 al 23 de abril de 2009.

La tabla 3 muestra un ejemplo (de emparejamiento con el vecino más cercano) que utiliza datos de los resultados de aprendizaje de niños de 6.º grado (o año) en una prueba estandarizada. La columna 1 muestra la puntuación en la prueba obtenida por los individuos del grupo de tratamiento, y las columnas 4 a 8 muestran la puntuación en la prueba de los 5 vecinos más cercanos a cada uno en el grupo de comparación. La puntuación media de los 5 vecinos se muestra en la columna 2, y la diferencia entre la puntuación en la prueba del individuo de tratamiento y este promedio se muestra en la columna 3. La estimación del impacto por diferencia única es el promedio de los valores de la columna 4.

Tabla 3. Cálculo de la estimación del impacto mediante puntuaciones de la propensión. Ejemplo en el que se emplean los datos de la puntuación en la prueba

Observados i)	Y_{1i}	$Y_{0i\ (media)}$	Y_{1i}-Y_{0i}	$Y_{0i\ 1)}$	$Y_{0i\ 2)}$	$Y_{0i\ 3)}$	$Y_{0i\ 4)}$	$Y_{0i\ 5)}$
	1)	2)	3)	4)	5)	6)	7)	8)
1	48,2	42,4	5,8	44,1	45,1	43,8	43,2	35,8
2	50,2	42,6	7,6	42,1	45,2	48,1	38,4	39,3
3	50,6	43,1	7,5	40,8	43,7	45,3	44,1	41,8
4	48,1	38,9	9,1	43,6	35,6	36,9	41,4	37,2
5	69,0	59,7	9,3	55,6	57,6	57,1	62,4	65,8
…	…	…	…	…	…	…	…	…
199	58,6	52,2	6,4	55,5	48,2	54,7	53,4	49,1
200	45,4	39,3	6,1	41,2	39,1	38,7	40,1	37,5
Promedio	52,9	45,5	7,4					

En la práctica, no es necesario hacer estos cálculos manualmente, ya que existen paquetes estadísticos (por ejemplo, Stata, SAS o R) para llevar a cabo el análisis.

¿Qué se necesita para llevar a cabo el emparejamiento por puntuación de la propensión?

El emparejamiento por puntuación de la propensión requiere datos tanto del grupo de tratamiento como de un grupo de comparación potencial. Ambas muestras deben ser más grandes que el tamaño de la muestra sugerido mediante el cálculo de la potencia estadística (es decir, el cálculo que indica el tamaño de la muestra necesario para detectar el impacto de una intervención) ya que se descartan las observaciones no incluidas en la región de soporte común. En general, el sobremuestreo debe ser mayor para el grupo de comparación potencial que para el grupo de tratamiento.

El emparejamiento por puntuación de la propensión puede efectuarse utilizando datos de encuestas, registros administrativos, etc. Los datos de los grupos de tratamiento y de comparación pueden provenir de diferentes conjuntos de datos, siempre que: 1) contengan datos sobre las mismas variables (es decir, definidos de la misma manera); y 2) los datos se recopilen durante el mismo periodo de tiempo. El último

requisito es particularmente importante para las variables estacionales, es decir, variables sensibles a las diferentes estaciones, como el peso para la edad.

Ventajas y desventajas del emparejamiento por puntuación de la propensión

Las dos principales ventajas del emparejamiento por puntuación de la propensión son que siempre es factible si se dispone de datos y que se puede llevar a cabo después de que una intervención haya finalizado, incluso en ausencia de datos de referencia (aunque no es lo ideal). Si no se dispone de datos de referencia, puede utilizarse la «memoria» para reconstruir las características previas a la intervención. Sin embargo, esto puede ser impreciso, y a la hora de decidir qué variables pueden recordarse con precisión debe emplearse el sentido común.

El principal inconveniente es que el emparejamiento por puntuación de la propensión se basa en emparejar individuos en función de características observables vinculadas a la probabilidad de participación prevista. Por tanto, si hay características «no observadas» que afectan la participación y cambian con el tiempo, las estimaciones serán sesgadas e influirán en los resultados observados. Otra limitación práctica del uso del emparejamiento por puntuación de la propensión es que se necesita la asistencia de un estadístico o de alguien capacitado para el uso de diferentes paquetes estadísticos.

Diseño de regresión discontinua

¿Qué es el diseño de regresión discontinua?

Este enfoque puede utilizarse cuando las personas que participan en la intervención que se evalúa deben cumplir un criterio previo, conocido como umbral. El umbral determina la idoneidad de participación en el programa o política y generalmente se basa en una variable continua que se evalúa en todos los individuos que son potencialmente aptos para participar. Por ejemplo, los estudiantes cuya calificación en una prueba es inferior a una puntuación determinada se inscriben en un programa de refuerzo educativo, o las mujeres por encima o por debajo de una cierta edad pueden participar en un programa de salud (por ejemplo, las mujeres mayores de 50 años pueden acceder a un programa gratuito de detección del cáncer de mama).

Claramente, las personas que se encuentren por encima y por debajo del umbral son diferentes, y el criterio (o criterios) del umbral podría estar correlacionado con el resultado y dar lugar a un sesgo de selección. El refuerzo educativo se proporciona para mejorar los resultados del aprendizaje y, por tanto, se escoge para el programa a los estudiantes con peores resultados. Las mujeres mayores son más propensas a padecer cáncer de mama, y en consecuencia son ellas las seleccionadas para la detección. De modo que si simplemente se compara a los individuos que participan en el programa con los que no participan, los resultados serán sesgados.

Sin embargo, aquellos justo a ambos lados del umbral no son muy diferentes. Si el umbral para la inclusión en un programa de refuerzo educativo es una puntuación de 60 en una prueba, los alumnos inscritos en el programa que hayan obtenido una puntuación de 58 a 59,9 no serán muy diferentes de los que obtienen una puntuación de 60 a 60,9 y no participan. La regresión discontinua se basa en una comparación de la diferencia en los resultados medios de estos dos grupos.

Cómo aplicar el diseño de regresión discontinua

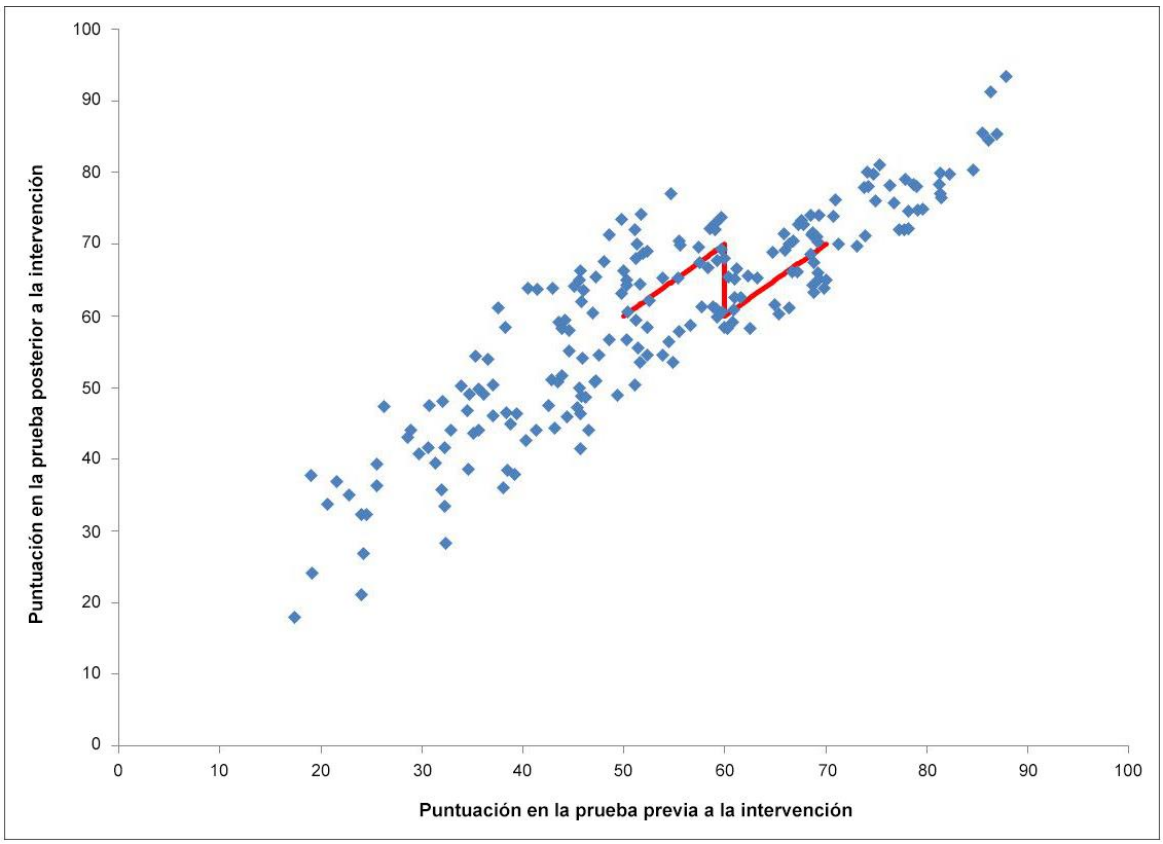
El primer paso es determinar el margen alrededor del umbral, lo cual se efectúa por medio de un enfoque iterativo. Inicialmente se puede configurar un margen reducido y comparar el equilibrio o la similitud entre los grupos de tratamiento y de comparación resultantes. Si el emparejamiento es bueno, se puede ampliar un poco el margen y volver a comprobar el equilibrio. Este proceso debe repetirse hasta que las muestras

comienzan a ser diferentes. Aunque el equilibrio se basa en las características observables, no hay ninguna razón para esperar un desequilibrio entre las características no observables (no ocurre lo mismo en el caso del emparejamiento por puntuación de la propensión, tal como se explicó anteriormente).

Una vez establecida la muestra, se ajusta una línea de regresión. Se trata de una línea trazada a través de los puntos de datos que representa el «ajuste óptimo» entre las variables estudiadas o que resume la «relación» entre las variables seleccionadas, es decir, cuando la línea es descendiente (desde la parte superior izquierda a la inferior derecha) indica una relación negativa o inversa; cuando es ascendiente (desde la parte inferior izquierda a la parte superior derecha) indica una relación positiva o directa. En este caso, la línea de regresión se traza sobre el resultado de interés seleccionado (por ejemplo, las puntuaciones en una prueba). La muestra para la regresión se limita a las observaciones situadas justo a cada lado del umbral. El gran reto del diseño de regresión discontinua suele residir en la falta de observaciones suficientes a cada lado del umbral para poder ajustar la línea de regresión.

En el gráfico 2 se muestra un ejemplo de un programa de refuerzo educativo. El criterio de selección para determinar la idoneidad de participación en el programa es la puntuación obtenida en una prueba previa a la intervención, con un umbral de 60. La variable «resultado» es la puntuación obtenida en una prueba posterior a la intervención. El diagrama de dispersión muestra que, como era de esperar, estas dos variables están relacionadas. Existe una relación positiva entre las puntuaciones previas y posteriores a la intervención. Los niños con una puntuación en la prueba previa a la intervención inferior a 60 asistieron a las clases de refuerzo. La muestra empleada para el análisis se tomó justo a cada lado del umbral (los niños incluidos obtuvieron puntuaciones previas a la intervención en el intervalo de 50 a 70, es decir, 10 unidades a cada lado del umbral). La línea de regresión trazada presenta un «salto»; se trata de la discontinuidad. El tamaño de este salto (10) es el impacto del programa, es decir, el programa de refuerzo educativo aumenta las puntuaciones en un promedio de 10 puntos.

Gráfico 2. Diseño de regresión discontinua



¿Qué se necesita para el diseño de regresión discontinua?

Es necesario contar con datos sobre la variable de selección y el indicador de resultado de todos los individuos considerados para la intervención, tanto si se los acepta como si no. Sin embargo, numerosos programas no conservan la información sobre los individuos no aceptados, lo cual puede dificultar el diseño de regresión discontinua.

Ventajas y desventajas del diseño de regresión discontinua

El diseño de regresión discontinua se ocupa de las características no observables de forma más convincente que otros métodos de emparejamiento cuasiexperimentales. También puede utilizar en gran medida datos administrativos, lo que reduce la necesidad de recopilar datos, aunque suele ser necesario reunir datos sobre los resultados de las personas no aceptadas en el programa.

Las limitaciones de esta técnica son que los criterios de selección o el umbral no siempre resultan claros y que la muestra puede no ser lo bastante grande para el análisis (como se señaló anteriormente). Además, el diseño de regresión discontinua produce un «efecto del tratamiento en la zona local». Es decir, la estimación del impacto es válida para los participantes cercanos al umbral, pero el impacto en los más alejados del mismo puede ser diferente (podría ser mayor o menor). No obstante, en la práctica, cuando se ha podido comparar este efecto «local» con el efecto «medio», no se han observado grandes diferencias. Esto indica que el diseño de regresión discontinua es un método aceptable para estimar los efectos de un programa o política.

Enfoques epidemiológicos

Los epidemiólogos aplican una serie de datos estadísticos tomados de las poblaciones tratadas y no tratadas, incluidos los mínimos cuadrados ordinarios y las regresiones logísticas en el caso de resultados dicotómicos (cumplen la condición = 1, no cumplen la condición = 0). Al utilizar estos métodos, es preferible: 1) emplear datos procedentes de grupos de tratamiento y de comparación bien emparejados, y 2) restringir el análisis de regresión a las observaciones de la región de soporte común. (No obstante, en la actualidad estos pasos por lo general no se llevan a cabo). Algunos estudios epidemiológicos presentan la diferencia de medias entre las observaciones tratadas y no tratadas, pero este enfoque no tiene en cuenta el posible sesgo de selección.

El análisis de supervivencia puede constituir un enfoque apropiado cuando los datos están censurados, es decir, que el periodo de exposición es incompleto debido al momento de recolección de los datos o a la muerte del participante en el estudio. En esas circunstancias, por lo general se emplea el modelo de riesgos proporcionales de Cox.

4. MÉTODOS CUASIEXPERIMENTALES PARA EL ANÁLISIS DE DATOS

Estimaciones del impacto por diferencia única

Las estimaciones del impacto por diferencia única comparan los resultados del grupo de tratamiento con los resultados del grupo de comparación una única vez tras la intervención (t +1, en la tabla 1).

Diferencia en diferencias

¿Qué es la «diferencia en diferencias»?

El método de la diferencia en diferencias, también conocido como de «diferencia doble», compara los cambios en los resultados *en el curso del tiempo* entre los grupos de tratamiento y de comparación a fin de estimar el impacto.

El método de la diferencia en diferencias proporciona una estimación del impacto más sólida que el método de la diferencia única, que solo compara la diferencia en los resultados entre los grupos de tratamiento y de comparación después de la intervención (en t +1). La aplicación del método de la diferencia en diferencias elimina la diferencia en los resultados entre los grupos de tratamiento y de comparación al inicio del estudio. No obstante, este método se utiliza mejor en conjunción con otros métodos de emparejamiento, como el emparejamiento por puntuación de la propensión o el diseño de regresión discontinua. Si el método de la diferencia en diferencias se emplea sin emparejamiento, los investigadores deben comprobar el «supuesto de tendencias paralelas», es decir, que la tendencia de los resultados en las zonas de tratamiento y de comparación haya sido similar antes de la intervención.

A continuación se muestra un ejemplo hipotético del método de la diferencia en diferencias. La tabla 4 muestra los datos del estado nutricional, medido según puntuaciones Z de peso para la edad (WAZ), de los grupos de tratamiento y de comparación antes y después de un programa de suplementación nutricional.

Tabla 4. Estado nutricional de los niños (WAZ) para los grupos de tratamiento y de comparación al inicio y al final del estudio

	Inicio	Final	Cambio
Tratamiento (Y_1)	-0,66	-0,48	+0,18
Comparación (Y_0)	-0,62	-0,58	+0,04
Diferencia		+0,10	+0,14

La magnitud del impacto estimada por los métodos de diferencia única y diferencia doble es muy distinta. El método de diferencia única (SD) calcula la diferencia entre las WAZ de los grupos de tratamiento y de comparación después de la intervención, es decir, SD = -0,48 – (-0,58) = 0,10. La estimación del método de la diferencia en diferencias (DID) es la diferencia entre las WAZ del grupo de tratamiento al inicio del estudio y después de la intervención menos la diferencia entre las WAZ del grupo de comparación al inicio del estudio y después de la intervención, es decir, DID = [-0,48 – (-0,66)] – [-0,58 – (-0,62)] = 0,18 – 0,04 = 0,14.

La estimación obtenida al aplicar el método de diferencia doble es mayor que la estimación del método de diferencia única, ya que el grupo de comparación tenía unas WAZ mejores que el grupo de tratamiento al inicio del estudio. El método de la diferencia en diferencias permite eliminar la diferencia inicial de WAZ entre los grupos de tratamiento y de comparación; la diferencia única no lo permite, y por tanto en este ejemplo da lugar a una subestimación del impacto del programa.

Cómo aplicar el método de la diferencia en diferencias

El primer paso consiste en identificar los indicadores de interés (resultados e impactos) que hay que medir y que afectan la intervención que se está evaluando. A continuación, las diferencias en los valores de los indicadores del grupo de tratamiento antes y después de la intervención se comparan con las diferencias en los mismos valores del grupo de comparación. Por ejemplo, con el fin de identificar los efectos de un programa de comida gratuita en el estado nutricional de los niños, se calcularía la diferencia de medias, tanto para el grupo de tratamiento como para el grupo de comparación, y luego se examinaría la diferencia entre ambos, es decir, la diferencia entre los cambios en el estado nutricional de los niños que participaron en la intervención y aquellos que no lo hicieron. Idealmente, los grupos de intervención y de comparación se habrán emparejado en función de las características clave empleando el emparejamiento por puntuación de la propensión, como se ha descrito anteriormente, para asegurarse de que, por lo demás, sean lo más similares posible.

Ventajas y desventajas del método de la diferencia en diferencias

La principal limitación del método de la diferencia en diferencias es que se basa en la suposición de que los indicadores de interés siguen la misma trayectoria temporal en los grupos de tratamiento y de comparación. Esta suposición se conoce como «supuesto de tendencias paralelas». Si el supuesto es correcto, la estimación del impacto de un programa efectuada con este método no resultaría sesgada. No obstante, si hay diferencias entre los grupos que cambian con el tiempo, este método no ayuda a eliminarlas.

En el ejemplo anterior, si las situaciones comparadas experimentaron cambios que influyeron en el estado nutricional de los niños —después su incorporación al programa de comida gratuita en otras situaciones— entonces el uso del método de la diferencia en diferencias no proporcionaría, por sí solo, una evaluación precisa del impacto. (Estos cambios pueden ocurrir, por ejemplo, a causa de programas de desarrollo que eleven el nivel de ingresos de los residentes, lo que implica que pueden permitirse proporcionar a sus hijos una dieta más nutritiva).

En resumen, la diferencia en diferencias es un buen enfoque para el cálculo de una estimación cuantitativa del impacto, pero, por lo general, este método por sí solo no es suficiente para hacer frente al sesgo de selección. Para evitar el sesgo de selección es necesario el emparejamiento, a fin de que los grupos de tratamiento y de comparación sean lo más parecidos posible.

Métodos basados en la regresión para calcular la estimación del impacto por diferencia única y doble

Las estimaciones del impacto por diferencia única y doble también pueden calcularse mediante la regresión de mínimos cuadrados ordinarios. Este enfoque se aplica a los mismos datos emparejados, incluida una variable indicadora del programa o la política en el lado derecho de la ecuación de regresión. Las variables que captan otros factores de confusión también pueden incluirse en la parte derecha para eliminar el efecto restante de cualquier desviación de estas variables en las zonas de tratamiento y de comparación sobre los resultados después del emparejamiento.

5. CUESTIONES ÉTICAS Y LIMITACIONES PRÁCTICAS

Cuestiones éticas

Los métodos cuasiexperimentales ofrecen opciones prácticas para llevar a cabo evaluaciones de impacto en situaciones reales. Mediante el uso de grupos preexistentes o autoseleccionados, por ejemplo personas que ya están participando en un programa, estos métodos evitan las preocupaciones éticas asociadas a la distribución aleatoria, como la retención o el retraso de un tratamiento potencialmente efectivo o la administración de un tratamiento menos eficaz a un grupo de participantes en el estudio (véase la Síntesis n.º 7 (Ensayos controlados aleatorios)).

Limitaciones prácticas

La falta de datos de buena calidad suele representar un obstáculo fundamental para el empleo de métodos cuasiexperimentales. Cualquier método que se aplique después de la finalización de un programa o política puede verse afectado sustancialmente por la falta de datos de referencia.

Debido a que los métodos cuasiexperimentales se basan en determinados supuestos (véase el cuadro 1), las conclusiones relativas a la causalidad extraídas de esos estudios son menos definitivas que las de un ensayo controlado aleatorio adecuadamente conducido. No obstante, en la mayoría de casos, si se llevan a cabo correctamente y se presentan con claridad (es decir, explicitando las limitaciones y cómo estas afectan los resultados), los métodos cuasiexperimentales suelen ser bien recibidos por los responsables de tomar decisiones.

6. ¿QUÉ OTROS MÉTODOS FUNCIONAN BIEN CON ESTE?

Como se destacó anteriormente, es recomendable utilizar diferentes métodos cuasiexperimentales de manera conjunta; por ejemplo, el método de la diferencia en diferencias puede combinarse con el de emparejamiento por puntuación de la propensión. Se recomienda usar los métodos cualitativos en combinación con los métodos cuasiexperimentales para entender mejor «por qué» un programa o política ha funcionado o no.

7. PRESENTACIÓN DE LOS RESULTADOS Y ANÁLISIS

Al redactar los resultados basados en una evaluación cuasiexperimental, es primordial proporcionar detalles sobre la metodología específica, incluida la recolección de datos. Dado que el éxito de estos métodos depende en gran medida de la calidad de los datos recogidos (o ya disponibles), debe ofrecerse algún tipo de garantía de calidad. También es importante brindar información acerca de la plausibilidad de los supuestos en que se basan estos métodos. Aunque algunos de los supuestos no pueden contrastarse directamente (por ejemplo, los supuestos de tendencias paralelas), los autores deben ofrecer argumentos claros sobre por qué creen que estos supuestos tienen fundamento.

Se recomienda que en la descripción de la metodología figuren detalles del método de muestreo, así como el enfoque con que se han construido los grupos de tratamiento y de comparación (incluidos el número de individuos, hogares o grupos participantes). El análisis y el informe de los resultados pueden efectuarse tanto para la totalidad de la muestra como para los subgrupos importantes (predefinidos) (por ejemplo, por edad o por sexo) para identificar y discutir los efectos diferenciales. A continuación es necesario vincular las constataciones a la teoría del cambio (véase la Síntesis n.º 2 (La teoría del cambio)), y emplearlas para

responder a las preguntas clave de evaluación; por ejemplo, ¿apoyan los resultados la teoría del cambio? Si no es así, ¿qué supuesto subyacente a la teoría del cambio no se cumplió?

Este tipo de análisis puede ayudar a los evaluadores a identificar recomendaciones para un programa o política en concreto, las cuales deberían formar parte de la conclusión del informe. En la mayoría de los casos, también sería útil incluir un debate sobre si los resultados pueden extrapolarse a diferentes entornos y en qué medida es posible hacerlo. Las conclusiones extraídas de los diseños cuasiexperimentales son causalmente válidas siempre y cuando se cumplan los supuestos relativos al método de emparejamiento empleado. También se debe contrastar y notificar la calidad del emparejamiento.

8. EJEMPLO DE BUENAS PRÁCTICAS

Un ejemplo de UNICEF en el que se emplean el método de la diferencia en diferencias y el de emparejamiento por puntuación de la propensión

UNICEF llevó a cabo una evaluación de impacto de Chile Solidario[3], un programa de transferencias monetarias condicionadas de Chile que se proponía mejorar varios resultados socioeconómicos de familias que vivían en la pobreza. Para evaluar el impacto del programa, los autores utilizaron datos de dos rondas de una encuesta nacional de panel (encuesta Casen).

Los autores usaron el método de la diferencia en diferencias y el de emparejamiento por puntuación de la propensión para obtener una estimación no sesgada del impacto. Las personas que recibieron la transferencia monetaria condicionada eran diferentes en algunos aspectos fundamentales a las que no recibieron la transferencia, por lo que el diseño de la evaluación de impacto tuvo que abordar el problema del sesgo de selección. Por ejemplo, los beneficiarios del programa eran más pobres, y tenían un nivel educativo y condiciones de vida peores que los que no participaron en él.

El empleo del emparejamiento por puntuación de la propensión permitió la construcción de un grupo de comparación compuesto por individuos similares a los individuos del grupo de tratamiento por lo que respecta a las características más observables. Los autores estimaron la ecuación de participación incluyendo los valores previos a la intervención de los ingresos procedentes del trabajo y de las rentas de los hogares, el número mayor de años de educación de los integrantes del hogar y el número de niños (menores de 14 años) en el hogar, además de tres características del entorno doméstico (suministro de agua, estado del techo y número de personas por habitación) y la región de residencia para dar cuenta de los factores específicos de la zona que podrían haber influido en las tasas de participación. Los hogares de tratamiento se emparejaron con los cuatro «vecinos» más cercanos (es decir, aquellos con las puntuaciones de la propensión más cercanas) no incluidos en el programa. No obstante, los autores no proporcionaron tablas ni cifras que mostraran la calidad del emparejamiento, lo que constituye un déficit de transparencia de la presentación.

Se calcularon una serie de estimaciones del impacto por diferencia única o doble, que demostraron que la transferencia monetaria condicional tuvo un impacto significativo en las familias puesto que ayudó a sacarlas de la pobreza extrema. Además, este estudio también concluyó que el programa contribuyó al aumento de la escolarización de los niños de 6 a 15 años y de su acceso a los servicios de salud pública.

[3] Martorano, Bruno y Marco Sanfilippo, «Innovative Features in Conditional Cash Transfers: An impact evaluation of Chile Solidario on households and children», *Innocenti Working Paper* No. 2012-03, Centro de Investigación de UNICEF Innocenti, Florencia, 2012. Véase http://www.unicef-irc.org/publications/pdf/iwp_2012_03.pdf.

9. EJEMPLOS DE POSIBLES PROBLEMAS

El mayor escollo potencial de los métodos cuasiexperimentales es el riesgo de obtener un emparejamiento de mala calidad. El grupo de comparación debe ser lo más similar posible al grupo de tratamiento antes de la intervención. Por tanto, es muy importante comprobar la calidad del emparejamiento mediante la elaboración de tablas de balance de los factores determinantes de los resultados de interés y de los resultados mismos.

Otro obstáculo potencial radica en la tendencia a centrarse en resultados significativos desde el punto de vista estadístico en detrimento de los no significativos. Deben comunicarse todos los resultados, y la discusión no ha de centrarse exclusivamente en los estadísticamente significativos.

En el informe de los resultados es tan importante analizar la magnitud de los efectos como su significación estadística. Por lo general, esto se pasa por alto, a pesar de que la significación estadística no necesariamente es suficiente para que una intervención interese a las autoridades o para que sea una opción rentable. También deben existir pruebas de que el efecto es suficientemente grande.

Todos los estudios cuantitativos se basan en el supuesto de que los datos son de alta calidad, por tanto deben realizarse controles de calidad de los mismos.

10. LECTURAS Y ENLACES CLAVE

Angrist, Joshua D. y Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Nueva Jersey, 2009, págs. 227-259.

Caliendo, Marco y Sabine Kopeinig, «Some Practical Guidance for the Implementation of Propensity Score Matching», *IZA Discussion Paper* No. 1588, Institute for the Study of Labor (IZA), Bonn, 2005. Véase http://ftp.iza.org/dp1588.pdf.

Gertler, Paul J., *et al.*, *Impact Evaluation in Practice*, Banco Mundial, Washington D. C., 2010, págs. 81-116. Véase http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf.

Kaplan, Josiah, «Propensity Scores», página web, BetterEvaluation, 2011, http://betterevaluation.org/evaluation-options/propensity_scores.

Khandker, Shahidur R., *et al.*, *Handbook on Impact Evaluation: Quantitative Methods and Practices,* Banco Mundial, Washington D. C., 2010, págs. 53-103. Véase http://bit.ly/1d2Ve8m.

Lee, David S. y Thomas Lemieux, «Regression discontinuity designs in economics», *NBER Working Paper* No. 14723, Oficina Nacional de Investigaciones Económicas, Cambridge, Massachusetts, 2009. Véase http://www.nber.org/papers/w14723.pdf?new_window=1.

Martorano, Bruno y Marco Sanfilippo, «Innovative Features in Conditional Cash Transfers: An impact evaluation of Chile Solidario on households and children», *Innocenti Working Paper* No. 2012-03, Centro de Investigación de UNICEF Innocenti, Florencia, 2012. Véase http://www.unicef-irc.org/publications/pdf/iwp_2012_03.pdf.

Qasim, Qursum y 3ie, «Regression Discontinuity», página web, BetterEvaluation, 2011, http://betterevaluation.org/evaluation-options/regressiondiscontinuity.

Ravallion, Martin, «Assessing the Poverty Impact of an Assigned Program», en F. Bourguignon y L. A. Pereira da Silva (eds.), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation*

Techniques and Tools, Volume 1, Oxford University Press, Nueva York, 2003. Véase http://origin-www.unicef.org/socialpolicy/files/Assessing_the_Poverty_Impact_of_an_Assigned_Programme.pdf.

Shadish, William R., *et al.*, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, 2002, págs. 103-243.

GLOSARIO

Ensayos controlados aleatorios	*Diseño de investigación o evaluación con dos o más grupos seleccionados de forma aleatoria (un grupo experimental y un grupo de control) en los que el investigador controla o introduce una intervención (por ejemplo, un nuevo programa o política) y mide su impacto en la variable dependiente al menos dos veces (medición anterior y posterior al ensayo). En concreto, los ensayos controlados aleatorios —que tienen su origen en el contexto clínico y se conocen como la «regla de oro» de la investigación médica y sanitaria— se suelen utilizar para responder a las preguntas de investigación de la evaluación, que tratan de evaluar la eficacia de las intervenciones de un programa o política en entornos de desarrollo.*
Estimación de variables instrumentales	*Técnica estadística que estima las relaciones causales cuando no es factible llevar a cabo un ensayo controlado aleatorio o cuando una intervención no llega a cada participante o unidad de un ensayo controlado aleatorio.*
Estudio de base	*Análisis que describe la situación previa a una intervención, a partir de la cual puede medirse el avance o pueden efectuarse comparaciones. (Definición del CAD de la OCDE, 2010)*
Grupo de comparación	*En un diseño de investigación cuasiexperimental, es el grupo de participantes o sujetos de la investigación que, a efectos de comparación, no recibe el tratamiento o la intervención dados al grupo de tratamiento o intervención. Por lo general, los sujetos del grupo de comparación no se distribuyen de forma aleatoria por su condición, como ocurriría con los sujetos del grupo de control en un estudio de diseño experimental. Véase: grupo de control, grupo de tratamiento.*
Grupo de tratamiento	*Los sujetos o participantes expuestos a la variable independiente; también llamado «grupo experimental» o «grupo de intervención».*
Indicador	*Medida verificable seleccionada por la dirección del programa o política a fin de tomar decisiones al respecto. Por ejemplo, el porcentaje de estudiantes que aprueban una prueba estandarizada.*
Modelo de riesgos proporcionales de Cox	*Técnica estadística o de elaboración de modelos que examina la supervivencia de un paciente y una serie de variables exploratorias. (Definición procedente de www.whatisseries.co.uk)*
Regresión	*Procedimiento estadístico para predecir valores de una variable dependiente sobre la base de los valores de una o más variables independientes. La regresión ordinaria utiliza los mínimos cuadrados ordinarios para encontrar la línea que mejor se ajusta, y obtiene coeficientes que predicen el cambio en la variable dependiente para una unidad de cambio de la variable independiente. (Definición de la Universidad de Strathclyde)*
Regresión de mínimos cuadrados ordinarios	*Técnica de elaboración de modelos lineales generalizada que puede utilizarse para modelar una variable de respuesta única registrada en al menos una escala de intervalo. La técnica puede aplicarse a una o varias variables explicativas y también a variables explicativas categóricas adecuadamente codificadas. (Definición de Hutcheson, 2011)*
Regresión logística	*Técnica de regresión que estima la probabilidad de que ocurra un suceso. (Definición de la Universidad de Strathclyde). Véase: regresión.*