Gac Sanit. 2011;**25(Supl 1)**:9-16

Diseños evaluativos en salud pública: aspectos metodológicos

Mª José López[a,b,c*], Marc Marí-Dell'Olmo[a,b,c], Anna Pérez-Giménez[a,b,c] y Manel Nebot[a,b,c,d]

[a]*Agencia de Salud Pública de Barcelona, España*
[b]*CIBER de Epidemiología y Salud Pública (CIBERESP), España*
[c]*Institut d'Investigació Biomèdica (IIB Sant Pau), Barcelona, España*
[d]*Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, España*

Palabras clave:
Evaluación
Salud pública
Diseños evaluativos

RESUMEN

La evaluación de las intervenciones de salud pública, en la cual rara vez es posible la aleatorización de individuos y habitualmente intervienen múltiples factores, implica numerosos retos metodológicos. Para afrontarlos hay que tener en cuenta determinados aspectos, como la elección de un diseño evaluativo apropiado y la realización de un análisis estadístico que considere los posibles confusores. El objetivo de este artículo es describir los diseños más frecuentes en la evaluación de intervenciones (políticas, programas o campañas) de salud pública, enumerando sus características, analizando sus principales ventajas y limitaciones, y haciendo una breve descripción del análisis estadístico más utilizado en cada uno de ellos.

Evaluative designs in public health: methodological considerations

Keywords:
Evaluation
Public health
Evaluative designs

ABSTRACT

Evaluation of public health interventions poses numerous methodological challenges. Randomization of individuals is not always feasible and interventions are usually composed of multiple factors. To face these challenges, certain elements, such as the selection of the most appropriate design and the use of a statistical analysis that includes potential confounders, are essential. The objective of this article was to describe the most frequently used designs in the evaluation of public health interventions (policies, programs or campaigns). The characteristics, strengths and weaknesses of each of these evaluative designs are described. Additionally, a brief explanation of the most commonly used statistical analysis in each of these designs is provided.

Introducción

La evaluación de intervenciones en salud pública es un proceso complejo que requiere la definición de indicadores adecuados, la elección de un diseño evaluativo apropiado y la realización de un análisis estadístico que considere los posibles confusores. Todo ello permitirá estimar el efecto de la intervención, cuantificando la magnitud del cambio en el indicador de resultado y descartando explicaciones alternativas a la intervención que hubieran podido influir en el cambio observado.

Los diseños tratados en este artículo pueden clasificarse en tres grandes grupos: diseños no experimentales, diseños cuasiexperimentales y diseños experimentales[1,2] (tabla 1). Los diseños no experimentales se caracterizan por la ausencia de grupo de comparación que no recibiría la intervención. Los diseños experimentales y cuasiexperimentales incluyen, como mínimo, un grupo de comparación, y la diferencia principal entre ambos diseños es el tipo de asignación de los individuos. Los diseños experimentales se caracterizan por una asignación individual aleatoria entre el grupo que recibe la intervención y el grupo de comparación. En los diseños cuasiexperimentales esta asignación se basa en criterios de conveniencia, aunque también es posible la asignación de grupos de manera aleatoria (ensayo comunitario). Dependiendo del tipo de asignación, así como del número de mediciones realizadas antes y después de la intervención, los tres diseños básicos pueden dividirse en una serie de diseños evaluativos que serán abordados a lo largo de este artículo.

Nuestro objetivo es describir los diseños más frecuentes en la evaluación de intervenciones (políticas, programas o campañas) de salud pública, enumerando sus características, analizando sus principales ventajas y limitaciones, y haciendo una breve descripción del análisis estadístico más utilizado en cada uno de ellos.

Diseños no experimentales

Los diseños no experimentales carecen de grupo comparación. Reciben también el nombre de diseños reflexivos, ya que habitualmente el valor de la variable resultado postintervención se compara con el valor de la variable para esos mismos individuos antes de la intervención. En este diseño asumimos que la población permanece "igual" respecto a otros aspectos que pudieran modificar los resultados (p. ej., otras intervenciones o cambios históricos), por lo que el cambio observado podría atribuirse a la intervención. Como conse-

*Autora para correspondencia.
Correo electrónico: mjlopez@aspb.cat (Mª José López)

0213-911/$ - see front matter

Tabla 1
Clasificación y descripción de los principales diseños evaluativos en salud pública

	Presencia de grupo de comparación	Momento de selección del grupo de comparación	Unidad de asignación	Tipo de asignación	Unidad de intervención	Unidad de medida	Número de medidas
Diseño no experimental (reflexivo)							
Antes-después (pre-post)	No	----------	----------	----------	Individuos o grupos	Individuos o grupos	Mínimo una antes y una después
Serie temporal	No	----------	----------	----------	Grupo	Grupo	Número necesario para controlar por tendencia y estacionalidad
Diseño cuasi-experimental							
Antes-después (pre-post)	Sí	Normalmente previo	Individuo o grupo	Conveniencia	Individuos o grupos	Individuos o grupos	Mínimo una antes y una después de la intervención
Ensayo comunitario	Sí	Previo	Grupos	Aleatoria	Individuos o grupos	Individuos o grupos	Mínimo una antes y una después de la intervención
Regresión discontinua	Sí	Previo	Individuo o grupo	Por un punto de corte establecido	Individuos o grupos	Individuos o grupos	Mínimo una antes y una después de la intervención
Serie temporal múltiple	Sí	Previo o posterior	Grupo	Conveniencia	Grupo	Grupo	Número necesario para controlar por tendencia y estacionalidad
Diseño experimental							
Post	Sí	Previo	Individuo	Aleatoria	Individuo	Individuo	Mínimo una después de la intervención
Antes-después (pre-post)	Sí	Previo	Individuo	Aleatoria	Individuo	Individuo	Mínimo una antes y una después de la intervención

cuencia, este tipo de diseño evaluativo es el más vulnerable a las amenazas a la validez interna, ya que aunque el cambio puede cuantificarse, la ausencia de grupo de comparación dificulta la atribución del efecto a la intervención. Entre los principales diseños no experimentales se encuentran el diseño antes-después y la serie temporal.

Diseño antes-después (o pre-post)

El diseño no experimental antes-después (o pre-post) requiere una o más medidas tomadas antes y después de la intervención en la población intervenida (fig. 1 A). Al no haber grupo de comparación, el valor obtenido tras la intervención puede compararse con el valor previo a ésta. El cambio entre las medidas tomadas antes y después de la intervención se utiliza como medida del efecto. La principal limitación de este diseño es la dificultad de poder atribuir el efecto observado a la intervención, ya que éste podría deberse a otros factores distintos al programa. Entre las ventajas destaca que consume poco tiempo y recursos, y que permite estimar el efecto de una intervención cuando no disponemos de grupo de comparación. Este diseño es especialmente apropiado cuando la cadena causal entre la intervención del programa y el resultado esperado es directa (no hay muchos factores intermedios), las condiciones medidas son estables a lo largo del tiempo (p. ej., no se ven afectadas por variaciones estacionales) y el periodo analizado es corto. Conviene tener presente que cuanto más tiempo pasa entre las mediciones antes y después de la intervención, mayor es la posibilidad de que otros factores distintos a la intervención hayan intervenido en el efecto encontrado.

El diseño antes-después (o pre-post) se aplica principalmente en el ámbito de la evaluación de políticas públicas o programas de amplia cobertura, en los cuales encontrar una población similar a la intervenida pero no afectada por la intervención no siempre es posible. Éste es el caso de las normativas políticas definidas para el conjunto de una región o un estado, como la Ley de medidas sanitarias frente al tabaquismo que entró en vigor en 2006, cuyo impacto sobre la exposición al humo ambiental de tabaco[3] se evaluó tomando medidas de las concentraciones de nicotina ambiental antes y después de la implantación de la ley en los mismos lugares de trabajo y hostelería. Los datos mostraron disminuciones significativas en los lugares de trabajo y en aquellos locales de hostelería en que se prohibió fumar tras la ley. Estos resultados sugerían que la ley había sido efectiva en la reducción de la exposición en los lugares de trabajo, pero que sería necesaria la ampliación de la prohibición a todos los lugares de hostelería.

En el caso de que la variable resultado se haya medido en los mismos individuos antes y después de la intervención (datos emparejados), para determinar si hay un cambio significativo pueden aplicarse distintos tests clásicos en función de la distribución de los datos. Si la variable medida es continua y sigue una distribución normal (antes y después de la intervención), puede aplicarse la prueba *t* para datos emparejados. Si éstos no siguen una distribución normal, es posible aplicar el test no paramétrico de Wilcoxon para datos emparejados. Si se dispone de una variable categórica, y quiere estimarse un cambio en la proporción, aplicaremos la prueba de ji cuadrado de McNemar. En caso de que los datos no sean emparejados, se aplicaría la prueba *t*, la U de Mann-Whitney y la de ji cuadrado, respectivamente. En caso de disponer de varias medidas preintervención o postintervención, deberían aplicarse los siguientes tests para datos emparejados: ANOVA de medida repetidas (distribución normal), Friedman (distribución no normal) y Q de Cochran (en caso de proporciones). Finalmente, para datos independientes (no emparejados) se aplican el test de ANOVA (distribución normal), el de Kruskal-Wallis (distribución no normal) y la ji cuadrado.

Serie temporal

El diseño no experimental de serie temporal es apropiado cuando hay datos temporales agregados disponibles que permiten describir la tendencia temporal mediante una serie de medidas recogidas antes y después de la intervención (fig. 1 B). En general, el diseño es más robusto (es decir, menos vulnerable a las amenazas a la validez interna) cuando se dispone de un número importante de mediciones en el tiempo. Aunque no se ha establecido un número mínimo de mediciones consensuado, debe haber un número suficiente entre los valores previos y posteriores a la intervención. Algunos autores proponen 10 o 15 observaciones por periodo, mientras que otros sugieren un total de 50 observaciones entre ambos periodos[4,5]. Como en otros diseños no experimentales, entre sus limitaciones destaca la dificultad de atribuir

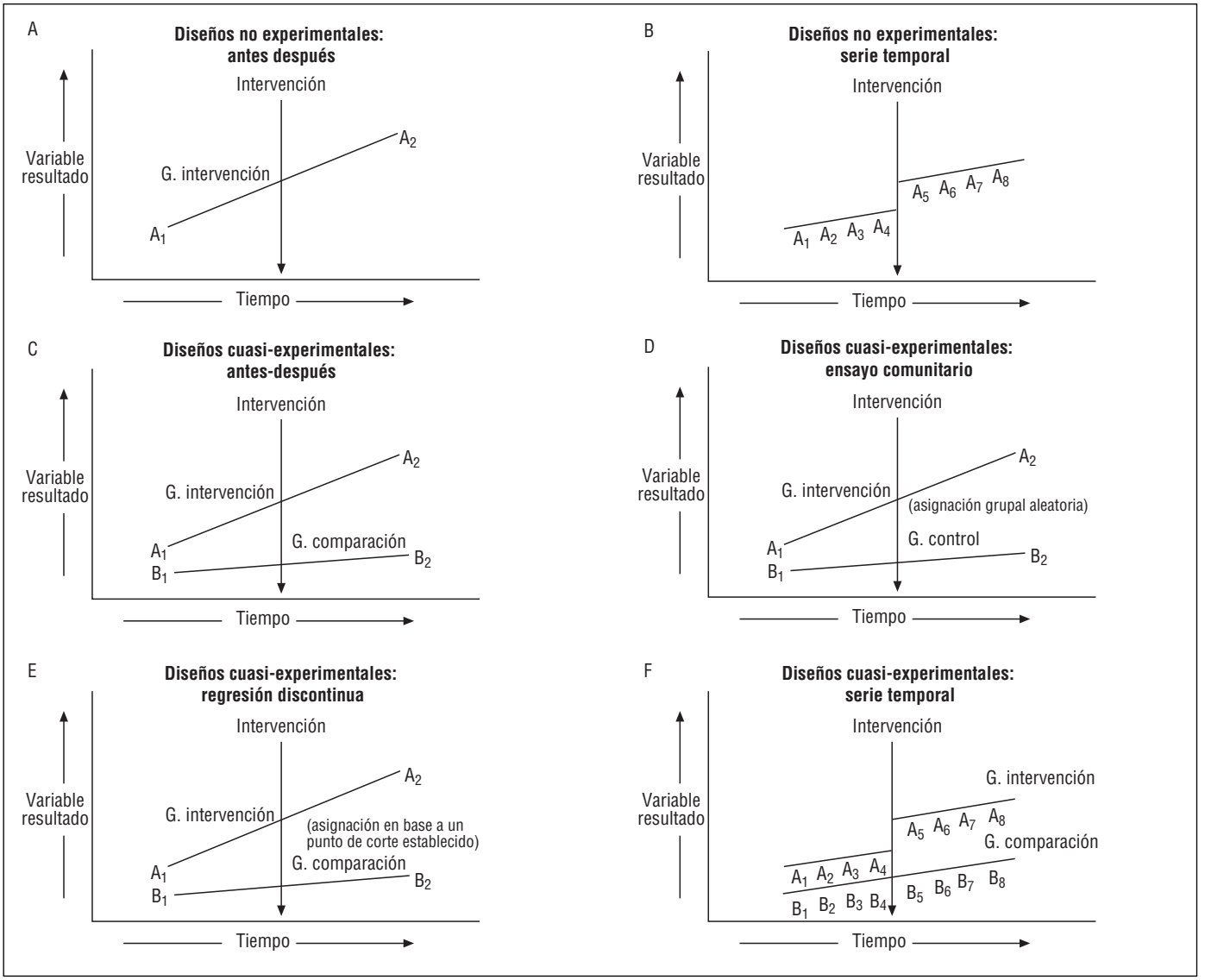

Figura 1. Principales diseños evaluativos (no experimentales y cuasiexperimentales) en salud pública.

una asociación causal entre la intervención y los cambios observados, ya que, al no disponer de grupo de comparación, éstos podrían deberse a otros factores distintos a la intervención. Entre sus fortalezas se encuentra la posibilidad de descartar los sesgos históricos, ya que permite controlar por determinados factores confusores tales como la tendencia temporal y la estacionalidad. Este diseño, al igual que el anterior, se ha utilizado principalmente para evaluar políticas públicas, como por ejemplo las relacionadas con el control de las lesiones por accidentes de tráfico. Así, en el estudio de Novoa et al[6] se evaluó el plan estratégico de seguridad vial (conjunto de medidas implementadas para promover la seguridad en las carreteras) comparando los datos de lesiones de tráfico de la Dirección General de Tráfico en el periodo preintervención (2000-2003) con los del periodo postintervención (2004-2006). Los resultados del estudio sugieren una importante reducción del riesgo de lesiones de tráfico, que podría ser atribuible a las medidas del plan estratégico.

Desde el punto de vista analítico puede considerarse que las series temporales están constituidas por tres componentes: tendencia, estacionalidad y componente residual[7]. De forma exploratoria, suele ser de utilidad realizar en primer lugar los gráficos que permitan observar visualmente si hay un cambio debido a la intervención. La representación gráfica más común consiste en la evolución de la serie temporal original (fig. 2 A), que puede desglosarse en las tendencias separadas para los periodos preintervención y postintervención (fig. 2 B), o bien en un gráfico de subseries estacionales. Por ejemplo, en la figura 2 C se representan las subseries temporales mensuales, describiendo el valor de la variable resultado del mismo mes para los distintos años (2001-2005). Para determinar el impacto de la intervención pueden emplearse modelos de regresión paramétricos, en los cuales es posible introducir variables confusoras de las que se disponga de valores para todo el periodo de la serie temporal. En caso de disponer de un mínimo de 50 mediciones, también podría utilizarse un modelo ARIMA (Auto Regresivo Integrado de Medias Móviles)[4].

Son tres los posibles patrones de cambio (fig. 3) entre los periodos antes y después[4,5]: *1)* cambio de nivel sin cambio de tendencia; *2)* cambio de tendencia sin cambio de nivel; y *3)* cambio de tendencia y de nivel. En caso de haber un cambio de nivel entre periodos sin un cambio de tendencia, que es uno de los patrones más habituales, el modelo de regresión paramétrico podría resumirse del siguiente modo[4,5,8]:

$$Y_t = \beta_0 + \beta_1 Int_t + \beta_2 T_t + \delta_1 Z_{1t} + \cdots + \delta_n Z_{nt} + \varepsilon_t$$

donde Y_t es la variable resultado para cada medida en el tiempo t (t = 1,···, T), Int_t es la variable intervención, que toma valor 1 para el pe-

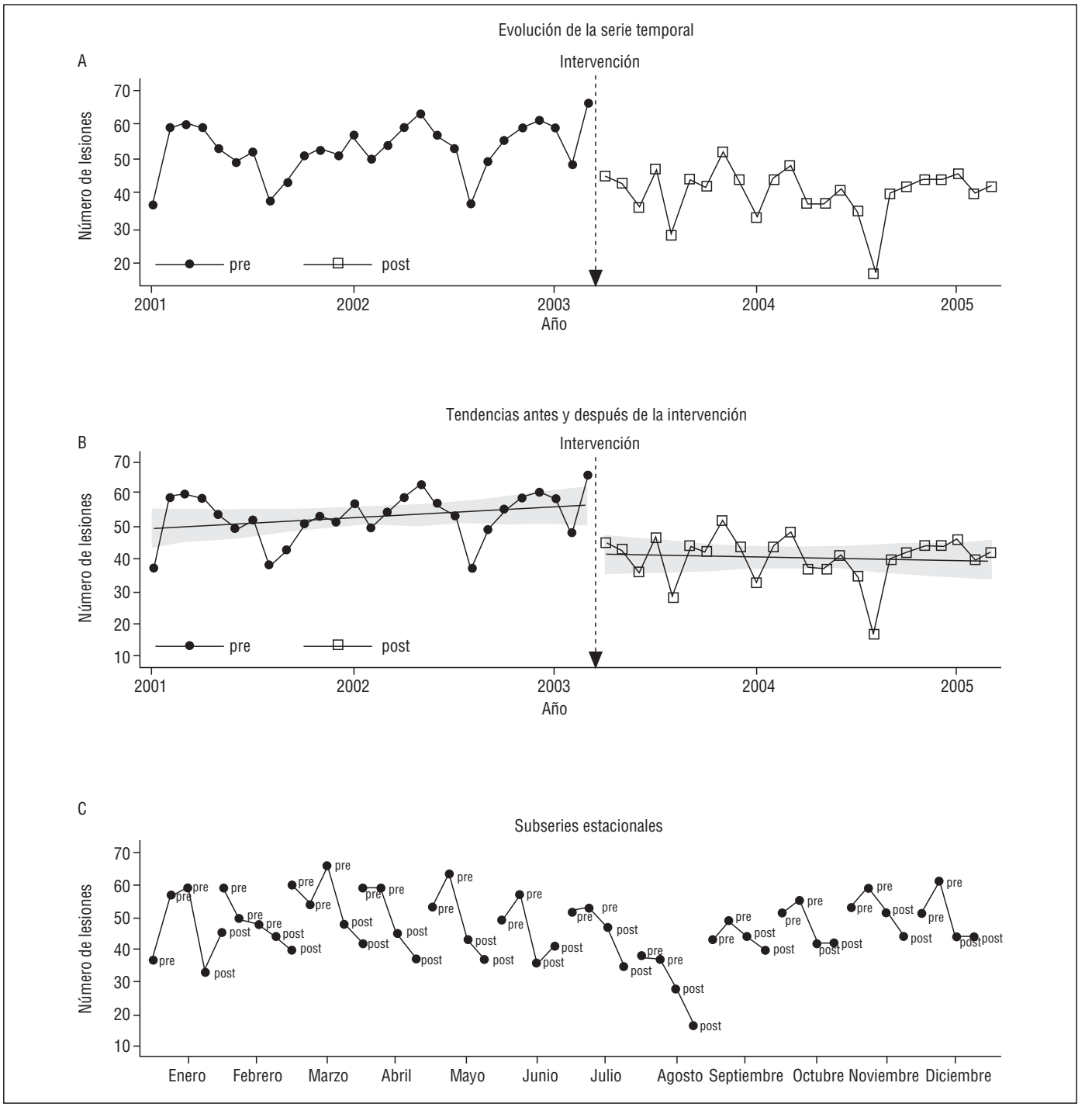

Figura 2. Análisis descriptivo de series temporales. Número de lesiones de tráfico (Barcelona, 2001-2005)[26].

riodo postintervención y 0 para el preintervención, y T_t es la variable que permite controlar la tendencia y toma valores de 1,···, T. Las variables Z_{1t} ,···, Z_{nt} son las *n* variables confusoras por las que queremos controlar la estimación del impacto de la intervención. Finalmente, si se observa estacionalidad, es posible controlarla introduciendo en el modelo diversas funciones (p. ej., funciones sinusoidales)[9]. En este modelo, el impacto de la intervención se estima a partir del parámetro β_1, que se interpreta como el cambio en media de la variable resultado en el periodo postintervención respecto al periodo preintervención, independientemente de que haya tendencia o estacionalidad en la serie.

En caso de que la intervención produzca un cambio de tendencia (patrones 2 y 3), será necesario incluir en el modelo la interacción de la variable intervención (Int_t) con la variable tendencia (T_t).

Diseños cuasiexperimentales

Los diseños cuasiexperimentales se caracterizan por una asignación no aleatoria de los individuos o grupos a un grupo de intervención y a un grupo de comparación, o bien una asignación aleatoria pero grupal (ensayo comunitario). Este diseño es muy común en el ámbito de la evaluación de programas preventivos, ya que habitualmente no es posible una asignación aleatoria individual. La validez de este tipo de diseños depende en gran medida de lo parecido que sea el grupo de comparación al grupo de intervención en todos los aspectos relevantes que podrían afectar a la variable resultado. Entre los principales diseños cuasiexperimentales destacan los diseños cuasiexperimentales antes-después (o pre-post) (fig. 1 C), los

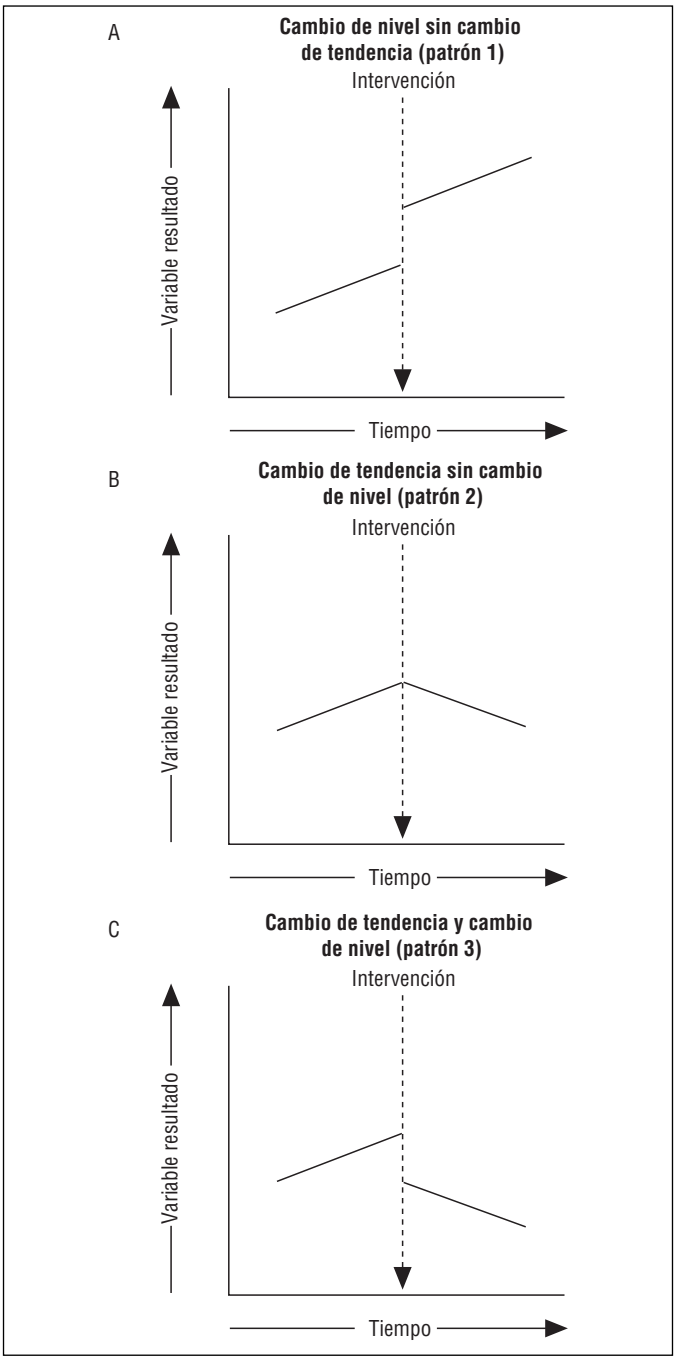

Figura 3. Patrones de series temporales.

ensayos comunitarios (fig. 1 D), el diseño de regresión discontinua (fig. 1 E) y las series temporales múltiples (fig. 1 F).

Diseño antes-después (o pre-post)

El diseño cuasi-experimental antes-después (o pre-post) es uno de los más frecuentes en la evaluación de intervenciones en salud pública. Es necesario tener, como mínimo, una medida antes de la intervención y otra después de ella, y comparar el cambio observado en la variable resultado entre el grupo de intervención y el grupo de comparación. La diferencia observada entre ambos cambios correspondería al efecto de la intervención si los dos grupos fueran equivalentes. Sin embargo, la dificultad para encontrar un grupo de comparación similar al de intervención limita la fuerza de este diseño para descartar que factores diferentes a la intervención puedan ser causa del efecto. Cuando haya dudas razonables de que los grupos sean equivalentes, habrá que intentar descartar el sesgo de selección controlando por las posibles variables confusoras. Este diseño es ampliamente utilizado en el ámbito de la evaluación de programas de prevención y promoción de la salud. Un ejemplo de este tipo de estudios es el de Robitaille et al[10], en el cual se evaluó un programa de prevención de las caídas en las personas de edad avanzada en Canadá. En este estudio se reclutó, mediante organizaciones que trabajaban con ancianos, un grupo de personas que recibieron una intervención que incluía varias sesiones en las cuales se les entrenaba con ejercicios que promovían el equilibrio, la flexibilidad y la fuerza. Asimismo, se escogió un grupo de comparación de características similares a las del grupo intervenido, y que posteriormente recibiría la misma intervención. A ambos grupos se les realizó una serie de pruebas antes y después de la intervención. Los resultados del estudio mostraron que las personas que habían realizado los ejercicios ofrecidos en la intervención presentaban mejoras en el equilibrio y la fuerza motora en relación al grupo de comparación, y por tanto concluyeron que la intervención era efectiva en la mejora de aquellas habilidades que podrían prevenir las caídas en las personas de edad avanzada.

En este tipo de diseños puede haber diversas variables confusoras que determinen la inclusión desigual de individuos en un grupo u otro. Por ello, el primer paso es identificar cuáles pueden ser estas variables y tenerlas en cuenta al elegir el grupo de comparación. El segundo paso, una vez seleccionados los grupos, es hacer una descripción comparativa de las posibles variables confusoras en ambos grupos, asumiendo que éstas sean conocidas, observables y mensurables. Para ello se aplicará el test de comparación que se corresponda a la distribución de la variable estudiada.

En la estimación del impacto de la intervención, una de las técnicas más comunes es el uso de modelos de regresión capaces de estimar el efecto de la intervención, controlando por las variables confusoras (Z). Este modelo se podría formular del siguiente modo:

$$Y_{i,t} = \beta_0 + \beta_1 Int_i + \beta_2 Y_{i,t-1} + \delta_1 Z_1 + \cdots + \delta_n Z_n + e_i$$

donde $Y_{i,t-1}$ y $Y_{i,t}$ son, para cada individuo *i*, el valor de la variable resultado antes y después de la intervención, respectivamente. La variable Int_i es la variable intervención, que habitualmente toma valor 1 para los individuos intervenidos y 0 para los individuos del grupo de comparación. Las variables $Z_1,\cdots, Z_n$ son las *n* variables confusoras. Finalmente, el impacto de la intervención se estima a partir del parámetro β_1, que se interpreta como la diferencia media de la variable resultado (en el post-test) en el grupo de intervención respecto al grupo de comparación, controlando por la variable resultado en el momento preintervención y por las variables confusoras $Z_1,\cdots, Z_n$.

Una alternativa a introducir las variables confusoras (*Z*) en el modelo es el cálculo del *Propensity Score* (PS)[11,12], que es la probabilidad condicionada que tiene cada individuo de la muestra de ser asignado al grupo de intervención dadas las variables confusoras $Z_1,\cdots, Z_n$[11]. El PS se obtiene habitualmente mediante regresión logística o análisis discriminante, donde las variables confusoras observadas son las variables independientes y pertenecer o no al grupo de intervención es la variable dependiente. El PS se aplica con el fin de equilibrar grupos no equivalentes mediante distintas técnicas: apareamiento (*matching*), estratificación, modelo de regresión (covarianza) o ponderación por el PS[13]. Como ejemplo de la utilización de esta técnica estadística, en el estudio de Lauby et al[14] se evaluó una intervención comunitaria que promocionaba el uso del preservativo entre las mujeres sexualmente activas. El PS se utilizó para controlar las diferencias entre las comunidades intervenidas y las del grupo de comparación con respecto a características sociodemográficas y comportamientos de riesgo.

Ensayos comunitarios

En el caso de los ensayos comunitarios, que incluyen asignación aleatoria de grupos, así como en los diseños que consideren varias me-

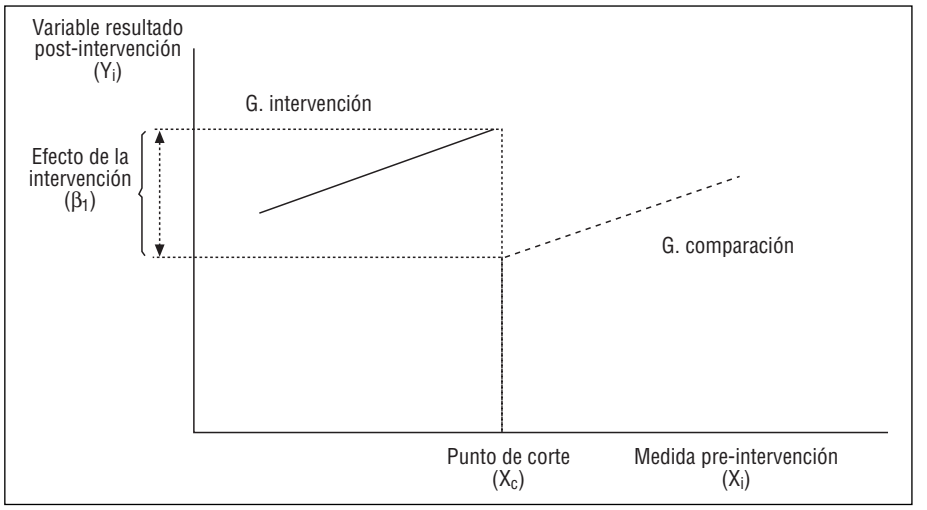

Figura 4. Diseño de regresión discontinua. Se representa la variable resultado postintervención (Y_i) en función de una medida preintervención (X_i). Se determina un punto de corte de una medida preintervención (X_c). Todos los individuos por debajo de este punto son intervenidos (línea continua), mientras que los que están por encima forman parte del grupo de comparación (línea discontinua). El efecto de la intervención (β_1) corresponde al cambio de nivel de la variable resultado en el punto de corte.

didas en el tiempo, es posible aplicar modelos que tengan en cuenta una estructura jerárquica, como las *generalized estimating equations* (GEE), los modelos mixtos o los modelos de análisis multinivel[15,16]. Los ensayos comunitarios son uno de los diseños más utilizados en la evaluación de programas de prevención y promoción en las escuelas. La principal ventaja de este tipo de diseños es que la asignación aleatoria de los grupos permite asumir que los principales factores confusores asociados al grupo se distribuirán por igual entre ambos. Además, la asignación de grupos permite minimizar la "contaminación" entre los individuos que reciben la intervención y los que no la reciben. Por ejemplo, si la mitad de los alumnos de un aula participaran en una intervención y el resto no, la contaminación entre ellos sería más probable que si estuvieran en aulas o escuelas separadas. Como desventaja, cabe señalar que los individuos de un mismo grupo pueden diferir en factores que afecten a la intervención. En tal caso, habría que tener en cuenta esa variabilidad individual con modelos estadísticos apropiados, como los ya comentados previamente. Faggiano et al[17] evaluaron la efectividad de un programa de prevención de abuso de sustancias en el que participaron 170 escuelas de siete países europeos asignadas aleatoriamente a grupos de intervención y de comparación. Para tener en cuenta la estructura jerárquica de los datos y el efecto del grupo se utilizó en el análisis un modelo multinivel, en el cual se incluyeron tres niveles: escuela, clase y estudiante. El estudio demostró la efectividad de la intervención en la disminución del abuso de alcohol y de cannabis.

Regresión discontinua

El diseño de regresión discontinua se caracteriza porque la asignación de individuos al grupo de intervención y al grupo de control se lleva a cabo basándose en un valor establecido (punto de corte) de una medida preintervención (fig. 4). Partiendo de una variable continua, todos los individuos que estén por encima de ese valor determinado se asignan a uno de los grupos (p. ej. al grupo de comparación) y todos aquellos por debajo de dicho valor al otro (en este caso el grupo de intervención). Si la intervención es efectiva, esperaremos que en el grupo intervenido varíe el valor en la variable medida tras la intervención, mientras que en el grupo de comparación no se observarán cambios. Este diseño es especialmente adecuado cuando queremos que la intervención se dirija a aquellos que más la necesitan, pues no requiere que asignemos al grupo de comparación individuos que potencialmente necesitarían la intervención. Como limitación se encuentra la dificultad de disponer, en muchas ocasiones, de una variable preintervención continua en la cual establecer un punto de corte que permita asignar individuos a ambos grupos. Como ejemplo, este diseño se utilizó en Chicago para evaluar la efectividad de los cursos de refuerzo ofrecidos a los alumnos con peores resultados en lectura y matemáticas. En este estudio, la asignación al grupo de intervención y al grupo de comparación se hizo según la puntuación obtenida en un test sobre estas asignaturas que se pasó a todos los alumnos. Los resultados demostraron un aumento en el rendimiento académico en los alumnos que atendieron el curso de refuerzo[18].

El modelo estadístico más sencillo para analizar este tipo de diseño sería el siguiente[19,20]:

$$Y_i = \beta_0 + \beta_1 Int_i + \beta_2(X_i - X_c) + e_i$$

donde Y_i es, para cada individuo *i*, el valor de la variable resultado en el momento posterior a la intervención. La variable Int_i es la variable intervención, que habitualmente toma un valor 1 para las individuos intervenidos y 0 para los del grupo de comparación. La variable X_i es la medida preintervención para cada individuo, y X_c es el valor de esta variable que se toma como punto de corte. Finalmente, el impacto de la intervención se estima a partir del parámetro β_1 (fig. 4), que se interpreta como la diferencia en media de la variable resultado en el grupo de intervención respecto el grupo de comparación.

Este modelo es correcto bajo dos supuestos: *1)* que la relación entre la variable dependiente y la medida preintervención es lineal; y *2)* que las pendientes de las rectas de regresión para ambos grupos son paralelas (fig. 4). En caso de no cumplirse estos supuestos, deberían aplicarse modelos más complejos[21].

Serie temporal múltiple

El diseño de serie temporal múltiple exige la existencia de datos temporales agregados de la variable resultado, disponibles en un grupo sujeto a la intervención y en un grupo de comparación. Este diseño compara las medidas observadas en los grupos de intervención y de comparación a lo largo del tiempo. Dentro de los cuasiexperimentales, se trata de uno de los diseños menos susceptibles a las amenazas a la validez interna, con la limitación de que suele ser difícil encontrar un grupo de comparación equivalente cuando la unidad de análisis es muy grande (como un país o una región). Su principal ventaja es que, al disponer de datos que muestran la tendencia, pueden descartarse los sesgos históricos. Asimismo, el grupo de comparación permite descartar que el impacto de la intervención se deba a otros factores, ya que se espera que la intervención produzca un efecto en la serie del grupo de intervención (entre las medidas recogidas antes y después de la intervención), pero que este efecto no se observe en la serie de comparación.

Este diseño se aplica principalmente en la evaluación de políticas públicas. Por ejemplo, Benavides et al[22] estudiaron el efecto de los planes de actuación preferente sobre la incidencia de lesiones por accidentes de trabajo. Los autores observaron que, para el conjunto de España, la tendencia de las lesiones por accidentes de trabajo por

causas mecánicas era ligeramente ascendente hasta el año 2000, y claramente descendente hasta 2004. El cambio en la tendencia coincidía con la puesta en marcha de estos planes en las comunidades autónomas, pero también se observó esta diferencia en comunidades que aún no los habían implementado. Por tanto, estos resultados no permitían atribuir exclusivamente a los planes de actuación preferente el descenso generalizado en la incidencia de lesiones por accidentes de trabajo no mortales a partir del año 2000 en España.

El análisis estadístico de series temporales múltiples puede abordarse con diferentes enfoques en función, entre otras cosas, del número de series temporales, de la cantidad de observaciones que se obtengan y del momento en que se realizan las intervenciones[4]. El enfoque más sencillo consiste en analizar cada una de las series temporales por separado y comparar si hay efecto de la intervención en cada una de ellas, o si el impacto es mayor en una serie u otra. Un enfoque más complejo consiste en introducir la información de varias series temporales en un único modelo. En caso de disponer de un patrón de cambio de nivel con igual tendencia antes y después de la intervención en ambos grupos, el modelo de regresión paramétrico se podría resumir del siguiente modo[5,8]:

$$Y_{it} = \beta_0 + \beta_1 Int_{it} + \beta_2 T_{it} + \beta_3 Grupo_{it} + \varepsilon_{it}$$

donde Y_{it} es la variable resultado en el grupo de intervención (i = 1) y el grupo de comparación (i = 2) para cada medida en el tiempo *t* (t = 1,···,T). Por lo tanto, disponemos de 2 × *T* observaciones en la base de datos. Int_{it} es la variable de intervención. Es importante destacar que esta variable toma valor 0 para todas las observaciones del grupo de comparación (i = 2) y para las observaciones preintervención del grupo de intervención (i = 1), mientras que toma valor 1 para las observaciones postintervención del grupo de intervención (i = 1). T_{it} es la variable que permite controlar la tendencia, y toma valores de 1,···,T tanto para el grupo de comparación como para el grupo de intervención. Finalmente, $Grupo_{it}$ toma valor 1 para las observaciones del grupo de comparación y 0 para las del grupo de intervención. En este modelo, el parámetro β_1 representa el impacto de la intervención. Al igual que en el diseño no experimental, es posible introducir en el modelo otras variables confusoras.

Diseños experimentales

Los diseños experimentales se caracterizan por una asignación aleatoria de los individuos a un grupo de intervención y a un grupo de comparación, que en este tipo de diseños se denomina grupo control. Si el tamaño de muestra es suficiente, ambos grupos serán, por efecto del azar, equivalentes, excepto en el hecho de que uno de ellos recibirá la intervención. Esto nos permitirá atribuir el efecto observado a la intervención, descartando otros posibles factores confusores. Sin embargo, en el ámbito de la evaluación en salud pública, este tipo de diseño es de los menos utilizados, debido tanto a conflictos de tipo ético como a dificultades logísticas o relacionadas con la falta de recursos. Los principales diseños experimentales se pueden clasificar en diseños post y diseños antes-después (o pre-post).

Diseño post

El diseño experimental post se caracteriza por no disponer de ninguna medición preintervención, ya que se asume que ambos grupos son perfectamente equivalentes antes de la intervención y, por lo tanto, los dos estarían sujetos al mismo grado de cambio inducido por factores externos al programa. De este modo, cualquier diferencia entre ellos en el resultado postintervención representaría el efecto de la intervención. La principal ventaja de este estudio es que no son necesarias las medidas preintervención. Sin embargo, la asunción de que en el momento preintervención ambos grupos tendrían el mismo valor para el criterio evaluado no siempre corresponde a la realidad, y esas posibles diferencias podrían afectar al cambio que atribuimos a la intervención. Este diseño es de los menos utilizados en evaluación en salud pública, aunque hay algunos ejemplos, como el estudio de Hilliard[23], que investigó el efecto de la musicoterapia en pacientes terminales asignando aleatoriamente a los pacientes a esta terapia, mientras que otros recibían los tratamientos paliativos tradicionales. El estudio mostró diferencias significativas en algunas variables, como por ejemplo el tiempo que transcurría antes de la muerte del paciente, que fue mayor en los que recibían la musicoterapia.

En cualquier caso, para descartar que las diferencias halladas entre el grupo de intervención y el grupo control no se deben al azar, debe aplicarse un test apropiado de significación estadística, como la prueba *t*, el análisis de la variancia (ANOVA), el análisis de la covarianza (ANCOVA) o tests no paramétricos cuando el supuesto de normalidad no se cumpla[24].

Diseño experimental antes-después (o pre-post)

En el diseño experimental antes-después (o pre-post) disponemos de una o más mediciones antes y después de la intervención. El cambio en la variable resultado en el grupo control representa lo que hubiera pasado a los individuos del grupo de intervención si no hubieran recibido la intervención. Por ello, si calculamos la diferencia del cambio en la variable resultado de ambos grupos obtenemos directamente el efecto del programa. Los diseños experimentales se consideran los menos vulnerables a las posibles amenazas a la validez interna, ya que se asume que el azar distribuye por igual las posibles variables confusoras, incluyendo las no conocidas o no observables. Sin embargo, en salud pública, debido a la complejidad de las intervenciones, este diseño es difícilmente aplicable en la mayoría de los casos. Un ejemplo de aplicación de este diseño es el estudio de evaluación de una intervención dirigida a mejorar la calidad de vida llevado a cabo por Marín et al[25], en el cual se asignó aleatoriamente a los individuos interesados a un grupo de intervención y a un grupo control que recibiría la intervención posteriormente. Se determinó en los individuos de ambos grupos la percepción del estado de salud y ciertos parámetros clínicos antes y después de la intervención. Los resultados del estudio mostraron una mejora del grupo de intervención respecto a los individuos del grupo control.

Además de los tests de comparación mencionados en el diseño post, es posible estimar modelos de regresión, que son aplicables con independencia de la distribución (normal o no) de la variable resultado, y además permiten estimar la magnitud del efecto de la intervención[3]. En este caso, el modelo de regresión sería el mismo que se ha utilizado en los diseños no experimentales, pero sin incluir las posibles variables de confusión.

Para algunos de los diseños expuestos a lo largo del manuscrito se ha especificado el modelo de regresión adecuado para detectar y cuantificar un efecto debido a la intervención. Con el fin de simplificar la notación, se ha incidido sobre todo en qué covariables deben introducirse en el modelo, y se ha considerado siempre que sólo habría un grupo de comparación y que la variable resultado es continua y sigue una distribución normal. Cabe mencionar que, en el caso de que la variable respuesta no fuera continua, deberían especificarse otros tipos de modelos de regresión, como por ejemplo un modelo de regresión logístico (variables dicotómicas) o un modelo de regresión de Poisson (recuentos). También es frecuente utilizar los modelos lineales generalizados (MLG), que permiten considerar, bajo un mismo marco, una gran variedad de distribuciones de la variable resultado.

Consideraciones finales

Aunque este artículo se centra en los diseños evaluativos más utilizados en salud pública y su análisis estadístico, hay muchos otros aspectos metodológicos relevantes a tener en cuenta en el ámbito de la evaluación, entre los que cabe destacar los relacionados con la muestra (técnicas de muestreo o tamaño de muestra) y con la elección de los

Tabla 2
Métodos para controlar los sesgos de selección según la fase de estudio

Método	Descripción	Fase del estudio	
		Diseño	Análisis
Aleatorización	Asignar al azar individuos al grupo de intervención y al grupo control	X	
Restricción	Limitar las características de los individuos incluidos en el estudio.	X	
Emparejamiento	Para cada individuo del grupo intervención, seleccionar uno o más individuos con las mismas características (excepto recibir la intervención) en el grupo de comparación	X	
Estratificación	Analizar los resultados según subgrupos (o estratos) de individuos de características similares		X
Ajuste por regresión	Introducir variables que pudieran ser confusoras en el modelo de regresión		X
Propensity Score	Calcular un índice (*propensity score*) que recoge la probabilidad condicionada que tiene cada individuo de ser asignado al grupo de intervención dadas unas determinadas variables confusoras		X
Análisis de sensibilidad	Estimar el impacto de la intervención bajo diversas condiciones del posible sesgo de selección		X

Adaptado de Fletcher et al[27]

indicadores adecuados. Asimismo, aparte de los diseños comentados existen otros que pueden ser útiles en determinados estudios, como el diseño Solomon (que permite corregir el efecto que la medida preintervención pueda tener en los individuos o grupos), el diseño factorial (que permite evaluar la efectividad de diversos componentes de una misma intervención) o la metaevaluación (que evalúa la efectividad a partir de la evidencia acumulada de evaluaciones previas).

Tal como se ha comentado a lo largo del artículo, el sesgo de selección es una de las principales amenazas a la validez interna. En algunos casos, este sesgo puede minimizarse directamente en el diseño del estudio mediante la aleatorización en la asignación de los individuos al grupo de intervención y al grupo control, restringiendo el estudio a sujetos de unas determinadas características o seleccionando para el grupo comparación personas que sean similares a cada una del grupo de intervención. Cuando este sesgo no se ha podido controlar en la fase de diseño, es posible minimizarlo durante la fase de análisis de los datos. Para ello, entre otras opciones, podemos estratificar los resultados, ajustar por variables confusoras o por un índice como el PS, o realizar un análisis de sensibilidad en el cual se recojan diversos escenarios alternativos (tabla 2).

Sin duda, la evaluación de intervenciones de salud pública, en la que rara vez es posible la aleatorización de individuos y habitualmente intervienen múltiples factores, implica numerosos retos metodológicos. Aun así, es absolutamente necesario intentar evaluar la efectividad de dichas intervenciones, haciendo uso de los recursos metodológicos disponibles que, en la mayoría de los casos, permitirán concluir si una intervención es efectiva y estimar en qué magnitud. Esta información debería ser una de las bases fundamentales en la inversión eficiente de recursos en intervenciones de salud pública, que se traduciría en una mejora de la calidad de vida y de la salud de la población.

Contribuciones de autoría

Mª José López y Marc Marí-Dell'Olmo lideraron la concepción del artículo y escribieron su primera versión. Todos los autores participaron en la escritura y revisión científica del manuscrito, y aprobaron la versión final para su publicación.

Financiación

Este artículo ha sido elaborado con el apoyo del Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya (AGAUR SGR 2009 1345).

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Rossi PH, Lipsey MW, Freeman HE. Evaluation: a systematic approach. 7th ed. Thousands Oaks (CA): Sage Publi; 2004.
2. Posavac E, Carey R. Program evaluation: methods and case studies. 7th ed. New Jersey: Pearson Education; 1992.
3. Nebot M, López MJ, Ariza C, et al. Impact of the Spanish smoking law on exposure to secondhand smoke in offices and hospitality venues: before-and-after study. Environ Health Perspect. 2009;117:344-7.
4. Arnau Gras J. Diseños temporales: técnicas de análisis. Barcelona: Edicions de la Universitat de Barcelona, 2001.
5. Langbein LI, Felbinger CL. Public program evaluation: a statisticall guide. New York: M.E. Sharpe; 2006.
6. Novoa AM, Pérez K, Santamarina-Rubio E, et al. Road safety in the political agenda: the impact on road traffic injuries. J Epidemiol Community Health. 2010; en prensa.
7. Tobías A, Sáez M, Galán I. [Graphic tools for the descriptive analysis of temporary series in medical research]. Med Clin (Barc). 2004;122:701-6.
8. Simonton DK. Cross-sectional time-series experiments: some suggested statistical analyses. Psychol Bull. 1977;84:489-502.
9. Stolwijk AM, Straatman H, Zielhuis GA. Studying seasonality by using sine and cosine functions in regression analysis. J Epidemiol Community Health. 1999;53:235-8.
10. Robitaille Y, Laforest S, Fournier M, et al. Moving forward in fall prevention: an intervention to improve balance among older adults in real-world settings. Am J Public Health. 2005;95:2049-56.
11. Luellen JK, Shadish WR, Clark MH. Propensity scores: an introduction and experimental test. Eval Rev. 2005;29:530-58.
12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41-55.
13. Expósito Ruiz M, Ruiz Bailén M, Pérez Vicente S, et al. Uso de la metodología propensity score en la investigación sanitaria. Rev Clin Esp. 2008;208:358-60.
14. Lauby JL, Smith PJ, Stark M, et al. A community-level HIV prevention intervention for inner-city women: results of the women and infants demonstration projects. Am J Public Health. 2000;90:216-22.
15. Braun TM. A mixed model-based variance estimator for marginal model analyses of cluster randomized trials. Biom J. 2007;49:394-405.
16. Shadish WR, Cook TD. The renaissance of field experimentation in evaluating interventions. Annu Rev Psychol. 2009;60:607-29.
17. Faggiano F, Vigna-Taglianti F, Burkhart G, et al. The effectiveness of a school-based substance abuse prevention program: 18-month follow-up of the EU-Dap cluster randomized controlled trial. Drug Alcohol Depend. 2010;108:56-64.
18. Jacob BA, Lefgren L. Remedial education and student achievement: a regression-discontinuity analysis. Review of Economics and Statistics. 2004; 86:226-44.
19. Linden A, Adams JL, Roberts N. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. J Eval Clin Pract. 2006;12:124-31.
20. Shadish WR, Cook TD. Experimental and quasi-experimental designs for generalized causal inference. Boston, New York: Houghton Mifflin Company; 2002.
21. Khandker SR, Koolwal GB, Samad H. Handbook on impact evaluation: quantitative methods and practices. Washington: World Bank Publications; 2009.
22. Benavides FG, Rodrigo F, García AM, et al. Evaluation of the effectiveness of preventive activities (Strategic Action Plans) on the incidence of non-fatal traumatic occupational injuries leading to disabilities in Spain (1994-2004). Rev Esp Salud Publica. 2007;81:615-24.
23. Hilliard RE. A post-hoc analysis of music therapy services for residents in nursing homes receiving hospice care. J Music Ther. 2004;41:266-81.
24. Dimitrov DM, Rumrill PD. Pretest-posttest designs and measurements of change. Work. 2003;20:159-65.
25. Marin GH, Homar C, Niedfeld G, et al. Evaluation of the state intervention project to improve quality of life and reduce the complications associated with aging: "Add health to your years". Gac Sanit. 2009;23:272-7.
26. Pérez K, Mari-Dell'olmo M, Tobías A, et al. Reducing road traffic injuries: effectiveness of speed cameras in an urban setting. Am J Public Health. 2007;97:1632-7.
27. Fletcher HR, Fletcher WS, Wagner HE. Epidemiología clínica: aspectos fundamentales. Madrid: Masson–Williams & Wilkins; 1989.