Gac Sanit. 2011;**25(Supl 1)**:3-8

Evaluación de la efectividad en salud pública: fundamentos conceptuales y metodológicos

Manel Nebot[a,b,c,d,*], Mª José López[a,b,d], Carles Ariza[a,b,d], Joan R. Villalbí[a,b,c,d] y Anna García-Altés[b,d,e]

[a]*Agencia de Salud Pública de Barcelona, Barcelona, España*
[b]*CIBER de Epidemiología y Salud Pública (CIBERESP), España*
[c]*Departamento de Ciencias Experimentales y de la Salud, Universitat Pompeu Fabra, Barcelona, España*
[d]*Institut d'Investigacions Biomèdiques Sant Pau (IIB Sant Pau), Barcelona, España*
[e]*Agència d'Informació, Avaluació i Qualitat en Salut (AIAQS), Barcelona, España*

Palabras clave:
Evaluación
Efectividad
Salud pública

RESUMEN

En los últimos años ha aumentado de forma notable el interés por la evaluación de las intervenciones en salud, especialmente en relación a su utilidad social y su eficiencia económica. Sin embargo, todavía estamos lejos de tener un grado suficiente de consenso en los aspectos básicos de la evaluación, como son la terminología, la finalidad y la metodología de trabajo. En este marco se revisan las principales definiciones y clasificaciones de la evaluación aplicada a los programas y políticas en salud pública. En relación a la evaluación de resultados, se presentan los principales diseños evaluativos y sus componentes, y se revisan las amenazas a la validez interna de los resultados de los diseños evaluativos débiles. Se analizan y discuten las características de las intervenciones de salud pública que limitan las opciones de evaluación con diseños tradicionales. Entre estas limitaciones destacan la complejidad de las intervenciones, que habitualmente tienen múltiples componentes, y la dificultad de establecer un grupo de comparación sin intervención, en especial mediante asignación aleatoria. Para finalizar, se describe una propuesta de evaluación a partir de diseños evaluativos débiles, consistente en la valoración de la adecuación y la plausibilidad. La adecuación estaría determinada por la existencia de un cambio observable en los indicadores de resultados, y podría ser suficiente para tomar decisiones bajo determinadas condiciones; otras veces sería necesario analizar la plausibilidad, o atribución de los resultados observados al programa.

Effectiveness assessment in public health: conceptual and methodological foundations

Keywords:
Evaluation
Effectiveness
Public health

ABSTRACT

In the last few years, interest has markedly increased in evaluating health programs, especially their social utility and economic efficiency. However, consensus on key issues in evaluation, such as terminology, goals and methods is still a long way off. In this context, we review the main definitions and classifications of evaluation applied to public health programs and policies. We describe the main evaluation designs and their components, focusing on outcome evaluation. Threats to the internal validity of the results of weak evaluation designs are also discussed. The characteristics of public health interventions that limit evaluation with traditional designs are also analyzed. These limitations include the complexity of interventions, usually with multiple components, and the difficulty of forming an equivalent control group with no intervention, especially through random assignment. Finally, a two-step approach to evaluation through weak designs, which takes into account adequacy and plausibility, is described. Adequacy consists of the observation of a change in the selected indicators after the intervention, and would be sufficient to take decisions under certain conditions; at other times, plausibility would need to be analyzed, defined as attribution of the results to the program or intervention.

*Autor para correspondencia.
Correo electrónico: mnebot@aspb.cat (M. Nebot)

0213-911/$ - see front matter

Introducción

Aunque la evaluación de los procedimientos y de los resultados siempre ha formado parte de las intervenciones sociales, es en la segunda mitad del siglo xx cuando el interés por la evaluación se hace explícito y da lugar a una disciplina específica, con una metodología de trabajo propia[1]. A finales de los años 1950, los programas de salud promovidos por las grandes agencias de salud y desarrollo (especialmente los programas de planificación familiar y nutrición) incorporan de manera formal indicadores de evaluación. Progresivamente, el desarrollo conceptual y metodológico de las ciencias sociales, y la utilización cada vez más frecuente de las encuestas y de nuevas técnicas cuantitativas de análisis, configuran una cierta especialización de la evaluación en el marco de las ciencias y los programas sociales. En la década de 1960 se produce un crecimiento espectacular de los estudios evaluativos y aparecen las primeras publicaciones dedicadas específicamente a la metodología de la evaluación. Es en esta época cuando se plantea por primera vez utilizar diseños experimentales y cuasiexperimentales en la evaluación de intervenciones no farmacológicas[2]. Desde entonces, el interés por la evaluación se ha generalizado entre los financiadores y los planificadores, los clientes y los usuarios de los programas, con la utilidad social como principal criterio de valoración. La aparición de la medicina basada en la evidencia, hace dos décadas, ha reforzado la importancia de demostrar la utilidad de las intervenciones de salud en términos de eficacia y efectividad[3]. Por otro lado, a finales de los años 1960 aparecieron los primeros estudios de evaluación económica de los programas de salud, que cobraron fuerza a mediados de la década de 1980, cuando se publicaron los primeros estudios que utilizaban indicadores de impacto, como los años de vida ajustados por calidad[4]. Cabe señalar también que algunos autores han criticado los enfoques centrados en resultados, señalando las limitaciones del modelo experimental y cuasiexperimental, y en especial las dificultades de controlar y reproducir los factores contextuales, que son determinantes en intervenciones sociales y de salud pública[5].

Globalmente, y a pesar del enorme desarrollo conceptual y metodológico de la evaluación (o quizás por ello), hay una notable confusión terminológica y conceptual, sobre todo a la hora de definir los criterios de efectividad[6]. En este artículo se revisan las bases conceptuales y metodológicas de la evaluación de la efectividad de las intervenciones de salud pública.

Conceptos y tipos de evaluación

Los términos utilizados con más frecuencia en el ámbito de la evaluación de intervenciones en salud pública, adaptados de los diccionarios de epidemiología[7] y salud pública[8], así como del manual *Evaluation: a systematic approach* de Rossi[1], se presentan en la tabla 1. Con respecto a la propia definición de evaluación en ciencias de la salud hay muchas propuestas, de las cuales una de las más apropiadas es probablemente la de Suchman[9], que considera la evaluación como «el juicio sobre el valor o utilidad de una intervención». Esta definición asume de forma implícita que lo que se juzga son los resultados de la intervención. En una visión más global, el *Diccionario de salud pública*[8] define la evaluación como «los esfuerzos dirigidos a determinar de forma sistemática y objetiva la efectividad y el impacto de las actividades realizadas para alcanzar objetivos de salud, teniendo en cuenta los recursos asignados». En esta definición se asume que hay diversas formas de abordar la evaluación, aunque ninguna sea por sí sola completamente satisfactoria.

Los principales tipos de evaluación, según el nivel, la finalidad y la perspectiva, se resumen en la tabla 2. Según el nivel en que la evaluación tiene lugar suele distinguirse entre evaluación táctica, que incluye la evaluación de la estructura, el proceso y los resultados, y evaluación estratégica (a veces también llamada evaluación de la pertinencia), que consiste en la valoración de los objetivos del programa o política de salud. La evaluación estratégica intenta responder a preguntas como: ¿la intervención o programa de salud se corresponde con las necesidades y con las prioridades? ¿responde a problemas relevantes desde el punto de vista de la sociedad y está planteado en la dirección

Tabla 1
Glosario de términos utilizados habitualmente en la evaluación de la efectividad en salud pública[1,7,8]

Accesibilidad	Grado en que se facilita la participación en el programa en su conjunto o en alguna de sus actividades o recursos
Cumplimiento (dosis, exhaustividad)	Medida de la cantidad de la intervención que ha sido aplicada; normalmente se expresa como el porcentaje del total del contenido previsto que ha sido efectivamente implementado
Eficacia	Grado en que una intervención produce un resultado beneficioso en los receptores del programa
Efectividad	Grado en que una intervención produce resultados beneficiosos en el conjunto de la población diana
Eficiencia	Efectos o resultados de una intervención en relación a los recursos empleados
Evaluación de impacto	Estudio evaluativo en el cual se valoran globalmente los resultados directos del programa, así como el impacto en las condiciones sociales en que el programa puede influir a largo plazo
Evaluación de impacto en salud	Análisis de las consecuencias y posibles implicaciones para la salud pública de iniciativas o procesos sociales o ambientales que no han sido diseñados primariamente como intervenciones de salud
Evaluación de proceso	Evaluación diseñada para determinar si el programa se administra de la forma planeada a la población diana
Evaluación de resultados	Evaluación diseñada para determinar si el programa ha alcanzado los objetivos previstos
Evaluación formativa	Evaluación que se realiza durante la fase de desarrollo de una intervención orientada a obtener información sobre el proceso y los mecanismos de acción con la finalidad de mejorarla y de explorar su factibilidad
Factibilidad	Viabilidad práctica de un estudio, programa o intervención
Fidelidad	Medida del grado en que los programas son aplicados de acuerdo al protocolo
Grupo de comparación	En diseños cuasiexperimentales, los individuos del grupo de comparación son los que no reciben la intervención que se administra a los individuos del grupo de intervención, en quienes se compararán los efectos observados
Grupo control	En un ensayo aleatorizado o comunitario, los individuos del grupo control son los que no reciben la intervención que se administra a los individuos del grupo de intervención, en quienes se compararán los efectos observados
Monitorización (del programa)	Documentación sistemática de aspectos de la ejecución del programa que permiten valorar si está siendo aplicado de la forma planificada o bien si cumple unos parámetros estándar determinados. La monitorización puede ser de proceso o de resultados
Población diana (población objetivo)	Conjunto de individuos o grupos (familias, comunidades, etc.) a los que se dirige el programa

Elaboración propia a partir de: Rossi et al[1], Last et al[7] y Porta[8].

Tabla 2
Principales enfoques y tipos de evaluación en salud pública[12,13,16]

Evaluación según el nivel	
De estructura	Adecuación de los recursos a las necesidades
De proceso	Adecuación de las actividades y de los servicios a los objetivos y al protocolo
De resultados	Consecución de los objetivos del programa
Estratégica	Evaluación de los objetivos (¿son pertinentes?)
Evaluación según la finalidad	
Formativa	Evaluación que se realiza en la fase de desarrollo de un programa (prueba piloto) para explorar su factibilidad y mejorarlo
Sumativa (de impacto)	Valoración de la eficacia o efectividad de un programa consolidado
Evaluación según la perspectiva	
De desarrollo	Análisis de la ejecución de las actividades y de los servicios
De gestión	Evaluación orientada a conocer y mejorar los programas y sus efectos
Experimental	Valoración de los resultados del programa en condiciones controladas
Económica	Estudio de la relación entre los costes y los resultados de la intervención

Elaboración propia a partir de: Overtveigt[12], Pineault[13] y Windsor[16].

adecuada? Idealmente, la evaluación estratégica tiene lugar en la fase previa a la implementación del programa, lo que se conoce como «pertinencia teórica». Sin embargo, no es infrecuente que se plantee con posterioridad, valorando los resultados alcanzados en el contexto de los problemas de salud y las prioridades.

Según la finalidad puede distinguirse entre evaluación formativa y evaluación sumativa o de impacto. Aunque en muchas ocasiones la evaluación formativa se confunde con la evaluación de proceso, la primera sería la que se realiza durante el período de desarrollo de una intervención, con la finalidad de detectar precozmente problemas o insuficiencias que puedan ser mejorados. La evaluación formativa es especialmente importante en los programas nuevos o adaptados de otros contextos, y finaliza cuando se considera que los contenidos y procedimientos del programa son estables y definitivos; sólo entonces puede procederse a la evaluación de los resultados y del impacto, lo que se conoce como evaluación sumativa[10]. La evaluación formativa debe diferenciarse de la evaluación de proceso, que se realiza para monitorizar la cobertura y la calidad de las intervenciones, y tiene especial relevancia en las intervenciones complejas y cuando los resultados en salud son a largo plazo[11].

También pueden distinguirse cuatro tipos de evaluación según la perspectiva[12]. Así, la evaluación en la perspectiva de desarrollo intenta ayudar a los profesionales y responsables a desarrollar y mejorar los tratamientos, los servicios y las políticas. La evaluación en la perspectiva de gestión está orientada a monitorizar y mejorar la ejecución de los programas y servicios, y a verificar que las actividades se han realizado de acuerdo al protocolo establecido. La evaluación experimental es la que trata de averiguar si una intervención ha tenido efecto, y las causas de este efecto. En consecuencia, la evaluación está diseñada para verificar o refutar una hipótesis, y el diseño evaluativo deber ser adecuado a este propósito. Finalmente, la evaluación económica se preocupa por los costes y beneficios económicos de la intervención.

Los distintos niveles de evaluación (estructura, proceso y resultados) de la evaluación táctica[13] y su relación con la evaluación económica se describen en la figura 1. La evaluación de la estructura se corresponde con la evaluación de los recursos, materiales y humanos. Tradicionalmente este tipo de evaluación ha tenido una importancia destacada en los servicios sanitarios. La evaluación del proceso, como ya se ha señalado, se corresponde con la valoración de los servicios y las actividades que se generan a partir de los recursos, mientras que la evaluación de los resultados consiste en verificar si estos servicios y actividades han permitido alcanzar los objetivos establecidos. La eficiencia sería la medida de la relación entre los recursos empleados y la consecución de los objetivos de las intervenciones, en el marco de la evaluación económica, que analiza los costes y los beneficios para optimizar la asignación de recursos, y resulta fundamental en un ámbito en el cual la demanda de servicios y el coste de éstos crecen a una velocidad muy superior a la de los recursos[14].

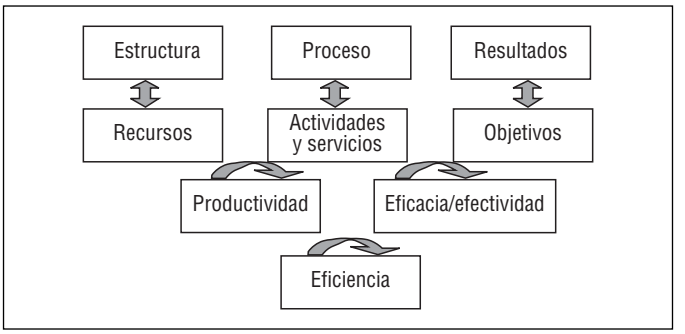

Figura 1. Componentes de la evaluación de los programas de salud y sus relaciones[13].

Evaluación de proceso y evaluación de resultados

La evaluación de proceso tiene como finalidad básica conocer la cobertura y la calidad de las intervenciones. En esencia, se trata de verificar si el programa ha alcanzado la población diana y los subgrupos relevantes, y si se han realizado las actividades previstas con la calidad necesaria. Es importante distinguir la evaluación de proceso, que puede ser una iniciativa o actividad puntual, de la monitorización, que implica disponer de forma estable de indicadores de proceso válidos. La monitorización mediante indicadores de proceso está indicada cuando los resultados de la intervención sólo pueden observarse a largo plazo, el proceso está suficientemente estandarizado y hay un buen conocimiento del modelo causal entre proceso y resultados. Por ejemplo, puede monitorizarse un programa de detección precoz del cáncer mediante los indicadores relativos a la cobertura y la detección de casos cuando la reducción de la mortalidad está sólidamente establecida por estudios específicos.

Los indicadores principales de la evaluación del proceso son la cobertura del programa, en el conjunto de la población diana y en los diversos subgrupos, y la calidad de la intervención, medidos por diversos indicadores (exhaustividad, fidelidad, satisfacción), que se comparan con estándares previos o con parámetros establecidos ad hoc por el protocolo del programa. La evaluación del proceso es fun-

Tabla 3
Diseños evaluativos básicos en evaluación de resultados[16]

Tipo	Definición	Ejemplo
Experimental	Diseño evaluativo en el cual la asignación de los individuos al grupo de intervención o al grupo de control es aleatoria (grupos equivalentes)	Ensayo clínico sobre educación sanitaria en pacientes hipertensos
Cuasiexperimental	Diseño evaluativo en el cual la asignación de individuos a los grupos de intervención y de comparación no es aleatoria	Comparación de los índices de salud dental en un condado donde se añade flúor al agua de consumo y en un condado similar sin fluorización
No experimental	Diseño evaluativo en el cual la medición del efecto se realiza únicamente en el grupo de intervención (sin grupo de comparación)	Evaluación de la concentración de humo ambiental de tabaco en locales cerrados antes y después de la ley de tabaquismo
Observacional	Evaluación del efecto mediante un estudio epidemiológico observacional (estudio de casos y controles o estudio de cohortes)	Evaluación de la efectividad de la vacuna BCG contra la tuberculosis mediante estudios de casos y controles

Elaboración propia a partir de: Windsor[16].

damental en los programas de promoción de la salud, especialmente si son complejos o innovadores. Sólo si un programa o intervención alcanza la población diana y es desarrollado conforme al protocolo establecido puede esperarse que sea efectivo, aunque una evaluación satisfactoria del proceso no garantiza por sí sola alcanzar los resultados esperados.

La evaluación de los resultados consiste esencialmente en verificar si se han alcanzado los objetivos establecidos. En salud pública es habitual que los resultados directos del programa (p. ej., obtener mayores tasas de cobertura vacunal o reducir la tasa de fumadores) no se traduzcan de forma inmediata en una reducción de la mortalidad o de la morbilidad, por lo que suele distinguirse entre los resultados directos o inmediatos y los resultados a largo plazo[15,16]. Con relación a los resultados, también suele distinguirse entre eficacia y efectividad, en función de si éstos se miden en los receptores del programa o de la intervención (eficacia) o en el conjunto de la población diana (efectividad).

Para evaluar los resultados se utilizan diversos diseños evaluativos, que se resumen en la tabla 3. Windsor et al[16] proponen utilizar tres categorías: diseños experimentales, diseños cuasiexperimentales y diseños no experimentales. A éstos pueden añadirse los estudios observacionales[17], que pueden utilizarse para estimar la efectividad aunque no se haya controlado la asignación a la intervención. En todos los diseños evaluativos hay al menos una intervención (X) y una medición de resultados realizada después de la intervención (O2) o medida postest. Los distintos diseños difieren en la presencia o ausencia de distintas características o atributos, básicamente el grupo de control y las distintas mediciones del indicador de resultados, incluyendo la medición anterior a la intervención, o medida pretest (fig. 2). La presencia de estas características será determinante para atribuir con mayor o menor seguridad los efectos observados a la intervención, lo que se conoce como validez interna de los resultados. Además, a estos atributos habría que añadirles el marco social, económico, cultural, político o normativo, que condiciona e influye en el desarrollo, la implementación y la efectividad de las intervenciones[5].

Como se ha señalado, junto a los estudios diseñados específicamente con una finalidad evaluativa, los estudios epidemiológicos observacionales también pueden utilizarse para evaluar la efectividad de las intervenciones en salud pública. Así, se han usado estudios de casos y controles para evaluar la efectividad de la vacunación con BCG para prevenir la tuberculosis[18], o para estimar la efectividad del cribado del cáncer cervical[19]. Este enfoque es especialmente útil para estudiar la protección por la vacuna en caso de brotes agudos de enfermedades vacunables, como el sarampión, ya que los factores de confusión son más fáciles de controlar y no hay que esperar largos períodos para realizar el estudio[20].

Aunque hay diseños más complejos que permiten analizar por separado otros componentes de la intervención (p. ej., puede estudiarse el efecto del test de forma separada del efecto de la intervención en el diseño conocido como «diseño Salomón»[16,21]), los diseños evaluativos más habituales en salud pública son los cuasiexperimentales, incluyendo las series temporales y el diseño antes-después.

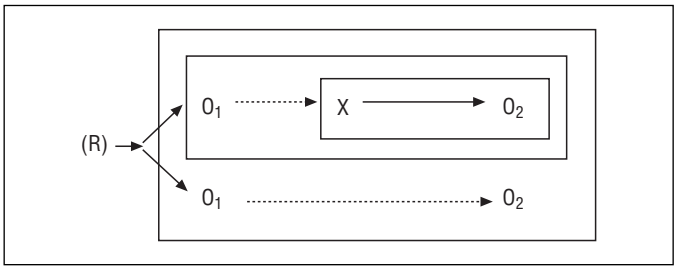

Figura 2. Elementos básicos del diseño evaluativo. R: aleatorización; X: intervención; O_1: observación previa a la intervención (pretest); O_2: observación posterior a la intervención (postest).

Validez interna de los diseños evaluativos

Una característica fundamental de los diseños evaluativos es su validez interna, definida como la capacidad del estudio para descartar posibles sesgos que puedan enmascarar explicaciones alternativas (por lo tanto, no debidas a la intervención) a los resultados observados[21,22]. La máxima validez interna es la del ensayo clínico aleatorizado, en el cual podemos tener la certeza de haber obtenido dos grupos iguales que sólo se diferencien (salvo por efecto del azar) en que uno de ellos ha sido sometido a la intervención. En la medida en que los grupos de control no sean equivalentes, o cuando no haya grupo de control, deberemos ser capaces de descartar estas explicaciones alternativas, que se conocen habitualmente como «amenazas a la validez interna», descritas por Cook y Campbell[21] y que se resumen en la tabla 4. El efecto de la *historia* es el sesgo que se produce cuando hay una tendencia previa en el mismo sentido de los efectos esperados, o bien cuando los efectos observados se deben a circunstancias o factores externos al programa, incluyendo otras intervenciones de salud. Por ejemplo, una campaña mediática de ámbito nacional alertando sobre los riesgos de conducir sin cinturón de seguridad puede explicar en parte el efecto observado tras un programa local con objetivos similares. Se trata de un sesgo que raramente tiene lugar en la investigación básica, en la cual es fácil controlar todos los factores externos, mientras que, por el contrario, es un problema habitual en las intervenciones de salud pública. El sesgo de *selección* se produce cuando el grupo de comparación no es equivalente, ya que sólo podremos controlar en el análisis los factores de confusión conocidos, observables y medidos, y en muchos casos no será suficiente para concluir que los resultados se deben a la intervención. Por ejemplo, si realizamos un programa de educación sanitaria para diabéticos en pacientes de atención primaria y comparamos los resultados con pacientes atendidos en el hospital; o si comparamos pacientes en un

Tabla 4
Amenazas a la validez interna de las intervenciones[21]

Historia	El efecto observado se debe a un factor externo a la intervención, que actúa durante o antes de la intervención, o bien forma parte de una tendencia anterior a la intervención
Selección	El efecto observado se debe a las diferencias entre los participantes en el grupo de intervención y el grupo de comparación
Maduración	El efecto observado se debe a que los participantes adquieren experiencia y conocimiento en el período entre el pretest y el postest
Efecto del test	El efecto observado se debe al aprendizaje que tiene lugar como consecuencia de la realización del pretest
Instrumentación	El efecto observado se debe al cambio en el instrumento o método de medición
Regresión a la media	El efecto observado se debe a una variación espontánea en los valores medios del parámetro o medida de efecto cuando la asignación de los participantes se asocia a valores extremos en la variable y ésta tiene variación aleatoria
Pérdidas	El efecto observado se debe a que las pérdidas en el seguimiento se distribuyen de forma desigual en los grupos de intervención y comparación

Elaboración propia a partir de: Cook y Campbell[21].

área urbana con otros pertenecientes a un área rural. La *maduración* es el sesgo que se produce cuando atribuimos a una intervención todo o parte del cambio observado, que en realidad se debe al proceso intrínseco de aprendizaje que se produce normalmente en los individuos; es un sesgo que puede afectar de forma especial a las intervenciones que se realizan en niños y adolescentes, o bien cuando transcurre mucho tiempo entre el pretest y el postest. El *efecto del test* sobre los resultados puede producirse tanto por la reiteración (los participantes acaban por "aprenderse" el test y sus respuestas) como por el efecto desencadenante de cambios de actitudes y de curiosidad sobre un tema concreto, que puede conllevar la búsqueda activa de información. En general, el efecto es más importante en el caso de temas poco conocidos o bien cuando son temas especialmente sensibles para los participantes, como sucede por ejemplo en investigaciones sobre el suicidio o sobre las prácticas sexuales. La *instrumentación* es el sesgo que se produce cuando el cambio aparente se debe en realidad a cambios en el instrumento de medida, ya sea en cambios instrumentales o de procedimiento (p. ej., en una técnica analítica o en el procedimiento para medir la presión arterial) o en definiciones operativas, como la definición de tabaquismo habitual. La *regresión a la media* es una distorsión en la medida del efecto que puede producirse cuando se seleccionan los participantes en función de alguna variable que tenga un componente aleatorio de variación. Por ejemplo, si se eligen los individuos con la presión arterial más alta para realizar una intervención de relajación con el fin de disminuir la presión arterial, hay que tener en cuenta que una parte de los individuos que hayan sido declarados hipertensos en la primera medida habrán normalizado espontáneamente sus valores de tensión arterial en medidas posteriores sin realizar ninguna intervención. Si no se tiene en cuenta este sesgo, puede atribuirse erróneamente el cambio a la intervención. Finalmente, las *pérdidas* pueden introducir un sesgo en la interpretación del efecto de una intervención si no se distribuyen por igual en los grupos de intervención y de comparación, ya que pueden producir grupos no comparables en el postest.

Limitaciones metodológicas en la evaluación en salud pública

Como se ha señalado, en muchas ocasiones la evaluación de intervenciones en salud pública depende de *diseños evaluativos débiles* (cuasiexperimental, no experimental u observacional) porque no es posible utilizar diseños experimentales como el ensayo clínico controlado[23,24]. Como consecuencia, puede resultar difícil descartar las amenazas a la validez interna de la evaluación, lo cual puede arrojar dudas sobre los resultados obtenidos y, por consiguiente, también sobre la utilidad social de la intervención. Junto a las limitaciones en la validez interna debidas a la falta de un grupo de comparación equivalente, también hay otras limitaciones tales como la complejidad de las intervenciones de salud pública y las limitaciones de los propios indicadores de medida de los efectos.

La falta de grupo de comparación equivalente puede deberse a diversas razones; a veces se trata de cuestiones éticas, como por ejemplo en las intervenciones sociales, en las cuales no es posible retrasar los beneficios de una intervención socialmente necesaria con el único propósito de la investigación[25]. En otros casos se debe a razones operativas, como ocurre en las intervenciones comunitarias, basadas en intervenciones normativas o sobre el ambiente, que se desarrollan sobre una comunidad natural con una base geográfica definida en la cual toda la población está potencialmente expuesta. Aunque puedan utilizarse comunidades vecinas como grupo de comparación, el riesgo de difusión de los efectos, especialmente en intervenciones basadas en ideas, es muy importante. Así, en el caso del programa comunitario de reducción del riesgo cardiovascular puesto en marcha en Karelia del Norte (Finlandia) en 1972, se observó una marcada disminución de la mortalidad y de sus principales factores de riesgo (presión arterial, tabaquismo, cifras de colesterol)[26,27], pero los resultados no pudieron atribuirse de forma concluyente al programa, debido a que en las áreas elegidas como grupo de comparación (inicialmente la vecina provincia de Kuopio) se observó también un descenso notable de la morbimortalidad y de los factores de riesgo. Aunque el programa puede considerarse un ejemplo de éxito de las intervenciones comunitarias de modificación del riesgo poblacional, la más que probable difusión a las provincias vecinas de los mensajes principales de la intervención, sin duda beneficiosos para la salud de la población, constituye un serio problema para la validez interna de la evaluación.

En otras ocasiones el obstáculo principal a la evaluación deriva de la complejidad de la intervención. Las intervenciones que por su naturaleza tienen múltiples componentes son difíciles de estandarizar y a veces es difícil atribuir los resultados a un componente particular. Esto sucede habitualmente con las intervenciones comunitarias, y de forma especial en aquellas dirigidas a promover cambios de conducta, ya que es habitual tratar de influir simultáneamente en varios de los determinantes de la conducta. En estos programas suele ser imposible enmascarar la asignación y seguir un protocolo muy cerrado, porque algunos de los componentes pueden tener que adaptarse a las circunstancias del contexto[28]. Muchas intervenciones aparentemente sencillas, como la fisioterapia en las lesiones de rodilla, implican una diversidad de elementos no siempre explícitos que con frecuencia interactúan entre ellos. En este ejemplo, incluso en el caso de que se haya estandarizado apropiadamente el tratamiento directo (una serie de ejercicios definidos en contenido, número y duración), hay otros aspectos que van a influir en los resultados aunque no estén normalmente protocolizados, como la actitud del terapeuta, su interacción emocional con el paciente, la información sobre los cuidados a realizar en el hogar y la actitud preventiva para evitar recaídas o nuevas lesiones, entre otros[29].

Finalmente, además de los factores descritos, las limitaciones de los indicadores de medida de los resultados pueden dificultar la evaluación de las intervenciones de promoción de la salud[22]. En muchas ocasiones, la principal dificultad en la evaluación de resultados se debe a que los cambios pueden ser difíciles de observar, bien porque sea complicado identificar los indicadores de resultado apropiados

(p. ej., si queremos promover una mayor actividad física) o porque los cambios tardan mucho en observarse, como sucede en los programas educativos para la prevención de la infección por el virus de la inmunodeficiencia humana en los adolescentes. Otras veces podemos estar ante una cadena causal compleja y esperar resultados múltiples, como por ejemplo en los programas multicomponente dirigidos a la prevención de la cardiopatía isquémica, que pueden afectar a diversos factores de riesgo y sus determinantes. En estas situaciones, es habitual que se utilicen los resultados a corto plazo por razones de visibilidad y responsabilidad social, aunque como consecuencia puede ocurrir que la evaluación definitiva basada en resultados a largo plazo quede postergada u olvidada.

Evidencias de efectividad: adecuación y plausibilidad

En el marco descrito, a menudo las evidencias de efectividad deberán basarse en la contribución de los *diseños evaluativos débiles*. Victora et al[23] y Habicht et al[24] proponen un razonamiento de efectividad en dos etapas, que describen como adecuación y plausibilidad, cuando no se puedan utilizar diseños experimentales. En esencia, estos autores señalan que en la evaluación de la efectividad de las intervenciones valoraremos los indicadores de cambio en función de la certeza con que podamos relacionarlos con la intervención. El primer paso (adecuación) consiste en valorar si se han alcanzado los objetivos previstos. Si se observan los cambios esperados, se considera que se cumple el criterio de adecuación (p. ej. cobertura vacunal del programa, reducción de la morbimortalidad), incluso en ausencia de unos objetivos formulados explícitamente. La evaluación de la adecuación no requiere, por lo tanto, un grupo de control ni un diseño evaluativo concreto, sino únicamente demostrar que se han producido los cambios previstos o deseables, ya sea entre los receptores del servicio o programa o en el conjunto de la población. Aunque no pueda establecerse una relación causal entre los cambios observados y el programa, en muchos casos la evaluación de la adecuación puede ser suficiente para tomar decisiones. El caso más claro es cuando no se observan los cambios esperados, situación que habitualmente nos permitirá concluir que el programa no ha funcionado. Si por el contrario se han observado cambios favorables, en determinadas condiciones pueden atribuirse al programa, y por tanto los responsables de la provisión de servicios pueden tomar la decisión de seguir apoyando el programa o generalizarlo, aplicando el principio de prevención[30] cuando no se disponga de otro tipo de evidencias y se den una serie de condiciones; entre éstas, la magnitud del cambio observado debe ser tan grande que haga muy improbables las explicaciones alternativas o amenazas a la validez interna. También es más plausible atribuir el cambio a la intervención si la cadena causal entre ella y sus efectos es simple (no hay muchas explicaciones alternativas posibles) y corta (los efectos son visibles poco tiempo después de la intervención). Un ejemplo podría ser la observación de un cambio importante en la siniestralidad por accidentes de tráfico inmediatamente después de la aplicación de normas de control o leyes regulatorias. Estas condiciones no se cumplen en muchas intervenciones en salud pública, por lo que será necesario adoptar una actitud más prudente cuando haya dudas razonables acerca de la posibilidad de explicaciones alternativas, cuando deban tomarse decisiones sobre programas a gran escala o especialmente costosos, o cuando se trate de un programa innovador y no se pueda descartar la existencia de potenciales confusores por el desconocimiento de los mecanismos de actuación. En todos estos casos deberemos combinar los criterios de adecuación con los de plausibilidad, lo que implica comparar los resultados con un grupo de comparación, para controlar las principales amenazas a la validez interna, utilizando el diseño evaluativo más apropiado a la situación. En definitiva, se trata de alcanzar un compromiso entre las exigencias de rigor requeridas a toda intervención de salud pública y las condiciones que puedan impedir la evaluación en condiciones experimentales. En todos los casos debemos asegurarnos de que los recursos se emplean de forma eficiente y de que las intervenciones son útiles para mejorar la salud de la población, finalidad última de todas las intervenciones de promoción de la salud.

Contribuciones de autoría

M. Nebot ha realizado el borrador y la versión final. Todos los autores han participado en la discusión y revisión del manuscrito.

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. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Skokie (IL): Rand McNally; 1966.
3. Sacket DL. Evidence-based medicine. Semin Perinatol. 1997;21:3-5.
4. Williams A. The nature, meaning and measurement of health and illness: an economic viewpoint. Soc Sci Med. 1985;20:1023-7.
5. Pawson R, Tilley N. Realistic evaluation. Sage: London (UK); 1997.
6. Nebot M. Evaluación en salud pública: ¿todo vale? Gac Sanit. 2007;21:95-6.
7. Porta M, Last JM. Dictionary of epidemiology. Oxford: Oxford University Press; 2008.
8. Last JM. A dictionary of public health. Oxford: Oxford University Press; 2007.
9. Suchman E. Evaluative research. New York: Rusell Sage Foundation; 1967.
10. Hawe P, Degeling D, Hall J. Evaluating health promotion. Sydney: MacLennan Petty Pty; 1990.
11. Scriven M. The methodology of evaluation. En: Weiss C, editor. Evaluating action programmes: readings in social action and education. Boston: Allyn and Bacon Publ; 1972. pp. 36-55.
12. Overtveigt J. Evaluation purpose, theory and perspective. En: Overtveigt J. Evaluating health interventions. Buckingham, UK: Open University Press; 1999. pp. 23-47.
13. Pineault R. La planificación sanitaria. Barcelona: Masson; 1981. pp. 25-67.
14. Haddick AC, Teutsch SM, Shaffer PA, et al. Prevention effectiveness. Introduction. New York: Oxford University Press; 1996. pp. 3-19.
15. Green LW, Kreuter MW. Health promotion planning. An educational and ecological approach. Palo Alto: Mayfield Publ Co; 1999.
16. Windsor R, Branowski T, Clark N, et al. Evaluation of health promotion, health education and disease prevention programs. Mountain View (CA): Mayfield Co; 1994.
17. Kleinbaum DG, Kupper LL, Morgenstern H. Types of epidemiolgic research. En: Epidemiologic research. Principles and quantitative methods. New York: Van Nostrand Reinhold; 1982. pp. 40-50.
18. Dantas OM, Ximenes RA, de Albuquerque M de F, et al. A case-control study of protection against tuberculosis by BCG revaccination in Recife, Brazil. Int J Tuberc Lung Dis. 2006;10:536-41.
19. Sasieni P, Castanon A, Cuzick J. Screening and adenocarcinoma of the cervix. Int J Cancer. 2009;125:525-9.
20. Luna Sánchez, A. Efectos de la cobertura vacunal previa en la dinámica de un brote de sarampión. Rev Esp Salud Pública. 1997;71:243-7.
21. Cook TD, Campbell DT. Quasi-experimentation. Design and analysis issues for field settings. Boston: Houghton Mifflin Co; 1979.
22. Green LW, Lewis FM. Measurement and evaluation in health education and health promotion. Palo Alto: Mayfield Publ. Co; 1986.
23. Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J P Health. 2004;94:400-5.
24. Habitch JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. Int J Epidemiol. 1999;28:10-8.
25. Thomson H, Hoskins R, Petticrew M, et al. Evaluating the health effects of social interventions. BMJ. 2004;328:282-5.
26. Vartiainen E, Jousilahti P, Alfthan G, et al. Cardiovascular risk factor changes in Finland, 1972-1997. Int J Epidemiol. 2000;29:49-56.
27. McAlister A, Puska P, Salonen JT, et al. Theory and action for health promotion: ilustrations from the North Karelia Project. Am J Public Health. 1982;72:43-50.
28. Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? BMJ. 1998;316:611-6.
29. Campbell M, Fitzpatrick R, Haines A, et al. Framework for design and evaluation of complex interventions to improve health. BMJ. 2000;321:694-6.
30. Nebot M. Health promotion evaluation and the principle of prevention. J Epidemiol Community Health. 2006;60:5-6.