BÁSICA

Cuestiones metodológicas en la evaluación de resultados terapéuticos

Methodological issues in assessment of treatment outcome

IRAURGI CASTILLO, I.

Módulo de Asistencia Psicosocial de Rekalde.
Profesor-Tutor de Psicología Experimental y Psicología Patológica en la UNED (Sede de Bergara).

RESUMEN: *Contexto:* la investigación clínica y la evaluación terapéutica son facetas afines del campo de conocimiento referido a la salud y la forma de atajar la enfermedad. Uno de los objetivos fundamentales de ambas disciplinas es determinar objetivamente, y con base empírica, la pertinencia, efecto e impacto de una actividad terapéutica respecto a un determinado problema de salud. No obstante, la actividad terapéutica cotidiana a menudo encuentra dificultades para evaluar sus actuaciones, en la mayoría de los casos debido a que el contexto de la práctica clínica es muy distinto al de la investigación.

***Objetivos:* en este artículo se revisan algunas de estas dificultades y se hacen una serie de sugerencias para encaminar la investigación de resultados terapéuticos.**

PALABRAS CLAVE: Métodos de investigación. Evaluación terapéutica. Diseños de investigación. Drogodependencias.

ABSTRACT: *Background:* the clinic investigation and therapeutic evaluation are related facets of the knowledge field refered to the health and the form of coping the disease. One of the fundamental objectives of both disciplines is to determine objectively, and with empirical base, the relevancy, effect and impact of a therapeutic activity with respect to a health problem. Nevertheless, the daily therapeutic activity often finds difficulties to evaluate their performances, in most cases due to the fact that the context of the practical clinic is very different to that of the investigation.

***Objective:* in this paper we review some of these difficulties and make a number of suggestions to advise the therapeutic results investigation.**

KEY WORDS: Investigation methods. Therapeutic assessment. Investigation designs. Drug dependence.

Correspondencia:

IOSEBA IRAURGI CASTILLO.
Módulo de Asistencia Psicosocial de Rekalde.
C/. Camilo Villabaso 24 lonja. 48002 Bilbao.
E-mail: iraurgi@euskalnet.net

Introducción

Resulta de perogrullo hacer observar a los lectores que la identificación de un problema de salud conduce necesariamente a la instauración de medidas preventivas, si la enfermedad y/o los efectos nocivos todavía no han aparecido, y a la aplicación de un tratamiento para paliar los efectos activos de la enfermedad, en el caso de que ésta esté presente. Tanto las primeras como las segundas, persiguen obtener un efecto derivado de la aplicación de la intervención, considerándose que esta relación reúne criterios de causalidad.

Según Jenicek[1] cualquier intervención en el área de la salud representa una causa en relación con el impacto esperado (efecto resultante) y ha de responder a cuatro cuestiones básicas: 1) ¿está la acción propuesta bien fundamentada, es decir, tiene sentido llevarla a cabo?, 2) ¿es adecuada la estructura de la intervención (¿cómo se organiza?), 3) ¿el proceso es aceptable, sucede como deseamos? y 4) ¿cuál es el resultado o impacto del tratamiento? Dar respuesta a estas cuatro cuestiones implica poner en marcha un proceso evaluativo de la intervención a aplicar que, de llevarse a cabo de la forma oportuna, permitirá a los agentes de salud contar con la información suficiente y necesaria

para optar por la medida terapéutica más adecuada a cada caso.

Consideraremos la evaluación como la actividad que de una forma sistematizada y objetiva pretende determinar la efectividad o impacto de una determinada actividad o intervención en función de un objetivo determinado; siendo el objetivo último de toda evaluación aportar evidencia empírica que sea de utilidad en el proceso de toma de decisiones. Como puede apreciarse, nuestra perspectiva de la evaluación terapéutica está más cercana al positivismo metodológico que a las nuevas corrientes pluralistas imperantes en el área de la evaluación donde se prioriza la emisión de juicios de mérito/valor sobre la utilización de procedimientos científicos[2]. La aplicación de una medida terapéutica persigue causar un efecto deseado, mejorar la calidad de la salud del afectado y, sobre todo, hacer más bien que mal *(primum non nocere),* y para ello debemos tener la máxima seguridad posible de la relación causa-efecto entre la intervención y el problema de salud para el cual se aplica.

Desde mi punto de vista, o al menos desde mis intereses en la evaluación de resultados terapéuticos, existen tres objetivos básicos en la evaluación sanitaria:

1. Determinar objetivamente, y con base empírica, la pertinencia, efecto e impacto de una actividad terapéutica respecto a un determinado problema de salud (estimación de la eficacia de la intervención para su generalización como recurso terapéutico),

2. Estimar el cambio del estado de salud producido en el usuario objeto de la intervención para valorar el éxito/fracaso terapéutico individual (criterios de decisión para el alta o para la modificación del tratamiento), y

3. Analizar las variables predictivas de la adecuación terapéutica, es decir, examinar los factores exógenos y/o endógenos a los usuarios receptores de la intervención que puedan explicar el éxito o fracaso terapéutico (permitirá la selección de los candidatos idóneos a un determinado tratamiento y la modificación o búsqueda de nuevos recursos para los casos fallidos).

Mi intención en este artículo es desarrollar algunas consideraciones metodológicas respecto al primer objetivo enunciado, dejando para ocasiones posteriores la presentación de estrategias de evaluación para los dos objetivos restantes. Revisaremos, en esta primera aproximación, conceptos fundamentales en la estructura de una evaluación terapéutica, haciendo especial hincapié en aspectos relacionados con la protocolización de la intervención, la elección del diseño de investigación/evaluación y la utilización de técnicas estadísticas como control secundario de la validez del estudio. No pretendo ser exhaustivo en el desarrollo de estos conceptos. Muchas publicaciones altamente especializadas ya han abordado este tema de forma meritoria, las cuales recogeremos en el epígrafe de bibliografía recomendada, para que el lector interesado pueda ahondar en el estudio de la evaluación terapéutica.

La evaluación de resultados terapéuticos: conceptos básicos

Cualquier evaluación de una intervención sanitaria es un proceso secuencial, paso a paso, que deberá basarse en un adecuado conocimiento de la patología diana, de la farmacología o técnica de intervención, así como de la metodología de investigación necesaria para su evaluación. Asimismo, ha de tenerse en cuenta que la evaluación de cualquier tipo de tratamiento[3] ha de considerar cinco tipos de datos: 1) la situación pretratamiento o, lo que es lo mismo, el momento evolutivo y la gravedad del proceso a tratar; 2) la eficacia relativa de las posibles formas de intervención terapéutica; 3) el impacto del tratamiento sobre el paciente en su conjunto y no sólo sobre la enfermedad; 4) la monitorización de los procesos tras la intervención, y 5) la efectividad comparativa del tratamiento o intervención utilizada.

La figura 1 representa de forma esquemática un modelo donde se pone en relación el proceso de intervención terapéutica con tres fases diferenciadas (el diagnóstico, la intervención en sí misma y el proceso de monitorización del seguimiento), cada una de las cuales puede ofrecer un conjunto de variables relacio-

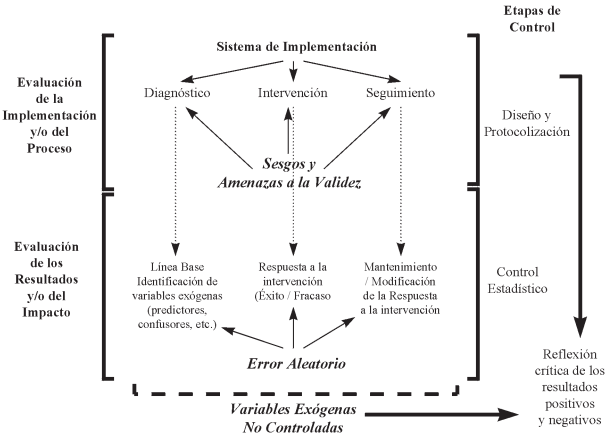

Figura 1. Fases y procesos implicados en la evaluación terapéutica.

nadas con los resultados a obtener y cubrirían los objetivos de la evaluación terapéutica antes descritos. Para dar respuesta al primer objetivo propuesto (comprobar la eficacia del tratamiento), podríamos pensar que nuestro interés ha de fijarse en la relación entre las variables intervención/tratamiento y los efectos perseguidos (valoración de éxito o fracaso de dicha intervención). Pero la realidad asistencial, nos muestra que los resultados son producto de todo un proceso terapéutico, y no de una de sus fases. Un buen diagnóstico nos permite conocer las condiciones basales del sujeto a tratar, fijando un punto de referencia a partir del cual poder evaluar sus progresos; nos permite, asimismo, identificar variables ajenas a la intervención que pudieran influir en los resultados finales; y, lo más importante, nos permite conocer las características de la enfermedad que se ajustan mejor a un determinado tratamiento. Es decir, un buen diagnóstico permite una prescripción ajustada al problema; sin un adecuado diagnóstico podríamos estar haciendo indicaciones con menos probabilidades de éxito. La intervención/tratamiento es por hipótesis la causa de los efectos perseguidos (resultados terapéuticos), y donde generalmente se ha focalizado toda la investigación evaluativa. Pero los resultados de una intervención suponen, en el ámbito sanitario, la recuperación de la salud o el mantenimiento de una situación menos dañina respecto a la situación que motivo la demanda. Por ello, el verdadero resultado terapéutico no ha de evaluarse inmediatamente después de la intervención, sino a mediano-largo plazo, haciéndose imprescindible la monitorización del seguimiento postratamiento. Se establece entre las tres fases una serie de relaciones de dependencia intrínseca que tiene como consecuencia la aparición de modelos multicausales que pueden dar explicación de los efectos obtenidos.

En la práctica investigadora, la búsqueda de relaciones causales entre variables se orienta, en la mayoría de los casos, en función de tres criterios: 1) *temporalidad*: que la variable que se supone causa ha de preceder al supuesto efecto, 2) *asociación*: debe existir una relación o covariación significativa entre la variable independiente (VI) o exposición (intervención) y la variable dependiente (VD) o respuesta, y 3) *ausencia de espureidad*: la relación entre estas variables no ha de ser aparente, es decir, no han de existir variables o factores que afecten la relación entre la(s) VI(s) y la(s) VD(s). El análisis de la validez interna de un diseño o investigación está directamente relacionado con el establecimiento de este tipo de relaciones causales, de modo que si el diseño permite de un modo efectivo concluir inequívocamente que dos o más variables están relacionadas o no, entonces dicho diseño tendrá validez interna.

Como se refleja en la figura 1, las relaciones entre las fases del proceso terapéutico y sus resultados muchas veces están afectadas por tres tipos de interferencias que conllevan problemas de validez: 1) los sesgos y amenazas a la validez del estudio (recogidas en la tabla I), 2) el error aleatorio: efectos debidos al azar y errores aleatorios de medición que afectan a la fiabilidad de las mediciones atenuando las relaciones causales, y 3) la existencia de variables exógenas no controladas: variables que están relacionadas a la vez con la intervención, con otras variables explicativas controladas y con los resultados, lo que supone una perturbación y falta de control de posibles explicaciones alternativas a los efectos hallados. El sistema para aumentar la validez interna de un diseño es precisamente transformar las posibles variables perturbadoras en controladas o aleatorias. Pero para ello, el primer paso ha de ser identificar esas posibles variables perturbadoras. Esta tarea se podrá realizar con mayor éxito en la medida que se tenga un conocimiento exhaustivo de las experiencias obtenidas en la investigación previa sobre el tema (revisión de la teoría actualizada) y se cuente con una buena protocolización de la fase de diagnóstico. Esta primera aproximación nos permitirá orientar de una manera más acertada el diseño oportuno para llevar a cabo nuestra evaluación. Es a través del diseño que podremos conseguir un mayor control de las variables identificadas a través de estrategias tales como: 1) la utilización de grupos de comparación y controles, con aleatorización de los sujetos participantes y asignación aleatoria a los grupos de tratamiento; 2) la utilización de diseños longitudinales que impliquen medidas repetidas —un mayor número de estimaciones a lo largo del tiempo aumenta la constancia de la medida e incorpora el control intragrupo—, 3) en caso de no poder utilizar la aleatorización es posible utilizar estrategias de control construido como las técnicas de bloqueo o apareamiento (formar grupos equivalentes de comparación en función de variables relevantes) o la utilización de comparaciones normativas (comparación de los resultados de la intervención con baremaciones o estándares establecidos previamente). Todo aquello que no pueda ser controlado mediante estrategias de diseño, pero de lo cual se tenga información identificada y recogida en el protocolo de evaluación, podrá ser controlado a través de métodos estadísticos (estratificación, modelos multivariados, etc.); ello, claro está, a costa de au-

Tabla I. Amenazas a la validez de los diseños.

Amenazas	Características
1. Validez Interna	**¿Es el Diseño lo suficientemente sensible como para detectar relaciones causales?**
a) Historia	Sucesos externos al tratamiento que pueden afectar a la(s) VD(s
b) Maduración	Cambios biológicos y psicológicos de los sujetos que afectarán a sus respuestas
c) Administración de pruebas	Los efectos del pre-test pueden alterar las respuestas en el postest independientemente del tratamiento
d) Instrumentación	Cambios en la instrumentación o en los observadores (dificultades de calibración)
e) Regresión estadística	Las puntuaciones extremas tienden a acercarse a la media en el postest a pesar del tratamiento
f) Selección	Diferencias en los sujetos anteriores al tratamiento
g) Mortalidad experimental	Pérdida selectiva de sujetos a lo largo del estudio
h) Desmoralización de los Controles	Los controles abandonan o interfieren al no recibir tratamiento o percibir diferenciación con los experimentales
i) Ambigüedad sobre la direccionalidad de la influencia causal	Problemas de interferencia en la dirección de la causalidad. Problema de la Confusión y/o Interacción
j) Difusión/Imitación de tratamientos	Los miembros de los grupos de tratamiento comparten las condiciones de tratamiento con cada uno de los demás o intentan copiar el tratamiento
k) Igualación compensatoria de tratamientos	Determinar que todos los sujetos, tanto del grupo experimental como del control, reciban un tratamiento que les proporcione efectos beneficiosos
l) Interacción de tratamientos intrasujeto	Los sujetos forman parte también de otros tratamientos (intrasujetos)
m) Interacción de administración de pruebas y tratamientos	La administración de las pruebas puede facilitar o inhibir el efecto del tratamiento. Efecto secuencial y/o de período
n) Interferencia de tratamientos múltiples	Interacción de los tratamientos anteriores con los posteriores
2. Validez Externa	**¿Pueden generalizarse los efectos y causas de un estudio a otros sujetos, situaciones o contextos?**
a) Representatividad de la muestra (validez de población)	Capacidad para generalizar el tratamiento a personas que no pertenezcan al grupo estudiado
b) Representatividad de los tratamientos	La elección arbitraria de los niveles de la VI no siempre representan a todos los posibles niveles que puede adoptar dicho valor
c) Efectos reactivos de la situación experimental (validez ecológica)	La artificialidad de la situación experimental puede llevar a los sujetos a responder de forma diferente a como lo harían en la vida normal. Capacidad de generalización del tratamiento a situaciones más allá de la estudiada
d) Interacción Historia-tratamiento (validez histórica)	Capacidad para generalizar el tratamiento a otras ocasiones temporales (pasado o futuro)
3. Validez de constructo	**¿Qué variables teóricas o implícitas están siendo estudiadas?**
a) Explicación preoperacional inadecuada	Escasa definición de los constructos
b) Empleo de operacionalizaciones únicas	Medida de una sola VD y/o medida de la VD mediante un solo método
c) Adivinación de hipótesis, Efecto Hawthorne	Los sujetos intentan adivinar la hipótesis experimental y actúan de la forma que creen que el investigador quiere que actúen
d) Recelo de evaluación	Los sujetos manifiestan cierto recelo ante la situación de evaluación
e) Expectativas del experimentador Efecto Rosenthal	Los experimentadores producen sesgos en el estudio a causa de sus expectativas en y durante el estudio
f) Confusión entre constructos y niveles de constructo	No se implementan todos los niveles del constructo y pueden presentarse de forma débil o no existir
4. Validez Estadística	**¿El estudio es sensible para detectar si las variables covarían?**
a) Baja potencia estadística	El Error de Tipo II aumenta cuando el valor de alfa es bajo y la muestra pequeña
b) Violación de los supuestos de las pruebas estadísticas	Todos los supuestos deben ser conocidos y comprobados cuando sea necesario
c) «Ir de pesca» y Tasa de Error de Tipo I	Se incrementa, a menos que se ajuste el número de contrastes posibles
d) Fiabilidad de medición	Fiabilidad baja implica más errores que constituyen un problema serio en los estadísticos inferenciales

VD(s): Variable(s) Dependiente(s); VI(s): Variable(s) Independiente(s)

mentar el tamaño de las muestras. Por tanto, nuestro esfuerzo ha de dirigirse a construir buenos diseños de investigación más que a confiar en exceso en las posibilidades de la estadística.

De una u otra manera, siempre existen variables que se escapan a nuestro control, y la única forma de «neutralizar» su efecto es a través de una reflexión crítica de los resultados obtenidos que permitan la proposición de nuevas hipótesis alternativas a contrastar en próximas evaluaciones/investigaciones.

Por otra parte, la evaluación terapéutica no ha de basarse exclusivamente en la evaluación de resultados; también ha de hacerse un esfuerzo por examinar la implementación o el proceso terapéutico, dado que este tipo de evaluaciones tiende a establecer protocolos de actuación cada vez más ajustados a las distintas manifestaciones del problema de salud. Al igual que se sugiere la conveniencia de utilizar protocolos de investigación[4] en los ensayos clínicos, sugerimos que de cara a la actividad evaluadora también sean utilizados. Parafraseando a J. Roca[4], los protocolos facilitan la planificación y organización de la investigación/evaluación de una forma lógica y eficiente, se constituye en un manual de operaciones que permite la coordinación de todo el equipo y la estandarización de las actividades y actuaciones a realizar en el proceso terapéutico. En la tabla II se reproduce un modelo de protocolo como guía de las actividades a llevar a

Tabla II. Componentes de un protocolo de evaluación.

Formulación del problema:
- Antecedentes y estado actual del tema.
- Hipótesis a probar.
- Objetivo general del estudio (propuesta cualitativa).
- Objetivos específicos (propuesta cuantitativa).

Definición clínica y en unidades cuantificables de las variables dependientes (problema de salud / enfermedad) e independientes (tipo de tratamiento, variables mediadoras, etc.), con criterios claros de inclusión y exclusión:
- Definición de la enfermedad en términos de su espectro, gradiente y curso.
- Definición operativa de la modalidad de tratamiento experimental.
- Definición operativa del tratamiento alternativo, de control o de referencia.
- Definición operativa y en unidades medibles de los efectos o resultados que se esperan de la intervención.
- Definición operativa y en unidades medibles de los efectos o resultados que se esperan de los tratamientos alternativos.

Inclusión de los sujetos:
- Población de pacientes, población de la que provienen.
- Plan de muestreo y tamaño de la muestra. Características de acceso.
- Criterios precisos de inclusión y exclusión de los sujetos.

Estructura del estudio/diseño:
- Tipo de diseño.
- Descripción minuciosa de las estrategias de control: aleatorización, técnicas de ciego, apareado/bloqueo, estratificación.
- Descripción de los sesgos posibles del estudio o errores posibles. Propuesta del tipo de control que se llevará a cabo antes, durante y después de la implementación.

Aspectos éticos:
- Consentimiento informado.
- Obrar según: *primum non nocere.*

Procedimiento para la evaluación:
- Planificación de las fases de la evaluación: evaluación de implementación, del proceso, de los resultados y del impacto.
- Descripción de las técnicas de recogida y registro de datos.
- Criterios de uniformidad en la recogida de datos.
- Estrategia programada para el análisis de los datos.
- Realización de análisis intermedios para detectar resultados no esperados (efectos adversos, etc.).
- Normas y procedimientos para interrumpir el tratamiento.
- Criterios para el manejo de las no respuestas, abandonos, etc.

Datos de línea base:
- Protocolización del diagnóstico.
- Información del estado inicial de los pacientes.
- Detección y registro de otros factores intervinientes (variables exógenas con carácter interactivo, confusor, predictivo, etc.).
- Otra información importante: comorbilidad, cotratamientos, etc.

Protocolización de la intervención, de las actividades terapéuticas:
- Descripción detallada, cualitativa y cuantitativamente, de la intervención terapéutica a evaluar.
- Descripción detallada apoyada en experiencias documentadas de las modalidades de tratamiento de referencia (controles).
- Descripción del programa de tratamiento: fases de implementación, temporalidad, etc.
- Planificación del seguimiento.

Planificación de la comunicación de los resultados:
- Distribución de responsabilidades en la elaboración del informe.
- Estructuración del informe según destinatario: pacientes, personal sanitario participante, gestores del centro de salud, responsables sanitarios de servicios centrales, etc.
- Presentación de resultados a la comunidad médica/científica: planificación de la publicación de resultados.

Adaptado de Jenicek[1], 1995.

cabo.

La elección del diseño

Según la clasificación de la FDA[5] (Food and Drug Administration) los estudios de evaluación encaminados a comprobar y demostrar la posible acción beneficiosa de un fármaco en humanos, y por extensión podríamos incluir a toda intervención terapéutica, se divide en cuatro fases (tabla III).

Cuando una intervención en el área de la salud es evaluada en una fase III mediante un ensayo clínico aleatorio, es decir, es llevada a cabo por equipos estructurados, entrenados y con suficientes recursos humanos y materiales, con criterios de selección y exclusión de pacientes, y con un seguimiento riguroso del proceso terapéutico, se está evaluando el procedimiento en condiciones óptimas, ideales o de laboratorio. Cuando se miden los efectos conseguidos en estas condiciones se está analizando la *«eficacia»* de la intervención, y puede asegurarse que los resultados obtenidos son debidos al efecto de la aplicación terapéutica; es decir, se obtiene una alta validez interna.

No obstante, cuando las técnicas terapéuticas avaladas por los resultados de un ensayo clínico se pretenden aplicar en la práctica clínica o en forma de programa sanitario (fase IV), a veces no se consigue el mismo nivel de eficacia. La explicación es que pueden existir factores diferenciadores importantes de la población donde se intenta aplicar la intervención correspondiente en relación a la que sirvió de base para el estudio de eficacia original: el grado de aceptación de la intervención por parte de las personas afectadas, la distribución de los factores de riesgo o de pronóstico implicados, el rendimiento de los profesionales, los recursos tecnológicos y las organizaciones, etc., pueden obstaculizar la obtención de los mismos resultados que en las condiciones de estudio. Por ello, al impacto real conseguido al implantar una intervención eficaz le llamamos *«efectividad»*. Al realizarse en condiciones reales, la generalización de los resultados es menos problemática que en el caso de la eficacia, pero a pesar de ello, la efectividad en un centro puede ser distinta de la efectividad en otro centro. Con la efectividad alcanzamos una mayor validez externa, claro está que a costa de perder la validez interna.

Nos encontramos ante la disyuntiva de realizar una evaluación en condiciones ideales frente a otras condiciones, llamaremos naturales, más próximas a nuestra realidad asistencial. La posibilidad de evaluar nuestros resultados terapéuticos mediante la aplicación de estrategias de ensayo clínico no es factible en la mayoría de los casos y mucho menos de una forma generalizada al conjunto de casos tratados. Por lo general, la evaluación mediante ensayo clínico es realizada en hospitales, dotados de mayor número de recursos y medidas de control de la intervención, y aplicados a muestras parciales del conjunto de casos potenciales. Efectivamente, es el mejor diseño posible para establecer y demostrar la existencia de causalidad entre la intervención y el efecto (eficacia del tratamiento), pero su implementación no es generalizable al conjunto de los recursos asistenciales. Nuestra sugerencia es que siempre que sea posible se realice la evaluación basándose en un diseño aleatorizado. Sería ingenuo negar que, en muchos casos, estudios iniciados como ensayos aleatorizados acaban incumpliendo este criterio, debiéndose

Tabla III. Fases de la evaluación clínica de los medicamentos.

Fase	Sujetos	Objetivo	Diseño
I	Sanos (voluntarios) muestras pequeñas	Respuesta biológica al medicamento: tolerancia, seguridad, etc.	Descriptivo
II	Enfermos seleccionados muestras pequeñas	Beneficios potenciales Efectos secundarios Definición dosis terapéuticas Estimación eficacia relativa	Observación de casos Ensayo terapéutico piloto
III	Enfermos seleccionados muestras suficientes	Eficacia del medicamento	Ensayo Clínico Aleatorio
IV	Enfermos No seleccionados muestras amplias	Efectividad del medicamento Efectos secundarios tardíos Nuevos efectos/indicaciones	Cuasi-experimentales Naturalísticos De caso único, etc.

Adaptado de Jenicek1, 1995.

analizar los datos mediante ajustes estadísticos. En cualquier caso, suponemos que un diseño aleatorizado, por degradado que esté, será más sólido para inferir causalidad que otras posibles alternativas pre-experimentales (estudios postintervención, diseños pre-post sin grupo control, estudios retrospectivos, etc.) o cuasi-experimentales, ya que la asignación a condiciones prevalentes en el post-test final seguirán estando basadas en parte en la aleatorización.

No obstante, como ya hemos comentado, cuando la evaluación se realiza en la práctica clínica cotidiana nos vemos obligados a optar por otras alternativas de diseño para la evaluación. Atendiendo a conceptos expresados previamente, dos consideraciones son de suma importancia a tener en cuenta; 1) para realizar una evaluación de resultados terapéuticos es imprescindible considerar la dimensión temporal del proceso, de forma que estamos obligados a optar por diseños de tipo longitudinal, preferiblemente, de tipo prospectivo, 2) si buscamos la eficacia relativa y efectividad de las posibles formas de intervención terapéutica, no podemos prescindir de la utilización de grupos de control o de intervenciones alternativas. Por tanto, los estudios basado en estrategias preexperimentales deberían ser utilizados con extrema cautela, dada su gran vulnerabilidad a los problemas de validez interna, y a ser posible no considerados en procesos de evaluación. Ahora bien, tampoco queremos ser inflexibles, y ante la disyuntiva de evaluar o no evaluar, siempre es preferible utilizar este tipo de diseños a no hacer nada. En cualquier caso, si bien los resultados obtenidos con este tipo de diseños no serán en absoluto concluyentes, si pueden ser base para generar hipótesis de trabajo terapéutico.

Existen diversas alternativas de diseños cuasiexperimentales (tabla IV) que pueden ser utilizados en la evaluación terapéutica, si bien existen algunos que, bajo nuestra perspectiva, son superiores para aproximarse a inferencias causales y son particularmente adecuados para el propósito evaluativo. A continuación trataremos de exponer tres de estos diseños: el cuasiexperimental de grupos no equivalentes, el diseño de línea base no causal construida y los diseños basados en series temporales.

Diseño de grupo(s) no equivalente(s)[6]

Este tipo de diseño (figura 2) comprende un grupo experimental o de intervención y otro(s) control(es), de los cuales ambos han sido evaluados en un pretest y un postest, pero no poseen equivalencia de muestreo. Es decir, los grupos constituyen entidades formadas naturalmente, tan similares como la disponibilidad lo permita. Esta situación reduce la potencia del diseño para establecer una relación causal, ya que hay dudas acerca de la equivalencia de los grupos antes de que se inicie la intervención, de ahí que se denomine diseño no equivalente. Por lo demás, el diseño puede seguir las mismas guías de actuación que en un ensayo clínico[1,7-9]: la asignación del tratamiento a uno u otro grupo se supone aleatoria y controlada por el experimentador, en ocasiones puede buscarse una mayor homogeneización de los grupos mediante técnicas de boqueo, emparejamiento, etc., también en ocasiones pueden aplicarse pruebas de

Tabla IV. Tipos de diseño en la evaluación terapéutica.

	Pre-experimentales	Cuasi-experimentales	Experimentales
Características			
Grupo control	No	Sí	Sí
Selección aleatoria de sujetos a grupos	No	No	Sí
Asignación aleatoria de tratamientos	No	Sí	Sí
Tipos de Diseños	• Postest de un grupo • Comparación postest con un grupo estático • Pretest-postest de un grupo o Estudios antes-después • Casos clínicos	• Grupo no equivalente • Controles apareados • Ensayos naturales • Series de casos consecutivos • Grupo de control histórico • Diseños compensados • De línea base no casual construida • Series temporales	• Ensayo clínico con grupos paralelos • Ensayo clínico cruzado • Ensayo secuencial • Diseños intrasujeto, N=1 • Diseños factoriales
Grado de control sobre las amenazas a la validez interna y Nivel de evidencia causal	Bajo	Moderado	Alto

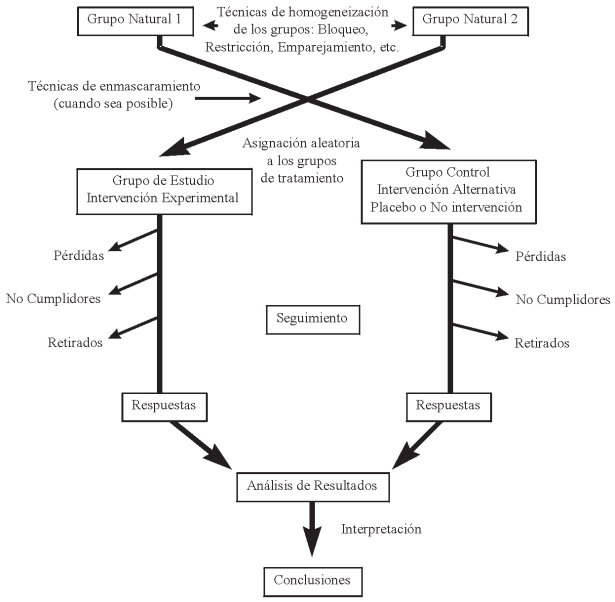

Figura 2. Esquema de un Diseño de Grupos No Equivalentes.

enmascaramiento (técnicas de ciego), el seguimiento del estudio y las estrategias de análisis y tratamiento de los casos perdidos son equivalentes, etc.

La amenaza más seria para la validez interna de este diseño es la selección, puesto que los grupos podrían diferir inicialmente en cuanto a características que podrían estar relacionadas con la variable dependiente. Con la inclusión del pretest es posible comparar los puntajes obtenidos en dicha prueba y ver si los grupos son equivalentes. Si lo son, no habrá que preocuparse tanto por su equivalencia, aunque sí un poco en tanto que podrían estar actuando otras variables no controladas que la aleatoriedad podría haber neutralizado. La identificación en el pretest de variables intervinientes que producen la falta de equivalencia de los grupos nos permite su ajuste posterior mediante técnicas de control estadístico como la utilización de análisis de covarianza, el MANOVA o la regresión múltiple.

El diseño de grupo control no equivalente es el diseño más comúnmente utilizado cuando no es posible la aleatorización de los sujetos participantes. Funciona de forma adecuada porque hay cierto control sobre la influencia de las variables extrañas, gracias a la utilización de un grupo control y al ajuste estadístico preprogramado. Aunque no se asegura la equivalencia de los grupos, se aproxima en alguna medida, y es la alternativa de elección ante la imposibilidad de llevar a cabo un ensayo clínico aleatorizado.

Diseño de línea base no causal construida

En ocasiones hemos de realizar una evaluación en una situación en la que no resulta disponible un grupo de control de sujetos no equivalentes. La imposibilidad de localizar o monitorizar un grupo control, la falta de recursos materiales y humanos para llevar a cabo un ensayo clínico, la limitación de tiempo o bien la incapacidad ética de negar a un grupo una terapia con efectos de mejora son razones frecuentes para no contar con un grupo control. Pero para realizar una evaluación adecuada hemos de encontrar un grupo de comparación alternativo y ello lo podemos conseguir a través de una línea base no causal basada en dos estrategias: 1) la utilización de un diseño de regresión-extrapolación, y 2) una comparación baremada en la que los sujetos son sometidos a pretest y postest y comparados con muestras tomadas de otras fuentes de datos basadas en la población.

Diseño de Regresión-Extrapolación

También llamado análisis de discontinuidad de la regresión[2,6], se basa en la comparación de la puntuación del grupo de tratamiento en el postest con su puntuación proyectada en el postest, basada en una tendencia madurativa lineal durante el tiempo transcurrido entre el pretest y el postest. Este diseño analiza los efectos incrementales del tratamiento por encima de los efectos proyectados, establecidos a partir de la maduración habida durante el proceso terapéutico. Al aplicar este diseño es importante tener en cuenta que a fin de realizar una proyección de la puntuación del postest, ha de quedar bien establecida la tendencia madurativa a través del tiempo.

Se requiere, por tanto, una medición de la variable dependiente (resultado) antes y después de la intervención y una medición pre-intervención de otra variable relacionada con la medida pretest que permita la formación de grupos, aunque éstos también se pueden constituir a partir de la variable resultado pretest en función del establecimiento de un punto de corte. Un requisito indispensable de este diseño es que las variables a considerar han de ser medidas en una escala continua, dado que la hipótesis que subyace a su aplicación se basa en modelos lineales.

Quizá la mejor forma de comprender este diseño sea dentro del contexto de un ejemplo, aunque la propuesta la formulemos de una forma hipotética. Supongamos que nuestro objetivo es evaluar la eficacia/efectividad de un programa de prevención de riesgos, educación para la salud y tratamiento anticu-

berculostático implementado en un Programa de Mantenimiento con Metadona (PMM). Como variable dependiente se ha considerado una estimación de la calidad de vida centrada en la salud. La revisión de la literatura nos muestra una estrecha relación entre esta variable de resultado y el tiempo de tratamiento en el PMM. Tras administrar el pretest al conjunto de participantes del programa, donde se evalúa la variable calidad de vida y se ha recogido el tiempo de permanencia en el programa, se ha examinado el diagrama de dispersión (Fig. 3A) generado por ambas variables calculado a través de un modelo de regresión, mostrando condiciones de linealidad. A partir de estos datos, se toma la decisión de establecer un punto de corte en función de la permanencia en el programa, estableciéndolo en los seis meses y que queda representado por una línea vertical en la figura. En función de este punto de corte se decide dar la intervención al grupo que queda a la izquierda (representado por el signo *) y dejar como control al de la derecha (representado por O).

La base del diseño consiste en comparar las dos rectas de regresión, la de las * y las de O, y ver si la relación pre-post observada en el grupo que recibe la intervención es la misma o se diferencia de la encontrada en el grupo que no la recibe. Ello se realiza extrapolando la regresión obtenida entre los O sobre la presentada por los *. Si coinciden (Fig. 3B) es que el programa no ha tenido efecto, dado que la extrapolación de O sería lo que obtendríamos si no se hubiera aplicado el programa. Por el contrario, si las proyecciones de las rectas de regresión no se igualan (Fig. 3C) estaremos ante un resultado positivo indicativo de que el programa ha funcionado. La diferencia expresada en la figura como 'a' es una estimación de la ganancia debida a la intervención.

Asimismo, a partir de la ecuación de regresión estimada con los datos del pretest se puede calcular para cada sujeto cuál es la ganancia en calidad de vida que se puede esperar por cada mes de estancia en el programa. Por otra parte, se calcula cuál es la ganancia en calidad de vida entre el pretest y postest debida al efecto de la aplicación del programa. A esta ganancia medida se resta la esperada, obteniéndose una aproximación más ajustada a la ganancia real.

En el ejemplo se ha utilizado un solo predictor, pero pueden analizarse situaciones en las que existan múltiples predictores. En estos casos, los datos y el fenómeno bajo estudio han de ser estables a través del tiempo para que la predicción de regresión sea ajustada. Si los datos son inestables, ha de haber una teoría muy sólida que explique la inestabilidad de los datos

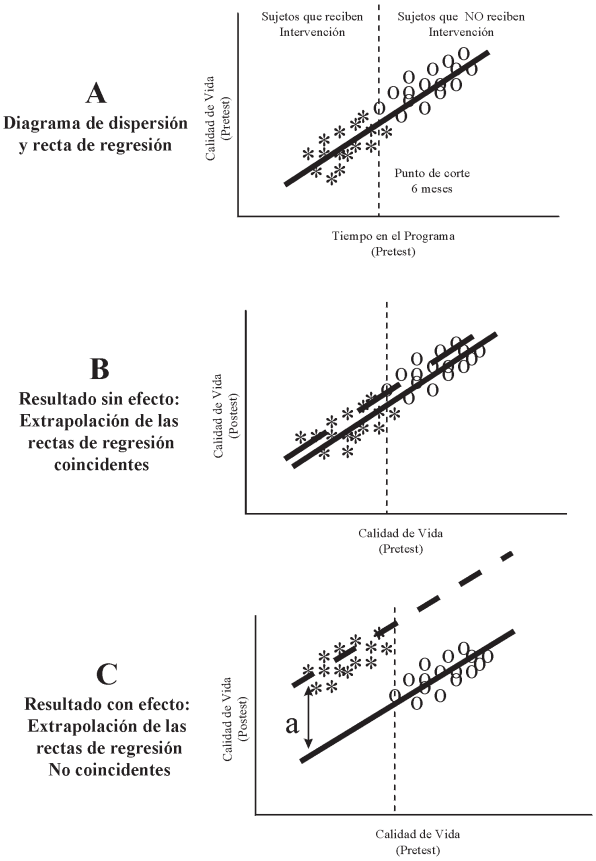

Figura 3. Diseño de Regresión-Extrapolación.

para que la predicción de regresión sea una técnica útil. Por otro lado, este tipo de diseño no sólo sirve para la evaluación de resultados, sino también para la evaluación del proceso terapéutico si se utiliza durante su implementación.

Diseños por comparación baremada[10]

Este es otro tipo de diseño de línea base no causal construida que puede utilizarse cuando no se dispone de un grupo control sin tratamiento. El modelo de referencia baremada toma la puntuación pretest media del grupo de tratamiento y lo refiere a tablas previamente baremadas, convirtiendo la puntuación en el pretest en un indicador como percentil. El percentil en el pretest es sustraído posteriormente del percentil en el postest, para obtener la evaluación del efecto del tratamiento.

Consideremos otro ejemplo. Supongamos que la respuesta a la intervención en un PMM hemos decidido medirla en unidades de calidad de vida utilizando, por ejemplo, una de las versiones del MOS-SF (Medi-

cal Outcome Survey-Sort Form). Los sujetos participantes en la intervención fueron evaluados con este cuestionario en el momento de iniciar el programa y tras 6 ó 12 meses de permanencia en el mismo. Para poder llevar a cabo este tipo de diseño, precisamos de una baremación de ese cuestionario realizada con una población extensa de usuarios de PMM. Supongamos que en la comunidad donde se lleva a cabo el programa se realizó un estudio de adaptación de esta escala a población toxicómana (en tratamiento o no, con consumos o abstinentes), para lo cual se utilizó una muestra de 2.000 personas. A partir de los resultados de este estudio se establecieron baremaciones de las puntuaciones en calidad de vida en función de diversas variables (sexo, edad, tiempo de consumo, tiempo de abstinencia, etc.). Una vez que contamos con la baremación existen diferentes posibilidades de evaluación. Una es la apuntada más arriba; se transforma la puntuación media en el pretest en unidades percentiles equivalentes en la muestra baremada. Por ejemplo, la media obtenida en el pretest es de 50 sobre 100, equivalente a un percentil 30 de la muestra baremada, lo que equivale a decir que la puntuación media de los sujetos que inician un PMM no superan las puntuaciones obtenidas por al menos el 70% de los sujetos toxicómanos. Del mismo modo, la puntuación en el postest (65 puntos) es equiparada a su puntuación en el baremo (percentil 85, por ejemplo), lo que equivale a decir que se ha conseguido llegar a un nivel de calidad de vida equiparable al del 85% de la población. Sustrayendo el percentil pretest al postest obtenemos un efecto del tratamiento equivalente a una mejora del 55%. Obsérvese que se podría haber calculado la diferencia entre la puntuación media pre y postest, indicando una ganancia de 15 puntos, lo cual es indicativo de una mejora en calidad de vida, pero no nos ofrece una magnitud real del efecto; una magnitud es grande o pequeña cuando puede ser comparada con un patrón de referencia, y éste, en nuestro caso, es la baremación del grupo normativo.

Otra posible utilización de la baremación es identificar sujetos en riesgo en función de la puntuación ofrecida. Teniendo en cuenta que puntuaciones bajas en la escala implican peor calidad de vida, tomaremos un punto de corte para diferenciar sujetos en riesgo. Este punto de corte viene dado como la puntuación resultante de restar una desviación típica a la media de la distribución (equivalente a una proporción de la población de aproximadamente el 16%), aunque también se pueden establecer criterios de tipo clínico. De esta forma, un sujeto con puntuación inferior a la dada por el punto de corte estaría reflejando una situación de riesgo que implicaría una respuesta urgente de intervención sobre el caso.

Este tipo de diseño conlleva ciertos supuestos (equivalencia en las unidades de cambio, equivalencia de la población, ...) que pueden ser compensados, pero el más importante es que ha de disponerse de una muestra de datos normativos, es decir, que exista un baremo. En algunas medidas de psiquiatría, por ejemplo, medidas de ansiedad y depresión, existen baremaciones publicadas como resultado de los estudios de adaptación de los instrumentos a la población española, pero en otros ámbitos de medida en toxicomanías (ASI, etc.) todavía no han sido publicados.

Diseños basados en series temporales

Como ya ha sido comentado previamente, son dos básicamente las estrategias para aumentar la validez interna de un diseño: 1) buscar grupos controles lo más equivalentes posibles, y 2) producir observaciones múltiples, es decir, medidas repetidas, de modo que aumentemos el control intragrupo y/o intrasujeto. Los diseños basados en series temporales interrumpidas consisten en tomar una serie de medidas del criterio o respuesta a lo largo de un determinado periodo de tiempo, interrumpir la serie con la aplicación del tratamiento y continuar con otra serie de medidas del criterio. El posible efecto del tratamiento puede ser estimado en función de la discontinuidad que presentan las medidas tomadas antes y después de su aplicación, ya que las medidas previas y posteriores a la presencia de la intervención suelen presentar una tendencia o estructura regular. Al analizar la discontinuidad de las series de medidas, y observando la orientación de la nueva tendencia tras la intervención, se pueden llegar a conclusiones válidas acerca de su efecto. El hecho de realizar múltiples medidas de la respuesta no sustituye adecuadamente al control experimental, pero minimiza el efecto de algunas variables intervinientes que pueden afectar a la validez interna. Asimismo, pueden utilizarse de forma aditiva otras estrategias para aumentar dicha validez[6]: 1) introducir un grupo de control no equivalente, 2) aumentar la frecuencia de las mediciones, 3) recoger en la serie temporal diferentes variables dependientes no equivalentes, 4) realizar series temporales con intervenciones alternantes, primero en un grupo y luego en otro, 5) con tratamientos múltiples, o 6) con retirada del tratamiento. Según Campbell y Stanley[6], con estos diseños se pueden controlar la mayoría de las fuentes de invalidez interna (tabla I), y sugieren su utilización en aquellas situaciones en las que se lleva acabo un registro

periódico de las respuestas de los sujetos como parte de un procedimiento regular de actuación (por ejemplo, la práctica de registros de conducta propio de la terapia conductual).

Relacionado con este tipo de diseños, dada la aplicación múltiple de medidas de resultado, queremos destacar por su idoneidad en la evaluación terapéutica los diseños llamados intrasujeto o de sujeto único (N=1), que por sus características habría que clasificarlo como diseño experimental[11] y no en la línea cuasi-experimental de los expuestos previamente. El procedimiento a seguir es el siguiente:

1. Especificar las características del sujeto. Ha de tenerse en cuenta que la capacidad de generalización de los resultados de un estudio de N=1 dependerá del conocimiento exacto de las características del sujeto empleado. Por ello, es necesario realizar una descripción lo más completa posible, tanto de la conducta o enfermedad en estudio en el momento de la intervención como de su historia y circunstancias.

2. Medir la conducta o fenómeno antes del tratamiento. Los cambios deben ser seguidos paso a paso utilizando medidas repetidas, lo que implica el empleo de operaciones claramente especificadas y repetibles por un mismo investigador y realizadas siempre bajo las mismas condiciones. Con ello se establece una línea base que proporcionará un punto de comparación del cambio producido tras la intervención, pero que para ser válida ha de cumplir dos condiciones: que sea suficientemente larga y estable. Se considerará adecuada cuando emerja una tendencia clara.

3. Implementación del tratamiento y registro repetido posterior de la conducta o fenómeno.

4. Con el fin de ampliar la generalización de los resultados de un experimento intrasujeto, éste debe repetirse, en primer lugar, con varios sujetos similares; a continuación con varios sujetos de otras características, y, finalmente, en otras situaciones o con otros terapeutas.

En la terminología del diseño intrasujeto, la línea base se especifica como A, y el tratamiento como B. En el esquema propuesto el diseño sería del tipo ABA, pero existen otras posibilidades de diseño en función de la combinación de las condiciones A y B. Así, el más sencillo sería el AB, donde se establece la línea base y se aplica el tratamiento; el ABAB, donde se alternan las condiciones de línea base y tratamiento de manera secuencial; o el AB_1AB_2A, donde se alternan dos tratamientos diferentes entre tres fases de medición de la conducta.

El texto de Arnau[12] constituye una excelente referencia para una aproximación más exhaustiva a estos diseños; asimismo Cajal[13] realiza una descripción detallada de este tipo de diseño aplicado al área de las toxicomanías, aunque con datos ficticios, donde el lector interesado podrá ampliar información al respecto.

Control mediante técnicas estadísticas

En la figura 1, al hacer referencia a las etapas de control, además de fijar la correspondiente a la protocolización y diseño de estudio, se proponía una segunda etapa de control estadístico para el ajuste de aquellas posibles perturbaciones ocasionadas por variables intervinientes identificadas pero que no se pudieron controlar en la fase de diseño.

Las disciplinas de la probabilidad y la estadística nos ofrecen herramientas útiles y efectivas para el tratamiento de los datos, pero como herramientas al servicio de la evaluación tienen un carácter funcional y no un sentido en sí mismas. Utilizar procedimientos estadísticos complejos que van más allá de lo que plantean las hipótesis, no sólo es ir en contra del principio de parsimonia, sino pretender cambiar el sentido de la evaluación. Una evaluación no es mejor por utilizar las últimas y más intrincadas técnicas estadísticas, sino por dar respuesta de forma sencilla y clara a la(s) hipótesis propuesta(s). La eficacia de la estadística depende de la calidad de los datos a analizar y de la correcta aplicación e interpretación de las pruebas empleadas, y no de las filigranas que se pueden llegar a hacer con ella.

A este respecto, quisiéramos hacer un comentario, ya apuntado por otros autores[7,14-15], sobre el sentido de la estadística. Es un error generalizado confundir la significación estadística con la significación clínica o científica. Un resultado puede ser estadísticamente muy significativo y carecer por completo de relevancia clínica. En el análisis estadístico, un valor de p pequeño ($p < 0,05$) sólo informa de la existencia de una diferencia entre los grupos o de una asociación entre variables, y de que muy probablemente esta diferencia no es debida al azar. Es decir, la expresión 'muy significativo' es un término estadístico que se utiliza para indicar que la hipótesis nula es muy poco verosímil, y nada tiene que ver con la importancia clínica, biológica o psicológica de la hipótesis[15]. En nuestro caso, la respuesta a la pregunta de si las diferencias halladas son debidas al efecto de la intervención terapéutica dependerá del diseño correcto del estudio, y no de la significación estadística encontrada. El verdadero in-

terés de la «p» es el de permitir descartar que la diferencia observada es fruto de la casualidad[7].

La expresión muy significativo tampoco tiene nada que ver con la magnitud del efecto ni con la intensidad de la relación entre las variables. Un estudio en el que se obtenga una $p < 0,001$ no quiere decir que la asociación encontrada sea más fuerte (o la diferencia más importante) que otro estudio en el que la «p» sea igual a 0,05; sólo quiere decir que es más improbable que su resultado sea debido al azar. Por ejemplo, en una muestra de 1000 sujetos se ha encontrado una asociación entre dos variables de $r = 0,104$ con un valor de $p < 0,001$, en cambio en una muestra de 10 sujetos la correlación entre dos variables ha sido de $r = 0,497$ y la probabilidad asociada de $p > 0,10$. En el caso de asociación entre variables la magnitud del efecto no viene dado por los valores de p, sino por el coeficiente de determinación (R^2-cuadrado de las correlaciones), que en el caso de la muestra de 1000 sujetos nos informa que la varianza común entre las variables en estudio es del 0,01% (que era estadísticamente muy significativa) y en la muestra de 10 sujetos del 0,247% (cuya significación estadística estaba por encima del nivel de confianza convencionalmente admitido ($\alpha \leq 0,05$). En definitiva, los valores de p no son una medida de la fuerza de la asociación.

Consecuencia de esta confusión es que los informes y publicaciones científicas están cargados de pruebas que confirman que los hallazgos son estadísticamente significativos, pero suele ser menos común que informen sobre el tamaño de los efectos obtenidos. Sobre las medidas del tamaño del efecto y las técnicas de estimación del cambio terapéutico entraremos en profundidad en un próximo artículo dedicado a dar respuesta al segundo de los objetivos de la evaluación que propusimos en la introducción, por lo que no entraré en detalles sobre el tema en estas líneas. No obstante, considero importante hacer notar a los lectores las siguientes observaciones. Un estudio que concluya que un efecto no es estadísticamente significativo puede estar cometiendo un error de tipo II o β, especialmente si falta potencia a causa de trabajar con muestras pequeñas; es decir, se dice que no hay efecto cuando en realidad si existe. Un modo de solucionar esta deficiencia es presentar los hallazgos indicando un rango de valores del efecto que son creíbles a partir de los datos recogidos en el estudio, y este rango no es otra cosa que el intervalo de confianza expresado al nivel de fiabilidad/confianza elegido (del 90, 95, 99%). Un intervalo de confianza expresa mejor la precisión de los resultados; cuanto más estrecho sea el intervalo, más preciso será el hallazgo. Por otra parte, el intervalo de confianza también nos ofrece información sobre la significación estadística del efecto, dado que si el intervalo no contiene el valor nulo, aquel propuesto por la hipótesis nula de no diferencias ($\mu_1 = \mu_2$) o no asociación ($\rho_1 = \rho_1$; RR = 1), nos está indicando que el rango de valores obtenidos corresponden a valores implícitos en la hipótesis alternativa. La indicación a los evaluadores terapéuticos es que en sus informes, además de presentar los resultados de las pruebas estadísticas, incorporen datos sobre la magnitud de los efectos, bien a partir de estimadores concretos (R^2, Eta^2, σ^2 de Yules, δ de Glass, 'd' de Cohen, la Puntuación de Cambio Precisa (PCP) de Jacobson y Truax, etc.) o bien a través de la presentación de los intervalos de confianza de los parámetros estudiados.

Respecto a las técnicas o pruebas estadísticas a utilizar en la investigación evaluativa, existe un amplio repertorio desarrollado de forma profusa por un buen número de manuales[15-16], por lo que no entraremos en la descripción pormenorizada de las mismas. No obstante, nos permitimos recoger en la tabla 5 aquellas que serían de elección cuando hemos de optar por técnicas multivariadas para el control de variables intervinientes en diseños cuasi-experimentales.

No queremos concluir este apartado sin hacer algunas observaciones que atañen a la elección de las pruebas estadísticas multivariadas a utilizar. Cada una de las técnicas recogidas en la tabla 5 supone el cumplimiento de ciertos requisitos que deben ser verificados antes de proceder con la prueba. El incumplimiento de esas exigencias conlleva una inadecuada indicación de la prueba obteniéndose de ella resultados inestables. Por ejemplo, procedimientos como la correlación de Pearson o la regresión múltiple se basan en la linealidad de las relaciones entre variables; el incumplimiento de este supuesto forzaría a una transformación de los datos en aras a obtener la linealidad o bien a la elección de otras técnicas.

Por otro lado, existe una tendencia en las ciencias de la salud a utilizar datos de tipo nominal (éxito *vs* fracaso terapéutico, la asignación de los grupos a evaluar, etc.) o a dicotomizar variables que en origen son de tipo continuo (escalas de depresión, p.ej., categorizadas como depresión grave o leve-moderada), forzando a utilizar procedimientos estadísticos no paramétricos, menos potentes que los paramétricos. Efectivamente, existen variables que no pueden ser medidas sino atendiendo a escalas nominales u ordinales, pero aconsejamos a los investigadores que en la medida de lo posible consideren variables, tanto de re-

Tabla V. Técnicas estadísticas de análisis multivariado.

Técnica	N.º VD	Tipo de VD	N.º VI	Tipo VI	Objetivo
Análisis de la covarianza	1	Cuantitativa	q	Cualitativas	Determinar si las diferencias entre las medias de la VD en los grupos establecidos por las combinaciones de los valores de las VIs son estadísticamente significativas.
MANOVA	p	Cuantitativa	q	Cualitativas	Determinar si las diferencias entre las medias de las VDs en los grupos establecidos por las combinaciones de los valores de las VIs son estadísticamente significativas.
MANOVA de medidas repetidas	p	Cuantitativas	—	—-	Determinar si las diferencias entre las medias de las VDs son estadísticamente significativas.
MANOVA intra e intersujetos	p	Cuantitativas	1	Cualitativas	Determinar si las diferencias entre las medias de las VDs en los grupos establecidos por los valores de la VI son estadísticamente significativas.
Regresión lineal	1	Cuantitativa	q	Cuantitativas y Dummy (Ficticias)	Estimar, mediante una función lineal de las VIs, el valor de la VD.
Análisis Discriminante	1	Cualitativa		Cuantitativas y Dummy (Ficticias)	Estimar, mediante funciones lineales de las VIs, la probabilidad de que cada individuo pertenezca a cada uno de los grupos establecidos por los valores de la VD.
Regresión Logística	1	Cualitativa-Dicotómica	q	Cuantitativas y Cualitativas	Estimar, mediante una función lineal de las VIs, la probabilidad de que cada individuo pertenezca a cada uno de los dos grupos establecidos por los valores de la VD.
Modelos de respuesta Probit	1	Cualitativa	q	Cuantitativas	Supuesto que los dos valores de la VD corresponden a la presencia o ausencia de respuesta frente a uno o más estímulos (VIs), estimar, mediante una combinación lineal de las VIs la probabilidad de la respuesta para los distintos niveles de las VIs.
Métodos Actuarial y de Kaplan-Meier	1	Tiempo que transcurre hasta que ocurre un desenlace	—	—	Estimar, en función del tiempo, la probabilidad de que ocurra un desenlace.
Regresión de Cox	1	Tiempo que transcurre hasta que ocurre un desenlace	q	Cuantitativas y Dummy (Ficticias)	Estimar, en función del tiempo, y mediante una función lineal de las VIs, la probabilidad de que ocurra un desenlace.
Modelos Loglineales	—	—	q	Cualitativas	Obtener un modelo lineal para los logaritmos de las frecuencias de la tabla de contingencia múltiple correspondiente al cruce de los valores de las q variables, con la finalidad de interpretar las relaciones entre ellas.
Series temporales, Modelos ARIMA	Medidas múltiples	Algún tipo de Indicador	—	—	Estudiar la evolución del indicador a lo largo del tiempo. Explicar la estructura de la serie y prever su evolución.

VD: Variable(s) Dependiente(s); VI: Variable(s) Independiente(s); MANOVA: Análisis Múltiple de la Varianza.

sultado como independientes, que puedan ser medidas en una escala continua. Si los intereses de la investigación aconsejan establecer categorías, siempre será factible hacerlo desde una variable continua atendiendo a puntos de corte que establezcan los límites de las categorías; el procedimiento inverso, pasar de una variable nominal a una continua, es imposible. Asimismo, hemos de considerar que la categorización de las variables supone una pérdida de información que puede ser vital para la explicación de determinados fenómenos. Una pérdida de información, siempre supone una pérdida de precisión.

Otra cuestión importante a tener en cuenta cuando utilizamos diseños cuasi-experimentales apoyados en pruebas multivariadas como forma secundaria de control es el número de variables que introducimos en los modelos. Un elevado número de predictores o variables intervinientes tienden a incrementar la probabilidad de un hallazgo significativo que es falso, es decir, a incurrir en el error de tipo I o error α. Una forma de controlar la inflación de α (nivel de confianza) es atender a análisis multivariados que contemplan interacciones entre variables o atendiendo a correcciones como la de Bonferroni (dividir el valor «p» por el número de comparaciones realizadas). Sin embargo, estas estrategias son de utilidad limitada para la mayoría de las investigaciones clínicas ya que se basan, en la mayoría de los casos, en un tamaño muestral inadecuado para estadísticas multivariadas (la recomendación de los metodólogos es incluir de 10 a 20 sujetos por variable interviniente en el modelo), teniendo como efecto una pérdida de poder dadas las drásticas reducciones de alfa. En estos casos, el control de la inflación de alfa nos impedirá identificar relaciones potencialmente importantes (error de tipo II). Consecuentemente, los investigadores interesados en la evaluación de resultados terapéuticos han de procurar contar con muestras suficientemente amplias, y por ello sería conveniente confluir en modelos comunes de evaluación que se realicen de forma multicéntrica, favoreciendo de este modo la ampliación del tamaño muestral.

Recomendaciones para la evaluación de resultados terapéuticos

1. Dado que los procesos terapéuticos están encuadrados dentro de una dimensión temporal, su evaluación no tiene por menos que basarse en estudios longitudinales si pretendemos que ésta sea efectiva y fidedigna. Asimismo, en tanto aspiramos a que los resultados de la evaluación puedan ser guías de actuación para intervenciones posteriores, los estudios que analizan esos resultados han de basarse en el mayor control posible de variables intervinientes y, por consiguiente, han de elegirse los diseños de tipo prospectivo. En definitiva, la primera condición necesaria, aunque no suficiente, para llevar a cabo una evaluación terapéutica eficaz con intención de generalizar la intervención es que sea encuadrada en diseños longitudinales prospectivos.

2. Con la evaluación terapéutica buscamos datos que nos permitan tomar decisiones: ¿estamos llevando a cabo la indicación más adecuada a las característica de esta persona y su problema?, ¿actúa mejor este tratamiento que otros alterativos?, en definitiva, ¿qué intervención es la más conveniente? Tomar una decisión para responder a estas preguntas supone tener datos comparados de los diferentes tratamientos alternativos para considerar aquél más eficaz, efectivo (útil) y/o eficiente en cada caso particular. Si la decisión se basa en datos comparados, entonces no podemos prescindir de la utilización de grupos de tratamiento alternativo o controles. Sólo a través de la comparación con éstos podremos estimar la eficacia relativa del tratamiento propuesto y la ganancia en salud obtenida con él. La segunda condición, también necesaria, es realizar nuestra evaluación considerando grupos de control.

3. Uno de los objetivos de la evaluación es obtener pruebas de la eficacia del tratamiento implementado, para lo cual hemos de extremar nuestro cuidado en obtener el mayor grado de validez interna en el estudio de evaluación. Como ha sido comentado previamente, el diseño que más se acerca a este requisito es el ensayo clínico aleatorio, y nuestra recomendación sería utilizarlo siempre que fuera posible. No obstante, la posibilidad de evaluar todas nuestras actuaciones a través de este diseño es, en la mayoría de los casos, inviable. La alternativa, es considerar diseños cuasi-experimentales, como algunos de los reseñados más arriba. La recomendación en el caso de utilizar esta vía, es replicar en la mayor medida posible el esquema de evaluación en distintos centros y contextos sanitarios. Un protocolo de evaluación bien programado que tenga en cuenta el conjunto de variables intervinientes, donde el procedimiento de actuación terapéutica esté bien pautado y que las técnicas de recogida de información se lleve de forma metódica, aplicado de forma multicéntrica y en contextos socioculturales diversos contribuirá no sólo a aumentar la validez externa o poder de generalización de los resultados, sino también a poder llevar a cabo un mayor control de la validez interna a través de un mayor poder estadístico

de los hallazgos. Por otra parte, al aumentar el número de participantes podemos obtener muestras grandes que permitan la baremación de resultados como estándares de comparación, y además cabría esperarse un mayor impacto positivo en las actividades médicas de los centros participantes.

4. Recordamos, en referencia a este último comentario, que la estadística no ha de utilizarse como herramienta al servicio de «nuestro» propósito de demostrar la bondad de «nuestra» actuación. La estadística es ciega, pero quien interpreta los resultado no. Como dice el *adagio*, «hay quien utiliza la estadística como el borracho la farola, más para apoyarse que para iluminarse». La estadística nos permite el control de las relaciones entre variables cuando se conoce la dirección de las posibles interacciones, y el único dato que nos ofrece es referente a la mayor o menor verosimilitud de los resultados hallados; nada nos dice sobre la relevancia clínica o científica de esos resultados. Nuestra recomendación a este respecto, es utilizar la estadística como lo que es: una herramienta que se pone al servicio de los objetivos preprogramados. Es decir, son nuestros objetivos e hipótesis los que determinan las técnicas estadísticas a emplear, y no la acomodación de éstas a los datos para verificar hipótesis a *posteriori*.

5. Por último, y quizá por ello la recomendación más importante. Una evaluación es un proceso que requiere ir de una fase a otra de una forma consecutiva. En primer lugar se precisa de un conocimiento sustantivo de lo que ahora se viene llamando el «estado del arte» del problema a evaluar. Este conocimiento orientará los objetivos e hipótesis del equipo investigador/evaluador, los cuales determinarán el conjunto de variables intervinientes en el fenómeno a estudiar. En primera instancia, el control de estas variables intervinientes vendrá determinado por las característica del diseño a elegir; es decir, todo aquello

Bibliografía

1. Jenicek M. Epidemiology: The Logic of Modern Medicine. Montreal: Epimed International; 1995.
2. Alvira F. Metodología de la evaluación de programas. Cuadernos Metodológicos, n.º 2. Madrid: Centro de Investigaciones Sociológicas; 1991.
3. Hunt S. Measuring health in clinical care and clinical trials. En Teeling-Smith G. (ed.): Measuring health: A practical approach, 7-20. London: John Wiley & Sons; 1988.
4. Roca J. Cómo y para qué hacer un protocolo. Med Clin (Barc) 1996;106:257-62.
5. Food and Drug Administration (FDA). General consideration for the clinical evaluation of drug. Washington: U.S. Governement Printing Office; 1977.
6. Campbell D, Stanley J. Diseños experimentales y cuasi-experimentales en la investigación social. Buenos Aires: Amorrortu; 1988.
7. Argimon JM, Jiménez-Villa J. Métodos de investigación aplicados a la atención primaria. Barcelona: Doyma; 1994.
8. Rubio-Terrés C. Diseño estadístico de ensayos clínicos. Med Clin (Barc) 1996;107:303-9.
9. Galende I, Sacristán JA, Soto J. Cómo mejorar la calidad de los ensayos clínicos. Med Clin (Barc) 1994;102:465-70.
10. Tallmadge GK. An empirical assessment of norm-referenced evaluation methodology. Journal Educational Meassurement 1982;19:97-112.
11. Cabello JB, Abraira V, Gómez-García J. El ensayo clínico para un solo paciente. Justificación, metodología y aportaciones bioéticas. Med Clin (Barc) 1997;109:592-8.
12. Arnau J. Diseños experimentales en psicología y educación. Vol. II. México: Trillas; 1985.
13. Cajal B. Análisis de datos longitudinales de un único sujeto. Adicciones 1997;9:109-27.
14. Guttman L. What is not what statistics. Statistician 1977;26:81-107.
15. Doménech JM. Métodos estadísticos en ciencias de la salud. Barcelona: Signo; 1996.
16. Stevens J. Applied multivariate statistics for the social science. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1992.
17. Conde JL. Evaluación de tecnologías médicas basada en la evidencia. Madrid. Agencia de Evaluación de Tecnologías Sanitarias. Instituto de Salud Carlos III. [Documento electrónico]. Diciembre de 1998. Disponible en: http://www.isciii.es/unidad/eat/doc/docconce.html.
18. Rodríguez-Pulido F, Sierra-López A. La investigación epidemiológica en las drogodependencias. Las Palmas; ICEPSS; 1995.
19. Rothman K. Epidemiología moderna. Madrid: Díaz de Santos; 1987.

Bibliografía recomendada. El lector interesado en la investigación clínica y evaluación terapéutica puede encontrar un material básico y con claridad expositiva en las referencias 1, 2, 3, 6, 7, 12, 14,17, 18 y 19.