De la opinión a la evidencia en ortopedia

MIGUEL MARÍA SÁNCHEZ MARTÍN
HOSPITAL CLÍNICO UNIVERSITARIO. FACULTAD DE MEDICINA. UNIVERSIDAD DE VALLADOLID.

Resumen. En el momento presente, la medicina basada en la evidencia es un imperativo para bien de la ciencia y de la mejor selección de tratamientos para nuestros pacientes. La realización de ensayos controlados aleatorizados precisa del conocimiento de principios de epidemiología y bioestadística. Además, obliga a utilizar información completa e imágenes que proporciona internet. Finalmente, es imprescindible que el cirujano moderno valore con precisión los datos de los artículos de revista basados en la evidencia.

From the opinion to the evidence in orthopedics

Summary. The based-evidence medicine is a necessity at present time, not only for science but in every way to choose the best traitement for our patients. To do a controlled aleatorized assay we need to know the principles of epidemiology and biostatistics. Besides, it needs to use complete information and images through the Internet. The surgeon must know with precision the data offered by journals' papers.

Correspondencia:
MM. Sánchez Martín
Regalado, 13, 6º
47002 - Valladolid
Telf. 983 300 121

Introducción

De manera tradicional, el cirujano ortopédico solamente necesitaba reunir las cualidades de hablar, leer, realizar exploraciones físicas y operaciones para mantenerse al más alto nivel. Sin embargo, el cirujano de hoy ya no puede ignorar por más tiempo la era electrónica, procedente de la tecnología digital. La aptitud de inclinarse sobre un ordenador, navegar algún sitio de la red, utilizar algún medio de investigación, comunicarse por correo electrónico y montar y llevar una presentación en power-point, se ha hecho esencial. Manejar bien estos elementos proporcionará un conocimiento más fácil y permitirá una comunicación más eficaz. La mayor parte de los cirujanos adquieren estas habilidades mediante experimentación o cambios uno a uno, cuyo resultado es una formación incompleta.

Reconociendo estas necesidades, hay varios avances como fotografía digital, dispositivos electrónicos, relación interactiva y consulta de revistas en páginas web y comunicación electrónica para publicación de artículos, que el cirujano debe conocer y practicar[1].

En los últimos años en la literatura ortopédica han aparecido numerosos nuevos términos en forma de acrónimos y abreviaturas, como CI, CONSORT, coste-utilidad, NNT, EBM, validez de enfrentarse al asunto, GCP, nivel de evidencia, coeficiente kappa, metaanálisis, análisis de poder, RCT..., que son mal conocidos por la mayoría de los cirujanos ortopédicos. Una reacción no infrecuente es ignorarlos, diciendo que "no representan nada en la práctica clínica". Otros se oponen a ello enérgicamente. Sin embargo, han llegado para quedarse. Representan parte de la evolución de la medicina, desde la opinión a la ciencia. Aunque no todos los cirujanos ortopédicos tienen que ser científicos, tenemos que ser capaces de entender la literatura clínica para ser capaces de sugerir los mejores tratamientos a nuestros pacientes. Esto requiere conocimiento de lo esencial de la investigación clínica moderna.

Se ha creado el Grupo de Trabajo Internacional de Cirugía Ortopédica basada en la evidencia (ieboswg a gmail.com) formado por cirujanos ortopédicos y epidemiólogos, con un entusiasta interés por la metodología de estudio y la ortopedia basada en la evidencia. El grupo se fundó en Canadá y luego se ha extendido a Holanda y los Estados Unidos de América con una expansión mundial

continuada. En 2006, el grupo inició un esfuerzo internacional para promover el enfoque basado en la evidencia de los dilemas ortopédicos, resumiendo la evidencia disponible en revisiones sistemáticas y escribiendo revisiones de formación para ayudar a los cirujanos ortopédicos a comprender mejor las herramientas precisas para la práctica basada en la evidencia. El grupo también promueve el diseño y dirección de una investigación de alta calidad, así como amplios estudios multicéntricos a escala internacional para contestar a importantes cuestiones de cirugía ortopédica. Últimamente, el grupo tiene como objetivo trasladar el paradigma de "la práctica basada en la opinión en práctica basada en la evidencia"[2].

CAPÍTULO I.
EPIDEMIOLOGÍA Y BIOESTADÍSTICA

Principios de epidemiología

La investigación epidemiológica se basa en la colección sistemática de observaciones relacionadas con el fenómeno de interés en una población definida. Estos datos se someten a cuantificación, es decir, medición de variables al azar, estimación de parámetros de población y comprobación estadística de hipótesis[3].

La epidemiología es la disciplina biomédica que se centra en la distribución y determinantes de enfermedad en grupos de individuos que se espera tengan algunas características, exposiciones o enfermedades en común. La epidemiología es el fundamento de la ciencia básica de salud pública considerada como estudio de la distribución y determinante societarias.

El cirujano ortopédico, en calidad de clínico científico, se preocupa de las causas de enfermedad y de determinar qué tratamiento es más beneficioso para su paciente, y diferenciarlo de otros perjudiciales[4].

Las observaciones y mediciones forman las unidades fundamentales de datos. La calidad de los datos se describe habitualmente utilizando estos cuatro términos: exactitud, precisión, fiabilidad y validez. *Exactitud* es el grado que representa la medición del valor cierto del atributo que se está mediendo[4]. Decir que un hombre es obeso puede ser una observación exacta; decir que pesa 140 kilos, es preciso. La precisión es la cualidad de definir justamente un detalle exacto. *Fiabilidad* es la medida de lo digno de confianza que es una observación, exactamente cuando se repite; se refiere al método de medir más que al atributo de ser medido. Fiabilidad no es sinónimo de repetitividad o reproducibilidad; es, más bien, un término más amplio que incluye el concepto de consistencia, que se refiere a lo ajustado de los hallazgos en diferentes muestras o poblaciones conforme una a otra bajo diferentes situaciones o momentos.

Una prueba se considera *válida* en términos epidemiológicos si mide lo que se pretende medir. Cuando los resultados obtenidos de un estudio están distorsionados por sesgo en el diseño del estudio o análisis de datos, el estudio pierde validez. Hay que hacer una distinción importante sobre los términos de validez interna y externa. La validez interna se refiere a interferencias sobre la población de individuos de interés restringido a partir de la cual se extrae una muestra para estudio; validez externa se refiere a interferencias sobre población externa por encima del interés del estudio restringido; por ejemplo, hay muchos métodos para controlar los factores que concurren en dismetría de las extremidades inferiores, Anderson y cols.[5] describen cuadros de predicción de lo que resta por crecer un miembro sobre la base de datos coleccionados en 100 niños (50 de cada sexo) en el Children's Hospital de Boston. Cincuenta y uno de estos niños eran normales y cuarenta y nueve habían sufrido poliomielitis, con afectación de una extremidad inferior, que no se incluyeron en el estudio. Moseley[6] opina que asumir que la longitud de las extremidades inferiores de todos los niños de una edad esquelética concreta tienen la misma proporción que la longitud de la longitud de los miembros de los mismos sujetos cuando alcancen la edad adulta, cualesquiera que sea su percentil o edad cronológica, no es fácil que sea cierto en niños de razas diferentes o en aquéllos que tienen hábitos familiares marcadamente diferentes. Para un ortopeda que viva en un continente distinto al de América del Norte, se puede cuestionar razonablemente la validez externa de los datos publicados por Anderson y cols.[5] y quiera repetir aquellos estudios en poblaciones más pertinentes antes de crear interferencias que podían utilizarse para tratar pacientes de su área local.

Diagnóstico de enfermedad

Antes de estudiar una enfermedad es necesario definirla y con frecuencia esto resulta en la definición de un caso que puede diferir de una definición clínica. Estudios en que se emplean definiciones diferentes pueden dar lugar a conclusiones diferentes, según que se pretenda un filtrado (screening), un diagnóstico o conocer su etiología.

Las definiciones que se basan en pruebas de identidad (prueba de Phalen en el síndrome del túnel carpiano, por ejemplo) deberían examinarse según cuatro parámetros: sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo (tabla 1). *Sensibilidad* se

	Estado actual de la enfermedad (Dx)		
Resultado de pruebas (T)	Positivo	Negativo	Total
Positivo	a	b	a + b
Negativo	c	d	c + d
Total	a + c	b + d	

$$\text{Sensibilidad} = \text{Probabilidad}\,(T^{+}/Dx^{+}) = \frac{a}{a+c}$$

$$\text{Especificidad} = \text{Probabilidad}\,(T^{-}/Dx^{-}) = \frac{d}{b+d}$$

$$\text{Valor predictivo}^{+} = \text{Probabilidad}\,(Dx^{+}/T^{+}) = \frac{a}{a+b}$$

$$\text{Valor predictivo}^{-} = \frac{d}{c+d}$$

Hay cuatro probabilidades condicionales: sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo. Las ecuaciones deberían interpretarse como en el ejemplo siguiente: sensibilidad= probabilidad de que el resultado de la prueba sea positivo, dado que dos cosas tratadas realmente tienen la enfermedad que la prueba se supone que detecta.

Tabla 1. Combinaciones posibles de si una enfermedad está o no presente y de si el resultado de una prueba diagnóstica es positivo o negativo.

refiere a la proporción de individuos en una población tratada que realmente tienen una enfermedad dada y se identifican como que la tienen. Sensibilidad se define como el número de resultados positivos ciertos dividido por la suma de resultados positivos ciertos y resultados falsos negativos. *Especificidad* se refiere a la proporción de individuos de una población tratada que no tiene esa enfermedad dada y que se identifican como que no la tienen. La especificidad se define como el número de resultados positivos ciertos divididos por la suma de los resultados verdaderos negativos y falsos positivos. En general, la sensibilidad está aumentada a expensas de la especificidad. Para pruebas de diagnóstico y de filtrado, la probabilidad de que una persona que tiene un resultado positivo ciertamente tenga una enfermedad dada se conoce como valor predictivo positivo, mientras que la probabilidad de que una persona que tiene un resultado ciertamente negativo no tenga enfermedad es un valor predictivo negativo. La sensibilidad, la especificidad y los valores predictivos son probabilidades condicionales. En conjunto, la definición de un caso concreto no debería ser útil como base para tomar decisiones terapéuticas o para determinar la causa de los síntomas, salvo que se incluyan pruebas específicas, por ejemplo, un electromiograma para el diagnóstico de síndrome del túnel carpiano.

Medidas epidemiológicas de enfermedad

En estudios epidemiológicos se comparan grupos de personas cuantitativamente respecto a alguna estructura de tiempo. La medición de enfermedad en un tiempo particular proporciona una tasa de prevalencia. *Prevalencia* es el número total de individuos que tienen una característica o enfermedad en un punto particular dividido por el número de los que están en situación de tener esta característica o enfermedad en este punto diseñado en tiempo[4]. La prevalencia depende del número de personas que han tenido la enfermedad o las características en el pasado y la duración de la enfermedad o característica. La medición de enfermedad durante un período de tiempo proporciona una tasa de incidencia. *La incidencia* es el número de casos nuevos de una enfermedad en una población definida dentro de un espacio de tiempo especificado, dividido por el número de los que pueden tener esta enfermedad o característica en este período de tiempo designado. Al intentar determinar si una exposición dada (como un traumatismo ocupacional repetido) influye en el desarrollo de una enfermedad dada (síndrome de túnel carpiano), una comparación de los individuos con y sin exposición aportan más información sobre si la enfermedad se debe a esta exposición, mientras que una determinación de prevalencia revela simplemente la tasa de enfermedad entre individuos que la tienen y que no han estado expuestos.

El concepto de asociación o dependencia estadística entre un factor y una enfermedad es fundamental para adscribir el factor como posiblemente causante. Una exposición o atributo que aumenta la probabilidad de una enfermedad es un factor de riesgo; por ejemplo, la variante ulnar negativa (cúbito corto) es una situación de riesgo para una enfermedad de Kienböck[7]. Los epidemiólogos calculan una medida de asociación como simple parámetro de resumen que estima la asociación entre una enfermedad y una determinada exposición. Un simple parámetro de resumen es una medición representativa en una población bajo estudio, justo como un medio matemático describe un grupo mediante representación de la media. De estas medidas de asociación se derivan dos conceptos: la probabilidad de un acontecimiento o enfermedad y la probabilidad (odds) de un acontecimiento o enfermedad. La probabilidad de un acontecimiento o de cualquier resultado de interés es la frecuencia relativa de este acontecimiento sobre un número infinito de ensayos al azar. La probabilidad (Pr) de un acontecimiento (E) es siempre mayor o igual a cero o menor o igual a 1, expresado como $0 \leq Pr(E) \leq 1$. Una probabilidad condicionada es la probabilidad de un acontecimiento (E) dado que otro acontecimiento (D) ha ocurrido, expresado como Pr(EyD) e igual a Pr(EyD)/ Pr(D).

Probabilidad (*odds*) es la ratio de la probabilidad de que ocurra un acontecimiento o de que no ocurra.

Las dos medidas de asociación frecuentemente utilizadas son el riesgo relativo y la odds ratio. El riesgo relativo indica el riesgo medio de enfermedad que se debe a la exposición dada en el grupo expuesto. Proporciona una estimación de la magnitud de asociación entre exposición y enfermedad; es la ratio del riesgo de enfermedad entre individuos expuestos con relación a los no expuestos. El riesgo relativo de un acontecimiento (E) dado por otro (D) es una probabilidad condicional expresada así: Pr(E/D) / Pr(E/no D). Si E y D son independientes, entonces el riesgo relativo es 1. Cuanto mayor sea la dependencia entre los acontecimientos más se alejará de 1 el riesgo relativo. El riesgo relativo puede calcularse sobre la base de estudios de corte transversal. La *odds ratio* es la ratio de la probabilidad de que una enfermedad ocurra entre individuos expuestos y no expuestos. La odds ratio se utiliza como medida del efecto en estudios -tal que estudios de caso control- en que las tasas de incidencia no pueden deducirse directamente. En estudios de caso control los participantes se seleccionan sobre la base de un estado de enfermedad; por tanto, no es posible calcular la tasa de enfermedad sobre la base de presencia o ausencia de exposición. La odds ratio y la ratio de riesgo son muy similares en ejemplos de enfermedad rara, pero bastante distintas cuando la prevalencia sobrepasa del cinco al diez por ciento.

Para facilitar el cálculo de estas medidas se presentan los datos epidemiológicos en tablas dos por dos (tabla 2) que pueden utilizarse, por ejemplo, para describir un estudio radiográfico de asociación entre variante ulnar y enfermedad de Kienböck (tabla 3). La odds ratio es la medida de asociación apropiada para este estudio de caso control, pero en este ejemplo no es muy diferente del riesgo relativo calculado[8]. Aunque los autores de este estudio encuentran marcada asociación (p= 0.0000) (la probabilidad de que se produzca el acontecimiento por casualidad sólo es al menos de 1/10.000) entre variante ulnar negativa y enfermedad de Kienböck, ellos cuidadosamente exponen "que la asociación no debe considerarse como relación etiológica primaria".

Interferencia casual

La preocupación de los epidemiólogos actuales es determinar la etiología de la enfermedad. La causa es un acontecimiento que, bien aisladamente o unida a otros elementos, produce una secuencia de acontecimientos que conducen a una consecuencia[9]. Con frecuencia se pregunta al ortopeda por la etiología de tal enfermedad, por ejemplo de nuevo la enfermedad de Kienböck, y resulta desesperante poder contestarlo, si bien, en cambio, podemos contestar a otras preguntas, tales como factores o eventualidades de riesgo, o el microtraumatismo repetido. El epidemiólogo quiere utilizar el término factor de riesgo en vez de causa que indique un atributo o exposición relacionado con un probable aumento de la

	Enfermedad	No enfermedad	Total
Exposición	a	b	a + b
No exposición	c	d	c + d
	a + c	b + d	

$$\text{Riesgo relativo} = \frac{a}{(a+b)} \Big/ \frac{c}{(c+d)}$$

$$\text{Odds ratio} = \frac{\frac{a}{(a+b)} \Big/ \frac{b}{(a+b)}}{\frac{c}{(c+d)} \Big/ \frac{d}{(c+d)}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

* La letra "a" representa a personas que tienen la enfermedad y se han expuesto a un factor de riesgo en estudio: dos medidas del efecto, el riesgo relativo y la odds ratio pueden calcularse utilizando este modelo de tabla.

Tabla 2. Tabla del dos por dos que muestra la relación entre la presencia de una enfermedad y la exposición a un factor de riesgo que pueda contribuir a la enfermedad*

	Enf. de Kienböck	No enf. de Kienböck	
Variante ulnar negativa	13	136	149
No variante ulnar negativa	2	340	342
	15	476	

$$\text{Riesgo relativo} = \frac{a}{(a+b)} \Big/ \frac{c}{(c+d)} = 14.9$$

95% de intervalo de confianza, 3.4 - 65.3; p = 0.0000

$$\text{Odds ratio} = \frac{\frac{a}{(a+b)} \Big/ \frac{b}{(a+b)}}{\frac{c}{(c+d)} \Big/ \frac{d}{(c+d)}} = a / b = ad = 16.25$$

95% de intervalo de confianza, 5.2 - 50.6; p= 0.0000

* El riesgo relativo, la odds ratio, 95% de intervalo de confianza y los valores p pueden calcularse manualmente (como se muestra para el factor riesgo y la odds ratio) o más fácilmente con un paquete ordenador estadístico (como se da para los intervalos de confianza y los valores p). El programa de ordenador interpreta los valores p sólo para cuatro decimales, de ahí que p= 0.0000 indica que la probabilidad del acontecimiento es al menos menor que 1 en 10.000.

Tabla 3. Ejemplo de utilización de formato de tabla dos por dos mostrado en la tabla 2: investigación de la relación entre un factor de riesgo postulado (variante ulnar negativa) y una enfermedad conocida (enfermedad de Kienböck)*

enfermedad. Además, para que sea considerado factor de riesgo debe preceder a la aparición de la enfermedad y la observación asociada no tiene que deberse a problemas de diseño del estudio o de análisis de datos.

La contingencia casual importante o primaria puede deberse al punto de vista de cada uno, como cuando, por ejemplo, un motorista a gran velocidad es atropellado por un automóvil y sufre múltiples fracturas y a continuación muere de síndrome de distress respiratorio del adulto; en la unidad de vigilancia intensiva, el anatomopatólogo puede atribuir la causa de muerte a edema pulmonar intersticial, el internista a asistolia cardíaca, el cirujano ortopédico a retraso en estabilizar las fracturas diafisarias y pelvianas, la Asociación de Madres contra la conducción embriagado (Mothers Against Drunk Driving= MADD) a la intoxicación por alcohol, y el funcionario que ejecuta la ley a exceso de velocidad. Cada uno selecciona la causa según su opinión; de ahí que haya que buscar la causa entre todas. Pero en la práctica, si el objetivo es prevenir la enfermedad, es más beneficioso orientar el factor causal remoto, en cierto modo, de la enfermedad. Así, es mejor disminuir, por ejemplo, la frecuencia de la tuberculosis en los países subdesarrollados mejorando las condiciones de vida que aplicando vacunación, quimioterapia o cirugía.

Los epidemiólogos se preocupan, de manera excepcional entre otros profesionales, más por teorías filosóficas que puramente técnicas del método experimental, ya que los experimentos juegan menor papel en el análisis de fenómenos que ocurren naturalmente. Los argumentos razonados pueden venir bien de general a particular (deductivos) o de particular a general (inductivos). La idea tradicional de la ciencia es que la inducción-formación de una hipótesis basada en la observación es fundamental para el método científico. El fallo de una hipótesis que deriva de inducción es que puede refutarse con la primera observación que pruebe una excepción. Por ejemplo, si por inducción creemos que el corazón siempre se encuentra colocado en el lado izquierdo del tórax, nuestra opinión puede ser destruida en un paciente con dextrocardia. En resumen, aparte disquisiones filosóficas, los epidemiólogos emplean ambos métodos, deductivos e inductivos. La variación entre puntos de vista con relación a la causalidad está enraizada en la variación entre varios puntos de vista filosóficos. No obstante, el establecimiento de la causa es un proceso en continua evolución.

Interferencia estadística

Es parte de la base epidemiológica, ya que las observaciones estudiadas están sujetas a fluctuaciones al azar. La comprobación de hipótesis es un procedimiento estadístico para determinar la probabilidad de que los datos coleccionados sean consecuentes con las hipótesis específicas que se estén investigando. La hipótesis es el estudio de la asociación entre, de nuevo, variante ulnar y enfermedad de Kienböck, por ejemplo, y puede señalarse como H_A (hipótesis alternativa) y establecer que hay una diferencia en la prevalencia de enfermedad de Kienböck entre personas que tienen una variante ulnar negativa comparada con aquéllos que la tienen neutra o positiva. Lo opuesto a H_A es una hipótesis H_O que estima que no hay diferencia en la prevalencia de enfermedad de Kienböck entre personas que tienen variante ulnar negativa comparada con las que la tienen neutra o positiva. Convencionalmente, el investigador busca negar ambas H_O; de ahí que se considere como hipótesis nula. H_A y H_O se excluyen mutuamente y exhaustivamente, es decir que una u otra pueden ser ciertas pero ambas pueden no serlo. Por ello, si H_O es negada, entonces H_A es afirmada. H_A y H_O se refiere a la población entera, incluso aunque los datos sean disponibles solamente de una muestra de pacientes (116).

La *prueba* significativa se basa en el cálculo de una prueba estadística (por ejemplo, el valor t, z o chi-square) y sobre algunos supuestos teóricos, uno de los cuales es que la hipótesis nula es cierta y otra que este muestro inciertamente es el azar. Basándose en estos supuestos, puede calcularse lo similar o no similar que el resultado de la muestra podría ser. Por tanto, la prueba estadística es un número que compara los valores observados y esperados del parámetro que se está midiendo bajo hipótesis nula. La prueba significativa abarca el fundamento de que diferentes valores de prueba estadística, como que los supuestos de la prueba, incluida la hipótesis nula, son ciertos y un raro acontecimiento ha surgido o uno de los supuestos no es cierto y, concretamente, la hipótesis nula es falsa[10]. El punto en que un valor de prueba estadística es bastante raro para que se garantice pueda ser rechazado de la hipótesis nula se determina por convención o pacto, pero es colocado en el valor que podría ocurrir no mayor de 5 a 1 por ciento. Este valor absoluto que debe superar en orden a la hipótesis nula para ser rechazado se denomina valor crítico. La probabilidad de que el valor de la prueba estadística sea demasiado pequeño para ser compatible con que HO sea cierto se conoce como valor de significación o valor alfa; se coloca de manera convencional a 0.05 ó 0.01 y se llama comúnmente valor p. En resumen, si la prueba estadística representa un acontecimiento menor de 5 (0.1) por ciento del tiempo bajo muestreo al azar si la hipótesis nula fuera cierta, enton-

ces el resultado se considera significativo y la hipótesis nula se rechaza a favor de H_A o hipótesis alternativa.

Si se han utilizado muchas pruebas significativas según la naturaleza y el tipo de distribución (discreta, continua, categoría, etc.) de los datos que se están analizando, con frecuencia un investigador selecciona la prueba apropiada, calcula la prueba estadística (por ejemplo, el valor t, z o chi-square) y el nivel significativo (el valor t) utilizando datos recogidos y rechaza o no consigue rechazar la hipótesis nula. Una advertencia consiste en estar atento a lo que Sackett[11] denomina "sesgos de dragado de datos". Cuando se analizan datos para todas las posibles asociaciones estadísticas sin una hipótesis específica, la posibilidad de que un investigador encuentre una asociación significativa, por suerte sólo aumenta en tanto en cuanto el número de pruebas estadísticas que son realizadas aumenta. Los análisis de datos sin planificar deberían identificarse como tales. Como Lang y Rhodes[12] mantienen: "si la expedición de pesca engancha una bota, el pescador no debería reivindicar que eran pescadores de botas".

Hay dos *tipos de errores* relacionados con la prueba significativa. Es posible rechazar la hipótesis nula cuando ha ocurrido un raro acontecimiento, incluso aunque la hipótesis nula sea realmente cierta. A esto se denomina error tipo I. En todos los casos en que una hipótesis nula es cierta se produciría un *error tipo I* ($100x\alpha$) por ciento del tiempo, en que α es igual al nivel significativo (usualmente 0.05 ó 0.01). En un *error tipo II*, la hipótesis nula es falsa pero la prueba estadística calculada no es significativa y, por tanto, está sentenciada a ser compatible aunque la hipótesis nula sea cierta. La hipótesis nula es, por tanto, aceptada como error. La relativa frecuencia con que un error tipo II se produce se simboliza con ß. Un experimento que se utiliza habitualmente está diseñado para controlar la probabilidad de ß, para que sea menor de 0.20. Poder es el complemento de un error tipo II o la probabilidad de que la hipótesis nula sea rechazada cuando es ciertamente falsa y es igual a 1-ß. Poder es una función del nivel de significado, de la fiabilidad de los datos de muestra (el grado de difusión de los datos o la desviación estándar) y del tamaño del efecto experimental. Un análisis de poder debe ser tenido en cuenta cuando no se encuentra diferencia significativa en los dos grupos que se están comparando en el estudio. Los grupos puede que no sean bastante grandes para permitir que se detecte una diferencia significativa (la hipótesis nula siendo que hay una diferencia) y el análisis de poder demostrará la probabilidad de que la hipótesis nula haya sido rechazada correctamente. La probabilidad de un error tipo I o II está en relación inversa para cualquier diseño experimental determinado y tamaño de muestra fijada. El poder puede aumentarse para mejorar el diseño experimental y aumentar el tamaño de la muestra.

Otra herramienta utilizada es la *estimación* en intervención estadística relacionada con cálculo de valor de parámetros específicos de población. Una estimación de 1 punto (media de muestra) puede utilizarse para estimar la media de la población. Un decaimiento en estimación de 1 punto es un fallo para hacer una manifestación de probabilidad con relación a lo ajustado de la estimación al parámetro de población[13]. El *intervalo de confianza* estima remediar este problema aportando un intervalo de estimaciones plausibles de la media de la población, así como una mejor estimación de su valor preciso. El intervalo de confianza que se elige convencionalmente del 95 al 90 por ciento es similar a la elección convencional de 0.05 ó 0.01 para el nivel de significación. La media del intervalo de confianza del 95 por ciento que, asumiendo la media de muestra seguiría una distribución aproximadamente normal, el 95 por ciento de todas las medias de muestras basadas en un tamaño de muestra dada, cabrían dentro de errores estándar ± 1.96 de la población media. Este número deriva de la distribución de la curva normal de las matemáticas. De forma similar, 99 por ciento de todas las medias de muestras basadas en un tamaño de muestra dado, caerán dentro de + 2.576 errores estándar de la media de la población (el 99% de intervalo de confianza)[14]. Por tanto, el intervalo de confianza es un campo de valores para un estudio variable que especifique la probabilidad (ordinariamente 95%) de que el valor cierto de la variable está incluido dentro del campo. El tamaño del intervalo de confianza da alguna idea de la precisión de la estimación del punto en una manera que no se ofrece por un valor p; cuanto más ajustado el intervalo de confianza, los datos serán más precisos. A medida que aumenta el tamaño de la muestra el tamaño del intervalo de confianza se reduce. A medida que la desviación estándar aumenta, reflejando la variabilidad aumentada entre observaciones individuales, el tamaño del intervalo de confianza aumenta. A medida que el nivel de confianza deseado aumenta (por ejemplo del 95 al 99 por ciento), el tamaño del intervalo de confianza aumenta. Por ejemplo, de acuerdo con lo definido en la muestra representada en la tabla 3, la odds de la enfermedad de Kienböck que ocurre asociada a variante ulnar negativa es cerca de 16 veces mayor que cuando ocurre asociada a variante ulnar positiva o neutra. Se puede asumir confiadamente que el 95 por ciento de tal muestra indica que se ha coleccionado dentro del intervalo de

confianza del 95 por ciento, de 5.2 a 50.6. Sin embargo, basar la confianza en el intervalo de confianza y aceptar esta estimación como afirmación de certeza sobre la población general requiere un salto de confianza característico de inducción. En vez de esto, es necesario utilizar un razonamiento deductivo e interpretar la estimación como una tentativa, como hipótesis no refutada[15].

El *razonamiento estadístico* se basa en el precepto de que los procesos naturales pueden describirse mediante modelos estocásticos (relativos al azar) y que el estudio de colecciones de individuos al azar permitirá identificar "tipos sistemáticos de significación científica"[16]. La certeza o falsedad de una hipótesis no puede inferirse desde una prueba de significación. Las tasas de error tipos I y II definen una región crítica para el resumen estadístico que representa una regla de decisión en cuanto a que la hipótesis nula va a ser o no rechazada. Goodman[17] advierte que una regla de decisión no dice nada acerca de si una hipótesis particular es cierta; sólo dice que si los investigadores se comportan de acuerdo a tal regla, en el largo recorrido rechazan si una hipótesis particular es cierta, no más, por ejemplo, que una vez entre 100 puede tener evidencia de que rechazarán la hipótesis suficientemente sólo cuando es falsa. Muchos datos a menudo son mantenidos de manera irrelevante por técnicas estadísticas sofisticadas. Susser[18] estima que una precisión sobre pruebas de significado es insuficiente para el análisis casual de datos. Una prueba de significación no hace ningún supuesto sobre la plausibilidad de la hipótesis nula. Los investigadores tienen que centrarse en la esencia de los asuntos que se están investigando y no deben llegar a menospreciar la mecánica del análisis de datos. El juicio debe tener prioridad sobre la deducción estadística cuando grandes fuentes de error no pueden cuantificarse en el análisis estadístico; aquéllos se conocen en conjunto como sesgos y precisan de consideración cuidadosa en el diseño, ejecución y análisis de todos los estudios.

Sesgos

El sesgo se define como "cualquier tendencia en la colección, análisis, interpretación, publicación o revisión de datos que puede conducir a conclusiones que son sistemáticamente diferentes de la verdad"[4]. El sesgo puede dar lugar a una estimación incorrecta de la asociación entre una exposición y el riesgo de una enfermedad. Tal asociación se considera auténtica si todos los intentos de explicarlo como debida a un sesgo han fallado. Hay tres grandes tipos de sesgos: de confusión, de selección y de información[19].

Confundir es una distorsión en una medida de efecto o consecuencia (como un riesgo relativo) que resulta de un efecto de otra variable (el que confunde) asociada con la enfermedad y exposición que se está estudiando. La confusión puede conducir a sobreestimación o infraestimación de la asociación de certeza entre una enfermedad y una exposición e incluso puede cambiar la dirección del efecto observado. Para que sea un factor, el que confunde, tiene dentro y fuera de ello que ser un factor de riesgo para la enfermedad en la población no expuesta y debe asociarse a exposición variable en la población de la que se derivan los casos. Además, no tiene que ser un paso intermedio en la vía causal entre la exposición y la enfermedad. Por ejemplo, en un estudio que investiga si el fumar es un factor de riesgo en accidentes de vehículos motorizados, el consumir alcohol debería ser considerado un confundidor debido a que el factor riesgo en este caso está asociado con fumar. Al introducir un nuevo procedimiento operatorio deben seleccionarse los pacientes con un buen riesgo con el procedimiento en tanto que los de mal riesgo pueden recibir el tratamiento estándar. A esto se denomina sesgo que confunde por la indicación[20]. La edad y el género son confundidores bien reconocidos.

Efecto modificador es un factor que cambia la magnitud de la medida de efecto (por ejemplo, riesgo relativo u odds ratio). La modificación del defecto difiere de la confusión; el último es un sesgo que el investigador trata de evitar o eliminar de los datos, mientras que el primero es una constante natural. Por ejemplo, el estado de inmunización contra la hepatitis sérica es un efecto modificador para las consecuencias de ser pinchado por una aguja que fue utilizada en una persona infectada de hepatitis. El estado de inmunización es un efecto modificador ya que las personas que están inmunizadas son menos propensas a contraer hepatitis que las que no lo están.

El *sesgo de relación* (también conocido como sesgo de detección y sesgo que desenmascara) se refiere a la distorsión en la estimación de un efecto debido a diferencias sistemáticas en características entre sujetos que están seleccionados para un estudio y los que no lo están. El sesgo de selección puede ocurrir cuando un procedimiento utilizado para identificar una enfermedad varía con el estado de exposición. Por ejemplo, este sesgo puede introducirse por un examinador que realiza una valoración clínica sin que sea ocultada a la enfermedad o al estado de exposición del sujeto. Los resultados de la evaluación pueden diferir si el que examina espera que la enfermedad esté presente en pacientes y ausente en los

controles. Una falta de respuesta a un cuestionario o la pérdida de pacientes para el seguimiento podrían no ser problemas serios si nuevamente resultan en una reducción del número de sujetos disponibles para el estudio. Sin embargo, pueden producir sesgos de relación si los que responden y los que no responden o los pacientes son seguidos, y aquéllos que se han perdido difieren con respecto a algunas características que se están estudiando.

Sesgos de información (también conocido como sesgo de observación y sesgo de mala clasificación) se refiere a distorsión en la estimación de un efecto que resulta de error en la medición de una exposición, enfermedad o de mala clasificación de sujetos con vistas al menos a una variable al describir imprecisión de medición. Pueden presentarse dos tipos de mala clasificación: no diferenciado, cuando la imprecisión es la misma para los dos grupos de estudio (por ejemplo, los pacientes y los controles en un estudio de casos control) y diferenciado (no al azar), cuando la imprecisión difiere entre grupos (por ejemplo, cuando una medida de exposición, tal como tareas repetitivas de trabajo, está determinado más precisamente entre pacientes que entre controles). La mala clasificación no diferenciada aumenta la similitud entre grupos expuestos y no expuestos. Cualquier asociación entre exposición y enfermedad será infravalorada, de manera que el efecto de estimación observado se dice que está sesgado hacia valor nulo de significación 1.0 que si la exposición y enfermedad no tienen asociación alguna, el riesgo relativo esperado debería ser igual a 1.0. La mala clasificación diferencial da lugar a una estimación de riesgo sesgada que bien puede pasarse (una sobrestimación) o quedarse corta (una infraestimación) del valor nulo. Por tanto, el riesgo de confusión es generalmente corregible en el estadio de análisis de estudio, pero los sesgos de selección e información puede que no sean corregibles.

Estrategias de diseño de un estudio para determinar relación entre exposición y enfermedad

En estudios ecológicos (correlación) se utilizan datos de poblaciones enteras para comparar características de enfermedad en grupos diferentes durante el mismo período de tiempo o en la misma población en diferentes momentos de tiempo. En un estudio ecológico, la población más que los individuos se designan como unidad de estudio y el estado de exposición está determinado sobre la base de un valor resumen para un grupo al que estas personas pertenecen. Por ejemplo, un investigador pudo examinar la relación entre el consumo de cigarrillos per cápita y las operaciones realizadas para tratamiento de pseudoartrosis ósea en cinco condados de California en 1966. Incluso si el estudio mostraba que los condados con mayor consumo tenían una tasa de operaciones más alta, el investigador no podría asegurar incluso que los sujetos que fumaban en estos condados ciertamente tenían mayor tasa de pseudoartrosis. Esta conclusión, si de verdad es errónea, se conoce como una falacia o sofisma ecológico debido a que la correlación entre dos variables ecológicas es con frecuencia diferente de la correspondiente correlación individual dentro de las mismas poblaciones[21]. El diseño de estudio ecológico no se utiliza a menudo en investigaciones de enfermedades del aparato locomotor.

Para investigar la etiología de la enfermedad se han descrito quince estudios de diseño, pero más corrientemente se agrupan en tres categorías de estudios de observación: transversales, de control de casos y cohortes (un grupo de personas)[19].

Los *estudios transversales*, también conocidos como estudios de prevalencia o encuestas sobre frecuencia de enfermedad, se utilizan para comprobar el estado de un individuo, con relación a la presencia o ausencia de exposición y enfermedad en el mismo punto de tiempo. No se puede determinar si la exposición precede o es consecuencia de la enfermedad. Los estudios de prevalencia se realizan en la población que sobrevive y, por ello, puede afectarse por sesgos de selección. Los diseños ecológicos y transversales son estrategias epidemiológicas descriptivas que tienen como objetivo formular hipótesis etiológicas. Sin embargo por sus limitaciones inherentes rara vez se utilizan como hipótesis de prueba.

Los *estudios de control de casos* y de cohorte son los dos tipos básicos de diseños de estudio de análisis observacional que tienen como objetivo comprobar hipótesis. El objetivo de un estudio observacional es llegar a las mismas conclusiones que podrían haberse derivado de un ensayo experimental[22]. En el estudio de control de casos se identifican los sujetos y se resuelven en base a si tienen la enfermedad que interesa (síndrome de túnel carpiano, por ejemplo) o no la tienen (controles) y la exposición pasada a factores de interés (por ejemplo, trauma repetitivo). Este diseño de estudio es bueno para investigar enfermedades raras pero muy susceptibles de seleccionar y recordar sesgos. Los recuerdos de sesgos se producen cuando un estudio confía en la memoria de pacientes para determinar el estado de exposición debido a que un paciente que tuvo una enfermedad es más fácil que recuerde posibles exposiciones que una persona sana. Los estudios de casos control son retrospectivos

pues empiezan después del comienzo de la enfermedad y los factores causales que se postulan se valoran de manera retrospectiva.

Una cohorte es un grupo de personas que se siguen durante un período de tiempo. En un *estudio de cohorte*, el grupo a estudiar se define en base al estado de exposición y es seguido en el tiempo con el objeto de valorar la ausencia de enfermedad. Todos los posibles sujetos deben estar libres de la enfermedad que se estudia en el tiempo en que se define el estado de enfermedad. Un estudio de cohorte puede ser prospectivo o retrospectivo. En un estudio prospectivo (coexistente) la enfermedad no había ocurrido aún al comienzo del estudio. En el estudio retrospectivo (histórico), tanto la exposición como la enfermedad ya han ocurrido. No debería ser ético valorar la hipótesis de que niveles bajos de exposición a la radiación acorten las expectativas de vida humana utilizando un estudio prospectivo. Sin embargo, si un grupo de personas (una cohorte) que ya ha estado expuesta puede identificarse, entonces, incluso si la exposición se hizo en el pasado, un estudio de cohorte retrospectiva podría ser realizable. Las ventajas de un estudio de cohorte incluyen una secuencia temporal de exposición y enfermedad que suele ser clara, una minimización del sesgo observacional para determinar la exposición, la capacidad de examinar efectos múltiples de una exposición y la utilidad cuando la exposición es rara. Los estudios de cohorte llevan tiempo, son costosos, no deseables para investigar enfermedades raras y posiblemente sesgadas con vistas a la pérdida de sujetos para seguimiento cuando éstos deben ser seguidos durante muchos años.

Los estudios de intervención (*ensayos clínicos*) son un tipo de estudio de cohorte en que los participantes se asignan por el investigador para recibir una de las exposiciones o tratamientos de estudio. La ventaja de este diseño es que se distribuyen a medias confundidores conocidos y desconocidos, igual entre los grupos de estudio. El sesgo de susceptibilidad puede ocurrir en ensayos clínicos si ambos grupos son desemejantes en términos de su estado inicial. Un ensayo clínico aleatorizado impide la posibilidad de sesgos de susceptibilidad[23]. Los problemas con ensayos operatorios aleatorizados en ortopedia, incluyendo las salidas éticas, de realización, resultados y filosóficas, se discuten en el trabajo de Keller.

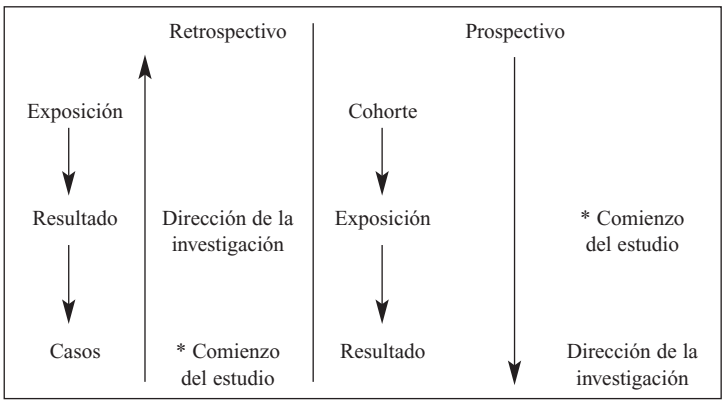

Figura 1. Comparación de diseños de estudios prospectivo y retrospectivo en base a la dirección de la investigación y el comienzo del estudio.

Diseños de estudio

La epidemiología es el estudio de la distribución y decisiones de la frecuencia de enfermedad[24]. La medicina basada en la evidencia y la valoración de los resultados del paciente irrumpió en la medicina clínica en los años ochenta y noventa como consecuencia de la influencia médica actual, la sociedad y la economía[25]. La epidemiología clínica proporciona la metodología para valorar la eficacia clínica.

Hay estudios de observación y experimentales. En los *estudios de observación* los investigadores observan grupos de pacientes sin asignación de la intervención, mientras que los investigadores de los *estudios experimentales* asignan el tratamiento. Los estudios experimentales con humanos se denominan ensayos. Los estudios de investigación pueden ser retrospectivos, indicando que la dirección de la pregunta es hacia atrás, a partir de casos, y ocurriendo los acontecimientos de interés antes del comienzo del estudio. De otra manera, los estudios pueden ser prospectivos, o sea que la dirección de la pregunta se hace hacia delante, a partir del principio de cohorte, ocurriendo los acontecimientos de interés después del comienzo del estudio. Los estudios transversales se utilizan para vigilar un punto en el tiempo y los longitudinales siguen a los mismos pacientes en múltiples puntos en el tiempo.

Todos los estudios de investigación son susceptibles de conclusiones no válidas debidas a sesgos, confusión o azar. Sesgo es el error sistemático -no al azar- en el diseño o conducción de un estudio. Ordinariamente, el sesgo no es intencionado; sin embargo, es penetrante e insidioso. Un estudio en cualquier fase puede corresponderse por diferentes tipos de sesgos, como selección del paciente (selección y calidad de miembros), realización del estudio (realización e información), seguimiento del paciente (no respuesta y traslado) y determinación de resultados (detección, rellamada, aceptación y entrevis-

ta). Los sesgos frecuentes en la literatura ortopédica son: de selección, cuando grupos no similares se comparan; sesgo de respuesta, cuando la tasa de seguimiento es baja; y sesgo de entrevistador, cuando el que investiga resuelve el resultado. Un confundidor es una variable que tiene asociaciones independientes con las variables del independiente (predictor) y el dependiente (resultado), potenciando así la distorsión de su relación. Por ejemplo, la asociación entre laxitud de rodilla y lesión del ligamento cruzado anterior puede confundirse en el sexo femenino ya que la mujer puede tener mayor laxitud de rodilla y mayor posibilidad de lesionar su ligamento cruzado anterior. Frecuentes situaciones de confusión en la investigación clínica son: género, edad, estado socioeconómico y enfermedades asociadas. La muerte puede conducir a conclusiones no válidas basadas en la probabilidad de que haya errores tipo I y II relacionados con el valor de p y poder.

Las consecuencias adversas de los sesgos, confusiones y azar pueden minimizarse mediante diseños de estudio y análisis estadístico. Los estudios prospectivos minimizan los sesgos de selección de pacientes, calidad de información, intentos para situaciones de rellamada preoperatoria y confundidores distribuidos con igualdad. El cegado (blinding) puede disminuir más los sesgos y las equiparaciones pueden disminuir la confusión. Los confundidores pueden a veces ser controlados post-hoc utilizando análisis estratificados o métodos multivariantes. Las consecuencias del azar pueden minimizarse mediante un tamaño adecuado de la muestra basado en cálculos de poder y utilización de apropiados niveles de significación en las pruebas de hipótesis. La capacidad de diseño del estudio para optimizar la validez mientras se reducen sesgos, confusión y azar se reconoce mediante adopción de niveles jerarquizados de evidencia basados en el diseño del estudio. Incluso, el modelo para probar la relación de causa-efecto se pone más alto para sugerir una asociación. La interferencia de causa requiere datos de apoyo a partir de estudios no observacionales, tales como un ensayo clínico aleatorizado, una explicación plausible biológicamente, un tamaño de efecto relativamente grande, reproductibilidad de hallazgos, una relación temporal entre causa y efecto y un gradiente biológico demostrado por una relación dosis/respuesta (tabla 6).

Los *diseños de estudios observacionales comprenden*: series de casos, estudios de control de casos, encuestas transversales y estudios de cohorte. Las series de casos son un recuento retrospectivo, descriptivo de un grupo de pacientes con características interesantes o series de pacientes que se han sometido a una operación. Una serie de casos que incluye a un paciente es la publicación de un caso. Este tipo de estudio es fácil de construir pudiendo aportar un foro de debate para presentar observaciones interesantes o inusuales. Sin embargo, muchas veces son anecdóticas y están sujetas a muchos posibles sesgos a falta de hipótesis y son difíciles de comparar con otras series, de ahí que a menudo se consideran como medio de generar hipótesis para estudios adicionales pero no concluyentes.

Un estudio de *control de casos* es aquél en que el investigador identifica a pacientes con resultado de interés (casos) y a pacientes sin resultado (controles), y entonces compara los dos grupos en términos de posibles situaciones de riesgo. Las consecuencias se publican utilizando la odds ratio. Son eficaces, sobre todo para evaluar afecciones o resultados inusuales y son relativamente fáciles de evaluar. Sin embargo, un grupo de control apropiado puede ser difícil de identificar siendo necesario utilizar historias médicas previas de alta calidad. Incluso, estos estudios son susceptibles de acompañarse de múltiples sesgos, en particular de selección y detección, basados en la identificación y control de casos. A menudo se utilizan encuestas transversales para determinar la prevalencia de la enfermedad e identificar asociaciones de pacientes en que coexiste una afección particular en un tiempo concreto en el tiempo. La prevalencia de una afección es el número de individuos afectados divididos por el número total de individuos en un punto en el tiempo. La incidencia, por el contrario, se refiere a un número de individuos afectados dividido por el número total de individuos durante un período de tiempo definido. De ahí que los datos de prevalencia se obtengan habitualmente de una muestra transversal que crea una proporción, mientras que los datos de incidencia suelen obtenerse de un estudio de cohorte prospectivo y un valor de tiempo está contenido en el denominador. Las encuestas también se realizan con frecuencia para determinar tipos de preferencias y de tratamiento. Como los *estudios transversales* representan una instantánea en el tiempo, pueden estar equivocados si la cuestión que se investiga se refiere al proceso de la enfermedad en el tiempo. La encuesta también puede presentar retos únicos en el sentido de tasa de respuestas adecuadas, muestras representativas y sesgos de aceptabilidad.

Un *estudio de cohorte* tradicional es aquél en que se identifica una población de interés y se sigue de manera prospectiva con el fin de determinar resultados y asociaciones con factores de riesgo. Los estudios de cohorte retrospectivos o históricos también pueden hacerse; en aquellos estudios de cohorte, los elementos de los mis-

mos se identifican basándose en sus historias clínicas, y el período de seguimiento se encuentra parcial o enteramente en el pasado. Los estudios de cohorte son óptimos para estudiar la incidencia, el curso y los factores de riesgo de una enfermedad ya que son longitudinales, indicando qué grupos de sujetos es seguido en el tiempo. Las consecuencias se publican frecuentemente en términos de riesgo relativo. Como son prospectivos pueden optimizar el seguimiento y la calidad de los datos y pueden minimizar los sesgos relacionados con la selección, información y medición. Además, tienen el tiempo/secuencia correcto para proporcionar fuerte evidencia con vista a las asociaciones. Sin embargo, estos estudios son costosos, exigentes logísticamente, a menudo precisan de un largo período de tiempo para su terminación y son ineficaces para valorar resultados o enfermedades poco corrientes.

Los *diseños de estudio experimental* pueden suponer utilizar controles concurrentes, controles secuenciales, ensayos (ir de un lado a otro) o controles históricos. El *ensayo clínico aleatorizado con controles* concurrentes (ECA) es el referente de la evidencia clínica ya que aporta las conclusiones más válidas (validez interna) para minimizar las consecuencias de sesgos y confusión. Para evitar ésta, la mejor manera de hacerla es una aleatorización rigurosa. La realización de un ensayo de control aleatorizado supone la construcción de un documento protocolizado que explícitamente establece la elegibilidad de criterios, tamaño de muestras, consentimiento informado, aleatorización, reglas para interrumpir o detener el ensayo, el cegado o enmascaramiento (blinding), medición, monitorización de complacencia, valoración de seguridad y análisis de datos. Debido a que la distribución es al azar, los sesgos de selección se reducen y los confundidores (conocidos y desconocidos) teóricamente se distribuyen por igual en los grupos. El estudio a doble ciego minimiza el sesgo de realización, entrevistador y de aceptabilidad; puede hacerse a cuatro niveles: participantes, investigadores, aplicación de intervención y asesores de resultados y analistas.

Los análisis que tratan la intención minimizan los sesgos de los que no responden y se trasladan de domicilio, mientras que la determinación del tamaño de la muestra asegura un adecuado poder. El principio de intención de tratar establece que deberán analizarse todos los pacientes dentro del grupo de tratamiento en el que ellos estaban distribuidos al azar con el fin de preservar los objetivos de la aleatorización.

Aunque el ensayo clínico aleatorizado es el compendio de diseños de investigación clínica, las desventajas de tales ensayos son el gasto y la logística de terminación. El incremento de pacientes y aceptación por el clínico puede ser difícil. Con la tecnología que avanza rápidamente se llega a aceptar muy pronto una técnica nueva que haga difícil aceptar un ensayo clínico aleatorizado existente o un ensayo clínico posible aleatorizado.

Los ensayos clínicos aleatorizados, éticamente requieren equilibrio clínico (igualdad de opiniones de tratamiento en opinión del clínico) para inscripción, detención de reglas interim para evitar daño y evaluar acontecimientos adversos y también el consentimiento informado. Finalmente, mientras que los ensayos clínicos aleatorizados tienen excelente validez interna se ha cuestionado su posibilidad de generalización (validez externa), debido a que el tipo de práctica y la población de pacientes incluidos en un ensayo clínico aleatorizado pueden ser demasiado constreñidos y no representativos.

Las consideraciones éticas son intrínsecas para diseño y dirección de estudios de investigación clínica. El consentimiento informado es de fundamental importancia y punto de atracción de muchas de las mesas de revisión institucional de actividad. Los investigadores deberían estar familiarizados con el Código de Nuremberg y la declaración de Helsinki ya que se refieren a los problemas éticos de riesgo y beneficios, protección de privacidad y respeto a la autonomía[26,27].

Comprobación de hipótesis

El objetivo de comprobar hipótesis es permitir generalizaciones desde una muestra a la población de donde procede: confirmar o refutar la aseveración de que los hallazgos observados no ocurren por casualidad solamente sino más bien debido a una auténtica asociación entre variables. Por negligencia, la hipótesis nula de un estudio afirma que no existe asociación marcada entre variables, en tanto que la hipótesis alternativa afirma que existe marcada asociación. Si los hallazgos del estudio no son significativos no puede rechazarse la hipótesis nula, mientras que si los hallazgos son significativos puede rechazarse la hipótesis nula, aceptándose las hipótesis alternativas.

Por ello, todos los estudios de investigación que se basan en una muestra hacen interferencia sobre la verdad de la población global. Construyendo una tabla de dos por dos de los posibles resultados de un estudio (tabla 4), se puede ver que la interferencia o dificultad de un estudio es correcta si no se encuentra asociación marcada, cuando no hay ninguna asociación verdadera o si se encuentra una asociación significativa cuando hay asociación verdadera. Sin embargo, un estudio puede tener

dos tipos de errores. El *error tipo I* o alfa (α) se produce cuando se encuentra asociación significativa, cuando no hay asociación verdadera, dando lugar a un estudio falso positivo que rechaza una hipótesis nula verdadera. El *error tipo II* o beta (ß) concluye de forma errónea que no hay ninguna asociación significativa, dando lugar a un estudio falso negativo que rechaza una hipótesis alternativa cierta.

El nivel alfa se refiere a la probabilidad de un error tipo I (α). Por convención, el nivel alfa de significación se establece a 0.05 que indica que se acepta el hallazgo de una asociación significativa si hay menos de 1 en 20 posibilidades de que la observación observada era debida al azar. Por ello, el *valor p*, que se calcula mediante una prueba estadística, es una medida del poder de la evidencia con tal de que los datos estén a favor de la hipótesis nula. Si el valor de p es menor que el nivel alfa, entonces la evidencia contra la hipótesis nula es bastante fuerte para ser rechazada y concluir que el resultado es significativo.

Los valores de p se utilizan frecuentemente en investigación clínica y se les da gran importancia por revistas y lectores; sin embargo, en bioestadística hay un fuerte movimiento para desestimar los valores de p debido a que un nivel de significación de p 0.05 es arbitraria, un punto tajante estricto puede ser desorientador (existe poca diferencia entre p= 0.049 y p= 0.051, pero sólo el primero se considera "significativo"), el valor de p no da información sobre el poder de la asociación y puede ser estadísticamente significativo sin que los resultados sean importantes clínicamente. Las alternativas a la tradicional dependencia sobre los valores de p incluyen el uso de niveles de significación de variables alfa basadas en las consecuencias de errores de tipo I y como 1 la probabilidad de un error tipo II (ß). Por convención, se establece como poder aceptable el 80%, lo cual indica que hay oportunidad del 20% de que el estudio no demostrará ninguna asociación significativa cuando hay una asociación verdadera. En la práctica, cuando un estudio demuestra asociación significativa, el posible error de preocupación es un error de tipo II (ß) expresado por poder, que en un estudio es el que demuestra ningún efecto significativo que puede ser ciertamente ningún efecto significativo pero el estudio fue infravalorado debido a que el tamaño de las muestras era demasiado impreciso. De ahí que, cuando un estudio no demuestra efecto significativo, el poder del estudio debería publicarse.

Los cálculos para analizar el poder difieren según los métodos estadísticos utilizados en el análisis; sin embargo, hay cuatro elementos: alfa, beta, tamaño del efecto y tamaño de la muestra (n). El *tamaño del efecto* es la diferencia que se quiera, capaz de detectar datos con alfa y beta. Se basa en el sentido clínico que sobre la amplia diferencia podría ser clínicamente significativa. Los tamaños de efecto a menudo se definen en términos de menor dimensión en base a la diferencia en valores medios divididos por la desviación estándar reunida de una comparación de dos grupos. Los tamaños de muestras pequeñas, los tamaños de efectos pequeños y las grandes varianzas disminuyen todas ellas el poder de un estudio. Un conocimiento de las posibilidades de poder es importante en investigación clínica y la publicación de valores de p sin utilizar el término "significativo". La utilización del 95% de intervalo de confianza en lugar de valores p ha ganado aceptación, ya que estos intervalos envían información con relación a la significación de los hallazgos (el 95% de intervalo de confianza no se sobrepasan si son significativamente diferentes), la magnitud de las diferencias y la precisión de la medición (indicada por la amplitud del 95% de intervalo de confianza). Mientras el valor p a menudo se interpreta como significativo o no, el intervalo de confianza del 95% proporciona una opción de valores que permite al lector interpretar las complicaciones de los resultados. Además, mientras que los valores de p no tienen unidades, los intervalos de confianza se presentan en unidades de la variable de interés, que ayuda al lector a interpretar los resultados. Por ejemplo, los autores de un estudio sobre el tiempo de estancia en el hospital de niños con artritis séptica de cadera tratados con una pauta clínica práctica, pueden establecer "que hay una estancia hospitalaria significativamente más corta en pacientes tratados de acuerdo a pauta" con la adición de "p= 0.003" si los valores de p se utilizan o "intervalos de confianza del 95%, 3.8 a 5.8 días para pacientes tratados de acuerdo a la pauta establecida y 7.3 a 9.3 días en pacientes no tratados de acuerdo a dicha pauta", si se emplean intervalos de confianza del 95%[28]. El acercamiento al valor p conlleva sólo significación estadística, mientras que el intervalo de con-

	Certeza	
Experimento	Ninguna asociación	Asociación
Ninguna asociación	correcto	error tipo II (ß)
Asociación	error tipo I (α)	correcto

* Valor de P = probabilidad de error tipo I (α). Poder = 1-probabilidad de error tipo II (ß).

Tabla 4. Prueba de hipótesis*

Experimento	Enfermedad positiva	Enfermedad negativa
Prueba positiva	a (cierto positivo)	b (falso positivo)
Prueba negativa	c (falso negativo)	d (cierto negativo)

* Sensibilidad= a/(a+c); especificidad= d/(b+d); precisión= (a+c)/(a+b+c+d); tasa de falso negativo= 1-sensibilidad; tasa de falso positivo= 1-especificidad; ratio de probabilidad (+)= sensibilidad/tasa de falso positivo; ratio de probabilidad (-)= tasa de falso negativo/especificidad; valor predictivo positivo= [(prevalencia) (sensibilidad)] / [(prevalencia)(sensibilidad)+(1-prevalencia)(1-especificidad)]; y valor predictivo negativo= [(1-prevalencia)(especificidad)] / [(1-prevalencia)(especificidad) + (prevalencia)(1-sensibilidad)].

Tabla 5. Realización de prueba diagnóstica*

fianza conlleva significación estadística (los intervalos de confianza no se superponen), significación clínica (la magnitud de los valores) y precisión (la amplitud de intervalos de confianza).

El poder es la probabilidad de encontrar una asociación significativa, si tal ciertamente existe, y se define para minimizar el empleo de recursos al planificar un estudio y asegurar su validez. Los cálculos del tamaño de muestra se realizan cuando se está planificando un estudio. Típicamente, el poder se establece en el 80%, alfa a 0.05, el tamaño del efecto y la varianza se estiman a partir de los datos piloto o de la literatura, y la ecuación se resuelve terminado el estudio -esto es, del análisis del poder post-hoc- es controvertido y desanimador.

Diagnóstico de realización

Una prueba diagnóstica puede situarse en cuatro situaciones posibles: 1) *cierto positivo* si la prueba es positiva y la enfermedad está presente; 2) *falso positivo*, si la prueba es positiva pero la enfermedad está ausente; 3) *cierto negativo*, si la prueba es negativa y la enfermedad está ausente; y 4) *falso negativo*, si la prueba es negativa y la enfermedad está presente (tabla 5). La sensibilidad de una prueba es el porcentaje (o proporción) de pacientes con la enfermedad que se clasifican como que tienen un resultado positivo de la prueba (el verdadero positivo). Una prueba con *sensibilidad* del 97% quiere decir que de 100 pacientes con la enfermedad, 95 tendrán una prueba positiva. Las pruebas de sensibilidad tienen una baja tasa de falso negativo. Un resultado negativo de prueba altamente sensible excluye o descarta la enfermedad (SNout). La *especificidad* de una prueba es el porcentaje (o proporción) de pacientes sin la enfermedad que se clasifican por tener un resultado negativo de la prueba (el verdadero negativo). Una prueba con especificidad del 91% supone que de 100 pacientes sin la enfermedad, 91 tendrán una prueba negativa. Las pruebas específicas tienen una baja tasa falso positiva. Un positivo resulta de pruebas altamente específicas incluyen la enfermedad (SPin). Sensibilidad y especificidad pueden combinarse en un único parámetro, la *ratio de posibilidad* (LR), que es la probabilidad de un verdadero positivo dividido por la probabilidad de un falso positivo. Sensibilidad y especificidad pueden establecerse en estudios en que los resultados de una prueba diagnóstica se comparan con aquéllos del referente de diagnóstico en los mismos pacientes -por ejemplo, comparando los resultados de resonancia magnética con los hallazgos artroscópicos[29].

Sensibilidad y especificidad son parámetros técnicos de realización de pruebas diagnósticas, y tienen importantes implicaciones para pautas de filtrado y práctica clínica[29,30]; sin embargo, son menos relevantes en casos clínicos típicos porque el clínico no sabe si el paciente tiene la enfermedad cuando el resultado es positivo (valor predictivo positivo = NPV). Ambos casos son probabilidades que requieren una estimación de la prevalencia de la enfermedad en la población y pueden calcularse utilizando ecuaciones que aplican el teorema de Bayes[32].

Existe un inherente cambalacheo entre sensibilidad y especificidad. Como existe típicamente cierto solapamiento entre grupos de enfermos con relación a una distribución de pruebas, el investigador puede seleccionar un criterio de positividad con una baja tasa de falso negativo (para optimizar la sensibilidad) o falso positivo (para optimizar la especificidad) (figura 2). En la práctica, los criterios de positividad se seleccionan en base a las consecuencias del diagnóstico falso positivo o falso negativo. Si las consecuencias del diagnóstico de un falso negativo pesan más que las consecuencias de un diagnóstico falso positivo de una afección (como artritis séptica de cadera en niños)[33], se escoge el criterio más sensible. La relación entre sensibilidad y especificidad de una prueba diagnóstica puede retratarse con el empleo de una curva de receptor que opera características (ROC). Un gráfico de éste muestra la relación entre la verdadera tasa positiva (sensibilidad) en el eje Y y la tasa de falso positivo (1-especificidad) sobre el eje X trazado en cada posible corte (figura 3). La realización del diagnóstico global puede evaluarse basándose en el área situada bajo el receptor que opera características[34].

Medidas de efecto

Las medidas de verosimilitud incluyen probabilidad y odds. *Probabilidad* es un número entre 0 y 1 que indica la facilidad con que va a ocurrir un acontecimiento en base al número de acontecimientos por el número de

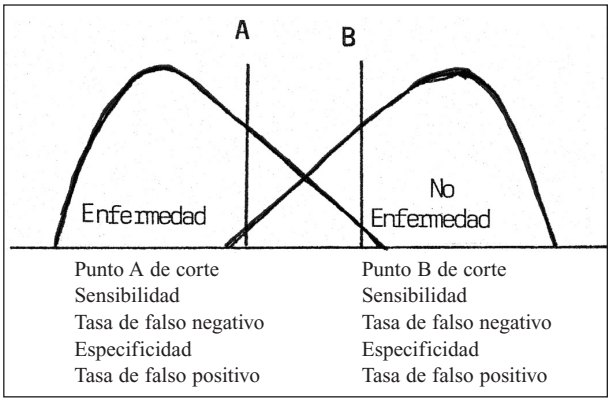

Figura 2. Selección de criterios de positividad. Como clásicamente hay superposición entre población enferma y no enferma en una amplitud de valores diagnósticos (eje de X), hay un cambalacheo intrínseco entre sensibilidad y especificidad. Cuando se identifican los resultados de la prueba positiva, como los que están a la derecha del punto A de corte, hay alta sensibilidad ya que la mayoría de los pacientes con la enfermedad están correctamente identificados como que tienen un resultado positivo. Sin embargo, hay casos que tienen incorrectamente identificado un resultado positivo (falsos positivos). Cuando los resultados de la prueba positiva se identifican como a la derecha del punto de corte B, hay menor sensibilidad ya que algunos pacientes con enfermedad son incorrectamente identificados como que tienen un resultado negativo (falsos negativos). Sin embargo, existe alta especificidad ya que la mayor parte de los pacientes sin enfermedad son identificados correctamente como que tienen un resultado negativo.

ensayos. La probabilidad de que salga cara al tirar una moneda al aire es 0.5. *Odds* es la ratio de probabilidad de que ocurra un acontecimiento o de que no ocurra. La odds de que salga cara al tirar la moneda es 1 (0.5/0.5), en que odds = probabilidad/(1-probabilidad).

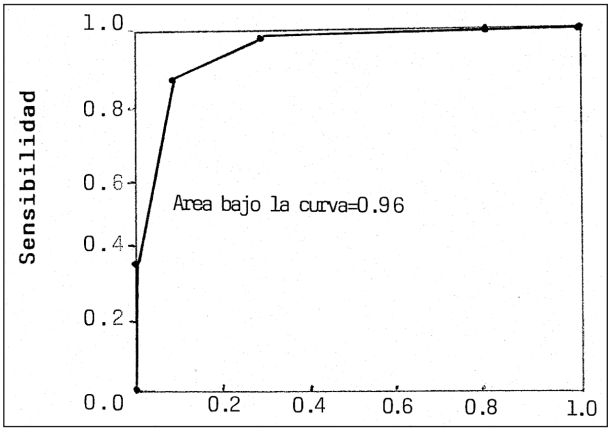

Figura 3. Características del receptor de operación (ROC) con su curva que rige la predicción clínica para diferenciar artritis séptica de sinovitis transitoria de la cadera en niños33. La tasa de falsos positivos (1-especificidad) se anota en el eje de X y la sensibilidad en el eje de Y. El área situada bajo la curva representa la realización diagnóstica global de una regla de predicción o prueba diagnóstica. Para una prueba perfecta, el área situada bajo la curva es de 1.0. Para pensar de manera aleatorizada, el área bajo la curva es de 0.5.

	Acontecimientos Adversos	Acontecimientos No Adversos
Grupo experimental	a	b
Grupo control	c	d

* Tasa de control de acontecimientos (CER)= c/(c+d); tasa de acontecimiento experimental (EER)= a/(a+b); odds de control de acontecimiento (CEO)= c/d; odds de acontecimiento experimental (EED)= a/b; riesgo relativo (RR)= EER/CER; odds ratio (OR)= EEO/CEO; reducción de riesgo relativo (RRR)= (EER-CER/CER); reducción de riesgo absoluto (ARR)= EER-CER; y número necesario de tratar (NNT)= 1/ARR.

Tabla 6. Consecuencias del tratamiento*

El *riesgo relativo* (RR) puede determinarse en estudio de cohorte prospectivo, donde el riesgo relativo es igual a la incidencia de la enfermedad en la cohorte expuesta dividido por la incidencia de la enfermedad en la cohorte no expuesta (tabla 6). Por ejemplo, si un estudio de cohorte prospectivo de esquiadores con deficiencia del ligamento cruzado anterior (LCA) muestra una proporción significativamente más alta de ulteriores lesiones de rodilla en esquiadores que no son tratados con una ortesis (12.7%) que en aquellos que son tratados con ella (2.0%), la ratio de riesgo es 6.4 (12.7/2.0)[35]. Esto puede interpretarse como un riesgo 6.4 veces mayor de ulterior lesión de rodilla en esquiadores con deficiencia de LCA que no es tratada con una ortesis, que en los que no se tratan con ortesis. Un estudio similar en un estudio de casos retrospectivos (en que no se puede determinar la incidencia) es la odds ratio (OR) que es la ratio de las odds de tener la enfermedad en un grupo de estudio de las odds de tener la enfermedad en el grupo control (tabla 6).

Los factores que incrementan con facilidad la incidencia, así como la prevalencia, morbilidad o mortalidad de una enfermedad se denominan *factores de riesgo*. El efecto de un factor que reduce la probabilidad de un resultado adverso puede cuantificarse mediante la reducción del riesgo relativo (RR), la reducción del riesgo absoluto (ARR) o el número que se necesita tratar (NNT) (tabla 6). El efecto de un factor que incrementa la probabilidad de un resultado adverso puede cuantificarse por el incremento del riesgo relativo (RRI), el incremento del riesgo absoluto (ARI) y el número necesario para perjudicar (NNH) (tabla 6).

Valoración de resultados

El proceso se refiere a la atención médica que recibe un paciente mientras los resultados se refieren al resultado de una atención médica. El énfasis del movimiento de valoración de resultados ha sido la valoración de los

resultados derivados del paciente. La medida de los resultados comprende: medidas genéricas, medidas de la afección específica y medidas de satisfacción del paciente[36].

Las *medidas genéricas* como el Short-Form-36 (SF-36) se utilizan para valorar el estado de salud o la calidad de vida relacionada con la salud, en base a la definición de salud de dominio múltiple de la Organización Mundial de la Salud (OMS).

Las *medidas de la afección específica*, como la puntuación de rodilla del Comité Internacional de Documentación de Rodilla (IKDC) o la puntuación de hombro de Constant, se utilizan para valorar aspectos de afecciones o aparatos concretos.

La *medida de la satisfacción* del paciente se utiliza para comprobar varios componentes de los cuidados y tiene aplicaciones diversas, como la evaluación de la calidad de la atención médica, aporte de cuidados sanitarios, modelos de cuidados centrados en el paciente y mejoras continuadas de la calidad.

El proceso de desarrollo de instrumentos de resultados comporta la identificación del montaje, proyectos, respuestas a escala, selección de casos, factores de construcción y creación de escalas. Un gran número de instrumentos de resultados se han desarrollado y utilizado sin valoración psicométrica formal de su fiabilidad, validez y correspondencia para cambiar. La fiabilidad se refiere a la repetitividad de un instrumento. La *fiabilidad* interobservadores e intraobservadores se refiere a la repetibilidad de un instrumento cuando se utiliza por diferentes observadores y por el mismo observador en diferentes puntos de tiempo respectivamente. La prueba de fiabilidad de repetir la prueba se puede valorar utilizando el instrumento para evaluar al mismo paciente en dos ocasiones diferentes sin un cambio de intervalo en la situación médica del paciente. Estos resultados suelen publicarse mediante estadísticas kappa o coeficiente de correlación intraclase. La *validez* se refiere a si un instrumento mide lo que separa para medir. La validez de contenido valora si un instrumento es representativo de la característica que se está midiendo de acuerdo a la opinión de consenso del experto (validez de cara). La validez de criterio valora la relación del instrumento con un instrumento aceptado de referencia. La validez de montaje valora si un instrumento sigue las hipótesis aceptadas y produce resultados de acuerdo con las expectativas teóricas. La correspondencia o conformidad para cambiar valora la manera en que los valores del instrumento cambian en el curso y tratamiento de la enfermedad.

Medicina basada en la evidencia

La medicina basada en la evidencia (MBE) supone el empleo concienzudo, explícito y juicioso de la mejor evidencia actual en la toma de decisiones sobre la atención al paciente individual[37]. La MBE integra la mejor evidencia de investigación con valores de experiencia clínica y de pacientes. Sus pasos propugnan convertir la necesidad de información en una cuestión contestable, lograr descubrir la mejor evidencia para contestar tal cuestión, evaluar con sentido crítico la evidencia con relación a su validez, impacto y aplicabilidad, e integrar el valor crítico con la experiencia clínica y los valores y circunstancias peculiares del paciente[38,39]. Los tipos de preguntas que se hacen en la MBE son cuestiones en primer término que pertenecen al conocimiento específico de la forma de tratar pacientes con una enfermedad particular. La evidencia se establece en grados en base al diseño del estudio, tal y como se expone en las Normas Editoriales de las Revistas, poniendo énfasis en los ensayos clínicos aleatorizados.

Una *revisión sistemática* es un resumen de la literatura médica donde se emplean métodos explícitos para realizar la búsqueda en la literatura y estudios de evaluación crítica. Un tipo especializado de aquélla es el *metaanálisis*, donde se emplean métodos cuantitativos para combinar los resultados de varios estudios independientes (habitualmente ensayos clínicos aleatorizados) para producir estadísticas sumariales. Por ejemplo, un estudio que revisa la literatura sistemáticamente (con criterios para estudios de inclusión y exclusión) para ser publicado -comparando osteosíntesis y artroplastia en el tratamiento de fracturas del cuello del fémur, y luego resume resultados y complicaciones- se considera como revisión sistemática. Por otra parte, se considera metaanálisis un estudio en que los investigadores revisan sistemáticamente la literatura (con criterios de estudios de inclusión y exclusión) y combinan luego los datos del paciente para realizar un nuevo análisis estadístico[40]. Las trayectorias clínicas o *pautas de práctica clínica* (CPG= Clinical Practice Guidelines) son algoritmos que se desarrollan en base a la mejor evidencia disponible para estandarizar procesos y optimizar resultados. Posiblemente también pueden disminuir errores de omisión y comisión, reducir variaciones de tipos de actuación práctica y disminuir costes.

Análisis de decisión es una herramienta metodológica que permite la evaluación cuantitativa de toma de decisiones bajo situaciones de inseguridad[41-43]. La exposición razonada que subyace explícita al análisis de deci-

sión es que una decisión debe hacerse a menudo bajo condiciones o circunstancias de inseguridad y que la teoría de la decisión racional optimiza el valor esperado. El proceso de análisis de decisión de valor esperado comporta la creación de un árbol de decisiones para estructurar el problema de decisión, determinación de probabilidad y utilidades de resultados (valores del paciente); *análisis de plegado atrás (fold-back)* para calcular el valor esperado de cada línea de decisión para determinar la estrategia óptima de toma de decisiones, y el análisis de sensibilidad para determinar la consecuencia de probabilidad de resultados variables y utilidad en toma de decisiones. El análisis de decisión puede identificar la estrategia de decisión óptima y cómo cambia esta estrategia con las variaciones de probabilidad de resultados o valores del paciente. Este proceso se utiliza explícita o implícitamente y se integra bien el modelo de médico/paciente más nuevo o de toma de decisiones compartidas.

Los *estudios de valoración económica* en medicina incluyen los siguientes de diseños: identificación del coste, análisis de coste-eficacia, análisis de coste-beneficio y coste-utilidad[44]. En los estudios de coste-identificación se identifican los costes de provisión de tratamiento; en el de coste-eficacia se valoran y publican los costes y resultados clínicos en forma de coste por resultado clínico. En el análisis de coste-beneficio se miden tanto los costes como los beneficios en unidades monetarias; y finalmente, en el análisis de coste-utilidad se miden ambos elementos y se publican en forma de coste por calidad y año de vida finiquitado.

Principios de bioestadística

La escala en que se mide una característica tiene implicaciones en la línea en que se resume y analiza la información. Los datos pueden ser categóricos, ordinales o continuos. Los datos categóricos indican tipos o categorías y se pueden considerar como cuentas o cómputos. Las categorías no representan un orden subordinado. Son ejemplos aquéllos que incluyen resultados de género y dicotómicos (si/no, éxito/fracaso). Los datos categóricos se denominan también nominales. Los datos categóricos se describen generalmente en términos de proporciones o porcentajes y se publican en tablas y diagramas de barras. Si hay un orden inherente entre categorías, entonces los datos son ordinales. El número representa un orden pero no necesariamente a escala. Ejemplos de ello son los estadios del cáncer y los grados de lesiones. Los datos ordinales también se describen generalmente en términos de proporciones o porcentajes y se presentan en tablas y diagramas de barras. Datos continuos son las observaciones sobre una solución de continuidad para las que las diferencias entre miembros tienen significado sobre una escala numérica. Ejemplos son: edad, peso y distancia. Cuando una observación numérica puede tener sólo valores numéricos enteros, la escala de medición se denomina discreta. Los datos continuos generalmente se describen en términos de desviación media y estándar y pueden expresarse en tablas o gráficos[45].

Los datos pueden resumirse en términos de medidas de tendencia central como media, mediana y modo (valor que ocurre con la mayor frecuencia) y en términos de medidas de dispersión, tales como amplitud, desviación estándar y percentiles. Los datos pueden caracterizarse por distribuciones diferentes: normales (de Gauss), oblicuas (sesgadas) y bimodales.

El análisis univariable o bivariable valora la relación de una variable única independiente o dependiente. Las pruebas estadísticas para comparar medias son variables continuas que están distribuidas normalmente, incluyen el test de Student para dos grupos independientes y el test t pareado para muestras pares. Para variables continuas o categóricas que no están distribuidas normalmente, los test estadísticos no paramétricos para comparar medianas incluyen el test Mann-Whitney (también conocido como test ranksum de Wilcoxon) para comparar dos grupos independientes en que los datos están distribuidos normalmente. El resultado principal es el test F, que incluye un valor p que indica si existe una diferencia significativa global. Para determinar si hay diferencias significativas entre grupos de individuos, los tests post-hoc se utilizan para realizar comparaciones múltiples emparejamientos entre grupos; estos test son los siguientes: Bonferroni, Tukey, Newman-Keuls, Scheffé, Fisher y Dunnett. El test de Krunkal-Wallis se utiliza para comparar medianas de tres o más grupos independientes en situaciones en que los datos no siguen una distribución normal. El test de Kruskal-Wallis es una alternativa no paramétrica para el análisis de varianza. Los análisis de varianza de medidas repetidas se utilizan para variables distribuidas normalmente en estudios que equiparan asuntos. El test no paramétrico para comparar medianas entre tres o más grupos equiparados se denomina test de Friedman.

Los test estadísticos utilizados para comparar proporciones de variables categóricas u ordinales incluyen el test de Pearson chi-square para dos o más grupos independientes y el test exacto de Fisher cuando la frecuencia de células esperadas es pequeña (cinco o menos). Para comparar muestras, el test de McNemar se utiliza

Tipo de datos	Nº de grupos	Grupos independientes	Muestras pareadas
Continuo			
Normal	2	Test de Student	Test t pareado
No normal	2	Test de Mann-Whitney U	Test Signed-rank Wilcoxon
Normal	3	Análisis de varianza	Análisis de varianza de medida repetidas
No normal	3	Test de Kruskal-Wallis	Test de Friedman
Ordinal	2	Test de Mann-Whitney U	Test Signed-rank Wilcoxon
	3	Test de Kruskal-Wallis	Test de Friedman
Nominal	2	Test exacto de Fisher	Test de McNemar
	3	Test chi-squiare Pearson	Test Q de Cochran
Supervivencia	2	Test log-rank	Regresión logística condicional

Tabla 7. Pruebas estadísticas para comparar grupos independientes y muestras pareadas.

para dos variables y el test Q de Cochran se utiliza para tres o más variables (tabla 7).

Los tests estadísticos que se utilizan para resolver asociaciones son la correlación momento/producto de Pearson (r) para variables continuas normalmente distribuidas, la correlación rank-order (rho) de Spearmen para variables no paramétricas y la rank-correlation de Kendall para variables ordinales.

El análisis de supervivencia se usa para analizar datos cuando el resultado de interés es el tiempo hasta que ocurre un acontecimiento. Se sigue a un grupo de pacientes para determinar si experimentan el acontecimiento de interés. El punto final en el análisis de supervivencia puede ser la muerte o un punto final clínico, como la revisión de una artroplastia total. Un paciente se escruta cuando el acontecimiento de interés no se produce en tal individuo durante el período de estudio. La supervivencia es el espacio de tiempo desde la entrada del paciente en el estudio hasta el acontecimiento de interés o hasta el momento de realizar el escrutinio. El tiempo de supervivencia para cada paciente rara vez se conoce cuando se construye una curva de supervivencia. En su lugar, el espacio de tiempo en que, por ejemplo, la artroplastia del paciente ha sobrevivido hasta aquí o el espacio de tiempo que sobrevivió antes de que se sepa que ha sido revisada; a menudo, los pacientes no han tenido un fallo al final del período de estudio pero se encuentra en situación de riesgo para un fallo en el futuro. Hay ejemplos en que la información es censurada porque el tiempo de supervivencia es observado parcialmente. Los datos de supervivencia son analizados típicamente utilizando el método de límite de producto de Kaplan Meier, en que la supervivencia (ausencia de acontecimiento) se calcula cada vez que se produce un acontecimiento pero no en el tiempo del escrutinio[46]. El análisis de Kaplan Meier se utiliza cuando se conoce la fecha del punto final. Los puntos finales que no se han alcanzado se tratan como escrutados en la fecha del último seguimiento del análisis. El análisis de supervivencia produce una tabla de vida que muestra el número de fallos que se producen en los intervalos de tiempo y el número de pacientes separados durante el intervalo. Una curva de supervivencia puede trazarse para ilustrar el porcentaje de pacientes libres de fallo (libres de acontecimiento) en el eje verti-

Variable continua: una variable con un potencialmente infinito número de valores posibles. Ejemplos de ello son la altura o la amplitud de movimiento.
Variable categórica: una variable en que los posibles valores constan de un número de categorías, por ejemplo género o clasificación de fractura. Un caso especial es cuando las categorías tienen un orden natural no ambiguo, una variable categórica denominada ordenada. Ejemplo de ello es el estadiaje A, B, C, D, de un tumor donde se conoce que existe progresión desde A a D.
Distribución normal (de Gauss): una distribución de frecuencia que se extrae a menudo como una curva familiar de buena forma, con las siguientes propiedades: continua, distribución simétrica con colas que llegan al infinito; media, modo y mediana son idénticos y su forma está determinada enteramente por la desviación media y estándar. La distribución normal subyace a todas las pruebas paramétricas; las pruebas no paramétricas se denominan libres de distribución.
Intervalo de confianza: intervalo computado con una probabilidad dada, a menudo del 95% de que el valor verdadero (o población) de una variable como media, proporción o tasa está contenida dentro del intervalo.
Hipótesis nula: hipótesis estadística de que dos o más distribuciones de población no difieren una de otra, o que una variable no tiene asociación con otra variable o grupo de variables.
Valor p: la probabilidad de que la prueba estadística podría ser tan extrema o más extrema que la observada si la hipótesis nula fuera cierta.
Análisis de regresión: método para encontrar el mejor modelo matemático para describir o vaticinar una variable dependiente en función de una o más variables independientes.

Tabla 8. Glosario de algunos términos estadísticos frecuentes[50]

cal, y el tiempo de seguimiento después de la operación quirúrgica en el eje horizontal. El 95 por ciento de los intervalos de confianza pueden construirse sobre la curva en puntos de tiempo seleccionados utilizando la fórmula de Greenwood[47]. La supervivencia para grupos diferentes puede compararse con el test long-rank para comparar la igualdad de las curvas[48].

El análisis de multivariables explora la relación entre múltiples variables. Regresión es el método de obtener una relación matemática entre una variable de resultado (Y) y una variable explicatoria (X) o un grupo de variables independientes (X_i's). La regresión lineal se utiliza cuando una variable de resultado es continua con el objeto de hallar la línea que predice mejor Y a partir de X. La regresión múltiple ajusta datos a un modelo que define Y como una función de dos o más variables explicatorias o predictivas. La regresión logística se utiliza cuando una variable de resultado es binaria o dicotómica y ha llegado a ser la forma más común de análisis multivariable para resultados no relacionados con el tiempo. Otros métodos de regresión son los datos de tiempo-acontecimiento (regresión de azar proporcional Cox) y datos de cuenta (regresión Poisson). La regresión de modelado se utiliza comúnmente para predecir resultados o para establecer asociaciones independientes (control para confusión y colinearidad) entre predictor o variables explicatorias. Por ejemplo, la regresión logística puede utilizarse para determinar predictores de artritis séptica frente a sinovitis transitoria de cadera en niños en base a un conjunto de variables existentes de tipo demográfico, de laboratorio y de imagen[33]. De la misma manera se puede utilizar la regresión lineal para resolver determinantes independientes de resultados del paciente medidos mediante instrumentos de resultado continuo[49]. Debido a que muchas variables suelen influenciar un resultado particular, se necesita utilizar un análisis multivariable cuyo objetivo es identificar, de entre las muchas variables del paciente, como las quirúrgicas observadas y recogidas y aquéllas más relacionadas con el resultado. La mayoría de análisis generan una gran cantidad de información y la interpretación adecuada requiere experiencia, por lo cual es ventajoso tener al lado a un colega experto en metodología estadística impuesto en análisis de variables múltiples.

Bibliografía:

1. Thomas NP, Hamblen DL. The computer, education and clinical practice (editorial). J Bone Joint Surg 2004; 86B: 1.
2. Editorial. Evidence-based orthopedics. Acta Orthopedica 2007; 78 (1): 1.
3. Szabo RM. Principles of epidemiology for the orthopaedic surgeon (current concept review). J Bone Joint Surg 1998; 80A: 111-120.
4. Last JM (ed). A dictionary of epidemiology. Ed 2. New York. Oxford University Press. 1988.
5. Anderson M, Green WT, Messner MB. Growth and predictions of growth in the lower extremities. J Bone Joint Surg 1963; 45A: 1-4.
6. Moseley CF. A straight-line graph for leg-length discrepancies. J Bone Joint Surg 1977; 59A: 174-179.
7. Hulten O. Über anatomische variationen der handgelenkknochen. Acta Orthop Scand 1928; 9: 155-196.
8. Gelberman RH, Salamon PB, Jurist JM, Posch JL. Ulnar variance in Kienböck's disease. J Bone Joint Surg 1975; 57A: 674-676.
9. Rothman KJ. Causes. Am J Epidemiol 1976; 104: 587-592.
10. Oakes M. Statistical inference. Chesnut Hill. Massachusetts. Epidemiology Resorces. 1990.
11. Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32: 51-63.
12. Lang TA, Rhodes R. Ten common statistical reporting errors in biomedical literature. CBE Views, 1996; 19: 82-83.
13. Kuzma JW. Basic statistics for the health sciencies. Ed 2. Mountain View. California. Mayfield Publishing. 1992.
14. Rosner F. Fundamentals of biostatistics. Ed 4. Belmont. California. Duxbury Press. 1995.
15. Maclure M. Popperian refutation in epidemiology. Am J Epidemiol 1985; 21: 343-350.
16. Zeger SL. Statistical reasoning in epidemiology. Am J Epidemiol 1991; 134: 1.062-1.066.
17. Goodman SN. P Values, hypothesis tests and likehood: implications for epidemiology of a negelcted historical debate. Am J Epidemiol 1993; 137: 485-496.
18. Susser M. Judgement and causal interference: criteria in epidemiologic studies. Am J Epidemiol 1977; 1: 1-15.
19. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiology research: principles and quantitative methods. Belmont. California. Lifetime Learning Publications. 1982.
20. Greenland S, Neutra R. Control of confounding in the assessment of medical tecnology. Internat J Epidemiol 1980; 9: 361-367.
21. Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev 1950; 15: 351-7.
22. Gray-Donald K, Kramer MS. Causalty inference in observational vs experimental studies. An empirical comparison. Am J Epidemiol 1988; 127: 885-892.
23. Keller RB, Rudicel SA, Liang MH. Outcomes research in orthopaedics. J Bone Joint Surg 1993; 75A: 1.562-74.
24. Hennekens CH, Buring JE. Epidemiology in medicine. Mayrent SL (ed). Boston. Little Brown. 1987.
25. Kocher MS, Zurakowski D. Clinical epidemiology and biostatistics: a primer for orthopaedic surgeons (current concepts review). J Bone Joint Surg 2004; 86A: 607-20.
26. Katz J. The Nuremberg Code and the Nuremberg trial. A reappraisal. JAMA 1996; 276: 1.662-6.
27. World Medical Association. Declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. JAMA 1999; 277: 925-6.
28. Kocher MS, Mandiga R, Murphy JM et al. A clinical practice guideline for treatment of septic arthritis in children: efficacy in improving process of care and effect on outcome of septic arthritis of the hip. J Bone Joint Sug 2003; 85: 994-9.
29. Kocher MS, Dicancio J, Zurakowski D, Micheli LJ. Diagnostic performance of clinical examination and selective magnetic resonance imaging in the evaluation of intraarticular knee disorders in children and adolescents. Am J Sports Med 2001; 29: 292-6.
30. Kocher MS. Ultrasonographic screening for developmental dysplasia of the hip: an epidemiologic analysis (part I). Am J Orthop 2000; 29: 929-33.
31. Kocher MS. Ultrasonographic screening for developmental dysplasia of the hip; an epidemiologic analysis (part II). Am J Orthop 2001; 30: 19-24.
32. Baron JA. Uncertainty in Bayes. Med Decis Making 1994; 14: 46-51.
33. Kocher MS, Zurakowski D, Kasser JR. Differentiating between septic arthritis and transient synovitis of the hip in children: an evidence-based clinical prediction algorithm. J Bone Joint Sug 1999; 81A: 1.662-1.670.
34. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC). Radiology 1982; 143: 29-36.
35. Kocher MS, Sterett WI, Briggs KK et al. Effect of functional bracing on subsequent knee injury in ACL-deficient proffesional skiers. J Knee Surg 2003;16: 87-92.
36. Kane RL. Outcome measures. En Kane RL (ed). Understanding health outcomes research. Gaithersburg, MD. Aspen. 1997. Pp. 17-18.
37. Sackett DL, Rosenberg WM. The need for evidence-based medicien. J R Soc Med 1995; 88: 620-624.
38. Evidence-based Medicine Working Group. A new approach to teaching the practice of medicine. JAMA 1992; 268: 2.420-5.
39. Sackett DL, Strauss SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine how to practice and teach evidence-based medicine (ed 2). Edinburgh. Churchill Livingstone. 2000.
40. Bhandari M, Devereaux PJ, Swiontkowski MF et al. Internal fixation compared with arthroplasty for displaced fractures of the femoral neck. A meta analysis. J Bone Joint Surg 2003; 85A: 1.673-1.681.
41. Pauker SG, Naglie IG. Decision analysis. N Engl J Med 1987; 316: 250-6.
42. Birkmeyer JD, Welch HG. A reader's guide to surgical decision analysis. J Am Coll Surg 1997; 184: 589-95.
43. Krahn MD, Naglie G, Maimark D et al. Primer on medical decision analysis: part 4-analysing the model and interpreting the results. Med Decis Making 1997; 17: 147-151.
44. Detsky AS, Naglie IG. A clinician's guide to cost-effectivess analysis. Ann Intern Med 1990; 113: 147-54.
45. Kocher MS, Zurakowski D. Clinical epidemiology and biostatistics: a primer for orthopaedic surgeons. J Bone Joint Surg 2004; 86A: 607-620.
46. Kaplan EL, Meier P. Nonparametric estimator from incomplete observations. Am Sgtatist Assoc 1958; 53: 457-81.
47. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York. Willey. 1980. Pp. 10-14.
48. Mantel N. Evaluation of survival data and two new rank order statistics arising in its considerations. Cancer Chemoter Rep 1966; 50: 163-70.
49. Kocher MS, Steadman JR, Briggs K et al. Determinants of patients satisfaction with outcome after anterior cruciate ligament reconstruction. J Bone Joint Surg 2002; 84A: 1.560-72.
50. Griffin D, Audige L. Common statistical methods in orthopaedic clinical studies. Clin Orthop 2003; 413: 70-9.