presentados por círculos azules y los no diabéticos, por círculos rojos. Observamos que, aunque las concentraciones de glucosa en sangre suelen ser más elevadas en los diabéticos que en los no diabéticos, no existe una concentración que separe claramente los dos grupos; existe cierto solapamiento entre diabéticos y no diabéticos en cada concentración de glucosa en sangre. Sin embargo, debemos seleccionar un punto de corte de modo que aquellos cuyos resultados se encuentren por encima de dicho punto de corte puedan considerarse «positivos» y puedan ser vueltos a explorar con más pruebas, y aquellos cuyos resultados se encuentren por debajo de dicho punto se consideren «negativos» y no sean programados para realizar pruebas adicionales.

Supongamos que se elige una concentración de corte relativamente elevada (fig. 5-3B). Claramente, muchos diabéticos no serán identificados como positivos; por otro lado, la mayoría de los no diabéticos serán identificados correctamente como negativos. Si representamos estos resultados en una tabla de 2 X 2, la sensibilidad de la prueba utilizando esta concentración de corte será del 25% (5/20) y la especificidad, del 90% (18/20).

¿Qué ocurre si se elige una concentración de corte baja (fig. 5-3C)? Muy pocos diabéticos serían mal diagnosticados. ¿Cuál es entonces el problema? Una gran proporción de los no diabéticos son identificados ahora como positivos por la prueba. Como se observa en la tabla de 2 X 2, la sensibilidad es ahora del 85% (17/20), pero la especificidad es únicamente del 30% (6/20).

La dificultad estriba en que en el mundo real no existe una línea vertical que separe los diabéticos de los no diabéticos, y, de hecho, se encuentran mezclados (fig. 5-3D); no son distinguibles ni con círculos rojos o azules (fig. 5-3E). Por tanto, si se usa una concentración de corte elevada (fig. 5-3F), a todos aquellos con resultados por debajo de la línea se les podrá asegurar que no tienen la enfermedad y no necesitan más seguimiento; si se usa una concentración de corte baja (fig. 5-3G), todos aquellos con resultados por encima de la línea serán vueltos a explorar con nuevas pruebas.

En la figura 5-4A se muestran datos reales sobre la distribución de las concentraciones de glucosa en sangre en diabéticos y en no diabéticos. Supongamos que quisiéramos realizar una prueba de cribado en esta población. Si decidimos establecer el punto de corte de modo que podamos identificar a todos los diabéticos (100% de sensibilidad), podríamos elegir una concentración de 80 mg/dl (fig. 5-4B). Sin embargo, el problema es que procediendo así también consideraremos positivos a muchos de los no diabéticos, es decir, la especificidad será muy baja. Por otro lado, si establecemos el punto de corte en 200 mg/dl (fig. 5-4C), todos los no diabéticos serán identificados como negativos (100% de especificidad), pero ahora podemos pasar por alto a muchos de los diabéticos verdaderos debido a que la sensibilidad será muy baja. Por tanto, entre sensibilidad y especificidad existe una compensación: si aumentamos la sensibilidad disminuyendo el punto de corte, disminuimos la especificidad; y si aumentamos la especificidad elevando el punto de corte, estamos reduciendo la sensibilidad. Como dijo un sabio: «Nadie da nada por nada.»

El dilema de decidir si se elige un punto de corte alto o bajo reside en el problema de los falsos positivos y los falsos negativos que resultan de la prueba. Es importante recordar que al realizar pruebas de cribado obtenemos grupos clasificados únicamente según los resultados de las pruebas de cribado, como positivos o negativos. Carecemos de información acerca del verdadero estado de su enfermedad, que, por supuesto, es el motivo para realizar el cribado. De hecho, los resultados de la prueba de cribado no proporcionan cuatro grupos, como se observa en la figura 5-5, sino dos grupos: un grupo de personas con resultados positivos en la prueba y otro grupo con resultados negativos. A los que obtuvieron resultados positivos se les notificarán los resultados de la prueba y se les pedirá que vuelvan para realizar pruebas adicionales. A las personas del otro grupo, con resultados negativos, se les notificará dicho resultado y, por tanto, no se les pedirá que vuelvan para realizar nuevas pruebas (fig. 5-6).

La elección de un punto de corte alto o bajo para realizar pruebas de cribado depende, por tanto, de la importancia que le otorguemos a los falsos positivos y los falsos negativos. Los falsos positivos se asocian con costes (emocionales y económicos), así como con la dificultad de «desetiquetar» a una persona que obtuvo resultados positivos y que posteriormente se concluyó que no presentaba la enfermedad. Además, los resultados falsos positivos suponen una carga importante al sistema de asistencia sanitaria, ya que un grupo numeroso de personas debe ser citado de nuevo para repetir pruebas, cuando sólo unas pocas presentarán la enfermedad. Por otro lado, los pacientes con resultados falsos negativos serán informados de que no tienen la enfermedad y no seguirán siendo revisados, por lo que posiblemente pueden pasarse por alto enfermedades graves en etapas tempranas tratables. Por tanto, la elección de los puntos de corte depende de la importancia relativa de la falsa positividad y la falsa negatividad para la enfermedad en cuestión.

USO DE PRUEBAS MÚLTIPLES

A menudo pueden realizarse varias pruebas de cribado en los mismos individuos, ya sea secuencialmente o simultáneamente. En esta sección se describen los resultados de estos abordajes.

Pruebas secuenciales (en dos etapas)

En las pruebas de cribado secuenciales o en dos etapas, por lo general se realiza primero una prueba menos cara, menos invasiva o menos incómoda, y aquellos en los que el resultado es positivo son vueltos a citar para realizar pruebas adicionales con una prueba más cara, más invasiva o más

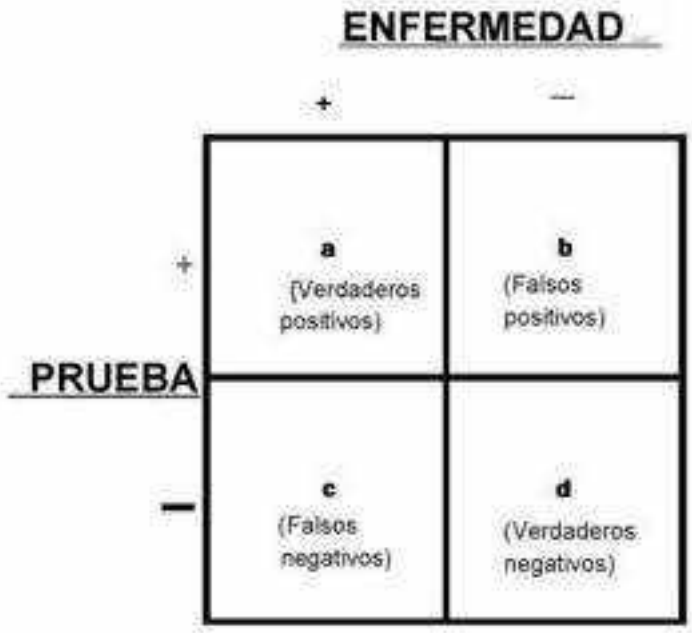


Figura 5-5. Diagrama en el que se muestran cuatro grupos posibles tras una prueba de cribado con una prueba dicotómica.

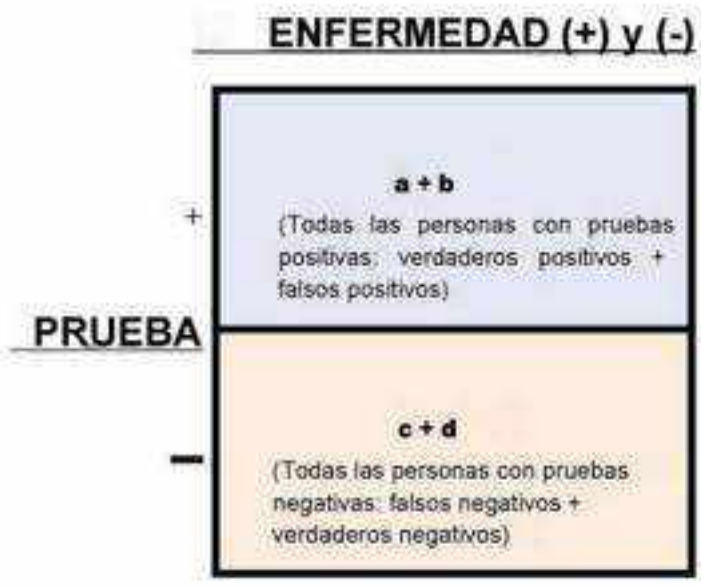


Figura 5-6. Diagrama que muestra los dos grupos de personas resultantes de una prueba de cribado con una prueba dicotómica: todas las personas con resultados positivos en la prueba y todas las personas con resultados negativos en la prueba.

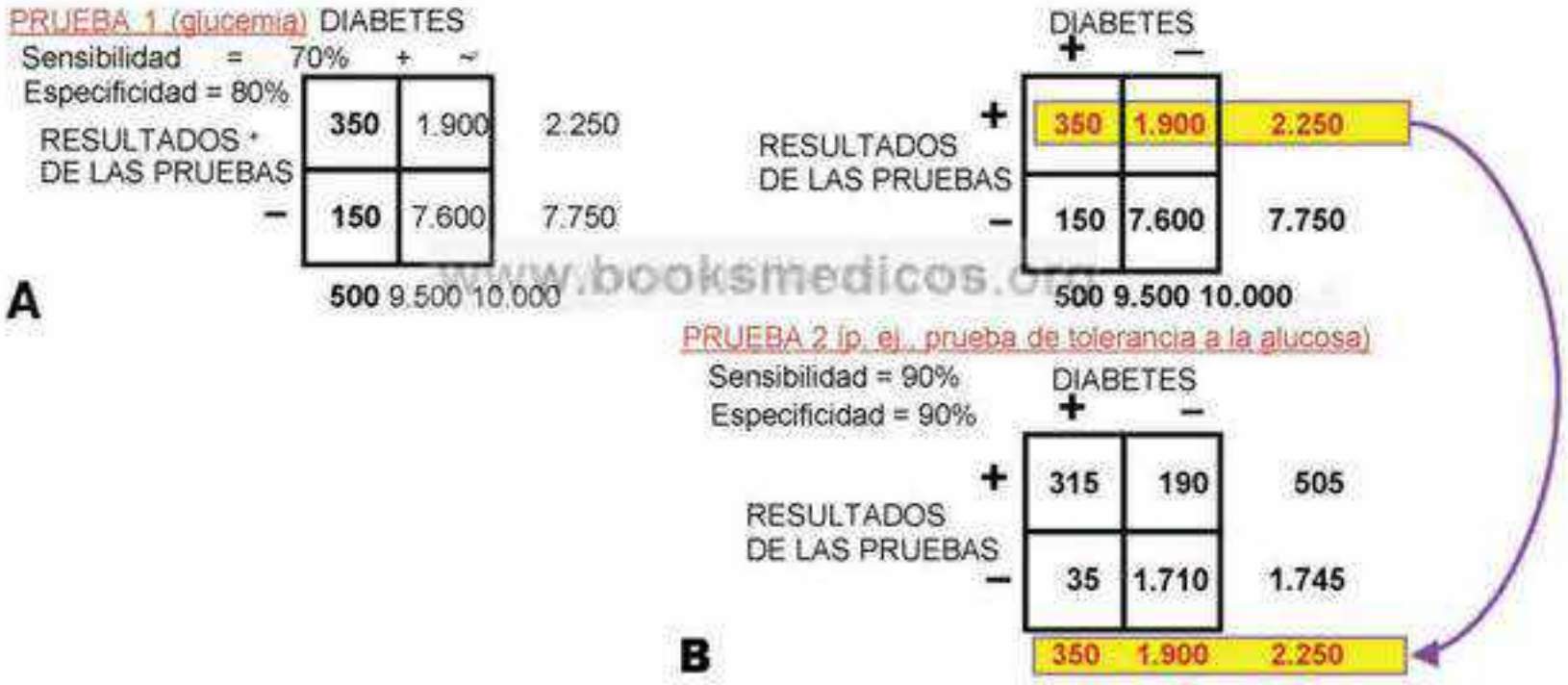


Figura 5-7. A-B, Ejemplo hipotético de un programa de cribado en dos etapas. A, Hallazgos de la prueba 1 en una población de 10.000 personas. B, Hallazgos de la prueba 2 en los participantes con resultados positivos en la prueba 1. (V. explicación en el apartado «Pruebas secuenciales (en dos etapas)».)

incómoda, que puede tener más sensibilidad y especificidad Es de esperar que, citando únicamente a aquellos positivos en la primera prueba de cribado para realizar pruebas adicionales, se reduzca el problema de los falsos positivos.

Consideremos el ejemplo hipotético de la figura 5-7A, en el que se realizan pruebas de cribado de diabetes en una población empleando una prueba con una sensibilidad del 70% y una especificidad del 80%. ¿Cómo se obtienen los datos mostrados en esta tabla? La prevalencia de la enfermedad en esta población es del 5%, por lo que 500 de cada 10.000 habitantes poseen la enfermedad. Con una sensibilidad del 70%, la prueba identificará correctamente a 350 de las 500 personas que tienen la enfermedad. Con una especificidad del 80%, la prueba identificará correctamente como no diabéticas a 7.600 personas de las 9.500 que no son diabéticas; sin embargo, 1.900 de estos 9.500 presentarán resultados positivos. Por tanto, un total de 2.250 personas obtendrán resultados positivos y serán vueltas a citar para realizar una segunda prueba. (Recuérdese que en la vida real no contamos con una línea vertical que separe a los diabéticos de los no diabéticos y no sabemos que sólo 350 de los 2.250 son diabéticos.)

Las 2.250 personas son vueltas a citar para realizar un cribado con una segunda prueba (como la prueba de tolerancia a la glucosa), que, para este ejemplo, asumimos que tiene una sensibilidad del 90% y una especificidad del 90%. En la figura 5-7B se muestra la prueba 1 conjuntamente con la prueba 2, que se realiza

sólo en las 2.250 personas con resultados positivos en la primera prueba de cribado y que han sido citados de nuevo para la segunda etapa del cribado.

Como 350 personas (de las 2.250) presentan la enfermedad y la prueba posee una sensibilidad del 90%, 315 de esas 350 serán identificadas correctamente como positivas. Como 1.900 (de las 2.250) no tienen diabetes y la especificidad de la prueba es del 90%, 1.710 de las 1.900 serán identificadas correctamente como negativas y 190 serán falsos positivos.

Ahora somos capaces de calcular la *sensibilidad* neto y la *especificidad neta* del uso de ambas pruebas secuencialmente. Tras completar ambas pruebas, 315 personas del total de 500 diabéticos en esta población de 10.000 habrán sido considerados correctamente positivos: 315/500 = *sensibilidad neta* del 63%. Por tanto, empleando ambas pruebas secuencialmente se produce una pérdida de sensibilidad neta. Para calcular la *especificidad neta,* hay que tener en cuenta que 7.600 individuos de los 9.500 de esta población que no son diabéticos fueron considerados correctamente negativos en la primera etapa del cribado y no fueron sometidos a más pruebas; en la segunda etapa del cribado 1.710 individuos más de los 9.500 no diabéticos fueron considerados correctamente negativos. Así, un total de 7.600 + 1.710 de los 9.500 no diabéticos fueron considerados correctamente negativos: 9.310/9.500 = *especificidad neta* del 98%. Por tanto, el uso de ambas pruebas secuencialmente ha resultado en una ganancia de *especificidad neta.*

Pruebas simultáneas

Centrémonos ahora en el uso de pruebas simultáneas. Asumamos que en una población de 1.000 personas, la prevalencia de una enfermedad es del 20%. Por tanto, 200 personas padecen la enfermedad, pero no sabemos quiénes son. Para identificar a las 200 personas que tienen esta enfermedad, realizamos pruebas de cribado en esta población de 1.000 personas utilizando 2 pruebas para esta enfermedad, la prueba A y la prueba B, al mismo tiempo. Asumamos que la sensibilidad y la especificidad de las dos pruebas son las siguientes:

Prueba A	Prueba B
Sensibilidad = 80%	Sensibilidad = 90%
Especificidad = 60%	Especificidad = 90%

Sensibilidad neta utilizando dos pruebas simultáneas

La primera pregunta que nos planteamos es: ¿cuál es la *sensibilidad neta* si se utilizan la prueba A y la prueba B *simultáneamente*? Para considerar a una persona positiva y, por tanto, poder incluirla en el numerador para calcular la sensibilidad neta de las dos pruebas utilizadas simultáneamente, dicha persona debe ser identificada como positiva por la prueba A, la prueba B o ambas.

Para calcular la sensibilidad neta, consideremos primero los resultados del cribado con la prueba A, cuya sensibilidad es del 80%: de las 200 personas que tienen la enfermedad, 160 son identificadas como positivas (tabla 5-3). En la figura 5-8A, la elipse representa a las 200 personas que tienen la enfermedad. En la figura 5-8B, el círculo rosa en el interior de la elipse representa a las 160 personas identificadas como positivas con la prueba A. Estas 160 personas son verdaderos positivos con la prueba A.

Consideremos a continuación los resultados del cribado con la prueba B, cuya sensibilidad es del 90%

TABLA 5-3. Resultados del cribado con la prueba A

Resultados del cribado	POBLACION: Enfermedad	POBLACION: No enfermedad
Positivo	**160**	320
Negativo	40	480
Total	200	800

Sensibilidad = 80% Especificidad = 60%

TABLA 5-4. Resultados del cribado con la prueba B

Resultados del cribado	POBLACIÓN: Enfermedad	POBLACIÓN: No enfermedad
Positivo	**180**	80
Negativo	20	720
Total	200	800

Sensibilidad = 90% Especificidad = 90%

(tabla 5-4). De las 200 personas que tienen la enfermedad, 180 son identificadas como positivas por la prueba B. En la figura 5-8C, la elipse representa de nuevo a las 200 personas que tienen la enfermedad. El círculo azul en el interior de la elipse representa a las 180 personas identificadas como positivas con la prueba B. Estas 180 personas son verdaderos positivos con la prueba B.

Con el fin de calcular el numerador para la sensibilidad neta, no podemos sumar simplemente el número de personas identificadas como positivas con la prueba A y el número de personas identificadas como positivas con la prueba B, pues algunas personas fueron identificadas como positivas con ambas pruebas. Estas personas se representan en lavanda en el área de solapamiento entre ambos círculos, y no queremos contarlas dos veces (fig. 5-8D). ¿Cómo determinamos cuántas personas fueron identificadas como positivas con ambas pruebas?

La prueba A posee una sensibilidad del 80% y, por tanto, identifica como positivas al 80% de las 200 personas que tienen la enfermedad (160 personas). La prueba B posee una sensibilidad del 90% y, por tanto, identifica como positivas al 90% de las mismas 160 personas que fueron identificadas por la prueba A (144 personas). Por tanto, cuando empleamos simultáneamente las pruebas A y B, 144 personas son identificadas como positivas con ambas pruebas (fig. 5-8E).

Recordemos que la prueba A identificó correctamente como positivas a 160 personas con la enfermedad. Como 144 de las mismas fueron identificadas por ambas pruebas, $160 - 144 = 16$ personas fueron identificadas correctamente *sólo* con la prueba A.

La prueba B identificó correctamente como positivas a 180 de las 200 personas con la enfermedad. Como 144 de las mismas fueron identificadas por ambas pruebas, $180 - 144 = 36$ personas fueron identificadas correctamente *sólo* con la prueba B. Por tanto, como se observa en la figura 5-8F, cuando se emplean simultáneamente las pruebas A y B, la

$$\text{sensibilidad neta} = \frac{16+144+36}{200} \frac{196}{200} = 98\%$$

Especificidad neta utilizando dos pruebas simultáneas

La siguiente pregunta que debemos plantearnos es: ¿cuál es la *especificidad neta* si se emplean las pruebas A y B *simultáneamente*?. Para poder incluir a una persona en el numerador para calcular la especificidad neta de las dos pruebas utilizadas simultáneamente, dicha persona debe ser identificada como *negativa* por *ambas pruebas*. Con el fin de calcular el numerador para la especificidad neta, necesitamos por tanto determinar cuántas personas presentaron resultados negativos en ambas pruebas. ¿Cómo hacemos esto?

La prueba A posee una especificidad del 60% y, por tanto, identifica correctamente al 60% de las 800 personas que no tienen la enfermedad (480 personas) (tabla 5-5). En la figura 5-9A, la elipse representa a las 800 personas que no tienen la enfermedad. El círculo verde en el interior de la elipse de la figura 5-9B representa a las 480 personas con resultados negativos en la prueba A. Éstos son los verdaderos negativos empleando la prueba A.

La prueba B posee una especificidad del 90% y, por tanto, identifica como negativas al 90% de las 800 personas que no tienen la enfermedad (720 personas) (tabla 5-6 y círculo amarillo de la fig. 5-9C). Sin embargo, para ser identificadas como negativas en pruebas simultáneas, sólo se considera que tienen resultados ne-

TABLA 5-5. Resultados del cribado con la prueba A

Resultados del cribado	POBLACION: Enfermedad	POBLACION: No enfermedad
Positivo	160	320
Negativo	40	**480**
Total	200	800

Sensibilidad = 80% Especificidad = 60%

TABLA 5-6. Resultados del cribado con la prueba B

Resultados del cribado	POBLACION: Enfermedad	POBLACION: No enfermedad
	180	80
	20	**720**
Total	200	800

Sensibilidad = 90% Especificidad = 90%

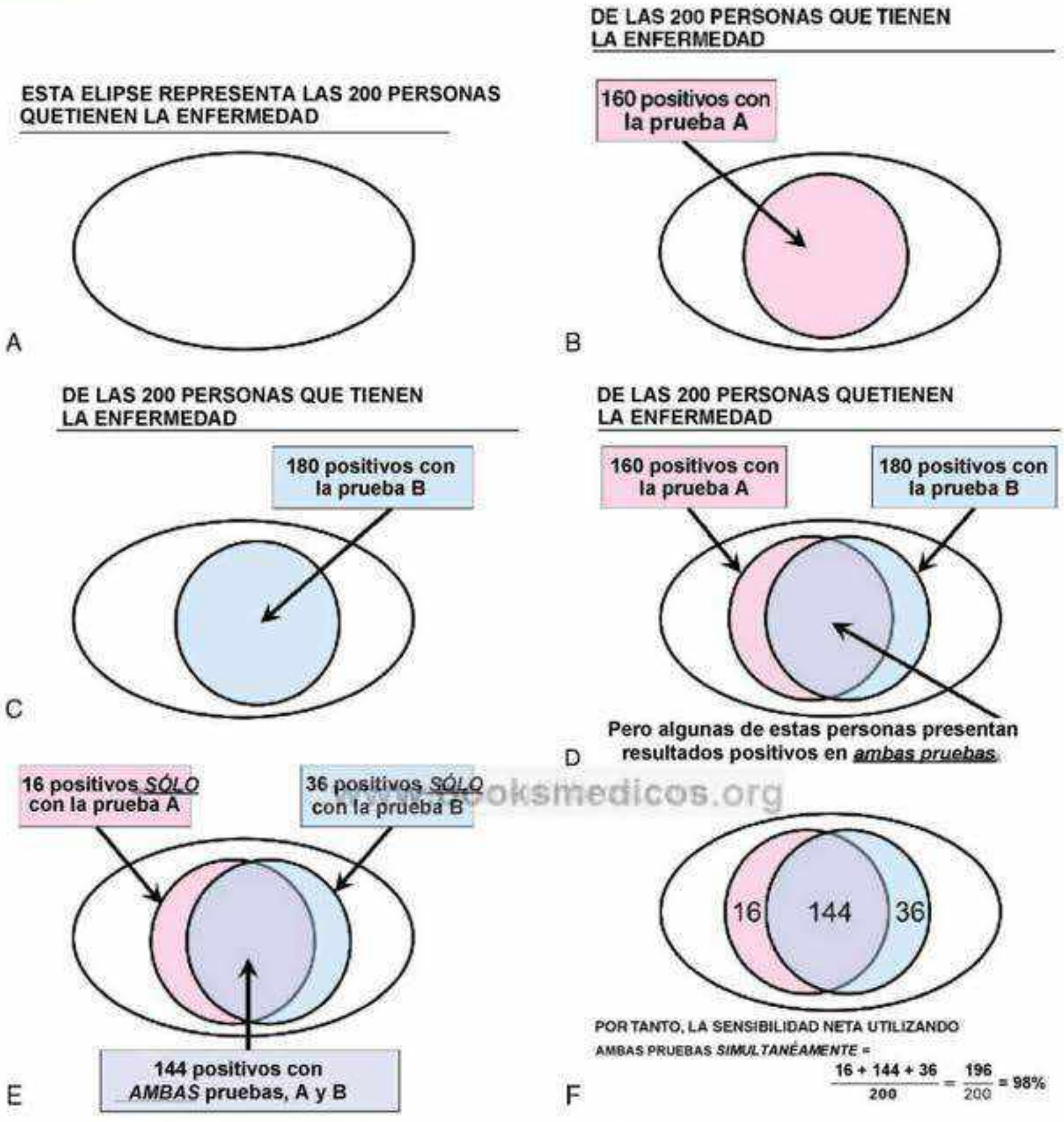


Figura 5-8. A-F, Sensibilidad neta: ejemplo hipotético de pruebas simultáneas. (V. explicación en el apartado «Sensibilidad neta utilizando dos pruebas simultáneas», pág. 96.)

gativos las personas con resultados negativos en ambas pruebas (fig. 5-9D). Estas personas se muestran en verde claro en el área de solapamiento entre los dos círculos. La prueba B también identifica como negativas al 90% de las mismas 480 personas identificadas como negativas por la prueba A (432 personas). Por tanto, como se muestra por los círculos que se solapan, cuando se utilizan simultáneamente las pruebas A y B, 432 personas son identificadas como negativas por ambas pruebas (fig. 5-9E). Así, cuando se emplean simultáneamente las pruebas A y B (fig. 5-9F), la

$$\text{especificidad neta} = \frac{432}{800} = 54\%$$

Por tanto, cuando se emplean dos pruebas simultáneas existe una ganancia neta de sensibilidad (del 80% utilizando la prueba A y el 90% utilizando la prueba B al 98% utilizando ambas pruebas simultáneamente). Sin embargo, existe una pérdida neta de especificidad (especificidad neta = 54%) respecto a cuando se utiliza cada prueba aisladamente (especificidad del 60% con la prueba A y del 90% con la prueba B).

Comparación de las pruebas simultáneas y secuenciales

En un contexto clínico, a menudo se utilizan múltiples pruebas simultáneamente. Por ejemplo, un paciente

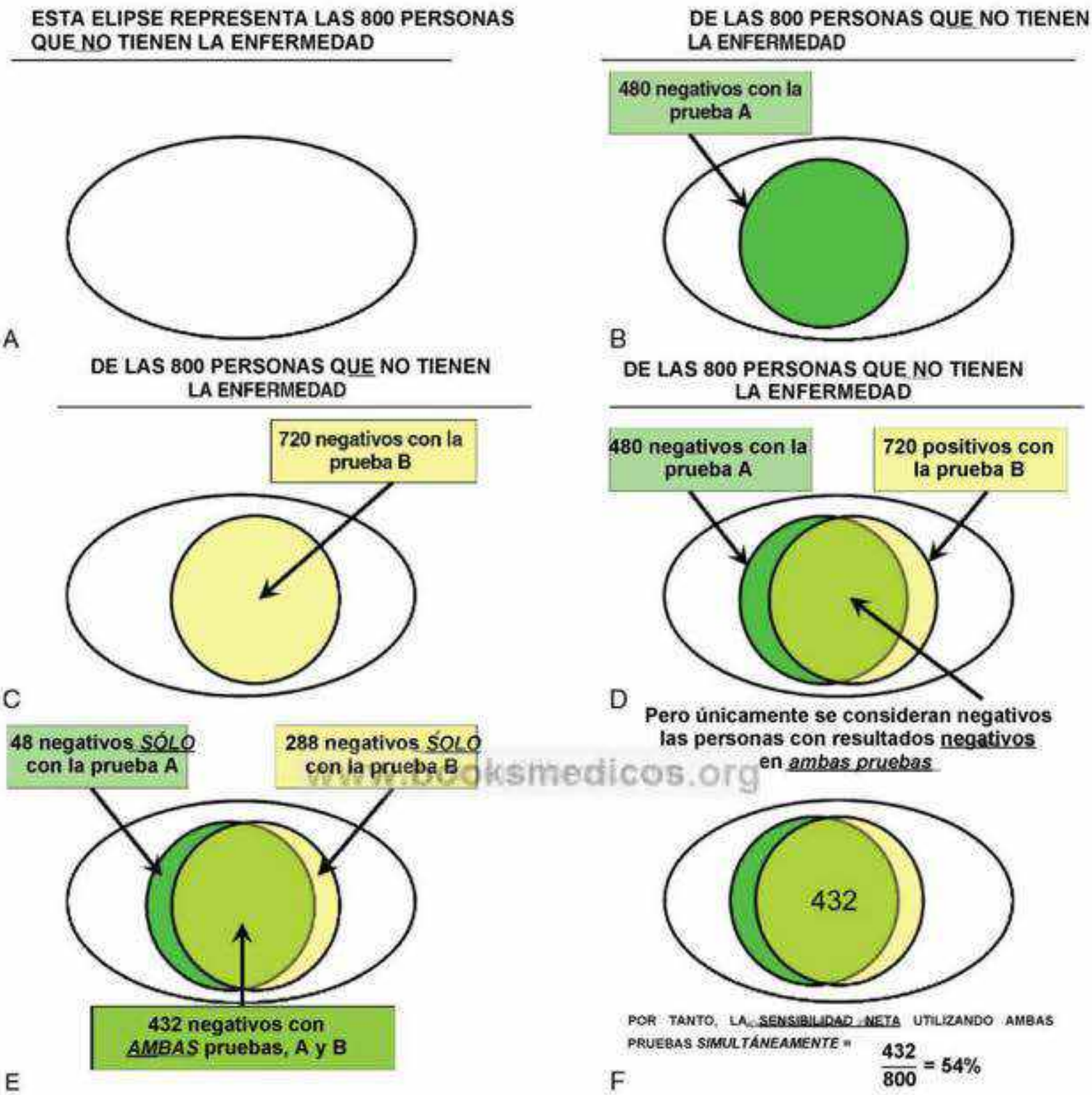
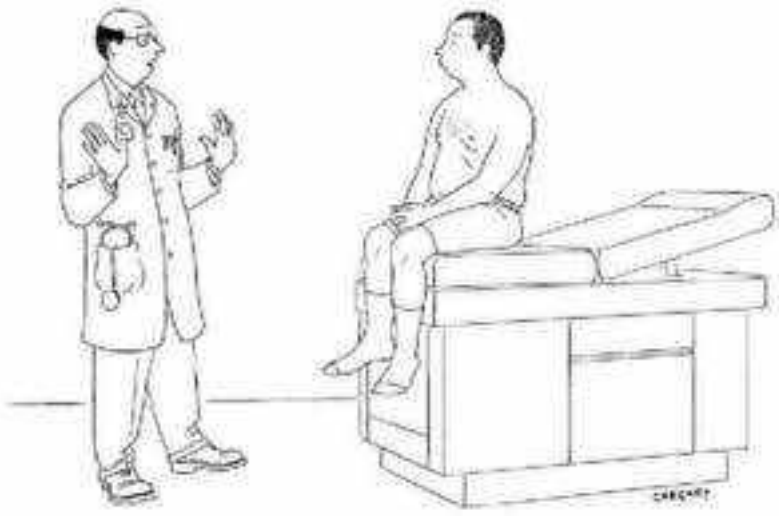


Figura 5-9. A-F, Sensibilidad neta: ejemplo hipotético de pruebas simultáneas. (V. explicación en el apartado «Sensibilidad neta utilizando dos pruebas simultáneas», pág. 97.)

ingresado en un hospital puede ser sometido a una batería de pruebas en el momento del ingreso. Cuando se utilizan múltiples pruebas simultáneamente para detectar una enfermedad específica, generalmente se considera que el resultado de la prueba en el paciente es «positivo» si ha obtenido un resultado positivo en una o varias de las pruebas. Se considera que el resultado de las pruebas del paciente es «negativo» si los resultados de todas las pruebas son negativos. Los efectos de este abordaje sobre la sensibilidad y la especificidad difieren de los que resultan de las pruebas secuenciales. Con las pruebas secuenciales, cuando volvemos a realizar pruebas a los identificados como positivos con la primera prueba, se produce una pérdida en la sensibilidad neta y una ganancia en la especificidad neta. Cuando se emplean las pruebas simultáneas, como un individuo identificado como positivo en *una o* en *múltiples* pruebas es considerado positivo, se produce una ganancia en la sensibilidad neta. Sin embargo, para ser considerada negativa, una persona debería obtener resultados negativos en *todas* las pruebas realizadas. Como resultado, se produce una pérdida en la especificidad neta.

En resumen, como hemos visto previamente, cuando se utilizan dos pruebas secuenciales, y los individuos que han obtenido resultados positivos en la primera prueba son vueltos a explorar con la segunda prueba, se produce una pérdida neta en la sensibilidad, pero una ganancia neta de la especificidad, en comparación a cuando cada prueba se realiza aisladamente. Sin embargo, cuando se utilizan dos pruebas simultáneamente, se produce una ganancia neta de sensibilidad y una pérdida neta de especificidad, en comparación a cuando las pruebas se realizan aisladamente.

Considerando estos resultados, la decisión de utilizar pruebas secuenciales o simultáneas a menudo se basa en los objetivos de las pruebas (la prueba se realiza con fines diagnósticos o de cribado) y en función de consideraciones prácticas relacionadas con el contexto en el que se realizan las pruebas, como la duración del ingreso hospitalario, los costes y el grado de invasividad de cada prueba, así como el grado de cobertura del seguro a terceros. En la figura 5-10 se muestra a un médico afrontando la sobrecarga de información percibida.



"Whoa—way too much information."

Figura 5-10. «¡Basta! Me está dando demasiada información.» (© The New Yorker Collection 2002. Alex Gregory from cartoonbank.com. Reservados todos los derechos.)

VALOR PREDICTIVO DE UNA PRUEBA

Hasta ahora nos hemos preguntado cómo es de buena la prueba para identificar a las personas que tienen la enfermedad y a las que no la tienen. Este punto es importante, especialmente cuando se realizan pruebas de cribado en poblaciones de la comunidad. En efecto, nos preguntamos: «Si realizamos un cribado en una población, ¿qué proporción de las personas que tienen la enfermedad serán identificadas correctamente?». Este aspecto es claramente una consideración de salud pública importante. En el contexto clínico, sin embargo, para el médico puede ser importante otra pregunta: si los resultados de la prueba son positivos en este paciente, ¿cuál es la probabilidad de que dicho paciente tenga la enfermedad? Éste es el denominado *valor predictivo positivo* (VPP) de la prueba. En otras palabras, ¿qué proporción de los pacientes con resultados positivos en la prueba tienen realmente la enfermedad en cuestión? Para calcular el valor predictivo positivo, dividimos el número de verdaderos positivos entre el número total de personas con resultados positivos (verdaderos positivos + falsos positivos).

Volvamos al ejemplo que se muestra en la tabla 5-1, en el que se realiza un cribado en una población de 1.000 personas. Como se observa en la tabla 5-7, la tabla de 2 X 2 muestra los resultados de una prueba de cribado dicotómica en dicha población. De las 1.000

TABLA 5-7. Valor predictivo de una prueba

Resultados del cribado	POBLACIÓN: Enfermedad	No enfermedad	Totales	
Positivo	80	100	180	Valor predictivo positivo = 80/180 = 44%
Negativo	20	800	820	Valor predictivo positivo = 800/820 = 98%
Totales	100	900	1,000	

personas, el resultado de la prueba es positivo en 180; de estas 180 personas, 80 tienen la enfermedad. Por tanto, el *valor predictivo positivo* es de 80/180 = 44%.

Sobre los resultados negativos de la prueba puede plantearse una pregunta paralela: «Si el resultado de la prueba es negativo, ¿cuál es la probabilidad de que este paciente no tenga la enfermedad?». Éste es el *valor predictivo negativo* (VPN) de la prueba. Se calcula dividiendo el número de verdaderos negativos entre el total de resultados negativos (verdaderos negativos + falsos negativos). Fijándonos de nuevo en el ejemplo de la tabla 5-7, la prueba arroja un resultado negativo en 820 personas, y de éstas, 800 no tienen la enfermedad. Por tanto, el *valor predictivo negativo* es de 800/820 = 98%.

Cada prueba realizada por un médico (historia clínica, exploración física, pruebas de laboratorio, radiografías, electrocardiogramas y otras intervenciones) se utiliza para facilitar la labor del médico para emitir un diagnóstico correcto. Lo que se quiere saber cuando se realiza una prueba a un paciente es: «Considerando este resultado positivo de la prueba, ¿cuál es la probabilidad de que el paciente tenga la enfermedad?».

A diferencia de la sensibilidad y la especificidad de la prueba, que pueden considerarse características de la prueba que se está utilizando, el valor predictivo positivo se ve afectado por dos factores: la prevalencia de la enfermedad en la población estudiada y, cuando la enfermedad es infrecuente, la especificidad de la prueba que se está empleando. En las siguientes secciones se analizan estas relaciones.

Relación entre el valor predictivo positivo y la prevalencia de la enfermedad

En la siguiente exposición del valor predictivo, el término *valor predictivo* se utiliza para denotar el valor predictivo *positivo* de la prueba.

La relación entre el valor predictivo y la *prevalencia de la enfermedad* puede verse en el ejemplo mostrado en la tabla 5-8. En primer lugar, dirijamos nuestra atención a la parte superior de la tabla. Asumamos que estamos utilizando una prueba con una sensibilidad del 99% y una especificidad del 95% en una población de 1.000 personas en la que la prevalencia de la enfermedad es del 1%. Como la prevalencia es del 1%, 100 de las 1.000 personas presentan la enfermedad y 9.900 no la presentan. Con una sensibilidad del 99%, la prueba identifica correctamente a 99 de las 100 personas que tienen la enfermedad. Con una especificidad del 95%, la prueba identifica correctamente como negativas a 9.405 de las 9.900 personas que no tienen la enfermedad. Por tanto, en esta población con una prevalencia del 1%, la prueba identifica como positivas a 594 personas (99 + 495). Sin embargo, de estas 594 personas, 495 (38%) son falsos positivos y, por tanto, el valor predictivo positivo es de 99/594, o de tan sólo el 17%.

Apliquemos ahora la misma prueba (con la misma sensibilidad y especificidad) a una población con una enfermedad de prevalencia más elevada (5%), como se observa en la parte inferior de la tabla 5-8. Realizando cálculos similares a los empleados en la parte superior de la tabla, el valor predictivo positivo es ahora del 51%. Por tanto, la mayor prevalencia en la población cribada ha causado un aumento importante del valor predictivo positivo utilizando la misma prueba. En la figura 5-11 se muestra la relación entre la prevalencia de la enfermedad y el valor predictivo. Claramente, la mayor parte de la ganancia del valor predictivo se produce cuando aumenta la prevalencia de la enfermedad en los casos en que ésta es más baja.

¿Por qué debe interesarnos la relación entre el valor predictivo y la prevalencia de la enfermedad? Como hemos visto, cuanto más elevada sea la prevalencia, mayor será el valor predictivo. Por tanto, un programa de cribado es más productivo y eficiente si se dirige a una población objetivo de alto riesgo. El cribado de una población completa para una enfermedad relativamente infrecuente puede suponer un gasto impor-

TABLA 5-8. Relación entre la prevalencia de la enfermedad y el valor predictivo positivo

EJEMPLO: SENSIBILIDAD == 99%, ESPECIFICIDAD 95%

Prevalencia de la enfermedad	Resultados de la prueba	Enfermos	No enfermos	Totales	Valor predictivo positivo
1%	+	99	495	594	$\frac{99}{594} = 17\%$
	−	1	9.405	9.406	
	Totales	100	9.900	10.000	
5%	+	495	475	970	$\frac{495}{970} = 51\%$
	−	5	9.025	9.030	
	Totales	500	9.500	10.000	

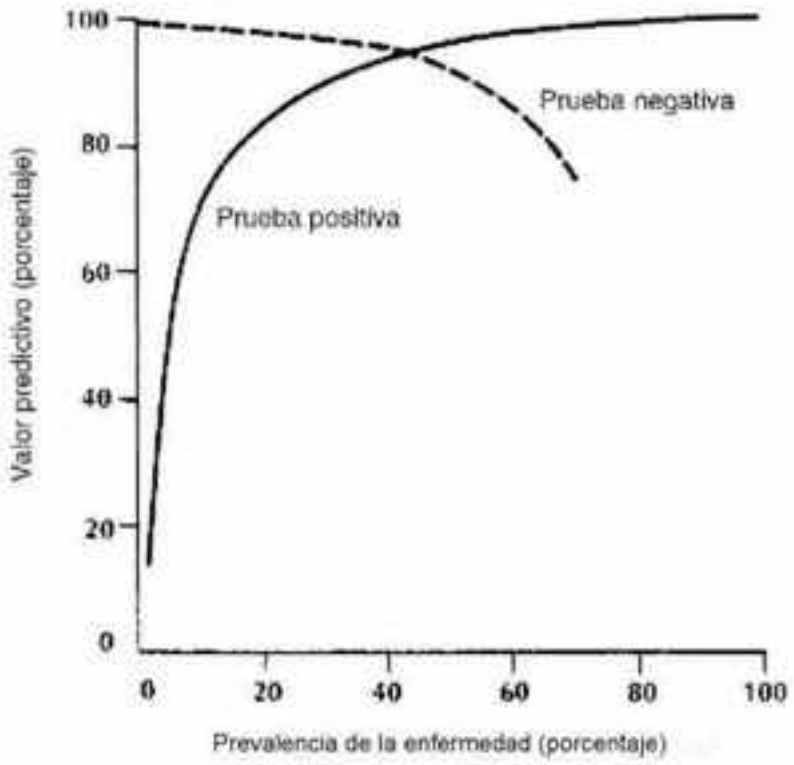


Figura 5-11. Relación entre la prevalencia de la enfermedad y el valor predictivo en una prueba con una sensibilidad del 95% y una especificidad del 95%. (De Mausner JS, Kramer S: Mausner and Bahn Epidemiology: An Introductory Text. Filadelfia, WB Saunders, 1985, pág. 221.)

tante de recursos y puede lograr la detección de pocos casos previamente no detectados en relación con la cantidad de esfuerzo empleado. Sin embargo, si puede identificarse un subgrupo de alto riesgo y el cribado puede centrarse en este subgrupo, es probable que el programa sea mucho más productivo. Además, una población de alto riesgo puede estar más motivada para participar en dicho programa de cribado y es más probable que adopte las acciones recomendadas si los resultados del cribado son positivos.

La relación entre valor predictivo y prevalencia de la enfermedad también muestra que los resultados de cualquier prueba deben interpretarse en el contexto de la prevalencia de la enfermedad en la población a la que pertenece el individuo. Un ejemplo interesante lo constituye la determinación de la concentración de a-fetoproteína (AFP) en el líquido amniótico para el diagnóstico prenatal de la espina bífida. En la figura 5-12 se muestra la distribución de las concentraciones de AFP en líquido amniótico en embarazos normales y en embarazos en los que el feto presentó espina bífida, que es un defecto del tubo neural. Aunque la distribución es bimodal, existe un tramo en el que la curva se solapa y en dicho tramo no siempre está claro a qué curva pertenecen la madre y el feto. Sheffield y cois.[1] revisaron los trabajos publicados y crearon poblaciones artificiales de 10.000 mujeres en las que se realizó un cribado de la AFP en el líquido amniótico para identificar fetos con espina bífida. Crearon dos poblaciones: una con alto riesgo de espina bífida y otra con riesgo normal.

En la tabla 5-9 se muestran los cálculos en las mujeres de alto y bajo riesgo. ¿Qué mujeres tienen un riesgo elevado de tener un hijo con espina bífida? Se sabe que las mujeres que han tenido previamente un hijo con un defecto del tubo neural poseen un riesgo mayor porque se sabe que el defecto se reproduce en los hermanos. En estos cálculos, el valor predictivo positivo fue del 82,9%. ¿Qué mujeres tienen un riesgo bajo pero aun así son sometidas a una amniocentesis? Las mujeres de mayor edad son sometidas a una amniocentesis debido a la posibilidad de tener un hijo con síndrome de Down o algún otro defecto asociado con el embarazo en madres de mayor edad. El riesgo de espina bífida, sin embargo, no se relaciona con la edad de la madre, por lo que estas mujeres no tienen un riesgo superior de tener un hijo con espina bífida. Los cálculos demuestran que, utilizando la misma prueba para la AFP que la empleada en las mujeres de alto riesgo, el valor predictivo positivo de la prueba es de tan sólo el 41,7%, considerablemente inferior al calculado en el grupo de alto riesgo.

Por tanto, vemos que la misma prueba puede tener un valor predictivo muy diferente cuando se realiza en

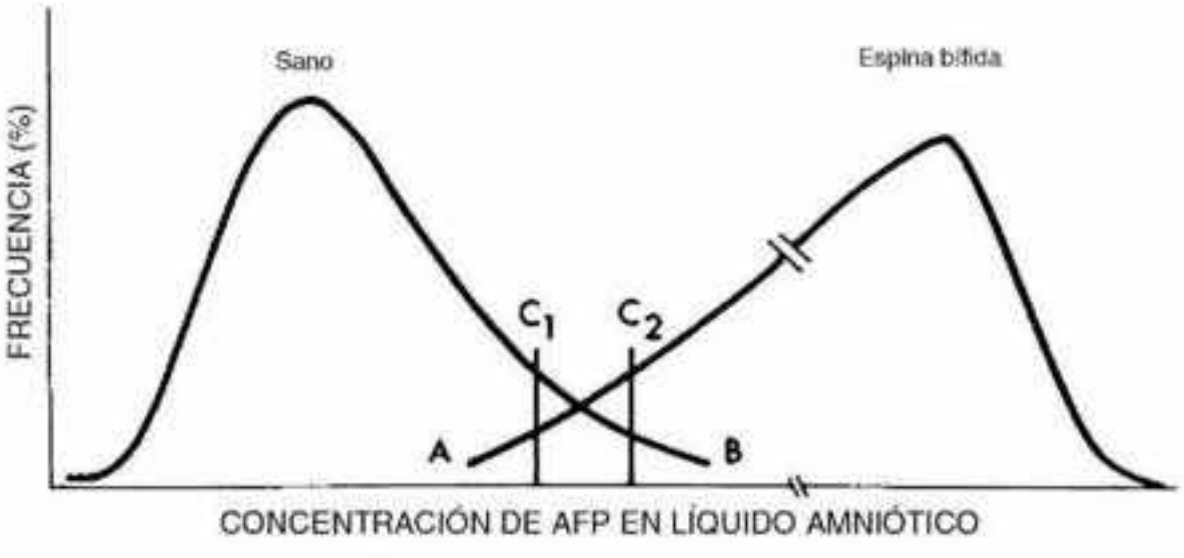


Figura 5-12. Concentración de a-fetoproteína (AFP) en el líquido amniótico de individuos sanos y de pacientes con espina bífida. (De Sheffield LJ, Sackett DL, Goldsmith CH, et al: A clinical approach to the use of predictive values in the prenatal diagnosis of neural tube defects. Am J Obstet Gynecol 145:319-324, 1983.)

TABLA 5-9. Cálculos de los valores predictivos para los defectos del tubo neural (DTN)* de la prueba de la oc-fetoproteína (AFP) en mujeres de alto riesgo y bajo riesgo

	RESULTADO DE LA GESTACIÓN		
Prueba de la AFP	DTN	Normal	Totales Valor predictivo (%)
Mujeres de alto riesgo Anormal	87	18	105 82,9
Normal	13	9.882	9.895 99,9
Totales	100	9.900	10.000
Mujeres de bajo riesgo Anormal	128	179	307 41,7
Normal	19	99.674	99.693 99,98
Totales	147	99.853	100.000

*Espina bífida o encefalocele.
De Sheffield LJ, Sackett DL, Goldsmith CH, et al: A clinical approach to the use of predictive values in the prenatal diagnosis of neural tube defects. Am J Obstet Gynecol 145:319-324,1983.

una población de alto riesgo (prevalencia elevada) o en una población de bajo riesgo (prevalencia baja). Las implicaciones clínicas de esta observación son claras: una mujer puede tomar la decisión de interrumpir un embarazo y un médico puede aconsejar a dicha mujer basándose en los resultados de la prueba. Sin embargo, el mismo resultado de la prueba puede interpretarse de modo diferente, dependiendo de si la mujer pertenece a un grupo de mujeres de alto o bajo riesgo, lo que se reflejará en el valor predictivo positivo de la prueba. Por tanto, el resultado de la prueba de modo aislado puede no ser suficiente para servir de guía sin tener en cuenta las otras consideraciones que acabamos de describir.

Los siguientes ejemplos reales destacan la importancia de este aspecto:

El líder de un sindicato de bomberos consultó a un cardiólogo universitario porque el médico de su unidad había leído un artículo en una revista médica de impacto que describía que cierto hallazgo electro car diográfico era muy predictivo de la existencia de cardiopatía coronaria grave, generalmente no reconocida. Basándose en este artículo, el médico de la unidad estaba apartando de tareas activas a muchos bomberos jóvenes, en buena condición física. El cardiólogo leyó el artículo y observó que el estudio se había efectuado en pacientes hospitalizados.

¿Cuál fue el problema? Como los pacientes hospitalizados poseen una prevalencia mucho mayor de cardiopatías que el grupo de bomberos jóvenes, el médico del cuerpo de bomberos había tomado erróneamente el elevado valor predictivo obtenido al estudiar una población con una gran prevalencia y lo había aplicado incorrectamente a una población de bomberos jóvenes de baja prevalencia, en los que la misma prueba habría arrojado un valor predictivo mucho más bajo.

Otro ejemplo:

Un médico visitó a su internista general para un examen médico anual rutinario, que incluía una exploración de heces para descartar sangre oculta. Una de las tres muestras de heces examinadas en la prueba fue positiva. El internista dijo a su paciente médico que el resultado no era significativo porque de manera regular encontraba muchos resultados falsos positivos en su ajetreada consulta. La prueba se repitió en tres nuevas muestras de heces y todas fueron ahora negativas. Sin embargo, percibiendo la preocupación persistente de su paciente, el internista remitió a su paciente médico a un gastroenterólogo. El gastroenterólogo dijo: «En mi experiencia, el hallazgo positivo en heces es grave. Dicho hallazgo casi siempre se asocia con trastornos gastrointestinales patológicos. Los resultados negativos posteriores no significan nada, porque podría tener un tumor que únicamente sangre intermitentemente.»

¿Quién tenía razón en este ejemplo? La respuesta es que tanto el internista general como el gastroenterólogo tenían razón. El internista emitió su valoración del valor predictivo basándose en su experiencia en su práctica médica general, una población con una prevalencia baja de enfermedades gastrointestinales graves. Por otra parte, el gastroenterólogo emitió su valoración del valor predictivo de la prueba basándose en su experiencia de pacientes remitidos, una consulta en la que la mayoría de los pacientes son remitidos debido a la posibilidad de que padezcan una enfermedad gastrointestinal grave (una población con una prevalencia elevada).

Relación entre el valor predictivo positivo y la especificidad de la prueba

En la siguiente exposición, el término *valor predictivo* se utiliza para referirse al valor predictivo *positivo* de la prueba.

Un segundo factor que afecta al valor predictivo de una prueba es la *especificidad* de la misma. Daremos ejemplos de esto primero en forma gráfica y luego en forma de tabla. En la figura 5- 13A-D se presentan en un diagrama los resultados del cribado de una población; sin embargo, las tablas 2 X 2 de estas figuras difieren de las presentadas en figuras anteriores. Cada celdilla se dibuja con su tamaño proporcional a la población que representa. En cada figura, las celdillas que representan a personas con resultados positivos en la prueba están coloreadas de azul; éstas son las celdillas que serán utilizadas para calcular el valor predictivo positivo.

En la figura 5-13A se muestra la población cribada que es utilizada en nuestro análisis: una población de 1.000 personas en la que la prevalencia es del 50%, es decir, 500 personas tienen la enfermedad y 500 no la tienen. Al analizar esta figura, también asumimos que la prueba de cribado que fue utilizada posee una sensibilidad del 50% y una especificidad el 50%. Como el resultado fue positivo en 500 personas, y 250 de éstas tienen la enfermedad, el valor predictivo es de 250/500 o del 50%.

Afortunadamente, la prevalencia de la mayoría de las enfermedades es mucho menor del 50%; por lo general tratamos con enfermedades relativamente infrecuentes. Por tanto, la figura 5-13B asume una prevalencia más baja, del 20% (aunque incluso esta cifra sería una prevalencia inusualmente alta para la mayoría de las enfermedades). Tanto la sensibilidad como la especificidad siguen siendo del 50%. Ahora sólo 200 de las 1.000 personas tienen la enfermedad y la línea vertical que separa a los enfermos de los no enfermos se ha desplazado a la izquierda. El valor predictivo ahora se calcula así: 100/500 = 20%.

Dado que estamos realizando un cribado en una población con una tasa de prevalencia más baja, ¿podemos mejorar el valor predictivo? ¿Cuál sería el efecto en el valor predictivo si aumentásemos la sensibilidad de la prueba? En la figura 5-13C se muestran los resultados cuando mantenemos la prevalencia del 20% y la especificidad del 50% pero aumentamos la sensibilidad

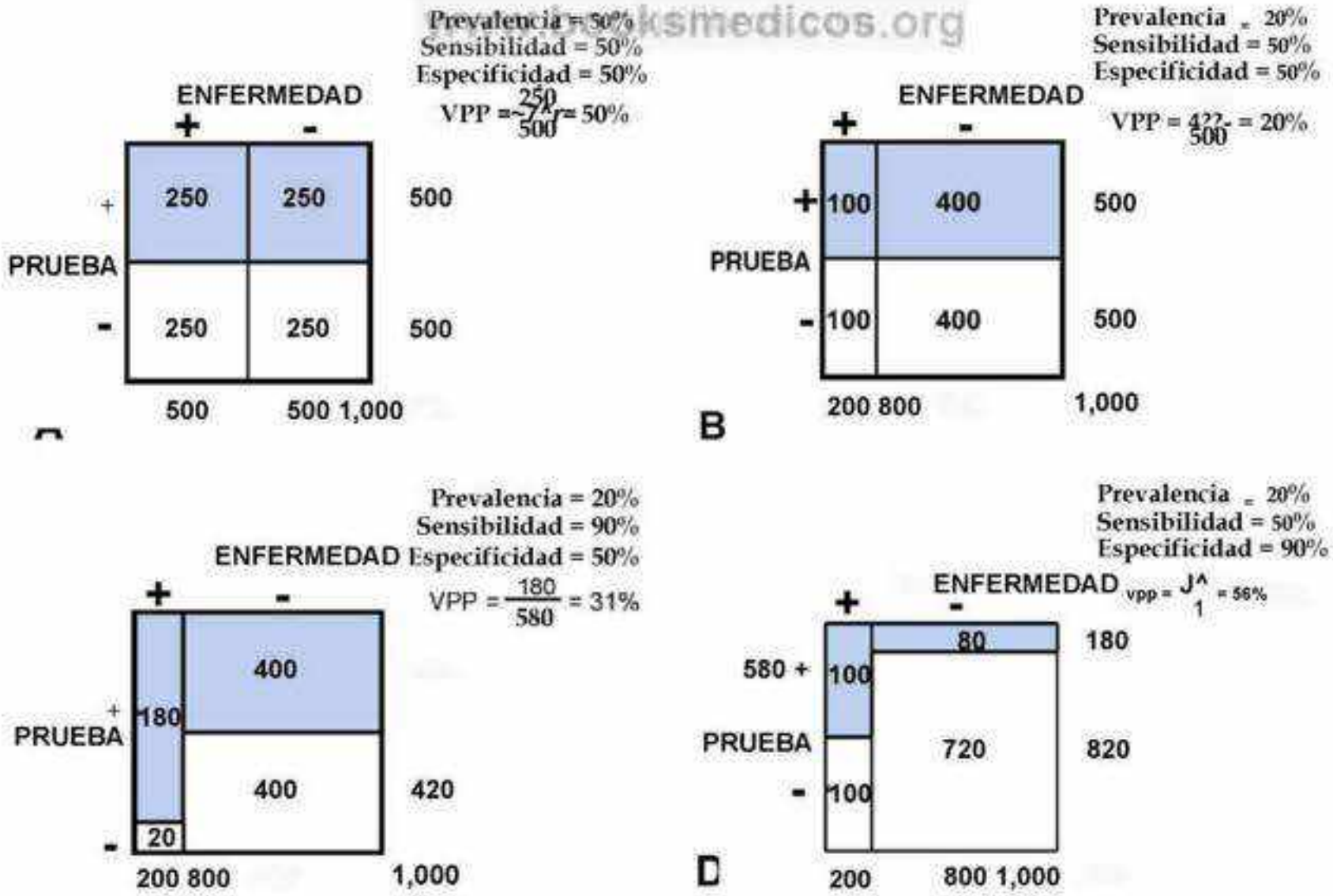


Figura 5-13. A-D, Relación entre la especificidad y el valor predictivo positivo (VPP). (V. explicación en el apartado «Relación entre el valor predictivo positivo y la especificidad de la prueba».)

al 90%. El valor predictivo es ahora 180/850 = 31%, un aumento modesto.

¿Y qué pasaría si en vez de aumentar la sensibilidad de la prueba aumentamos su especificidad? En la figura 5-13D se muestran los resultados cuando se mantiene la prevalencia al 20% y la sensibilidad al 50% pero aumentamos la especificidad al 90%. El valor predictivo ahora es de 100/180 = 56%. Por tanto, el aumento de la especificidad produce un mayor aumento del valor predictivo que el logrado con el mismo aumento de la sensibilidad.

¿Por qué la especificidad tiene mayor influencia sobre el valor predictivo que la sensibilidad? La respuesta es clara si observamos estas figuras. Como estamos tratando con enfermedades infrecuentes, la mayor parte de la población se encuentra a la derecha de la línea vertical. Por tanto, cualquier cambio a la derecha de la línea vertical afecta a un mayor número de personas que un cambio comparable a la izquierda de la línea. Así, un cambio en la especificidad produce un mayor efecto sobre el valor predictivo que un cambio comparable en la sensibilidad. Si estuviéramos tratando con una enfermedad de gran prevalencia, la situación sería diferente.

El efecto de los cambios en la especificidad sobre el valor predictivo también se observa en la tabla 5-10, en una forma similar a la utilizada en la tabla 5-8. Como se observa en este ejemplo, incluso con un 100% de sensibilidad, un cambio en la especificidad del 70% al 95% ejerce un efecto espectacular sobre el valor predictivo positivo.

FIABILIDAD (REPETIBILIDAD) DE LAS PRUEBAS

Consideremos otro aspecto de la valoración de las pruebas diagnósticas y de cribado: si una prueba es fiable o repetible. ¿Los resultados obtenidos podrían reproducirse si se repitiese la prueba? Claramente, con independencia de la sensibilidad y la especificidad de una prueba, si los resultados de la prueba no son reproducibles, el valor y la utilidad de la prueba son mínimos. El resto del presente capítulo analizará la fiabilidad o repetibilidad de las pruebas diagnósticas y de cribado. Los factores que contribuyen a la variación entre los resultados de la prueba se analizan en primer lugar: variación intraindividual (variaciones en un mismo individuo), variación intraobservador (variación en la lectura de los resultados de la prueba por el mismo observador) y variación interobservador (variación entre varias personas que analizan los resultados de la prueba).

Variación intraindividual

Los valores obtenidos al medir muchas características humanas a menudo varían a lo largo del tiempo, incluso durante un periodo corto de tiempo. En la tabla 5-11 se muestran los cambios en las mediciones de la presión arterial a lo largo de un período de 24 horas en tres personas. La variabilidad a lo largo del tiempo es considerable. Este hecho, así como las condiciones en las que se realizan ciertas pruebas (p. ej., tras una comida o tras realizar ejercicio, si se realiza en casa o en la consulta del médico), claramente pueden arrojar diferentes resultados en la misma persona. Por tanto, a la hora de valorar los resultados de cualquier prueba, es importante considerar las condiciones en las que se realizó la prueba, incluida la hora del día.

Variación intraobservador

En ocasiones se producen variaciones entre dos o más lecturas de los mismos resultados de la prueba valorada por un mismo observador. Por ejemplo, un radiólogo que interprete el mismo grupo de radiografías en dos ocasiones diferentes puede interpretar una o más de las radiografías de modo diferente la segunda vez. Las pruebas y las exploraciones se diferencian según el grado con el que entran en juego factores subjetivos en las conclusiones del observador; cuanto mayor sea el grado de subjetividad en las lecturas, mayor será la probabilidad de que se produzca una variación intraobservador en las mismas (fig. 5-14).

TABLA 5-10. Relación entre la especificidad y el valor predictivo positivo

EJEMPLO: PREVALENCIA = 10%, SENSIBILIDAD = 100%

Especificidad	Resultados de la prueba	Enfermos	No enfermos	Totales	Valor predictivo
70%	+	1.000	2.700	3.700	
	–	0	6.300	6.300	1.000/3.700 = 27%
	Totales	1.000	9.000	10.000	
95%	+	1.000	450	1.450	
	–	0	8.550	8.550	1.000/1.450
	Totales	1.000	9.000	10.000	

TABLA 5-11. Ejemplos que muestran la variación de las mediciones de presión arterial durante un periodo de 24 horas

Presión arterial (mmHg) Mujer de 27 años	Mujer de 62 años	Varón de 33 años
Basal 110/70	132/82	152/109
Más baja 86/47	102/61	123/78
Más alta 126/79	172/94	153/107
Ocasional 108/64	155/93	157/109

De Richardson DW, Honour AJ, Fenton GW, et al: Variation in arterial pressure throughout the day and night. Clin Sci 26:445,1964.

*"This is a second opinion. At first, I thought you had something else."

Figura 5-14. «Ésta ya es una segunda opinión. Al principio pensaba que tenía otra cosa.» Una visión de las segundas opiniones. (© The New Yorker Collection 1995. Leo Cullum from cartoonbank.com. Reservados todos los derechos.)

Variación interobservador

Otra consideración importante es la variación entre observadores. Dos examinadores a menudo no obtienen el mismo resultado. El grado de concordancia o discordancia entre observadores es un aspecto importante, ya sea si consideramos una exploración física, pruebas de laboratorio u otras técnicas de evaluación de características humanas. Necesitamos, por tanto, ser capaces de expresar el grado de concordancia en términos cuantitativos.

Porcentaje de concordancia

En la tabla 5-12 se muestra un esquema para examinar la variación entre observadores. Dos observadores fueron encargados de clasificar cada resultado de una prueba en una de las siguientes cuatro categorías: anormal, sospechoso, dudoso y normal. Este diagrama podría aplicarse, por ejemplo, a las lecturas realizadas por dos radiólogos. En este diagrama, las lecturas del observador 1 se presentan en formato de tabulación cruzada con las del observador 2. El número de lecturas en cada celdilla viene indicado por una letra del alfabeto. Así, A radiografías fueron consideradas anormales por ambos radiólogos. C radiografías fueron consideradas anormales por el radiólogo 2 y dudosas por el radiólogo 1. M radiografías fueron consideradas anormales por el radiólogo 1 y normales por el radiólogo 2.

Como se observa en la tabla 5-12, para calcular el porcentaje de concordancia global, sumamos los números de todas las celdillas en las que concordaron las interpretaciones de ambos radiólogos (A + F + K + P), dividimos dicha suma entre el número total de radiografías interpretadas y multiplicamos el resultado por 100 para obtener un porcentaje. En la figura 5-15A se muestra el uso de este abordaje para una prueba cuyos resultados posibles son «positivos» o «negativos».

TABLA 5-12. Variación por observador o instrumento: porcentaje de concordancia

	Lectura n.° 1			
Lectura n.° 2	***Anormal***	***Sospechosa***	***Dudosa***	***Normal***
Anormal	A	+ B	C	D
Sospechosa	E	F +	G	H
Dudosa	I	J	K +	L
Normal	M	N	O	P

$$\text{Porcentaje de concordancia} = \frac{A + F + K + P}{\text{Lecturas totales}} \times 100$$

Por lo general, la mayoría de las personas en las que se realizan pruebas obtienen resultados negativos. Esto se expone en la figura 5-15B, en la que el tamaño de cada celdilla guarda proporción con el número de personas en la misma. Probablemente exista una concordancia importante entre los dos observadores acerca de estos individuos, negativos o normales (celdilla d). Así, cuando se calcula el porcentaje de concordancia para todos los sujetos del estudio, su valor puede ser alto debido únicamente al elevado número de hallazgos claramente negativos (celdilla d) en los que concuerdan los observadores. El valor alto puede ocultar, por tanto, una gran falta de concordancia entre los observadores en la identificación de los sujetos que son considerados positivos por al menos un observador.

Un abordaje de este problema, expuesto en la figura 5-15C, es no tener en cuenta a los sujetos etiquetados como negativos por ambos observadores (celdilla d) y calcular el porcentaje de concordancia utilizando como denominador únicamente a los sujetos considerados anormales por al menos uno de los observadores (celdillas a, b y c) (fig. 5-15D).

Así, en las observaciones pareadas en las que al menos uno de los hallazgos de cada par fue positivo, es aplicable la siguiente ecuación:

$$\text{Porcentaje de concordancia} = \frac{a}{a+b+c} \times 100$$

Estadístico kappa

El porcentaje de concordancia entre dos observadores a menudo es valioso para valorar la calidad de sus observaciones. El grado de concordancia entre dos observadores, como, por ejemplo, dos médicos o dos enfermeras, a menudo es un índice importante de la calidad de la asistencia sanitaria que se está proporcionando. Sin embargo, el porcentaje de concordancia entre dos observadores no depende completamente de la calidad de su formación o su experiencia. En el grado de concordancia también influye de manera importante el hecho de que, aunque los dos observadores utilicen criterios completamente diferentes para identificar a sujetos como positivos o negativos, cabría esperar que los observadores coincidieran en las observaciones realizadas, al menos en algunos de los participantes, únicamente debido al azar. Lo que realmente queremos saber es cuánto mejor es su grado de concordancia que el que resultaría debido únicamente al azar. La respuesta a esta pregunta presumiblemente nos dirá, por ejemplo, hasta qué punto la formación y

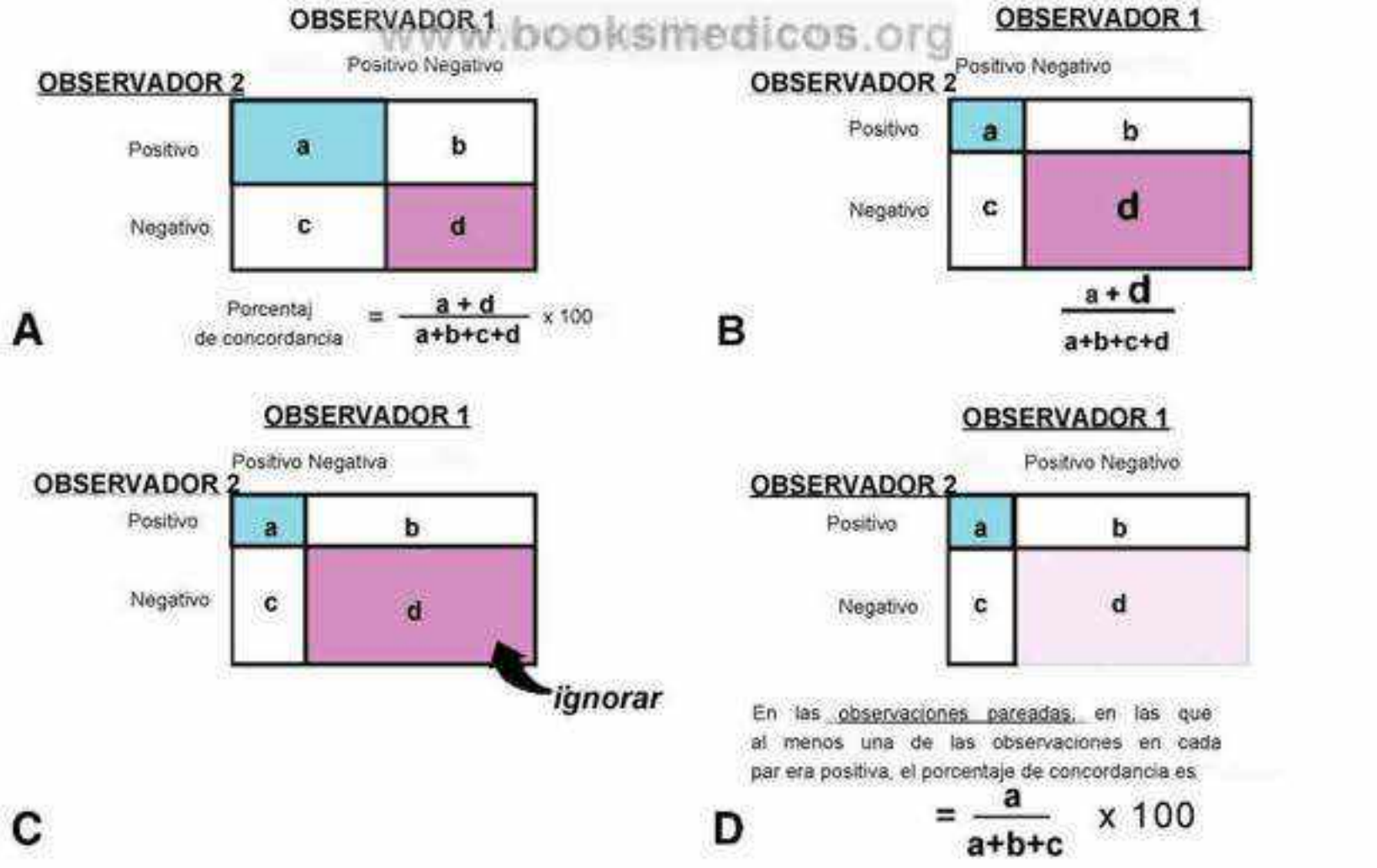


Figura 5-15. A-D. Cálculo del porcentaje de concordancia entre dos observadores. A, Porcentaje de concordancia cuando se examinan observaciones pareadas entre el observador 1 y el observador 2. B, Porcentaje de concordancia cuando se examinan observaciones pareadas entre el observador 1 y el observador 2, teniendo en cuenta que la celdilla d (concordancia en los negativos) es muy grande. C, Porcentaje de concordancia cuando se examinan observaciones pareadas entre el observador 1 y el observador 2, ignorando la celdilla d. D, Porcentaje de concordancia cuando se examinan observaciones pareadas entre el observador 1 y el observador 2, utilizando únicamente las celdillas a, b y c para el cálculo.

la práctica de los observadores mejoraron la calidad de sus observaciones de modo que el porcentaje de concordancia entre ellos aumentó más de lo que cabría esperar únicamente debido al azar.

Esto puede demostrarse intuitivamente en el siguiente ejemplo: usted es el jefe de un servicio de radiología que un día carece de suficiente personal y todavía tiene pendiente la interpretación de un gran número de radiografías de tórax. Para solucionar el problema, sale a la calle y le pide a algunos residentes del vecindario, sin formación en biología ni en medicina, que interpreten las radiografías que no están informadas y que valoren si son positivas o negativas. La primera persona ojea el montón de radiografías y las interpreta aleatoriamente como positiva, negativa, negativa, positiva, etc. La segunda persona hace lo mismo, siguiendo el mismo patrón, pero de manera completamente independiente respecto a la primera. Dado que ambas personas no poseen conocimientos, criterios o estándares para interpretar radiografías, ¿concordarán sus valoraciones sobre una radiografía específica? La respuesta es claramente afirmativa; en algunos casos coincidirán, únicamente debido al azar.

Sin embargo, si queremos saber cómo de bien han interpretado las radiografías dos observadores, podríamos preguntarnos: «¿Hasta qué punto coinciden sus interpretaciones *más allá de lo que cabría esperar únicamente por el azar*?». En otras palabras, ¿hasta qué punto la concordancia entre los dos observadores supera el grado de concordancia que resultaría únicamente por el azar? Un abordaje para responder a esta pregunta es calcular el estadístico kappa, propuesto por Cohén en 1960[2]. En esta sección analizaremos primero el fundamento del estadístico kappa y las preguntas para cuyas respuestas se diseñó el estadístico kappa. A continuación se expone un cálculo detallado del estadístico kappa para que sirva de ejemplo para los lectores intrépidos. Incluso aunque usted no siga los cálculos detallados que se presentan, es importante asegurarse de que ha comprendido el significado del estadístico kappa, pues se utiliza con frecuencia en la medicina clínica y en el ámbito de la salud pública.

Fundamento del estadístico kappa. Con el fin de comprender kappa, nos planteamos dos preguntas. La primera: «¿Cuánto mejor es la concordancia entre las interpretaciones de los observadores de lo que cabría esperar únicamente por el azar?». Esto puede calcularse como el porcentaje de concordancia observado menos el porcentaje de concordancia que cabría esperar únicamente por el azar. Éste es el numerador de kappa:

(Porcentaje de concordancia observado)
— (Porcentaje de concordancia esperado únicamente por el azar)

Nuestra segunda pregunta es: «¿Cuánto es lo máximo que los dos observadores podrían haber mejorado su concordancia sobre la concordancia que cabría esperar sólo por el azar?». Claramente, el máximo de concordancia sería el 100% (concordancia total: los dos observadores coinciden completamente). Por tanto, lo máximo que podemos esperar que sean capaces de mejorar (el denominador de kappa) sería:

100-(Porcentaje de concordancia
esperado únicamente por el azar)

Kappa expresa el grado en el que la concordancia observada supera a la que cabría esperar únicamente por el azar (es decir, el porcentaje de concordancia observado menos el porcentaje de concordancia esperado únicamente por el azar) [numerador] relativo al máximo que se podría esperar que los observadores mejorasen su concordancia (es decir, 100% menos el porcentaje de concordancia esperado únicamente por el azar) [denominador].

Por tanto, kappa cuantifica el grado en el que la concordancia observada lograda por los observadores supera a la que cabría esperar únicamente por el azar, y lo expresa como la proporción de la mejoría máxima que podría producirse más allá de la concordancia esperada únicamente por el azar. El estadístico kappa puede definirse por la siguiente ecuación:

$$\text{Kappa} = \frac{\left(\begin{matrix}\text{Porcentaje}\\ \text{de concordancia}\\ \text{observado}\end{matrix}\right) - \left(\begin{matrix}\text{Porcentaje}\\ \text{de concordancia}\\ \text{esperado únicamente}\\ \text{por el azar}\end{matrix}\right)}{100\% - \left(\begin{matrix}\text{Porcentaje de concordancia}\\ \text{esperado únicamente por el azar}\end{matrix}\right)}$$

Cálculo del estadístico kappa: un ejemplo. Para calcular el numerador de kappa, primero debemos calcular el grado de concordancia que podría esperarse únicamente por el azar. A modo de ejemplo, consideremos los datos comunicados sobre la clasificación histológica del cáncer de pulmón que se centró en la reproducibilidad de las decisiones de los anatomopatólogos a la hora de clasificar subtipos de carcinoma de pulmón de células no microcíticas[3]. En la figura 5-16A se muestran datos que comparan los hallazgos de los dos anatomopatólogos en la clasificación de 75 casos.

La primera pregunta es: «¿Cuál es la concordancia observada entre los dos anatomopatólogos?». En la figura 5-16B se muestran las lecturas del anatomopatólogo A en la parte inferior de la tabla y las del anatomopatólogo B en la parte derecha. El anatomopatólogo A identificó 45 (60%) del total de 75 muestras como grado II y 30 (40%) como grado III. El anatomopatólogo B

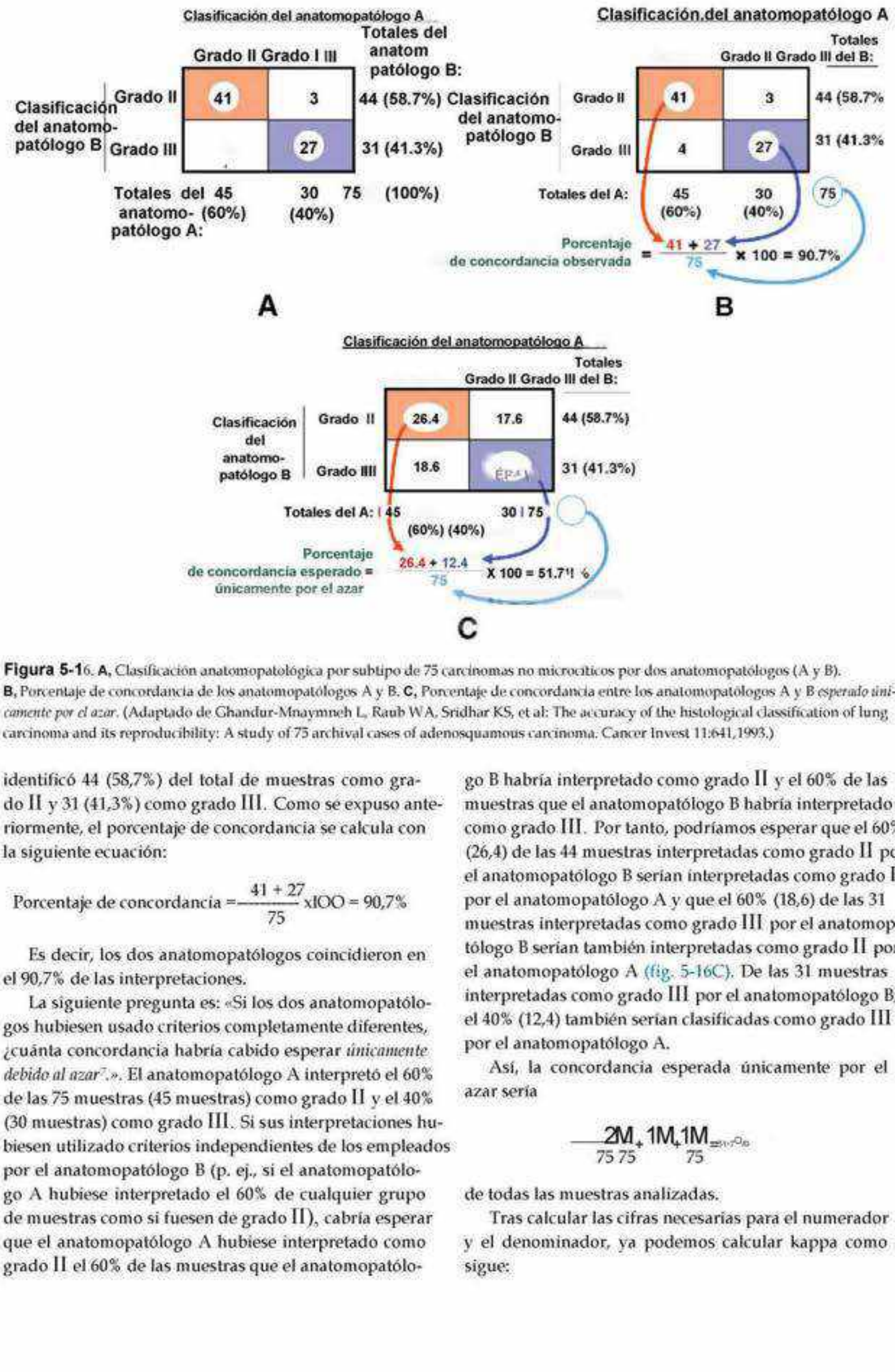


Figura 5-16. **A,** Clasificación anatomopatológica por subtipo de 75 carcinomas no microcíticos por dos anatomopatólogos (A y B). **B,** Porcentaje de concordancia de los anatomopatólogos A y B. **C,** Porcentaje de concordancia entre los anatomopatólogos A y B *esperado únicamente por el azar*. (Adaptado de Ghandur-Mnaymneh L, Raub WA, Sridhar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641,1993.)

identificó 44 (58,7%) del total de muestras como grado II y 31 (41,3%) como grado III. Como se expuso anteriormente, el porcentaje de concordancia se calcula con la siguiente ecuación:

$$\text{Porcentaje de concordancia} = \frac{41 + 27}{75} \times 100 = 90{,}7\%$$

Es decir, los dos anatomopatólogos coincidieron en el 90,7% de las interpretaciones.

La siguiente pregunta es: «Si los dos anatomopatólogos hubiesen usado criterios completamente diferentes, ¿cuánta concordancia habría cabido esperar *únicamente debido al azar*?». El anatomopatólogo A interpretó el 60% de las 75 muestras (45 muestras) como grado II y el 40% (30 muestras) como grado III. Si sus interpretaciones hubiesen utilizado criterios independientes de los empleados por el anatomopatólogo B (p. ej., si el anatomopatólogo A hubiese interpretado el 60% de cualquier grupo de muestras como si fuesen de grado II), cabría esperar que el anatomopatólogo A hubiese interpretado como grado II el 60% de las muestras que el anatomopatólogo B habría interpretado como grado II y el 60% de las muestras que el anatomopatólogo B habría interpretado como grado III. Por tanto, podríamos esperar que el 60% (26,4) de las 44 muestras interpretadas como grado II por el anatomopatólogo B serían interpretadas como grado II por el anatomopatólogo A y que el 60% (18,6) de las 31 muestras interpretadas como grado III por el anatomopatólogo B serían también interpretadas como grado II por el anatomopatólogo A (fig. 5-16C). De las 31 muestras interpretadas como grado III por el anatomopatólogo B, el 40% (12,4) también serían clasificadas como grado III por el anatomopatólogo A.

Así, la concordancia esperada únicamente por el azar sería

$$\frac{2M_+ \; 1M_+ 1M_-}{75\;75} = 51{,}7\%$$

de todas las muestras analizadas.

Tras calcular las cifras necesarias para el numerador y el denominador, ya podemos calcular kappa como sigue:

$$\text{Kappa} = \frac{\left(\begin{array}{c}\text{Porcentaje}\\ \text{de concordancia}\\ \text{observado}\end{array}\right) - \left(\begin{array}{c}\text{Porcentaje}\\ \text{de concordancia}\\ \text{esperado únicamente}\\ \text{por el azar}\end{array}\right)}{100\% - \left(\begin{array}{c}\text{Porcentaje de concordancia}\\ \text{esperado únicamente por el azar}\end{array}\right)}$$

$$= \frac{90{,}7\% - 51{,}7\%\ 39\%}{100\% - 51{,}7\%\ 48{,}3\%} = \frac{}{} = 0{,}81$$

Landis y Koch[4] sugieren que un kappa mayor de 0,75 representa una concordancia excelente más allá del azar, un kappa menor de 0,40 representa una concordancia baja y un kappa entre 0,40 y 0,75 representa una concordancia de intermedia a buena. Fleiss[7] ha estudiado la significación estadística de kappa. Existe gran controversia acerca del uso apropiado de kappa, un tema estudiado por MacLure y Willet[6].

RELACIÓN ENTRE VALIDEZ Y FIABILIDAD

Para finalizar este capítulo, comparemos la validez y la fiabilidad utilizando una representación gráfica.

La línea horizontal de la figura 5-17 es una escala de los valores para una variable determinada, como la concentración de glucosa en sangre, en la que se indica el valor real. Los resultados obtenidos con la prueba se muestran mediante la curva. La curva es estrecha, lo que indica que los resultados son bastante fiables (repetibles); desafortunadamente, sin embargo, se agrupan lejos del valor real, por lo que no son válidos. En la figura 5-18 se muestra una curva que es ancha y, por tanto, poco fiable. Sin embargo, los valores obtenidos se agrupan alrededor del valor real, por lo que son válidos. Claramente, lo que querríamos lograr son resultados válidos y fiables (fig. 5-19).

Es importante destacar que en la figura 5-18, en la que la distribución de los resultados es una curva ancha centrada sobre el valor real, describimos los resultados como válidos. Sin embargo, los resultados son válidos sólo para un grupo (es decir, tienden a agruparse alrededor del valor real). No hay que olvidar que lo que puede ser válido para un grupo o una población puede no serlo para un individuo en un contexto clínico. Cuando la fiabilidad o repetibilidad de una prueba es baja, la validez de la prueba para un individuo concreto también puede ser mala. Por tanto, es importante tener en cuenta la distinción entre validez grupal y validez individual a la hora de valorar la calidad de las pruebas diagnósticas y de cribado.

CONCLUSIÓN

Este capítulo ha estudiado la validez de las pruebas diagnósticas y de cribado analizando la sensibilidad y la especificidad, el valor predictivo y la fiabilidad o repetibilidad. Claramente, con independencia de la sensibilidad y la especificidad de una prueba, si sus resultados no pueden repetirse, la prueba es poco útil. Por tanto, todas estas características deben tenerse en cuenta cuando se valora una prueba, junto con la finalidad para la que se quiere utilizar dicha prueba.

Figura 5-17. Gráfico de los resultados de una prueba hipotética que son fiables, pero no válidos.

Figura 5-18. Gráfico de los resultados de una prueba hipotética que son válidos, pero no fiables.

Figura 5-19. Gráfico de los resultados de una prueba hipotética que son válidos y fiables.

BIBLIOGRAFÍA

1. Sheffield LJ, Sackett DL, Goldsmith CH, et al: A clinical approach to the use of predictive values in the prenatal diagnosis of neural tube defects. Am J Obstet Gynecol 145:319, 1983.
2. Cohen J: A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37,1960.
3. Ghandur-Mnaymneh L, Raub WA, Sridhar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641,1993.
4. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159,1977.
5. Fleiss JL: Statistical Methods for Rates and Proportions, 2nd ed. New York, John Wiley & Sons, 1981.
6. MacLure M, Willett WC: Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126:161,1987.

Véanse las preguntas de repaso en las páginas 114-115.

APÉNDICES DEL CAPÍTULO 5

El texto del capítulo 5 se centra en la lógica que respalda el cálculo de la sensibilidad, la especificidad y el valor predictivo. En el Apéndice 1 se resumen medidas de validez para las pruebas de cribado para detectar la ausencia o la presencia de una enfermedad determinada; primero se dedican una páginas en el texto a las medidas y a la interpretación de cada medida. Los que prefieran ver las fórmulas de cada medida pueden consultar la columna derecha de esta tabla; no obstante, no son esenciales para comprender la lógica que respalda el cálculo de cada medida.

En el Apéndice 2 se resumen los tres pasos necesarios para calcular el estadístico kappa.

Apéndice 1 del capítulo 5. Medidas de la validez de una prueba y su interpretación

Medidas de validez de una prueba	Números de página	Interpretación	Fórmula
Sensibilidad	90	La proporción de los que *tienen* la enfermedad en los que la prueba es *positiva*	VP / VP + FN
Especificidad	90	La proporción de los que *no tienen* la enfermedad en los que la prueba es *negativa*	VN/VN + FP
Valor predictivo positivo	100-101	La proporción de los que tienen resultados *positivos* que tienen la enfermedad	VP/VP + FP
Valor predictivo negativo	100-101	La proporción de los que tienen resultados *negativos* que NO tienen la enfermedad	VN / VN + FN
Sensibilidad neta	95-96	La proporción de los que *tienen* la enfermedad y obtienen resultados *positivos* en AMBAS pruebas (1 y 2)	(Sensibilidad de la prueba 1) x (Sensibilidad de la prueba 2)
Especificidad neta	95-96	La proporción de los que *no tienen* la enfermedad y obtienen resultados *negativos* en la prueba 1 o en la prueba 2	(Especificidad de la prueba 1 + Especificidad de la prueba 2) - (Especificidad de la prueba 1 x Especificidad de la prueba 2)
Sensibilidad neta	96-97	La proporción de los que *tienen* la enfermedad y obtienen resultados *positivos* en la prueba 1 o en la prueba 2	(Sensibilidad de la prueba 1 + Sensibilidad de la prueba 2) - (Sensibilidad de la prueba 1 x Sensibilidad de la prueba 2)
Especificidad neta	97-98	La proporción de los que *no tienen* la enfermedad y obtienen resultados *negativos* en la prueba 1 y en la prueba 2	(Especificidad de la prueba 1) x (Especificidad de la prueba 2)

Abreviaturas: FN, falsos negativos; FP, falsos positivos; VN, verdaderos negativos; VP, verdaderos positivos.

Apéndice 2 del capítulo 5. Los tres pasos necesarios para calcular el estadístico kappa (κ)

Componentes	Pasos
NUMERADOR ¿En qué cuantía la concordancia observada es mejor que la que cabría esperar únicamente por el azar?	PASO1: (Porcentaje de concordancia observado) − (Porcentaje de concordancia esperado únicamente por el azar)
DENOMINADOR ¿Cuál es la máxima mejoría que podrían haber logrado los observadores sobre la concordancia esperada únicamente por el azar?	PASO 2: 100% − (Porcentaje de concordancia esperado únicamente por el azar)
$\frac{\text{NUMERADOR}}{\text{DENOMINADOR}}$ = ESTADÍSTICO KAPPA (κ) Del máximo aumento en la concordancia esperado más allá del debido únicamente al azar que podría haberse producido, ¿qué proporción se ha producido realmente?	PASO3: $\kappa = \frac{(\text{Porcentaje de concordancia observado}) - (\text{Porcentaje de concordancia esperado únicamente por el azar})}{100\% - (\text{Porcentaje de concordancia esperado únicamente por el azar})}$

En las páginas 107-110 se expone una explicación detallada de kappa y un ejemplo de su cálculo.

PREGUNTAS DE REPASO DEL CAPÍTULO 5

Las preguntas 1,2 y 3 se basan en la siguiente información:

Se realizó una exploración física como cribado del cáncer de mama en 2.500 mujeres con adenocarcinoma de mama demostrado mediante biopsia y en 5.000 mujeres controles de edad y raza similares. Los resultados de la exploración fueron positivos (es decir, se palpó una masa) en 1.800 casos y en 800 de las mujeres controles, todas las cuales carecían de signos de cáncer en la biopsia.

1. La sensibilidad de la exploración física fue:_____
2. La especificidad de la exploración física fue:_____
3. El valor predictivo positivo de la exploración física fue:_____

La pregunta 4 se basa en la siguiente información:

Una prueba de cribado se utiliza del mismo modo en dos poblaciones similares, pero la proporción de resultados falsos positivos entre los que obtienen resultados positivos en la población A es menor que entre los que obtienen resultados positivos en la población B.

4. ¿Cuál es la explicación probable de este hallazgo?
   a. Es imposible determinar la causa de esta diferencia.
   b. La especificidad de la prueba es menor en la población A.
   c. La prevalencia de la enfermedad es menor en la población A.
   d. La prevalencia de la enfermedad es mayor en la población A.
   e. La especificidad de la prueba es mayor en la población A.

La pregunta 5 se basa en la siguiente información:

Se realizó una exploración física y una audiometría a 500 personas en las que se sospechaban problemas auditivos; fueron encontrados en 300 de las mismas. Los resultados de la exploración fueron los siguientes:

Exploración física

	PROBLEMAS AUDITIVOS	
Resultado	**Presentes**	**Ausentes**
Positivo	240	40
Negativo	60	160

Audiometría

	PROBLEMAS AUDITIVOS	
Resultado	**Presentes**	**Ausentes**
Positivo	270	60
Negativo	301	40

5. En comparación con la exploración física, la audiometría es:
   a. Igual de sensible y específica.
   b. Menos sensible y menos específica.
   c. Menos sensible y más específica.
   d. Más sensible y menos específica.
   e. Más sensible y más específica.

Las pregunta 6 se basa en la siguiente información:

Dos pediatras quieren estudiar una nueva prueba de laboratorio que identifica las infecciones estreptocócicas. El Dr. Kidd utiliza la prueba de cultivo estándar, que posee una sensibilidad del 90% y una especificidad del 96%. El Dr. Childs utiliza la prueba nueva, que posee un 96% de sensibilidad y un 96% de especificidad.

6. Si realizamos el cultivo en 200 pacientes con ambas pruebas, ¿cuál de las siguientes afirmaciones es correcta?
   a. El Dr. Kidd identificará correctamente a más personas con infección estreptocócica que el Dr. Childs.
   b. El Dr. Kidd identificará correctamente a menos personas con infección estreptocócica que el Dr. Childs.
   c. El Dr. Kidd identificará correctamente a más personas sin infección estreptocócica que el Dr. Childs.
   d. Se necesita conocer la prevalencia de la infección estreptocócica para determinar qué pediatra identificará correctamente a un mayor número de personas con la enfermedad.

Las preguntas 7 y 8 se basan en la siguiente información:

En Nottingham, Inglaterra, se está llevando a cabo un estudio de cribado de cáncer de colon. Se estudiarán individuos de 50-75 años con la prueba Hemoccult. En esta prueba se estudia la presencia de sangre en una muestra de heces.

7. La prueba Hemoccult posee una sensibilidad del 70% y una especificidad del 75%. Si la prevalencia del cáncer de colon en Nottingham es de 12/1.000, ¿cuál es el valor predictivo positivo de la prueba?

8. Si el resultado de la prueba Hemoccult es negativo, no se realizan nuevas pruebas. Si el resultado de la prueba Hemoccult es positivo, se volverá a analizar una segunda muestra de heces del individuo con la prueba Hemoccult II. Si el resultado en esta segunda muestra también es positivo, el individuo será remitido para realizar un estudio más extenso. ¿Cuál es el efecto sobre la sensibilidad neta y la especificidad neta de este método de cribado?
   a. Tanto la sensibilidad neta como la especificidad neta aumentan.
   b. La sensibilidad neta se reduce y la especificidad neta aumenta.
   c. La sensibilidad neta no cambia y la especificidad neta aumenta.
   d. La sensibilidad neta aumenta y la especificidad neta disminuye.
   e. El efecto sobre la sensibilidad neta y la especificidad neta no puede determinarse a partir de estos datos.

Las preguntas 9-12 se basan en la siguiente información:

Se pidió a dos médicos que clasificasen 100 radiografías de tórax como anormales o normales independientemente. La comparación de su clasificación se expone en la siguiente tabla:

Comparación entre la clasificación de las radiografías de tórax por el médico 1 y el médico 2

	Médico 2		
Médico 1	*Anormal*	*Normal*	**Total**
Anormal	40	20	60
Normal	10	30	40
Total	50	50	100

9. El porcentaje de concordancia simple entre los dos médicos respecto al total es:_____

10. El porcentaje de concordancia entre los dos médicos, excluyendo las radiografías clasificadas como normales por ambos médicos es:_____

11. El valor de kappa es:_____

12. Este valor de kappa, ¿qué grado de concordancia representa?
   a. Excelente.
   b. Intermedio-bueno.
   c. Bajo.

Capítulo 6

La historia natural de la enfermedad: formas de expresar el pronóstico

Objetivos de aprendizaje

- Comparar cinco formas diferentes de describir la historia natural de la enfermedad: tasa de letalidad, supervivencia a 5 años, supervivencia observada, mediana de supervivencia y supervivencia relativa.
- Describir dos abordajes para calcular la supervivencia observada a lo largo del tiempo: el abordaje de la tabla de vida y el método Kaplan-Meier.
- Ilustrar el uso de tablas de vida para estudiar cambios de la supervivencia.
- Describir cómo las mejoras en los métodos diagnósticos disponibles pueden afectar a la estimación del pronóstico (migración de estadios).

Hasta ahora hemos aprendido cómo las pruebas diagnósticas y de cribado permiten la diferenciación entre individuos sanos y enfermos. Una vez que se identifica que una persona tiene una enfermedad, la pregunta que surge es: «¿Cómo podemos caracterizar la historia natural de la enfermedad en términos cuantitativos?». Dicha cuantificación es importante por varios motivos. En primer lugar, es necesario describir la gravedad de una enfermedad para establecer prioridades en los servicios clínicos y en los programas de salud pública. En segundo lugar, los pacientes a menudo plantean preguntas acerca del pronóstico (fig. 6-1). En tercer lugar, dicha cuantificación es importante para establecer una línea basal de la historia natural, de modo que, a medida que se disponga de nuevos tratamientos, los efectos de estos tratamientos puedan compararse con el resultado esperado sin los mismos. Además, si se dispone de diferentes tipos de tratamientos para una cierta enfermedad, como tratamientos médicos o quirúrgicos, o dos tipos diferentes de intervenciones quirúrgicas, queremos ser capaces de comparar la eficacia de las diferentes modalidades terapéuticas. Por tanto, para poder realizar dicha comparación, necesitamos medios cuantitativos para expresar el pronóstico en grupos que reciben diferentes tratamientos.

Este capítulo expone algunas de las formas de describir el pronóstico de un grupo de pacientes en términos cuantitativos. Por tanto, este capítulo estudia la historia natural de la enfermedad (pronóstico). En capítulos posteriores se analiza cómo se puede intervenir en la historia natural de la enfermedad para mejorar el pronóstico: en los capítulos 7 y 8 se estudia cómo se utilizan los ensayos clínicos aleatorizados para seleccionar el fármaco u otro tratamiento más apropiado y en el capítulo 18 se estudia cómo puede detectarse una enfermedad en un momento más temprano de lo habitual en su historia natural para maximizar la eficacia del tratamiento.

Para estudiar el pronóstico, comencemos con una representación esquemática de la historia natural de la enfermedad en un paciente, según se muestra en la figura 6-2.

El punto A marca el comienzo biológico de la enfermedad. A menudo, este punto no puede identificarse porque se produce de manera subclínica, quizá como un cambio subcelular, como una alteración del ADN. En algún punto en la progresión del proceso de la enfermedad (punto P), podrían obtenerse pruebas patológicas de la enfermedad si éstas se buscaran. Posteriormente, el paciente presenta los signos y los síntomas de la enfermedad (punto S) y, algún tiempo después, el paciente puede buscar asistencia médica (punto M). A continuación, el paciente puede ser diagnosticado (punto D), tras lo que puede pautarse un tratamiento (punto T). La evolución posterior de la enfermedad podría resultar en la cura, el control de la enfermedad (con o sin discapacidad) o incluso la muerte.

¿En qué momento comenzamos a cuantificar el tiempo de supervivencia? De modo ideal, preferiríamos hacerlo desde el comienzo de la enfermedad. Por lo general, esto no es posible porque el momento del comienzo biológico en un individuo es desconocido. Si quisiéramos contar desde el momento en el que comienzan los síntomas, introduciríamos una gran variabilidad subjetiva al medir la duración de la supervivencia. Por lo general, para estandarizar los cálculos, la duración de la supervivencia se mide desde el momento del diagnóstico. Sin embargo, incluso con el uso de este

"How much time do I have, Doc?"

Figura 6-1. «¿Cuánto tiempo me queda, doctor?». Preocupación acerca del pronóstico. (© The New Yorker Collection 2001. Charles Barsotti from cartoonbank.com. Reservados todos los derechos.)

punto de comienzo, se produce variabilidad porque los pacientes difieren en el momento en el que buscan asistencia médica. Además, algunas enfermedades, como ciertos tipos de artritis, son indolentes y se desarrollan lentamente, de modo que puede que los pacientes no sean capaces de detallar el comienzo de los síntomas o el punto en el tiempo en el que solicitaron asistencia médica. Además, cuando la supervivencia se cuenta desde el momento del diagnóstico, todo paciente que haya fallecido antes de ser diagnosticado es excluido del recuento. ¿Cómo afectaría este problema a nuestras estimaciones sobre el pronóstico?

Una pregunta relacionada importante es: «¿Cómo se realiza el diagnóstico?». ¿Existe una prueba patognomónica clara para la enfermedad en cuestión? Con frecuencia no disponemos de dicha prueba. En ocasiones, una enfermedad puede ser diagnosticada tras el aislamiento de un microorganismo infeccioso, pero, como las personas pueden ser portadoras de microorganismos sin estar realmente infectadas, no siempre sabemos si el microorganismo aislado es la causa de la enfermedad. En algunas enfermedades preferiríamos alcanzar el diagnóstico mediante confirmación tisular, pero con frecuencia existe variabilidad en la interpretación de las muestras de tejido por diferentes anatomopatólogos. Un problema adicional es que, en ciertos problemas de salud, como las cefaleas, las lumbalgias y la dismenorrea, puede no existir un diagnóstico tisular específico. Por tanto, cuando decimos que la supervivencia se mide desde el momento del diagnóstico, la franja temporal no siempre está clara. Estos aspectos deben tenerse en cuenta cuando avancemos en el análisis de los diferentes abordajes para estimar el pronóstico.

El pronóstico puede expresarse en función de las muertes debidas a la enfermedad o en función de los que sobreviven a la enfermedad. Aunque en la siguiente exposición empleamos ambos abordajes, el punto final empleado para los propósitos de nuestro análisis es la muerte. Como la muerte es inevitable, no nos referimos a morir frente a no morir, sino a prolongar el intervalo hasta que se produce la muerte. Se pueden utilizar otros puntos finales, como el intervalo desde el diagnóstico hasta la recurrencia de la enfermedad o desde el diagnóstico hasta el momento en el que aparece afectación funcional, discapacidad o cambios en la calidad de vida del paciente, todos los cuales pueden verse afectados por la invasividad de los tratamientos disponibles o por el grado de mejoría alcanzable en algunos de los síntomas, incluso aunque no pueda aumentarse la esperanza de vida del paciente. Todas estas son medidas importantes, pero no se tratan en este capítulo.

TASA DE LETALIDAD

La primera forma de expresar el pronóstico es la *tasa de letalidad* (se expuso en el cap. 4). La tasa de letalidad se define como el número de personas que mueren por una enfermedad dividido entre el número de personas

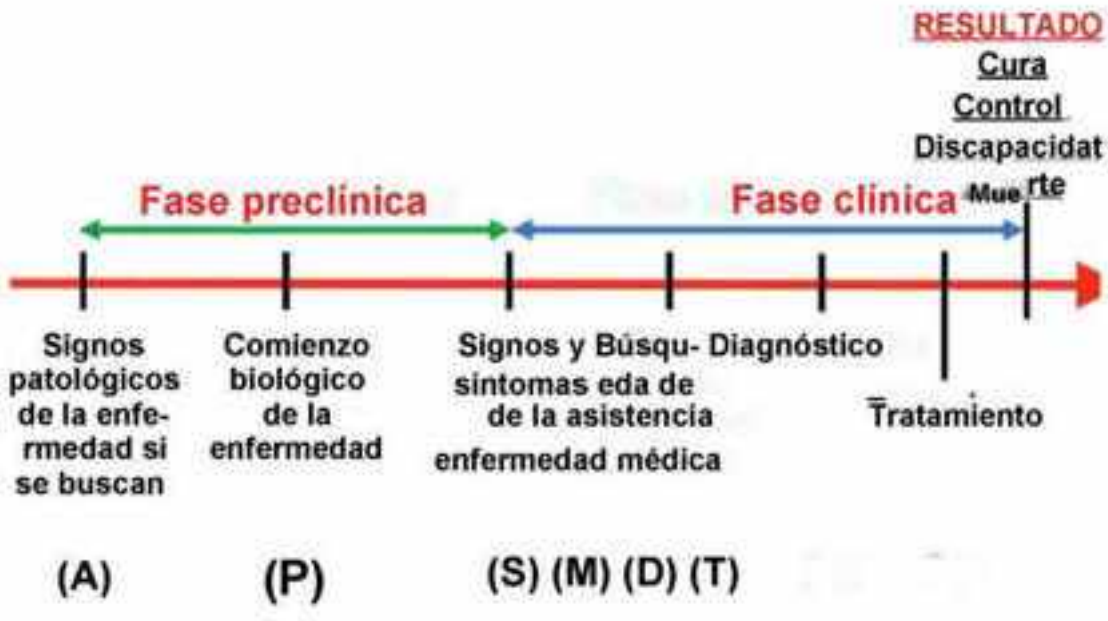


Figura 6-2. La historia natural de la enfermedad en un paciente.

que tienen la enfermedad. Cuando una persona tiene una enfermedad, ¿cuál es la probabilidad de que muera de dicha enfermedad? Obsérvese que el denominador de la tasa de letalidad es el número de personas que tienen la enfermedad. En esto se diferencia de la *tasa de mortalidad*, en la que el denominador incluye a cualquier persona con riesgo de morir de la enfermedad: tanto personas que tienen la enfermedad como personas que (todavía) no tienen la enfermedad, pero que podrían presentarla.

La tasa de letalidad no incluye ninguna mención explícita del tiempo. Sin embargo, el tiempo es expresado implícitamente, porque la tasa de letalidad suele usarse en enfermedades agudas en las que la muerte, si se produce, ocurre relativamente pronto tras el diagnóstico. Por tanto, si se conoce la historia natural habitual de la enfermedad, el término *tasa de letalidad* se refiere al período tras el diagnóstico durante el que cabría esperar que el paciente falleciera.

La tasa de letalidad es apropiada para enfermedades agudas de corta duración. Para las enfermedades crónicas en las que la muerte puede producirse muchos años tras el diagnóstico y la posibilidad de morir de otras causas se vuelve más probable, la tasa de letalidad es una medida menos útil. Por tanto, usamos diferentes abordajes para expresar el pronóstico en dichas enfermedades.

PERSONAS-AÑOS

Una forma útil de expresar la mortalidad es mediante el número de muertes dividido entre las personas-años a lo largo de los que se observa un grupo. Como los individuos a menudo son observados durante diferentes períodos de tiempo, la unidad usada para contar el tiempo de observación es personas-años. (Las personas-años se abordaron en el cap. 3, págs. 42-45.) El número de personas-años para dos personas, cada una de las cuales es observada durante 5 años, es igual al de 10 personas, cada una de las cuales es observada durante 1 año, es decir, 10 personas-años. Los números de personas-años pueden sumarse y el número de acontecimientos, como las muertes, pueden calcularse para el número de personas-años observado.

Un problema de utilizar las personas-años es que se asume que cada persona-año es equivalente al resto de personas-años (es decir, que el riesgo es el mismo en cualquier persona-año observado). Sin embargo, puede que esto no sea así. Consideremos la situación de la figura 6-3, que muestra dos ejemplos de 10 personas-años: dos personas observadas durante 5 años y cinco personas observadas durante 2 años. ¿Son equivalentes?

Supongamos la situación que se muestra en la figura 6-4: observamos que el período de mayor riesgo

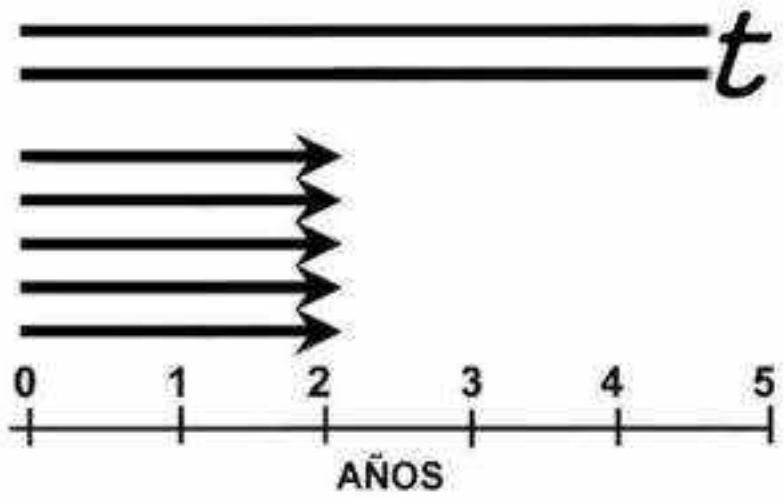


Figura 6-3. Dos ejemplos de 10 personas-años: dos personas, cada una de ellas observada durante 5 años, y cinco personas, cada una de ellas observada durante 2 años.

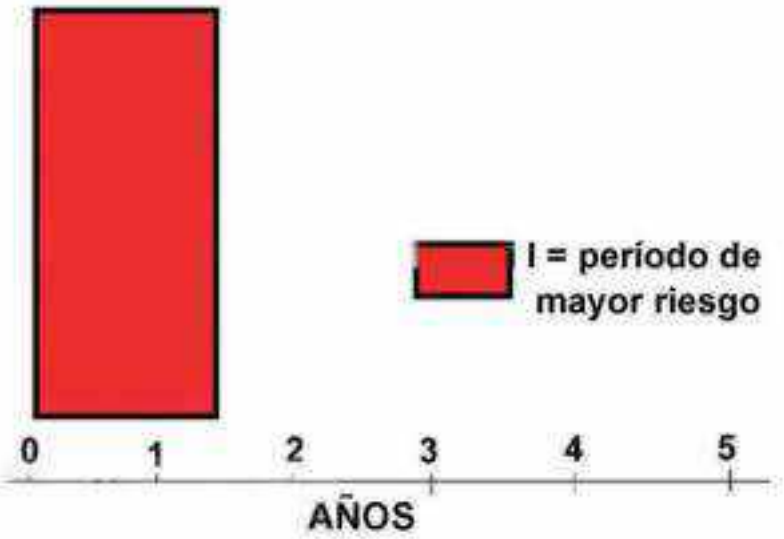


Figura 6-4. El momento de mayor riesgo es desde poco después del diagnóstico hasta aproximadamente 20 meses después del mismo.

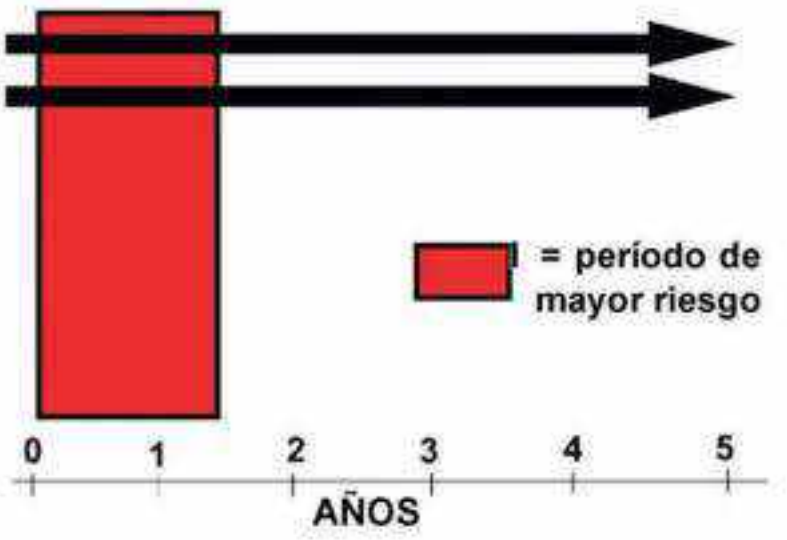


Fig U ra **6-5.** Dos personas, cada una de ellas observada durante 5 años, y la relación con el período de mayor riesgo.

de morir es desde poco tiempo después del diagnóstico hasta aproximadamente 20 meses después del diagnóstico. Claramente, la mayor parte de las personas-años del primer ejemplo, es decir, dos personas observadas durante 5 años, se encontrarán fuera del período de mayor riesgo (fig. 6-5). Por el contrario, la mayor parte de los intervalos de 2 años de las 5 personas mostradas en

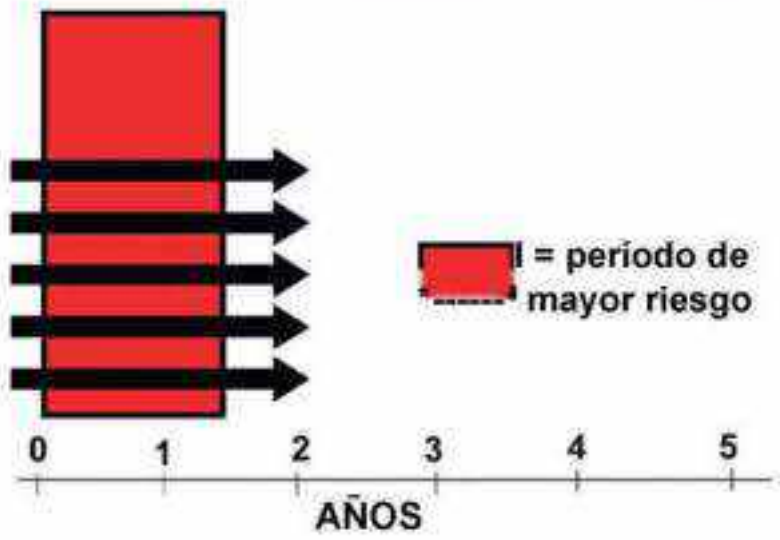


Figura 6-6. Cinco personas, cada una de ellas observada durante 2 años, y la relación con el período de mayor riesgo.

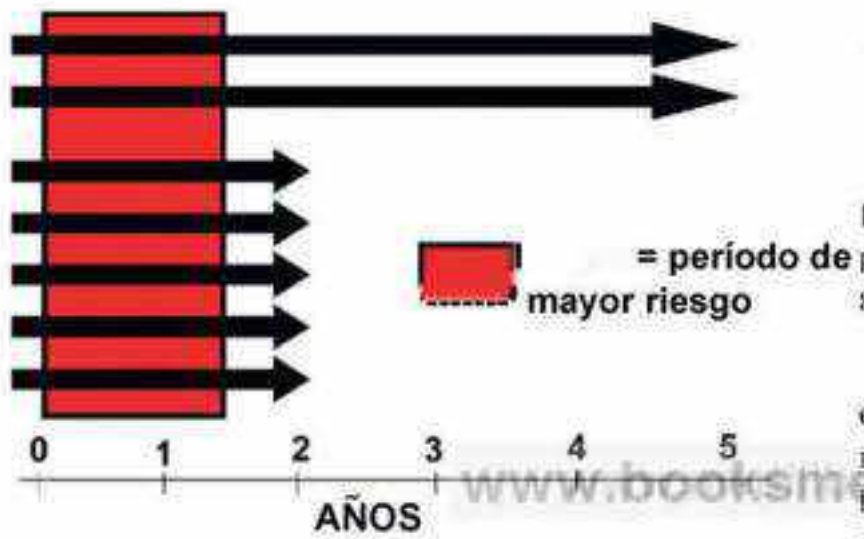


Figura 6-7. Dos ejemplos de 10 personas-años en los que el período de mayor riesgo es desde poco después del diagnóstico hasta aproximadamente 20 meses tras el mismo.

el segundo ejemplo tendrán lugar durante el período de mayor riesgo (fig. 6-6). Por tanto, cuando comparamos los dos ejemplos (fig. 6-7), cabría esperar más muertes en el ejemplo de las cinco personas observadas durante 2 años que en el ejemplo de las dos personas observadas durante 5 años. A pesar de este aspecto, las personas-años resultan útiles como denominadores de tasas de acontecimientos en muchas situaciones, como en ensayos clínicos aleatorizados (v. caps. 7 y 8) y en estudios de cohortes (v. cap. 9).

SUPERVIVENCIA A CINCO AÑOS

La *supervivencia a cinco años* es otra medida empleada para expresar el pronóstico. Este término se utiliza con frecuencia en la medicina clínica, especialmente para evaluar tratamientos del cáncer.

La supervivencia a 5 años es el porcentaje de pacientes que están vivos 5 años después del comienzo del tratamiento o 5 años después del diagnóstico. (Aunque a menudo se habla de la supervivencia a 5 años como una tasa, realmente es una proporción.) A pesar del uso

Figura 6-8. El problema de la supervivencia a 5 años en una población cribada: I. Situación sin cribado.

Figura 6-9. El problema de la supervivencia a 5 años en una población cribada: II. Detección más temprana de la enfermedad gracias al cribado.

extendido del intervalo de 5 años, se debe precisar que no hay nada mágico acerca del mismo. Ciertamente, en la historia natural de una enfermedad no se produce ningún cambio biológico significativo de forma abrupta a los 5 años que justifique su uso como punto final. Sin embargo, la mayoría de las muertes por cáncer se producen durante este período tras el diagnóstico, por lo que la supervivencia a 5 años se ha utilizado como índice de éxito del tratamiento del cáncer.

Un problema con el uso de la supervivencia a 5 años se ha vuelto más importante en los últimos años con el empleo de programas de cribado. Estudiemos un ejemplo hipotético: en la figura 6-8 se muestra la cronología de una mujer con cáncer de mama de comienzo biológico en el año 2000. Como la enfermedad era subclínica en esa fecha, se encontraba asintomática. En 2008 notó un bulto en la mama que la llevó a consultar a su médico, que realizó el diagnóstico. La paciente fue sometida posteriormente a una mastectomía. En 2010 falleció por un cáncer metastásico. Si utilizamos como medida la supervivencia a 5 años, que se emplea con frecuencia en oncología como medida del éxito del tratamiento, esta paciente no ha sido un «éxito» porque sólo sobrevivió 2 años tras el diagnóstico.

Imaginemos ahora que esta mujer vivía en una comunidad en la que existía una campaña agresiva de cribado del cáncer de mama (cronología inferior en la fig. 6-9). Al igual que antes, el comienzo biológico de la enfermedad tuvo lugar en el año 2000, pero en 2005 se identificó una masa muy pequeña en su mama

por medio del programa de cribado. Fue intervenida quirúrgicamente en 2005, pero falleció en 2010. Como sobrevivió 5 años tras el diagnóstico y el tratamiento, sería identificada como un éxito terapéutico en términos de la supervivencia a 5 años. Sin embargo, esta supervivencia aparentemente más prolongada es un artefacto. La muerte siguió ocurriendo en 2010; la vida de la paciente no fue más prolongada tras la detección y el tratamiento más tempranos. Lo que ha ocurrido es que el intervalo entre el diagnóstico (y el tratamiento) y su muerte aumentó por el diagnóstico más precoz, pero no se retrasó la fecha de su muerte. (El intervalo entre el diagnóstico más temprano en 2005, hecho posible por el cribado, y el momento habitual de diagnóstico más tardío en 2008 se denomina *adelanto en el momento del diagnóstico*. Este concepto se aborda en detalle en el capítulo 18 en el contexto de la evaluación de los programas de cribado.) Es engañoso concluir que, teniendo en cuenta la supervivencia a 5 años de la paciente, el resultado del segundo escenario es mejor que el del primero, porque no se ha producido un cambio en la historia natural de la enfermedad, como refleja el año en el que se produjo la muerte. De hecho, el único cambio que ha tenido lugar es que, cuando se realizó el diagnóstico 3 años antes (2005 frente a 2008), la paciente recibió cuidados médicos para su cáncer de mama, con todas las dificultades acompañantes, durante 3 años adicionales. Así pues, cuando se realizan pruebas de cribado, puede observarse una supervivencia a 5 años más alta, no porque los pacientes vivan más tiempo sino únicamente porque el diagnóstico se ha realizado más precozmente. Este tipo de sesgo potencial (conocido como *sesgo por adelanto en el momento del diagnóstico)* debe tenerse en cuenta cuando se valora cualquier programa de cribado antes de poder concluir que el cribado es beneficioso para aumentar la supervivencia.

Otro problema con la supervivencia a 5 años es que, si queremos fijarnos en la experiencia de supervivencia de un grupo de pacientes que fueron diagnosticados hace menos de 5 años, claramente no podemos utilizar este criterio, porque en estos pacientes se necesitan 5 años de observación para calcular la supervivencia a 5 años. Por tanto, si queremos valorar un tratamiento que fue iniciado hace menos de 5 años, la supervivencia a 5 años no es una medida apropiada.

Figura 6-10. Curvas de supervivencia a 5 años en dos poblaciones hipotéticas.

Un último aspecto relacionado con la supervivencia a 5 años se muestra en la figura 6-10. En esta figura observamos curvas de supervivencia de dos poblaciones, A y B. La supervivencia a 5 años es de aproximadamente el 10%. Sin embargo, las curvas que dan lugar a la misma supervivencia a 5 años son bastante diferentes. Aunque la supervivencia a 5 años sea la misma en ambos grupos, la mayoría de las muertes en el grupo A no se produjeron hasta el quinto año, mientras que la mayoría de las muertes en el grupo B se produjeron en el primer año. Así, a pesar de supervivencias a 5 años idénticas, la supervivencia durante los 5 años es claramente mejor para los pacientes del grupo A.

SUPERVIVENCIA OBSERVADA

Fundamento de la tabla de vida

Otro abordaje consiste en utilizar la supervivencia real observada a lo largo del tiempo. Para este fin, empleamos una *tabla de vida*. Examinemos el marco conceptual que subyace en el cálculo de las tasas de supervivencia usando una tabla de vida.

En la tabla 6-1 se muestra un estudio hipotético de los resultados del tratamiento en pacientes tratados

TABLA 6-1. Estudio hipotético de los resultados del tratamiento de pacientes tratados de 2000 a 2004 y seguidos hasta 2005 (ninguna pérdida de seguimiento)

Año de tratamiento	Número de pacientes tratados	NÚMERO DE VIVOS EN EL ANIVERSARIO DEL TRATAMIENTO				
		2001	2002	2003	2004	2005
2000	84	44	21	13	10	8
2001	62		31	14	10	6
2002	93			50	20	13
2003	60				29	16
2004	76					43

TABLA 6-2. Reestructuración de los datos de la tabla 6-1 mostrando la supervivencia tabulada por años desde el inicio del tratamiento (ninguna pérdida de seguimiento)

		NÚMERO DE VIVOS AL FINAL DEL AÑO				
Año de tratamiento	**N.° de pacientes tratados**	**1.er año**	**2.° año**	**3.er año**	**4.° año**	**5.° año**
2000	84	44	21	13	10	8
2001	62	31	14	10	6	
2002	93	50	20	13		
2003	60	29	16			
2004	76	43				

de 2000 a 2004 y seguidos hasta 2005. (Simplemente mirando esta tabla, usted se dará cuenta de que el ejemplo es hipotético, porque el título indica que ningún paciente fue perdido durante el seguimiento.)

Para cada año de calendario de tratamiento, la tabla muestra el número de pacientes que reciben el tratamiento y el número de pacientes vivos en cada año de calendario tras el inicio de dicho tratamiento. Por ejemplo, de los 84 pacientes que iniciaron el tratamiento en el año 2000, 44 estaban vivos en 2001, un año después de comenzar el tratamiento; 21 estaban vivos en 2002, y así sucesivamente.

Los resultados de la tabla 6-1 son de todos los datos disponibles para valorar el tratamiento. Si queremos describir el pronóstico en estos pacientes tratados utilizando todos los datos de la tabla, evidentemente no podemos emplear la supervivencia a 5 años, porque todo el grupo de 375 pacientes no ha sido observado durante 5 años. Podríamos calcular la supervivencia a 5 años a partir únicamente de los 84 pacientes que iniciaron el tratamiento en 2000 y fueron observados hasta 2005, porque fueron los únicos observados durante 5 años. Sin embargo, esto nos obligaría a descartar el resto de los datos, lo que sería inapropiado, dado el esfuerzo y los gastos involucrados en la obtención de los datos, y también debido a la luz adicional que la experiencia de supervivencia de esos pacientes arrojaría sobre la eficacia del tratamiento. La pregunta es: ¿cómo podemos utilizar *toda* la información de la tabla 6-1 para describir la experiencia de supervivencia de los pacientes de este estudio?

Para utilizar todos los datos, reestructuramos los datos de la tabla 6-1 como se muestra en la tabla 6-2. En esta tabla, los datos se muestran como el número de pacientes que comenzaron el tratamiento en cada año del calendario y el número de aquellos vivos en cada aniversario del inicio del tratamiento. Los pacientes que comenzaron el tratamiento en 2004 fueron observados únicamente durante un solo año, porque el estudio finalizó en 2005.

Con los datos en este formato, ¿cómo utilizamos la tabla? En primer lugar, preguntémonos: «¿Cuál es la probabilidad de sobrevivir 1 año tras el inicio del tratamiento?». Para responder a esta pregunta, dividimos el número total de pacientes que estaban vivos el primer año después del inicio del tratamiento (197) entre el número total de pacientes que comenzaron el tratamiento (375) (tabla 6-3).

TABLA 6-3. Análisis de la supervivencia de los pacientes tratados de 2000 a 2004 y seguidos hasta 2005 (ninguna pérdida de seguimiento): I

		NÚMERO DE VIVOS AL FINAL DEL AÑO				
Año de tratamiento	**N.° de pacientes tratados**	**1.er año**	**2.° año**	**3.er año**	**4.° año**	**5.° año**
2000	84	44	21	13	10	8
2001	62	31	14	10	6	
2002	93	50	20	13		
2003	60	29	16			
2004	76	43				
Totales	375	197				

P_1 = Probabilidad de sobrevivir el 1.er año $= \frac{197}{375} = 0{,}525$

TABLA 6-4. Análisis de la supervivencia de los pacientes tratados de 2000 a 2004 y seguidos hasta 2005 (ninguna pérdida de seguimiento): II

		NÚMERO DE VIVOS AL FINAL DEL AÑO				
Año de tratamiento	**N.° de pacientes tratados**	**1.er año**	**2.° año**	**3.er año**	**4.° año**	**5.° año**
2000	84	44	21	13	10	8
2001	62	31	14	10	6	
2002	93	50	20	13		
2003	60	29	16			
2004	76	43				
Totales		197	71			

$$P_2 = \text{Probabilidad de sobrevivir el 2.° año} = \frac{71}{197\text{-}43} = 0{,}461$$

La probabilidad de sobrevivir el primer año (P_1) es:

$$P_1 = \frac{197}{375} = 0{,}525$$

A continuación nos preguntamos: «¿Cuál es la probabilidad de que, tras sobrevivir el primer año tras iniciar el tratamiento, el paciente sobreviva el segundo año?». En la tabla 6-4 observamos que 197 personas sobrevivieron el primer año, pero de 43 de ellos (los que iniciaron el tratamiento en 2004) no tenemos más información porque fueron observados durante sólo 1 año. Como 71 sobrevivieron el segundo año, calculamos la probabilidad de sobrevivir el segundo año si el paciente sobrevivió el primer año (P_2) del siguiente modo:

$$P_2 = \frac{71}{197\text{-}43} = 0{,}461$$

En el denominador restamos los 43 pacientes de los que no tenemos datos durante el segundo año.

Siguiendo este patrón, nos preguntamos: «Dado que una persona ha sobrevivido hasta el final del segundo año, ¿cuál es la probabilidad de que sobreviva hasta el final del tercer año?».

En la tabla 6-5 observamos que 36 sobrevivieron el tercer año. Aunque 71 habían sobrevivido el segundo año, no disponemos de más información sobre la supervivencia de 16 de ellos porque fueron incorporados tarde al estudio. Por tanto, restamos 16 a 71 y calculamos la probabilidad de sobrevivir el tercer año, teniendo en cuenta la supervivencia al final del segundo año (P_2), del siguiente modo:

$$P_3 = \frac{36}{71\text{-}16} = 0{,}655$$

Seguidamente nos preguntamos: «Si una persona sobrevive hasta el final del tercer año, ¿cuál es la probabilidad de que sobreviva hasta el final del cuarto año?».

TABLA 6-5. Análisis de la supervivencia de los pacientes tratados de 2000 a 2004 y seguidos hasta 2005 (ninguna pérdida de seguimiento): III

		NÚMERO DE VIVOS AL FINAL DEL AÑO				
Año de tratamiento	**N.° de pacientes tratados**	**1.er año**	**2.° año**	**3.er año**	**4.° año**	**5.° año**
2000	84	44	21	13	10	8
2001	62	31	14	10	6	
2002	93	50	20	13		
2003	60	29	16			
2004	76	43				
Totales			71	36		

$$P_3 = \text{Probabilidad de sobrevivir el 3.er año} = \frac{36}{71\text{-}16} = 0{,}655$$

Como se observa en la tabla 6-6, un total de 36 personas sobrevivieron el tercer año, pero carecemos de información para 13 de ellos. Como 16 sobrevivieron el cuarto año, la probabilidad de sobrevivir el cuarto año, si la persona había sobrevivido el tercer año (P_4), es:

$$P_4 = \frac{16}{36-13} = 0{,}696$$

Por último, realizamos la misma operación para el quinto año (tabla 6-7). Observamos que 16 personas sobrevivieron el cuarto año, pero carecemos de más información para 6 de ellos.

Como 8 personas estaban vivas al final del quinto año, la probabilidad de sobrevivir el quinto año, cuando se ha sobrevivido el cuarto año (P_5), es:

$$P_5 = \frac{8}{16-6} = 0{,}800$$

Utilizando todos los datos que hemos calculado, nos preguntamos: «¿Cuál es la probabilidad de sobrevivir los 5 años?». En la tabla 6-8 se muestran todas las probabilidades que hemos calculado de sobrevivir cada año individual.

Ahora podemos responder a esta pregunta: «Si una persona es incorporada al estudio, ¿cuál es la probabilidad de que sobreviva 5 años tras iniciar el tratamiento?». La probabilidad de sobrevivir 5 años es el producto de las probabilidades de sobrevivir cada año, mostradas en la tabla 6-8. Por tanto, la probabilidad de sobrevivir 5 años es:

$$= P_1 x P_2 x P_3 x P_4 x P_5$$
$$= 0{,}525 \times 0{,}461 \times 0{,}655 \times 0{,}696 \times 0{,}800$$
$$= 0{,}088 \text{ o } 8{,}8\%$$

Las probabilidades de sobrevivir diferentes periodos de tiempo se muestran en la tabla 6-9. Estos cálculos pueden presentarse gráficamente en una curva de supervivencia, como se observa en la figura 6-11. Obsérvese que estos cálculos utilizan todos los datos que hemos

TABLA 6-6. Análisis de la supervivencia de los pacientes tratados de 2000 a 2004 y seguidos hasta 2005 (ninguna pérdida de seguimiento): IV

		NÚMERO DE VIVOS AL FINAL DEL AÑO				
Año de tratamiento	**N.° de pacientes tratados**	**1.er año**	**2.° año**	**3.er año**	**4.° año**	**5.° año**
2000	84	44	21	13	10	8
2001	62	31	14	10	6	
2002	93	50	20	13		
2003	60	29	16			
2004	76	43				
Totales				36	16	

P_4 = Probabilidad de sobrevivir el 4.° año $= \frac{16}{36-13} = 0{,}696$

TABLA 6-7. Análisis de la supervivencia de los pacientes tratados de 2000 a 2004 y seguidos hasta 2005 (ninguna pérdida de seguimiento): V

		NÚMERO DE VIVOS AL FINAL DEL AÑO				
Año de tratamiento	**N.° de pacientes tratados**	**1.er año**	**2.° año**	**3.er año**	**4.° año**	**5.° año**
2000	84	44	21	13	10	8
2001	62	31	14	10	6	
2002	93	50	20	13		
2003	60	29	16			
2004	76	43				
Totales					16	8

P_5 = Probabilidad de sobrevivir el 5.° año $= \frac{8}{16-6} = 0{,}800$

TABLA 6-8. Probabilidad de supervivencia en cada año del estudio

P_1 = Probabilidad de sobrevivir el 1.er año $= \frac{197}{375} = 0{,}525 = 52{,}5\%$

P_2 = Probabilidad de sobrevivir el 2.° año dada la supervivencia al final del 1.er año $= \frac{71}{197\text{-}43} = 0{,}461 = 46{,}1\%$

P_3 = Probabilidad de sobrevivir el 3.er año dada la supervivencia al final del 2.° año $= \frac{36}{71-16} = 0{,}655 = 65{,}5\%$

P_4 = Probabilidad de sobrevivir el 4.° año dada la supervivencia al final del 3.er año $= \frac{16}{36-13} = 0{,}696 = 69{,}6\%$

P_5 = Probabilidad de sobrevivir el 5.° año dada la supervivencia al final del 4.° año $= \frac{8}{16-6} = 0{,}800 = 80{,}0\%$

TABLA 6-9. Probabilidades acumuladas de sobrevivir diferentes períodos de tiempo

Probabilidad de sobrevivir 1 año = $P_1 = 0{,}525 = 52{,}5\%$
Probabilidad de sobrevivir 2 años = $P_1 \times P_2 = 0{,}525 \times 0{,}461 = 0{,}242 = 24{,}2\%$
Probabilidad de sobrevivir 3 años = $P_1 \times P_2 \times P_3 = 0{,}525 \times 0{,}461 \times 0{,}655 = 0{,}159 = 15{,}9\%$
Probabilidad de sobrevivir 4 años = $P_1 \times P_2 \times P_3 \times P_4 = 0{,}525 \times 0{,}461 \times 0{,}655 \times 0{,}696 = 0{,}110 = 11{,}0\%$
Probabilidad de sobrevivir 5 años = $P_1 \times P_2 \times P_3 \times P_4 \times P_5 = 0{,}525 \times 0{,}461 \times 0{,}655 \times 0{,}696 \times 0{,}800 = 0{,}088 = 8{,}8\%$

Figura 6-11. Curva de supervivencia para un ejemplo hipotético de pacientes tratados de 2000 a 2004 y seguidos hasta 2005.

obtenido, incluidos los datos de los pacientes que no fueron observados durante los 5 años del estudio. Como resultado, el uso de los datos es económico y eficiente.

Cálculo de una tabla de vida

Fijémonos ahora en los datos de este ejemplo en la forma de tabla estándar en la que suelen presentarse para calcular una tabla de vida. En el ejemplo que acabamos de analizar, las personas de las que no se disponían datos para los 5 años del estudio fueron las que se incorporaron tiempo después de que el estudio hubiese comenzado, por lo que no fueron seguidas durante el período total de 5 años. En prácticamente todos los estudios de supervivencia, sin embargo, también se pierden individuos durante el período de seguimiento. Puede ocurrir que se pierdan o que declinen seguir participando en el estudio. Para calcular la tabla de vida, las personas de las que carecemos de datos durante el período completo de seguimiento (bien porque el seguimiento no fue posible o porque se incorporaron al estudio una vez que éste ya había comenzado) se denominan «pérdidas» (o perdidos durante el seguimiento).

En la tabla 6-10 se muestran los datos de este ejemplo con información sobre el número de muertes y pérdidas en cada intervalo. Las columnas se numeran únicamente para tener una referencia. En la fila directamente inferior a los números de las columnas se muestran los términos empleados con frecuencia en los cálculos de las tablas de vida. Las cinco filas siguientes de la tabla proporcionan los datos de los 5 años del estudio.

Las columnas son las siguientes:

Columna (1): el intervalo desde el comienzo del tratamiento.
Columna (2): el número de individuos del estudio que estaban vivos al comienzo de cada intervalo.
Columna (3): el número de individuos del estudio que murieron durante dicho intervalo.
Columna (4): el número que se «perdió» durante el intervalo, es decir, el número de individuos del estudio que no fueron seguidos durante todo el período del estudio, porque se perdieron durante el seguimiento o porque se incorporaron al estudio una vez que el mismo ya había comenzado.

TABLA 6-10. Reestructuración de datos en formato estándar para calcular una tabla de vida

(1) Intervalo desde el comienzo del tratamiento	(2) Vivos al comienzo del intervalo	(3) Muertos durante el intervalo	(4) Perdidos durante el intervalo
x	4	dx	w_x
1.er año	375	178	0
2.° año	197	83	43
3.er año	71	19	16
4.° año	36	7	13
5.° año	16	2	6

La tabla 6-11 incorpora columnas adicionales a la tabla 6-10. Estas columnas muestran los cálculos y son las siguientes:

Columna (5): el número de personas que tienen efectivamente riesgo de morir durante el intervalo. Se supone que las pérdidas de seguimiento (perdidos) durante cada intervalo de tiempo han ocurrido uniformemente durante todo el intervalo. (Esta suposición es más probable que se cumpla cuando el intervalo es corto.) Por tanto, asumimos que tenían riesgo durante la mitad del intervalo. Así, para calcular el número de personas con riesgo durante cada intervalo restamos la mitad de los perdidos durante dicho intervalo, como se indica en el encabezado de la columna 5.

Columna (6): la proporción que murió durante el intervalo se calcula dividiendo:

$$\frac{\text{El número que falleció durante el intervalo (columna 3)}}{\text{El número que efectivamente tenía riesgo de morir durante el intervalo (columna 5)}}$$

Columna (7): la proporción que no murió durante el intervalo, es decir, la proporción de los que estaban vivos al inicio del intervalo y que sobrevivieron dicho intervalo = 1,0 - proporción que murió durante el intervalo (columna 6).

Columna (8): la proporción que sobrevivió desde el punto en el que se incorporaron al estudio hasta el final de este intervalo (supervivencia acumulada). Se obtiene multiplicando la proporción de los que estaban vivos al inicio de este intervalo y los que sobrevivieron este intervalo por la proporción que había sobrevivido desde la incorporación hasta el final del intervalo previo. Así, cada una de las cifras de la columna 8

TABLA 6-11. Cálculo de una tabla de vida

(1) Intervalo desde el comienzo del tratamiento	(2) Vivos al comienzo del intervalo	(3) Muertos durante el intervalo	(4) Perdidos durante el intervalo	(5) Número efectivo de expuestos al riesgo de morir durante el intervalo: Col (2) - % [Col (4)]	(6) Proporción que falleció durante el intervalo: Col (3) / Col (5)	(7) Proporción que no falleció durante el intervalo: 1 - Col (6)	(8) Proporción acumulada que sobrevivió desde la incorporación al final del intervalo: supervivencia acumulada
x	L	dx	w_x	r_x	q_x	px	Px
1.er año	375	178	0	375,0	0,475	0,525	0,525
2.° año	197	83	43	175,5	0,473	0,527	0,277
3.er año	71	19	16	63,0	0,302	0,698	0,193
4.° año	36	7	13	29,5	0,237	0,763	0,147
5.° año	16	2	6	13,0	0,154	0,846	0,124

informa de la proporción de personas que iniciaron el estudio que sobrevivió hasta el final de este intervalo. Esto se demostrará calculando las dos primeras filas de la tabla 6-11.

Fijémonos en los datos del primer año. (En estos cálculos, redondearemos los resultados en cada paso y utilizaremos las cifras redondeadas para el próximo cálculo. En realidad, sin embargo, cuando se calculan las tablas de vida, se utilizan las cifras no redondeadas para calcular cada intervalo posterior y, al final de todos los cálculos, todas las cifras se redondean con el fin de presentar los resultados.) Había 375 individuos incorporados al estudio que estaban vivos al comienzo del primer año tras su incorporación (columna 2). De éstos, 178 murieron durante el primer año (columna 3). Todos los individuos fueron seguidos durante el primer año, por lo que no hubo pérdidas (columna 4). Por tanto, 375 personas tenían efectivamente riesgo de morir durante este intervalo (columna 5). La proporción que murió durante este intervalo fue 0,475: 178 (el número que murió [columna 3]) dividido entre 375 (el número que tenía riesgo de morir [columna 5]). La proporción que no falleció durante el intervalo es 1,0 — [la proporción que falleció (1,0 - 0,475)] = 0,525 (columna 7). Para el primer año tras la incorporación, ésta también es la proporción que sobrevivió desde la incorporación hasta el final del intervalo (columna 8).

A continuación fijémonos en los datos del segundo año. Es importante que comprendamos estos cálculos, ya que sirven de modelo para calcular cada año sucesivo en la tabla de vida.

Para calcular el número de individuos vivos al comienzo del segundo año, comenzamos con el número de vivos al comienzo del *primer* año y restamos a ese número la cifra de muertos y perdidos durante dicho año. Por tanto, al comienzo del segundo año, 197 individuos estaban vivos al comienzo del intervalo (columna 2 [375 - 178 - 0]). De éstos, 83 murieron durante el segundo año (columna 3). Se produjeron 43 pérdidas de individuos que habían sido observados durante sólo 1 año (columna 4). Como se ha expuesto anteriormente, restamos la mitad de las pérdidas, 21,5 (43/2), a los 197 que estaban vivos al inicio del intervalo; el resultado son 175,5 personas que tenían efectivamente riesgo de morir durante este intervalo (columna 5). La proporción que murió durante este intervalo (columna 6) fue 0,473, es decir, 83 (el número que murió [columna 3]) dividido entre 175,5 (el número con riesgo de morir [columna 5]). La proporción que no murió durante el intervalo es 1,0 - la proporción que murió (1,0 - 0,473) = 0,527 (columna 7). La proporción de individuos que sobrevivieron desde el comienzo del tratamiento hasta el final del segundo año es el producto de 0,525 (la proporción de los que habían sobrevivido desde el comienzo del tratamiento hasta el final del primer año, es decir, el comienzo del segundo año) por 0,527 (la proporción de personas que estaban vivas al comienzo del segundo año y sobrevivieron hasta el final del segundo año) = 0,277 (columna 8). Por tanto, un 27,7% de los individuos sobrevivieron desde el comienzo del tratamiento hasta el final del segundo año. Fijándonos en la última entrada de la columna 8, observamos que el 12,4% de todos los sujetos que iniciaron el estudio sobrevivieron hasta el final del quinto año.

Analice los años restantes de la tabla 6-11 para asegurarse de que entiende los conceptos y los cálculos.

EL MÉTODO KAPLAN-MEIER

A diferencia del abordaje que acabamos de exponer, en el método Kaplan-Meier[1] no se utilizan intervalos predeterminados, como 1 mes o 1 año. Con este método identificamos el punto exacto en el tiempo en el que se produjo cada muerte, de modo que cada muerte termina el intervalo previo y comienza un nuevo intervalo (y una nueva fila en la tabla de Kaplan-Meier). El número de personas que murieron en dicho punto se utiliza como numerador y el número de vivos hasta ese punto (incluidos los que murieron en ese punto en el tiempo) se emplea como denominador, después de restar los perdidos producidos antes de ese punto.

Fijémonos en el pequeño estudio hipotético que se muestra en la figura 6-12. Seis pacientes fueron estudiados, de los que cuatro murieron y dos fueron perdidos durante el seguimiento («perdidos»). Las muertes se produjeron 4,10,14 y 24 meses después de

Figura 6-12. Ejemplo hipotético de un estudio de seis pacientes analizados con el método Kaplan-Meier.

TABLA 6-12. Cálculo de la supervivencia empleando el método Kaplan-Meier*

(1) Tiempos hasta las muertes desde el inicio del tratamiento (meses)	(2) Números de vivos en cada momento	(3) Número de fallecidos en cada momento	(4) Proporción de fallecidos en ese momento: Col (3) / Col (2)	(5) Proporción que sobrevivió en ese momento: 1 — Col (4)	(6) Proporción acumulada que sobrevivió hasta ese momento: supervivencia acumulada
4	6	1	0,167	0,833	0,833
10	4	1	0,250	0,750	0,625
14	3	1	0,333	0,667	0,417
24	1	1	1,000	0,000	0,000

*Véase el texto y la figura 6-12 en relación con las pérdidas.

la incorporación en el estudio. Los datos se organizan como se muestra en la tabla 6-12:

Columna (1): los tiempos hasta las muertes desde el momento de la incorporación (tiempo en el que se inició el tratamiento).
Columna (2): el número de pacientes que estaban vivos y eran seguidos en el momento de esa muerte, incluidos los que murieron es ese tiempo.
Columna (3): el número de muertos en ese tiempo.
Columna (4): la proporción entre los que estaban vivos y eran seguidos (columna 2) y los que murieron en ese tiempo (columna 3) [columna 3 / columna 2].
Columna (5): la proporción de los que estaban vivos y sobrevivieron (1,0 - columna 4).
Columna (6): supervivencia acumulada (la proporción de los que participaron desde el inicio y sobrevivieron hasta ese punto).

Fijémonos en la primera fila de la tabla. La primera muerte se produjo a los 4 meses, cuando 6 pacientes estaban vivos y eran seguidos (v. fig. 6-12). En ese punto se produjo una muerte (columna 3), para una proporción de 1/6 = 0,167 (columna 4). La proporción de los que sobrevivieron en ese momento es de 1,0 - columna 4, o 1,0 — 0,167 = 0,833 (columna 5), que también es la supervivencia acumulada en ese punto (columna 6).

La siguiente muerte tuvo lugar 10 meses después de la incorporación inicial de los 6 pacientes en el estudio, y los datos para este tiempo se observan en la siguiente fila de la tabla. Aunque antes de esta muerte sólo se había producido otro fallecimiento, el número de vivos y seguidos es de sólo 4 porque también se había producido una pérdida antes de este punto (no se muestra en la tabla pero puede verse en la fig. 6-12). Por tanto, se produjo una muerte (columna 3) y, como se observa en la tabla 6-12, la proporción que murió es $^1/_4$ o 0,250 (columna 4). La proporción que sobrevivió es 1,0 - columna 4, o 1,0 — 0,250 = 0,750 (columna 5). Por último, la proporción acumulada de supervivientes (columna 6) es el producto de la proporción que sobrevivió hasta el final del intervalo previo (hasta justo antes de la muerte previa), mostrada en la columna 6 de la primera fila (0,833), por la proporción que sobrevivió desde ese momento hasta justo antes de la segunda muerte (segunda fila en la columna 5: 0,750). El producto es 0,625, es decir, un 62,5% de los que iniciaron el estudio sobrevivieron hasta este punto. Revise las siguientes dos filas de la tabla para asegurarse de que ha entendido los conceptos y los cálculos.

Los valores calculados en la columna 6 se representan como se observa en la figura 6-13. Obsérvese que los datos se representan escalonadamente en vez de en una pendiente suave, ya que, tras la disminución de la supervivencia resultante de cada muerte, la supervivencia permanece sin cambios hasta que tiene lugar el siguiente fallecimiento.

Cuando se dispone de información acerca del momento exacto de la muerte, el método Kaplan-Meier claramente hace pleno uso de la misma, porque los datos se usan para definir los intervalos. Aunque el método es adecuado para estudios con pocos pacientes, hoy día existen programas computarizados fácilmente disponibles que hacen que este método sea aplicable también a grupos de datos extensos. Muchos de los estudios de los trabajos publicados comunican en la actualidad datos de supervivencia empleando el método Kaplan-Meier. Por ejemplo, en el año 2000 Rosenhek y cois, publicaron un estudio de pacientes con estenosis aórtica grave pero asintomática[2]. Un aspecto no resuelto era si los pacientes con enfermedad asintomática debían ser sometidos a un recambio valvular aórtico. Los investigadores examinaron la historia natural de esta enfermedad para valorar la supervivencia global

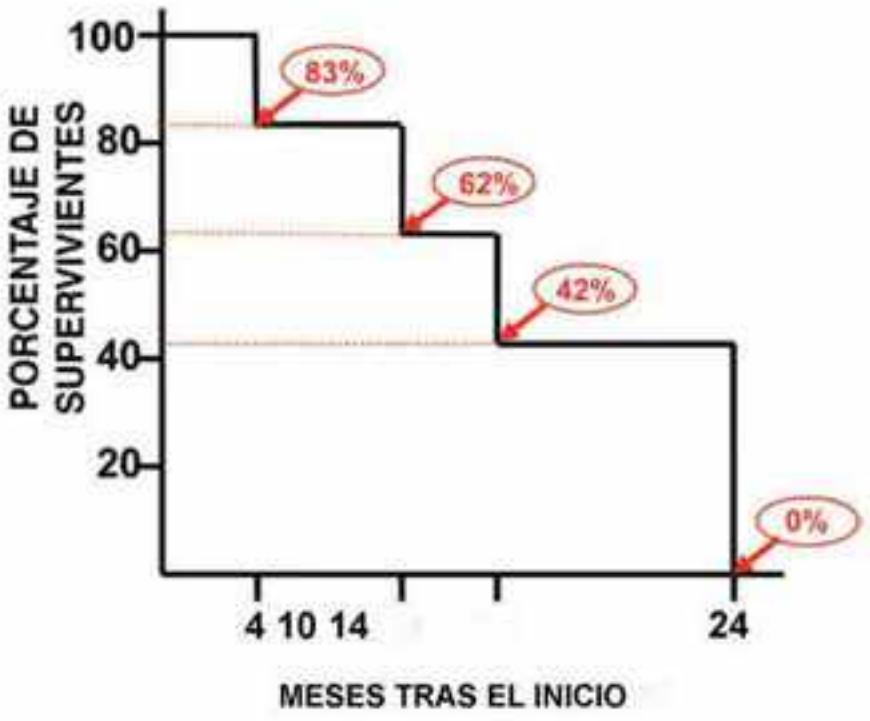


Figura 6-13. Gráfico de Kaplan-Meier del estudio de supervivencia hipotético de seis pacientes mostrados en la figura 6-12. Los porcentajes en rojo indican las proporciones acumuladas de supervivientes tras las muertes mostradas en la figura 6-12 y se han tomado de la columna 6 de la tabla 6-12. (V. explicación del método Kaplan-Meier en las págs. 126-128.)

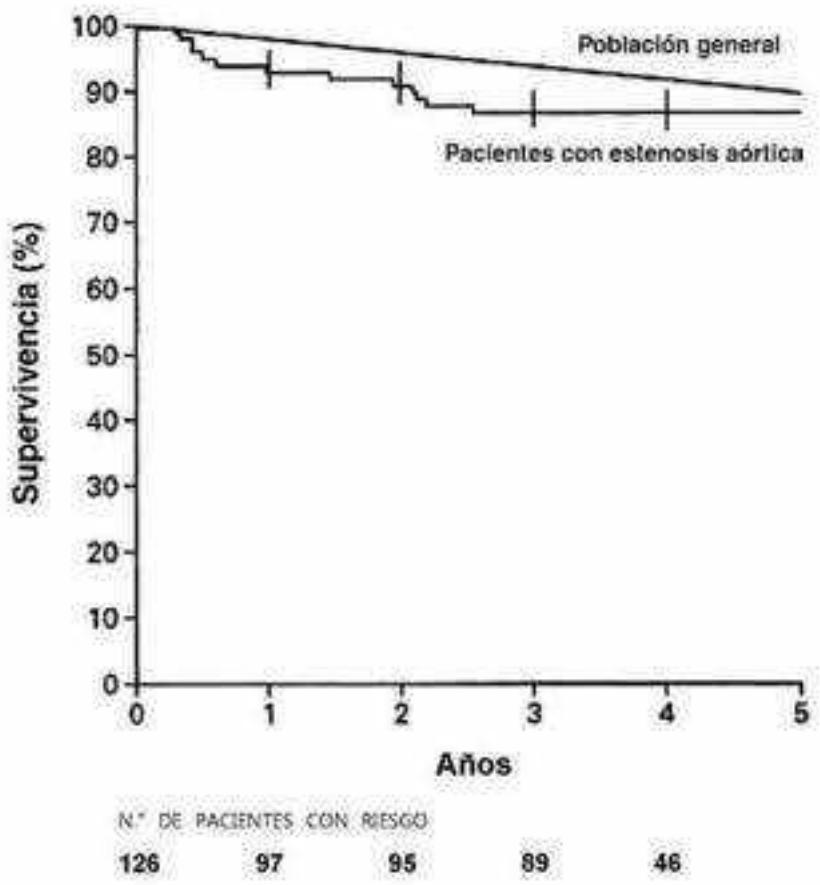
de estos pacientes e identificar factores predictivos del resultado. En la figura 6-14A se muestra su análisis de Kaplan-Meier de la supervivencia de 126 pacientes con estenosis aórtica comparados con personas de sexo y edades similares de la población general. Aunque la supervivencia era ligeramente inferior en los pacientes con estenosis aórtica, la diferencia no fue significativa. Cuando analizaron varios factores de riesgo, encontraron que la calcificación moderada o grave de la válvula aórtica era un factor predictivo importante de complicaciones cardíacas posteriores y de un pronóstico muy malo (v. fig. 6-14B). La supervivencia sin complicaciones fue mucho peor en los pacientes con calcificación moderada o grave que en los pacientes con calcificación leve o ausente. Los autores concluyeron que dichos pacientes debían ser sometidos precozmente a un recambio valvular en vez de retrasar la cirugía hasta que se desarrollen los síntomas.

SUPOSICIONES EMPLEADAS AL USAR TABLAS DE VIDA

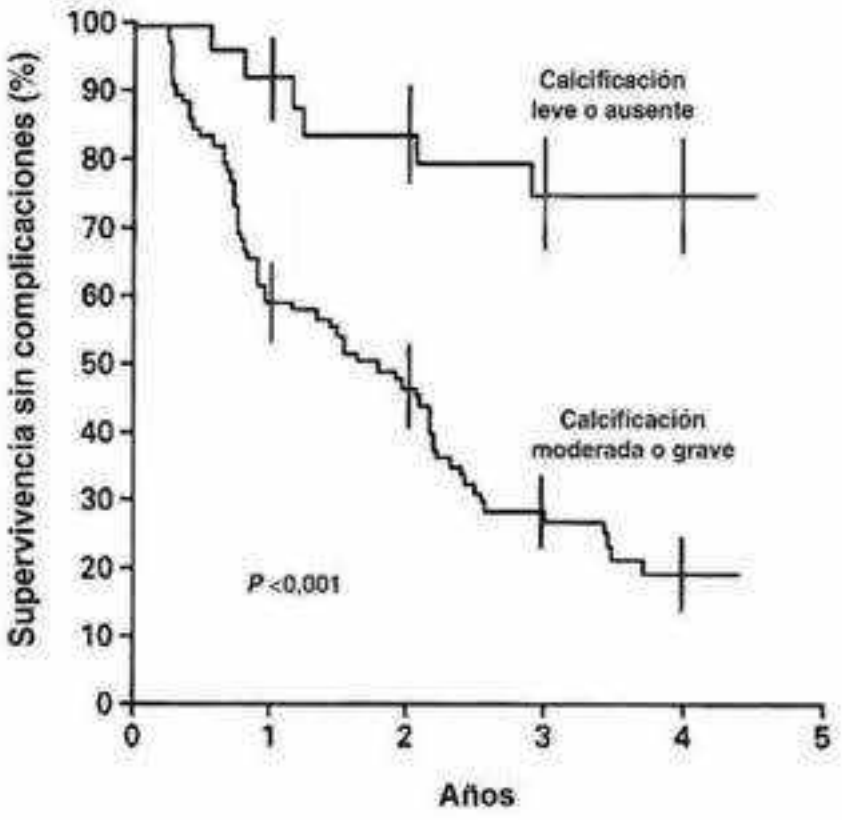
Cuando se utilizan tablas de vida estamos suponiendo dos aspectos importantes. En primer lugar, suponemos que no se han producido cambios seculares (temporales) en la eficacia del tratamiento o en la supervivencia a lo largo del tiempo de calendario. Es decir, asumimos que durante el período del estudio no se han producido mejorías en el tratamiento y que la supervivencia en un año de calendario del estudio es la misma que en otro año de calendario del estudio. Claramente, si el estudio se realiza a lo largo de muchos años, esta suposición puede no ser válida, porque afortunadamente los tratamientos mejoran con el paso del tiempo. Si creemos que la eficacia del tratamiento puede haber cambiado durante el período del estudio, podríamos examinar los datos iniciales separadamente de los datos más tardíos. Si encontramos diferencias, podríamos analizar separadamente los períodos iniciales y tardíos.

La segunda suposición se relaciona con el seguimiento de las personas incorporadas al estudio. En prácticamente todos los estudios reales se pierde el seguimiento de algún paciente. Esto puede ocurrir por diversos motivos. Algunos pueden morir y no pueden ser seguidos. Algunos pueden cambiar de residencia o buscar asistencia médica en otro centro. Algunos pueden perderse porque la enfermedad desaparece y se encuentran bien. En la mayoría de los estudios desconocemos los motivos reales de las pérdidas de seguimiento. ¿Cómo podemos abordar el problema de las personas que perdemos durante el seguimiento y de las cuales no tenemos, por tanto, más información sobre su supervivencia? Como disponemos de datos basales de estas personas, podríamos comparar sus características con las de las personas que continuaron en el estudio, pero el problema sigue presente. Si se pierde el seguimiento de una gran proporción de la población del estudio, los hallazgos del estudio serán menos válidos. El reto es minimizar las pérdidas de seguimiento. En cualquier caso, la segunda suposición asumida en las tablas de vida es que la experiencia de supervivencia de las personas de las que se perdió el seguimiento es la misma que la experiencia de los que continuaron el seguimiento. Aunque esta suposición se asume con el fin de realizar los cálculos, la realidad es que su validez a menudo puede ser cuestionable.

Aunque el término *tabla de vida* puede sugerir que estos métodos son útiles únicamente para calcular la supervivencia, en realidad no es así. La muerte no tiene por qué ser el punto final de estos cálculos. Por ejemplo, la *supervivencia* puede calcularse como el tiempo que transcurre hasta la aparición de hipertensión, de una



A



B

N.° DE PACIENTES CON RIESGO					
Calcificación leve o ausente	25	23	20	17	9
Calcificación moderada o grave	101	48	38	21	7

Figura 6-14. A, Análisis de Kaplan-Meier de la supervivencia global de 126 pacientes con estenosis aórtica asintomática, pero grave, en comparación con personas de sexo y edad similares de la población general. Este análisis incluye las muertes perioperatorias y postoperatorias de los pacientes que precisaron sustitución valvular durante el seguimiento. B, Análisis de Kaplan-Meier de supervivencia libre de complicaciones de 25 pacientes con calcificación de la válvula aórtica leve o ausente en comparación con 101 pacientes con calcificación moderada o grave. Las barras verticales indican errores estándar. (De Rosenhek R, Binder T, Porenta G, et al: Predictors of outcome in severe, asymptomatic aortic stenosis. N Engl J Med 343:611-617, 2000.)

recurrencia de un cáncer o el tiempo de supervivencia sin efectos adversos del tratamiento. Además, aunque podemos fijarnos en un sola curva de supervivencia, a menudo el mayor interés reside en la comparación entre dos o más curvas de supervivencia, como las de los tratados y no tratados en un ensayo clínico aleatorizado. Al realizar dichas comparaciones, existen métodos estadísticos disponibles para determinar si una curva es significativamente diferente de otra.

Ejemplo de utilización de una tabla de vida

Las tablas de vida se emplean en prácticamente todas las áreas clínicas. Son una forma estándar de expresar y comparar la supervivencia. Examinemos algunos ejemplos. Uno de los grandes triunfos de la pediatría en las últimas décadas ha sido el tratamiento de la leucemia infantil. Sin embargo, la mejoría ha sido mucho mayor en los blancos que en los negros, y los motivos de estas diferencias no están claros.

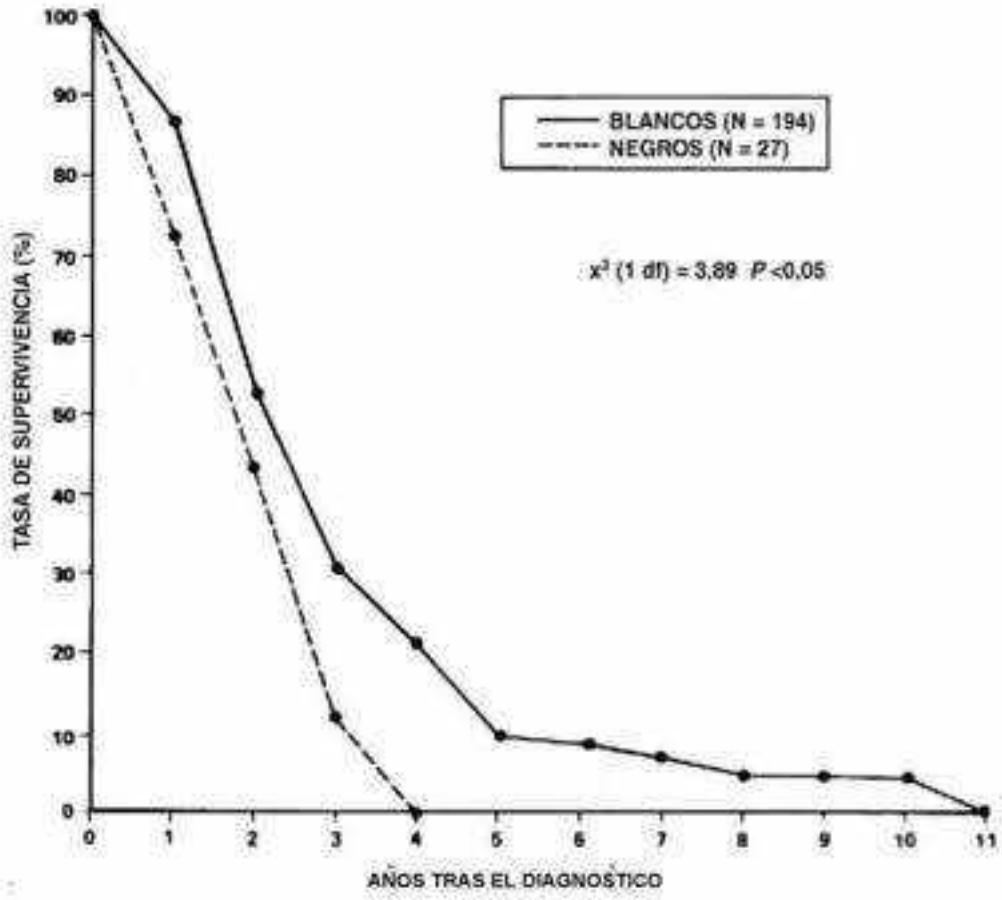


Figura 6-15. Supervivencia de niños de 0 a 19 años con leucemia linfocítica aguda por raza, área metropolitana de Baltimore, 1960-1975. (De Szklo M, Gordis L, Tonascia J, et al: The changing survivorship of white and black children with leukemia. Cancer 42:59-66, 1978. Copyright © 1978 American Cancer Society. Reproducido con autorización de Wiley-Liss, Inc., una filial de John Wiley & Sons, Inc.)

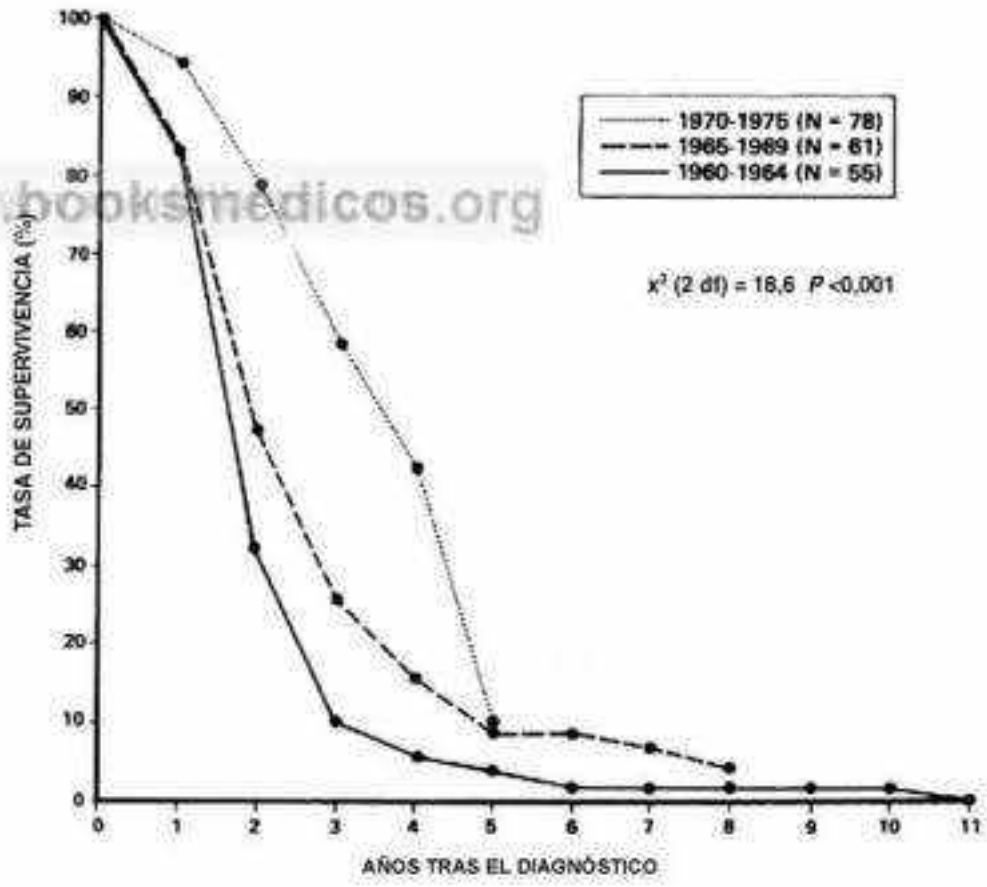


Figura 6-16. Cambios temporales en la supervivencia de niños blancos de 0 a 19 años con leucemia linfocítica aguda, área metropolitana de Baltimore, 1960-1975. (De Szklo M, Gordis L, Tonascia J, et al: The changing survivorship of white and black children with leukemia. Cancer 42:59-66, 1978. Copyright © 1978 American Cancer Society. Reproducido con autorización de Wiley-Liss, Inc., una filial de John Wiley & Sons, Inc).

En un periodo en el que las tasas de supervivencia de la leucemia aguda infantil estaban aumentando rápidamente, se llevó a cabo un estudio para explorar las diferencias raciales en la supervivencia. Los datos de este estudio se muestran en las figuras 6-15 a 6-17[1]. Las curvas se basan en tablas de vida que fueron realizadas empleando el abordaje expuesto anteriormente.

En la figura 6-15 se muestra la supervivencia de los niños blancos y negros con leucemia en Baltimore a lo largo de un período de 16 años. Ningún niño negro sobrevivió más de 4 años, pero algunos niños blancos sobrevivieron hasta 11 años en este periodo de observación de 16 años.

¿Qué cambios tuvieron lugar en la supervivencia durante los 16 años del estudio? En las figuras 6-16 y 6-17 se muestran los cambios en la mortalidad por leucemia a lo largo del tiempo en los niños blancos y negros, respectivamente. El período de 16 años fue dividido en tres períodos: de 1960 a 1964 *(línea continua)*, de 1965 a 1969 *(línea discontinua)* y de 1970 a 1975 *(línea de puntos)*.

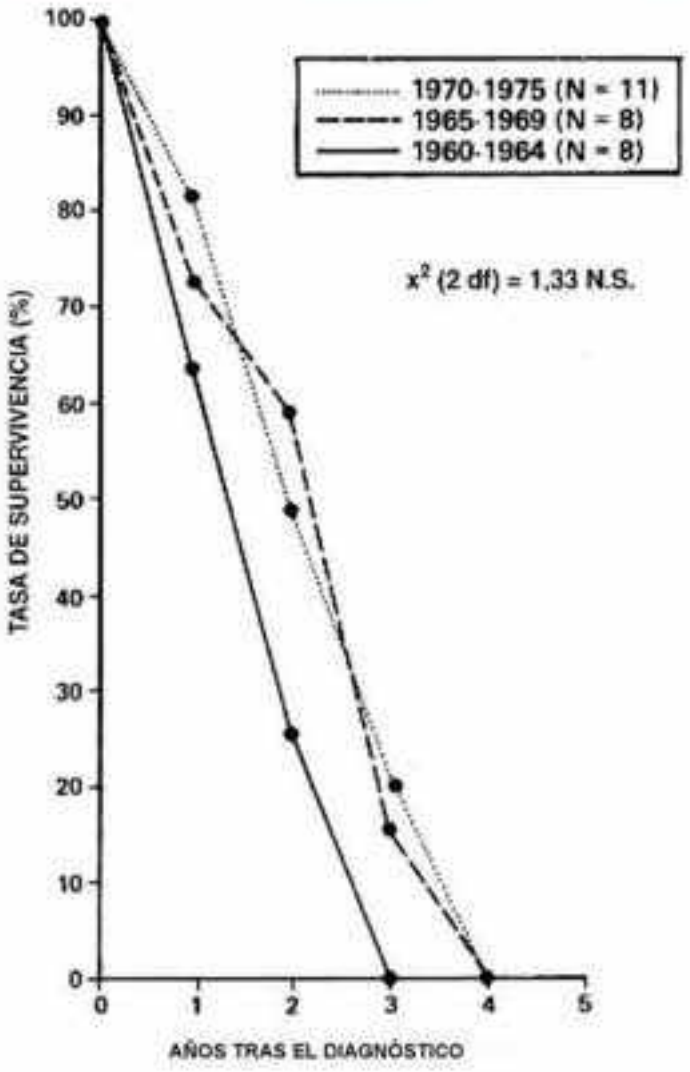


Figura 6-17. Cambios temporales en la supervivencia de niños negros de 0 a 19 años con leucemia linfocítica aguda, área metropolitana de Baltimore, 1960-1975. (De Szklo M, Cordis L, Tonascia J, et al: The changing survivorship of white and black children with leukemia. Cancer 42:59-66, 1978. Copyright © 1978 American Cancer Society. Reproducido con autorización de Wiley-Liss, Inc., una filial de John Wiley & Sons, Inc.)

En los blancos (v. fig. 6-16), la supervivencia aumentó en cada período sucesivo. Por ejemplo, si examinamos la supervivencia a 3 años fijándonos en el punto de 3 años en cada curva sucesiva, observamos que la supervivencia mejoró del 8% al 25% y al 58%. Por el contrario, en los negros (v. fig. 6-17) se produjo una mejoría más leve de la supervivencia a lo largo del tiempo; las curvas de los dos períodos tardíos de los 5 años casi se superponen.

¿Qué explica esta diferencia racial? En primer lugar, debemos tener en cuenta los pequeños números involucrados y la posibilidad de que las diferencias pudieran haberse debido al azar. Asumamos, sin embargo, que las diferencias son reales. Durante las últimas décadas se han producido varios avances en el tratamiento de la leucemia a través de terapias combinadas, como la radiación del sistema nervioso central añadida a la quimioterapia. ¿Por qué existen entonces diferencias raciales en la supervivencia? ¿Por qué las mejoras terapéuticas que han sido tan efectivas en los niños blancos no han tenido un beneficio comparable en los niños negros? Análisis posteriores del intervalo desde el momento en el que la madre notó los síntomas hasta el momento del diagnóstico y el tratamiento indicaban que las diferencias en la supervivencia no parecían ser debidas a un retraso de los padres negros en buscar u obtener asistencia médica. Como la leucemia aguda es más grave en los negros y se encuentra más avanzada en el momento del diagnóstico, la diferencia racial podría reflejar las diferencias biológicas de la enfermedad, como una forma más agresiva y rápidamente progresiva de la enfermedad. La explicación definitiva no está clara todavía.

EFECTOS APARENTES SOBRE EL PRONÓSTICO DE LAS MEJORAS DIAGNÓSTICAS

Hemos analizado la suposición asumida al usar una tabla de vida de que *no se ha producido mejora en la eficacia del tratamiento* a lo largo del tiempo de calendario durante el período del estudio. Otro aspecto del cálculo y la interpretación de las tasas de supervivencia es el posible efecto de las *mejoras en los métodos diagnósticos* a lo largo del tiempo de calendario.

Un ejemplo interesante fue comunicado por Feinstein, Sosis y Wells[4]. Compararon la supervivencia en una cohorte de pacientes con cáncer de pulmón tratados por primera vez en 1977 con la supervivencia en una cohorte de pacientes con cáncer de pulmón tratados de 1953 a 1964. La supervivencia a seis meses fue superior en el segundo grupo tanto para la totalidad del grupo como para los subgrupos creados según el estadio de la enfermedad. Los autores encontraron que la aparente mejora en la supervivencia se debió en parte a la *migración de estadios*, un fenómeno que se muestra en la figura 6-18A-C.

En la figura 6-18A, los pacientes con cáncer son divididos en estadios «buenos» y «malos» en función de si tenían metástasis detectables en 1980. Algunos pacientes que habrían sido asignados al estadio «bueno» en 1980 puede que tuvieran micrometástasis en ese momento que habrían pasado desapercibidas (fig. 6-18B). Sin embargo, en el año 2000, a medida que mejoraron las técnicas diagnósticas, muchos de estos pacientes habrían sido asignados al estadio «malo», porque sus micrometástasis ahora se habrían identificado utilizando las nuevas técnicas diagnósticas ya disponibles (fig. 6-18C). Si esto se hubiera producido, parecería que la supervivencia por estadio habría mejorado incluso aunque no hubiese aumentado la eficacia del tratamiento durante este tiempo.

Consideremos un ejemplo hipotético que ilustra este efecto de la migración de estadios. En la figura 6- 19A-C se muestra un estudio hipotético de la tasa de letalidad en 300 pacientes con cáncer en dos períodos de tiempo, 1980 y 2000, *suponiendo que no se han producido mejoras en la eficacia del tratamiento disponible entre los dos períodos*. Asumiremos, como se muestra en la

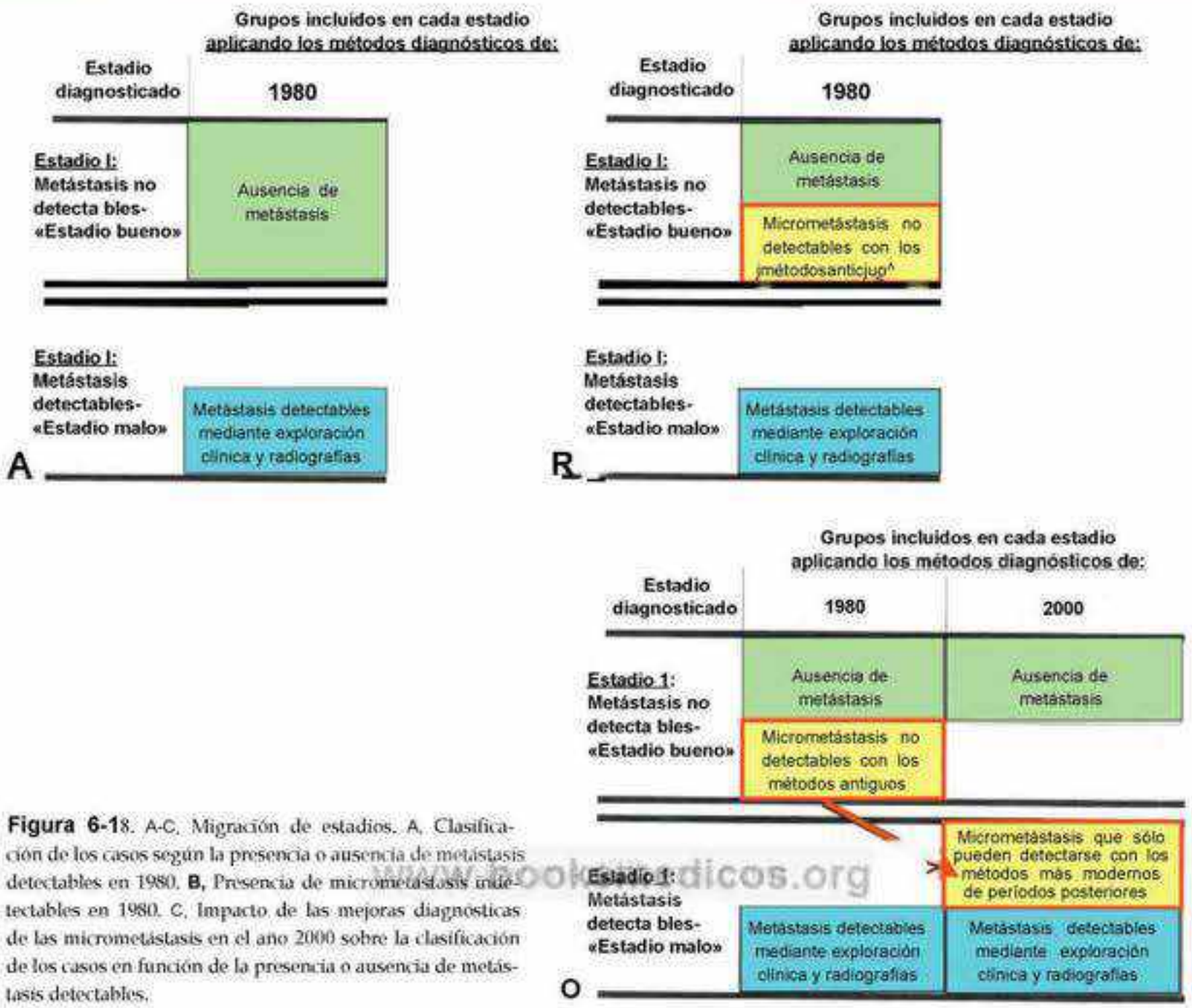


Figura 6-18. A-C, Migración de estadios. A, Clasificación de los casos según la presencia o ausencia de metástasis detectables en 1980. **B,** Presencia de micrometástasis indetectables en 1980. C, Impacto de las mejoras diagnósticas de las micrometástasis en el año 2000 sobre la clasificación de los casos en función de la presencia o ausencia de metástasis detectables.

figura 6- 19A, que en ambos periodos de tiempo la tasa de letalidad es del 10% para los pacientes sin metástasis, del 30% para los pacientes con micrometástasis y del 80% para los pacientes con metástasis. Fijándonos en la figura 6-19B, observamos que, en 1980,200 pacientes fueron clasificados en el estadio I. Cien de estos pacientes no tenían metástasis y 100 presentaban micrometástasis ocultas. La tasa de letalidad en estos casos era del 10% y el 30%, respectivamente. En 1980,100 pacientes presentaban claramente metástasis evidentes y fueron clasificados en el estadio II; su tasa de letalidad era del 80%.

Como resultado de las mejoras en las técnicas diagnósticas en el año 2000, se detectaron micrometástasis en los 100 pacientes afectados, y estos pacientes fueron clasificados en el estadio II (fig. 6-19C). Como el pronóstico de los pacientes con micrometástasis es peor que el de los otros pacientes del estadio I, y como, en el periodo tardío del estudio, los pacientes con micrometástasis ya no son incluidos en el grupo de estadio I (porque han migrado al estadio II), la tasa de letalidad de los pacientes del estadio I parece haber disminuido desde el 20% en el período inicial al 10% en el periodo tardío. Sin embargo, aunque el pronóstico de los pacientes que migraron del estadio I al estadio II fue peor que el de los otros pacientes en estadio I, el pronóstico de estos pacientes seguía siendo mejor que el de los otros pacientes en el estadio II, que tenían metástasis de mayor tamaño, de diagnóstico más fácil y una tasa de letalidad del 80%. Por tanto, la tasa de letalidad de los pacientes en estadio II también parece haber mejorado, habiendo disminuido desde el 80% en el período inicial hasta el 55% en el periodo tardío, *incluso en ausencia de mejora en la eficacia del tratamiento.*

Las mejoras aparentes en la supervivencia tanto en los pacientes en estadio I como en los pacientes en estadio II se deben sólo al cambio de clasificación de los pacientes con micrometástasis en el período tardío. Si nos fijamos en la última línea de la figura, observamos que la tasa del letalidad del 40% para el total de los 300 pacientes no ha cambiado desde el período inicial hasta el período tardío. Únicamente han cambiado las tasas de letalidad específicas de estadio aparente. Por tanto, es importante excluir la posibilidad de que se haya producido migración de estadios antes de atribuir

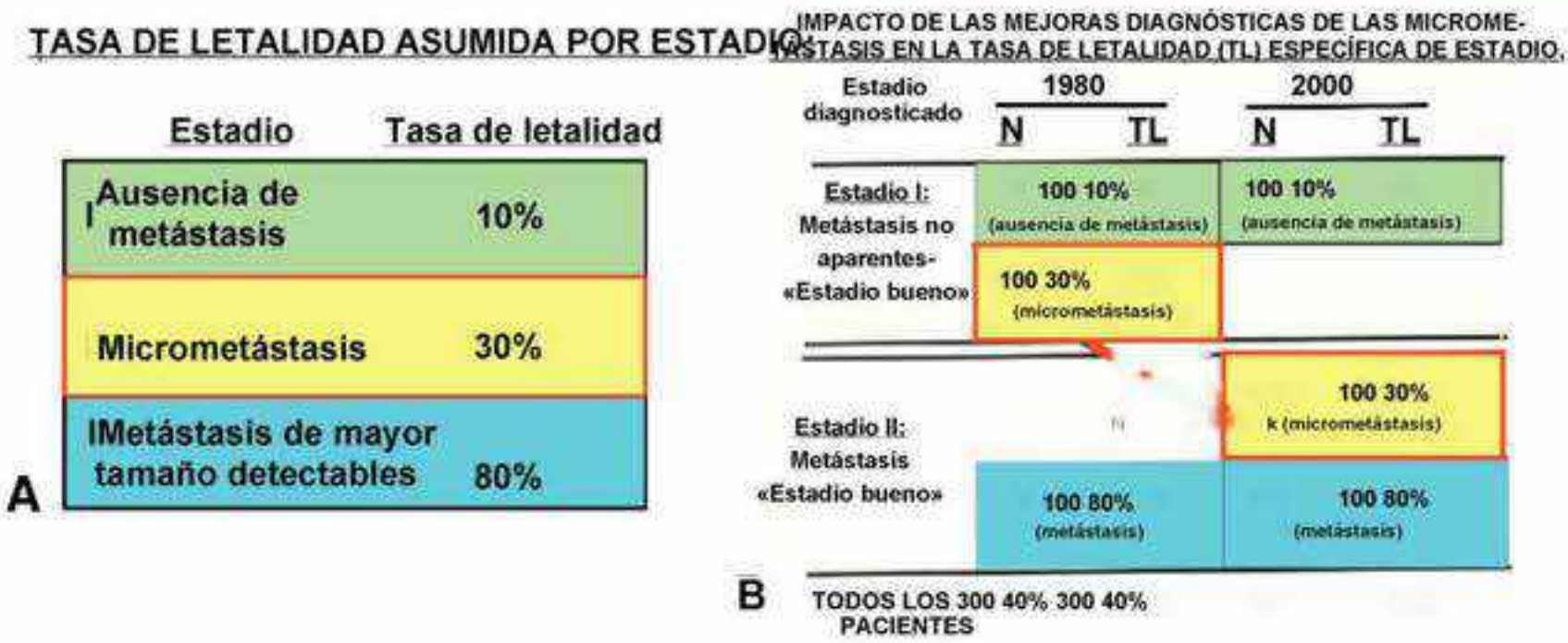


Figura 6-19. A-C, Ejemplo hipotético de migración de estadios. A, Tasa de letalidad asumida por estadio. **B**, Impacto de las mejoras diagnósticas de las micrometástasis en la tasa de letalidad (TL) específica de estadio. C, Mejoras aparentes en la supervivencia específica de estadio como resultado de la migración de estadios incluso sin mejoras en la eficacia del tratamiento.

la mejora aparente del pronóstico a la mayor eficacia de la asistencia médica.

A la migración de estadios los autores la denominan «fenómeno de Will Rogers», en referencia a Will Rogers, un humorista americano durante la época de la depresión económica de la década de 1930. En esa época, debido a las dificultades económicas, muchos residentes de Oklahoma abandonaron su estado y emigraron a California. Rogers comentó: «Cuando los habitantes de Oklahoma abandonaron su estado y emigraron a California, aumentó el nivel medio de inteligencia en ambos estados.»

MEDIANA DE SUPERVIVENCIA

Otra forma de expresar el pronóstico es mediante la *mediana de supervivencia*, que se define como el periodo de tiempo en el que sobrevive la mitad de la población del estudio. ¿Por qué deberíamos emplear la mediana de supervivencia en vez del tiempo medio de supervivencia, que es la media de los tiempos de supervivencia? La mediana de supervivencia ofrece dos ventajas sobre la supervivencia media. En primer lugar, se ve menos afectada por los extremos, mientras que la media se ve muy afectada incluso por un solo valor extremo. Una o dos personas con un tiempo de supervivencia muy prolongado podrían afectar significativamente a la media, incluso aunque todos los otros tiempos de supervivencia fuesen mucho más cortos. En segundo lugar, si utilizáramos la supervivencia media, deberíamos observar todas las muertes del estudio antes de poder calcular la media. Sin embargo, para calcular la mediana de supervivencia, sólo debemos observar las muertes de la mitad del grupo.

SUPERVIVENCIA RELATIVA

Consideremos la supervivencia a 5 años para un grupo de varones de 30 años con cáncer colorrectal. ¿Qué supervivencia a 5 años esperaríamos que tuvieran si no padeciesen un cáncer colorrectal? Claramente, sería casi del 100%. Por tanto, estamos comparando la supervivencia observada en varones jóvenes con cáncer colorrectal con una supervivencia de casi el 100% que es la esperada en los que no padecen cáncer colorrectal. ¿Qué pasaría si consideramos un grupo de varones de

80 años con cáncer colorrectal? En una población de esta edad no esperaríamos nada próximo a una supervivencia a 5 años del 100%, incluso aunque no padeciesen un cáncer colorrectal. Querríamos comparar la supervivencia observada en varones de 80 años con cáncer colorrectal con la supervivencia esperada en varones de 80 años sin cáncer colorrectal. Así, en todo grupo de personas con una enfermedad, queremos comparar su supervivencia con la supervivencia que cabría esperar en ese grupo de edad aunque *no* tuviese la enfermedad. Ésta es la denominada *supervivencia relativa*.

La supervivencia relativa se define, por tanto, como el cociente entre la supervivencia observada y la supervivencia esperada:

$$\text{Supervivencia relativa} = \frac{\text{Supervivencia observada en personas con la enfermedad}}{\text{Supervivencia esperada si la enfermedad no estuviese presente}}$$

¿Tiene alguna importancia la supervivencia relativa?

En la tabla 6-13 se muestran datos de supervivencia relativa y supervivencia observada en pacientes con cáncer de colon y recto, desde 1990 hasta 1998. Cuando nos fijamos en los grupos de edad más avanzada, que presentan altas tasas de mortalidad por otras causas, existe una gran diferencia entre la supervivencia observada y la supervivencia relativa. Sin embargo, en las personas jóvenes, que generalmente no se mueren de otras causas, la supervivencia observada y la supervivencia relativa en el cáncer de colon y recto no difieren de modo significativo.

Otra forma de ver la supervivencia relativa es examinando las curvas hipotéticas de supervivencia a 10 años en varones de 80 años, que se muestran en la figura 6-20A-D. Como referencia, en la figura 6-20A se muestra una curva de supervivencia perfecta del 100% (la curva horizontal de la parte superior) a lo largo de los 10 años del período del estudio. En la figura 6-20B se añade una curva de supervivencia observada, es decir, la supervivencia real observada en este grupo de pacientes con la enfermedad a lo largo de un período de 10 años. Como se observa en la figura 6-20C, la supervivencia esperada en este grupo de varones de 80 años es claramente menor del 100% porque en este grupo de edad las muertes por otras causas son importantes. La supervivencia relativa es el cociente entre la supervivencia observada y la supervivencia esperada. Como la supervivencia esperada se aleja de la supervivencia perfecta

TABLA 6-13. Supervivencia observada y relativa (%) a cinco años para el cáncer de colon y recto: programa SEER (Surveillance, Epidemiology, and End Results Study), 1990-1998

Edad (años)	Supervivencia observada (%)	Supervivencia relativa (%)
<50	60.4	61,5
50-64	59.4	63.7
65-74	53.7	63.8
>75	35.8	58,7

Adaptada de Edwards BK, Howe HL, Ries LAG, et al: Annual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on U.S. cancer burden. Cancer 94:2766-2792,2002.

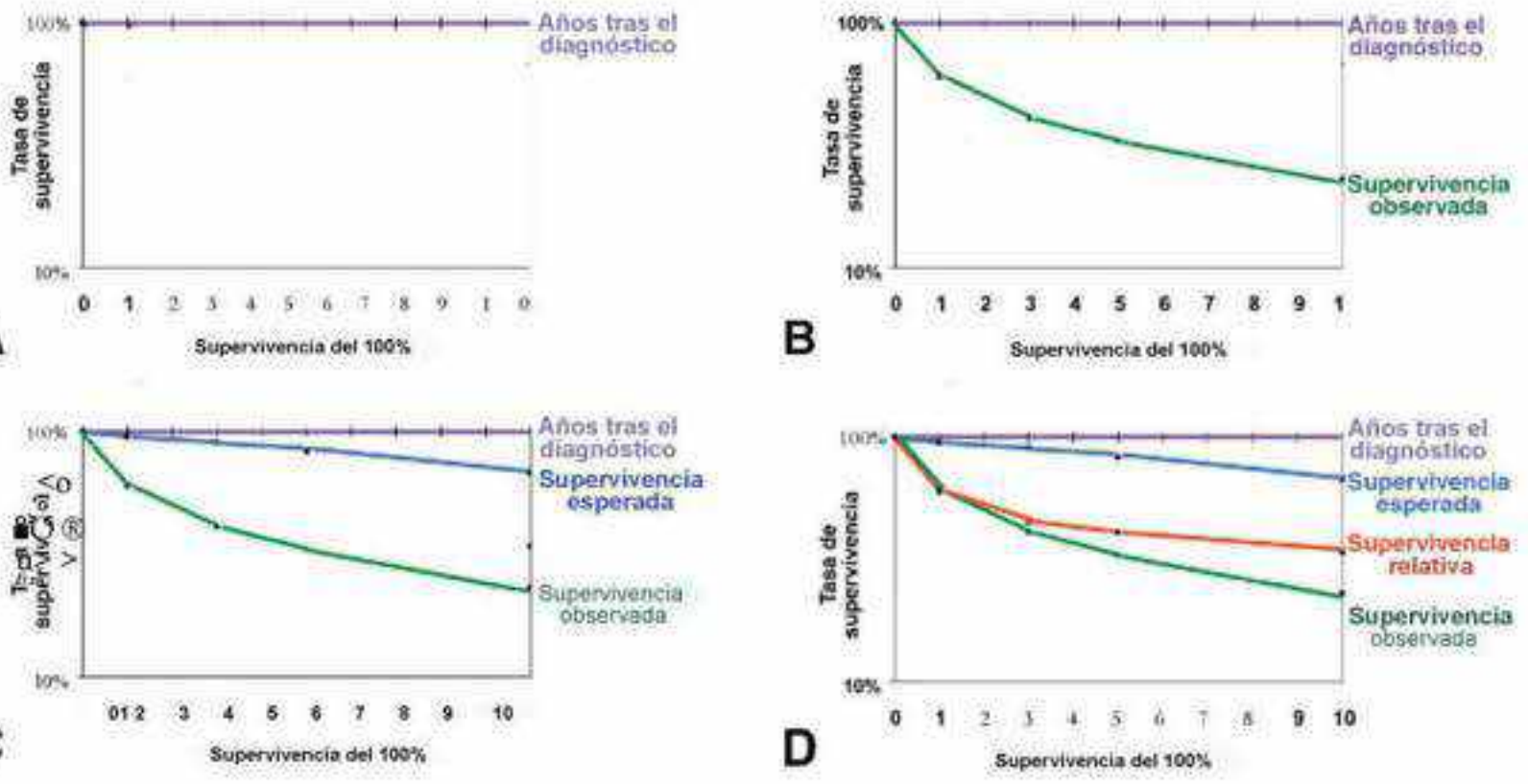


Figura 6-20. A-D, Supervivencia relativa. A, Supervivencia del 100% a lo largo de 10 años. B, Supervivencia observada. C, Supervivencia observada y esperada. D, Supervivencia observada, esperada y relativa.

(100%), y la supervivencia esperada es el denominador para estos cálculos, la supervivencia relativa será mayor que la supervivencia observada (fig. 6-20D).

GENERALIZACIÓN DE LOS DATOS DE SUPERVIVENCIA

Un último aspecto relacionado con la historia natural y el pronóstico de la enfermedad es la cuestión de qué pacientes son seleccionados para el estudio. Fijémonos en un ejemplo.

Las convulsiones febriles son frecuentes en los lactantes. Los niños por lo demás sanos a menudo sufren convulsiones asociadas con la fiebre elevada. La duda se plantea acerca de si estos niños deberían tratarse con un régimen de fenobarbital u otra medicación anticonvulsivante a largo plazo. Es decir, ¿las convulsiones febriles son un signo premonitorio de una epilepsia futura o se trata simplemente de un fenómeno asociado con la fiebre en los lactantes, en cuyo caso es poco probable que los niños sufran posteriormente convulsiones no febriles?

Para tomar una decisión lógica acerca del tratamiento, la pregunta que nos debemos plantear es: «¿Cuál es el riesgo de que un niño que ha presentado una convulsión febril sufra posteriormente convulsiones no febriles?». En la figura 6-21 se muestran los resultados de un análisis de Ellenberg y Nelson de los estudios publicados[5].

Cada punto indica el porcentaje de niños con convulsiones febriles que posteriormente desarrollaron convulsiones no febriles en un estudio diferente. Los autores dividieron los estudios en dos grupos:

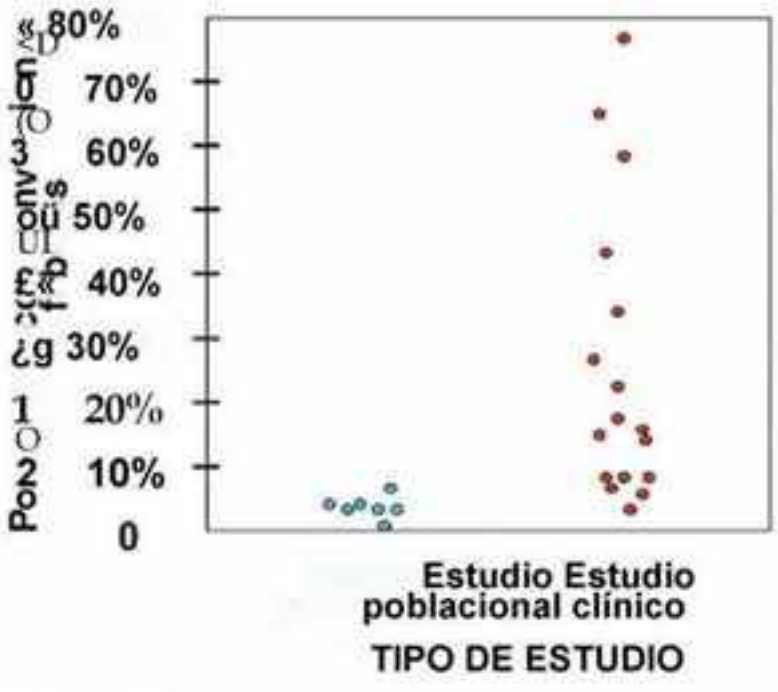


Figura 6-21. Porcentaje de niños que sufrieron convulsiones no febriles tras uno o más episodios de convulsiones febriles, por diseño de estudio. (Adaptado de Ellenberg JH, Nelson KB: Sample selection and the natural history of disease: Studies on febrile seizures. JAMA 243:1337-1340,1980.)

estudios poblacionales y estudios clínicos basados en clínicas pediátricas o de epilepsia. Los resultados de diferentes estudios clínicos muestran un riesgo considerable de sufrir posteriormente convulsiones no febriles. Sin embargo, los resultados de los estudios poblacionales muestran poca variación en el riesgo, y los resultados de todos estos estudios suelen agruparse alrededor de un nivel de riesgo bajo.

¿Por qué deberían diferenciarse los dos tipos de estudios? ¿Qué resultados creería usted? Es probable que cada una de las clínicas tuviera diferentes criterios de selección y diferentes patrones de remisión. Por tanto, los diferentes riesgos observados en los diferentes estudios basados en clínicas son probablemente resultado de la selección de poblaciones diferentes en cada una de las clínicas. Por el contrario, en los estudios poblacionales, este tipo de variación debida a la selección se ve reducida o eliminada, lo que explica el agrupamiento de los datos y el hallazgo resultante de que el riesgo de convulsiones no febriles es muy bajo. El punto importante es que puede resultar muy tentador analizar historiales de pacientes hospitalarios y generalizar los hallazgos para todos los pacientes en la población general. Sin embargo, éste no es un abordaje válido porque los pacientes que acuden a una cierta clínica u hospital a menudo no son representativos de todos los pacientes de la comunidad. Esto no significa que los estudios realizados en un solo hospital o en una sola clínica carezcan de valor. De hecho, hay mucho que aprender de los estudios realizados en un solo hospital. Sin embargo, estos estudios son especialmente tendentes a sesgos de selección, y esta posibilidad siempre debe tenerse en cuenta cuando se interpretan los hallazgos de dichos estudios y su potencial para generalizar sus resultados.

CONCLUSIÓN

Este capítulo ha expuesto cinco formas de expresar el pronóstico (tabla 6-14). El mejor abordaje depende del tipo de datos disponibles y de la finalidad del análisis de los datos. En los capítulos 7 y 8 nos ocuparemos de cómo utilizar los ensayos clínicos aleatorizados para seleccionar los mejores medios de intervención para prevenir y tratar las enfermedades humanas.

TABLA 6-14. Cinco formas de expresar el pronóstico

1. Tasa de letalidad
2. Supervivencia a 5 años
3. Supervivencia observada
4. Mediana de supervivencia
5. Supervivencia relativa

BIBLIOGRAFÍA

1. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53:457-481,1958.
2. Rosenhek R, Binder T, Porenta G, et al: Predictors of outcome in severe, asymptomatic aortic stenosis, *N Engl J Med* 343:611-617,2000.
3. Szklo M, Gordis L, Tonascia J, et al: The changing survivorship of white and black children with leukemia, *Cancer* 42:59-66,1978.
4. Feinstein AR, Sosin DM, Wells CK: The Will Rogers phenomenon: Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer, *N Engl J Med* 312:1604-1608,1985.
5. Ellenberg JH, Nelson KB: Sample selection and the natural history of disease: Studies on febrile seizures, *JAMA* 243:1337-1340, 1980.

PREGUNTAS DE REPASO DEL CAPÍTULO 6

La pregunta 1 se basa en la información proporcionada en la siguiente tabla:

Año de tratamiento	N.° de pacientes tratados	N.° DE PACIENTES VIVOS EN CADA ANIVERSARIO DEL INICIO DEL TRATAMIENTO		
		1.°	2.°	3.°
2007	75	60	56	48
2009	63	55	31	
2010	42	37		
Total	180	152	87	48

Ciento ochenta pacientes fueron tratados de la enfermedad X desde 2007 a 2009, y su evolución fue seguida hasta 2010. Los resultados del tratamiento se presentan en la tabla. Ningún paciente se perdió durante el seguimiento.

1. ¿Cuál es la probabilidad de sobrevivir 3 años?

   _____

2. Una suposición importante en este tipo de análisis es que:
   a. El tratamiento ha mejorado durante el período del estudio.
   b. La calidad del mantenimiento de los datos ha mejorado durante el período del estudio.
   c. No se han producido cambios en la eficacia del tratamiento durante el período del estudio.
   d. Cada año se incorporaron al estudio un número igual de varones y mujeres.
   e. Ninguna de las anteriores.

3. ¿Cuál de los siguientes es un buen índice de la gravedad de una enfermedad aguda de corta evolución?
   a. Tasa de mortalidad específica de causa.
   b. Supervivencia a 5 años.
   c. Tasa de letalidad.
   d. Razón de mortalidad estandarizada.
   e. Ninguno de los anteriores.

4. Se dispone de una prueba diagnóstica que detectará cierta enfermedad 1 año antes de lo que se detecta habitualmente. ¿Qué es lo *más probable* que le suceda a la enfermedad 10 años después de la aparición de la prueba? (Suponga que la detección precoz no ejerce ningún efecto sobre la historia natural de la enfermedad. Suponga también que no se han producido cambios en los certificados de defunción durante los 10 años.)
   a. La tasa de prevalencia de período disminuirá.
   b. La supervivencia aparente a 5 años aumentará.
   c. La tasa de mortalidad ajustada por edad disminuirá.
   d. La tasa de mortalidad ajustada por edad aumentará.
   e. La tasa de incidencia disminuirá.

5. ¿Cuál de las siguientes afirmaciones sobre la supervivencia relativa es verdadera?
   a. Se refiere a la supervivencia de los parientes de primer grado.
   b. Suele ser más parecida a la supervivencia observada en las poblaciones de edad avanzada.
   c. Suele ser más parecida a la supervivencia observada en las poblaciones jóvenes.
   d. Generalmente se diferencia de la supervivencia observada en una cantidad constante, independientemente de la edad.
   e. Ninguna de las anteriores.

Las preguntas 6 a 8 se basan en los datos de la tabla que se muestra abajo. Los datos se obtuvieron de un estudio de 248 pacientes con SIDA que recibieron un nuevo tratamiento y fueron seguidos para determinar la supervivencia. La población del estudio fue seguida durante 36 meses.

Nota: realice los cálculos en la tabla con cuatro decimales (es decir; 0,1234), pero para la respuesta final use tres decimales (p. ej., 0,123 o 12,3%).

6. En las personas que sobrevivieron el segundo año, ¿cuál es la probabilidad de morir en el tercer año?
   _____

7. ¿Cuál es la probabilidad de que una persona incorporada al estudio sobreviva hasta el final del tercer año? _____

8. Antes de comunicar los resultados de este análisis de supervivencia, los investigadores compararon las características basales de las 42 personas de las que se perdió el seguimiento antes de que acabara el estudio con las de los participantes que finalizaron el seguimiento. ¿Cuál fue el motivo de esta comparación?
   a. Comprobar si la aleatorización fue exitosa.
   b. Estudiar si se produjeron cambios en el pronóstico a lo largo del tiempo.
   c. Comprobar si los que continuaron en el estudio representan a la población total del estudio.
   d. Determinar si los resultados de los que continuaron en el estudio son los mismos que los de la población general.
   e. Comprobar si existen factores de confusión en los grupos expuestos y no expuestos.

Supervivencia de pacientes con SIDA tras el diagnóstico

(1) Intervalo desde el comienzo del tratamiento (meses)	(2) Vivos al comienzo del intervalo	(3) Muertos durante el intervalo	(4) Perdidos durante el intervalo	(5) Número efectivo expuestos al riesgo de morir durante el intervalo: Col (2) - % [Col (4)]	(6) Proporción que murió durante el intervalo: Col (3) / Col (5)	(7) Proporción que no murió durante el intervalo: 1 - Col (6)	(8) Proporción acumulada que sobrevivió desde la incorporación al final del intervalo: supervivencia acumulada
x	*L*	*dx*	w_x	I'_x	*?x	Px	*Px*
1-12	248	96	27				
13-24	125	55	13				
25-36	57	55	2				

Capítulo 8

Ensayos aleatorizados: algunos aspectos adicionales

Objetivos de aprendizaje

- Definir conceptos clave del diseño de estudios epidemiológicos en el contexto de los ensayos aleatorizados: tamaño de la muestra, error de tipo I, error de tipo II, potencia, generalización (validez externa) y validez interna.
- Calcular e interpretar la eficacia en un ensayo aleatorizado.
- Describir el diseño y los resultados de cinco ensayos aleatorizados importantes.
- Definir las cuatro fases principales de los ensayos aleatorizados utilizadas por la agencia estadounidense del medicamento (FDA) para evaluar nuevos fármacos en Estados Unidos.
- Introducir algunas consideraciones éticas relacionadas con los ensayos aleatorizados.
- Analizar el motivo del requerimiento del registro del inicio de un nuevo ensayo aleatorizado.

TAMAÑO DE LA MUESTRA

En una reunión científica celebrada hace algunos años, un investigador presentó los resultados de un estudio que había realizado para valorar un nuevo fármaco para ovejas. «Tras administrar el fármaco», comentó, «un tercio de las ovejas mejoró considerablemente, otro tercio no experimentó ningún cambio y un tercio se escapó.»

Esta historia introduce una de las preguntas planteadas con mayor frecuencia por los médicos que realizan ensayos de nuevos fármacos o, de hecho, por cualquiera que realice estudios evaluadores: ¿cuántos sujetos se deben estudiar? El momento de responder esta pregunta es *antes* de realizar el estudio. Con demasiada frecuencia se realizan estudios, se invierten grandes sumas de dinero y otros recursos, y sólo después de que el estudio se ha completado es cuando los investigadores descubren que desde el inicio contaban con muy pocos sujetos para obtener resultados significativos.

La cuestión de cuántos sujetos se necesitan para un estudio no se basa en la mística. Esta sección presenta la lógica sobre cómo abordar la cuestión del tamaño de la muestra. Comencemos este análisis sobre el tamaño de la muestra con la figura 8-1.

Tenemos dos vasijas con cuentas; cada una contiene 100 cuentas, unas azules y otras blancas. Las vasijas son opacas, de modo que (a pesar de su aspecto en la figura) no podemos ver los colores de las cuentas del interior de las mismas. Queremos saber si la distribución de las cuentas por color es diferente en la vasija A y en la vasija B. En otras palabras, ¿hay una proporción mayor (o menor) de cuentas azules en la vasija A que en la vasija B?

Para contestar a esta pregunta, tomemos una muestra de 10 cuentas de la vasija A en una mano y una muestra de 10 cuentas de la vasija B en la otra. En función de la distribución de color de las 10 cuentas en cada mano, intentaremos alcanzar una conclusión acerca de la distribución de color de las 100 cuentas en cada una de las vasijas.

Asumamos que (como se muestra en la fig. 8-2) en una mano tenemos 9 cuentas azules y 1 cuenta blanca de la vasija A y en la otra mano tenemos 2 cuentas azules y 8 cuentas blancas de la vasija B. ¿Podemos concluir que el 90% de las cuentas de la vasija A son azules y que el 10% son blancas? Claramente, no. Es posible, por ejemplo, que de las 100 cuentas de la vasija A, 90 sean blancas y 10 azules, pero, por *azar*, nuestra muestra de 10 cuentas consta de 9 azules y 1 blanca. Esto es posible, pero muy poco probable. De modo similar, con respecto a la vasija B, no podemos concluir que el 20% de las cuentas son azules y el 80% son blancas. Es concebible que 90 de las 100 cuentas sean azules y 10 sean blancas, pero, por *azar*, la muestra de 10 cuentas contiene 2 azules y 8 blancas. Esto es posible, pero, de nuevo, muy improbable.

Basándonos en las distribuciones de las muestras de 10 cuentas en cada mano, ¿podemos concluir que las distribuciones de las 100 cuentas en las dos vasijas son diferentes? Teniendo en cuenta las muestras en cada mano, ¿podría ocurrir, por ejemplo, que la distribución de cuentas en cada vasija fuese de 50 azules y 50 blancas?

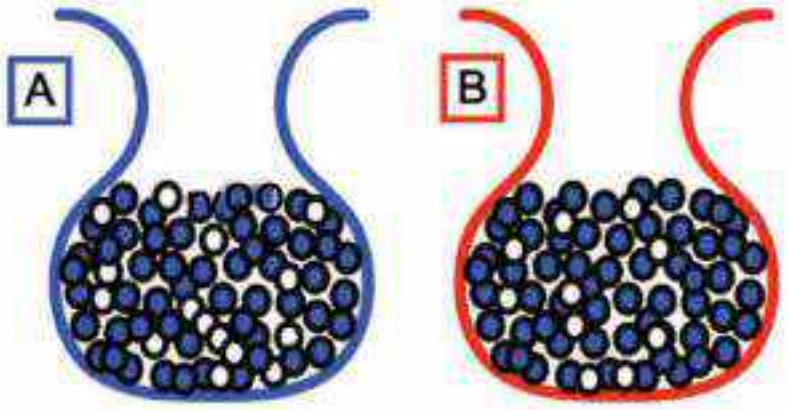


Figura 8-1. Dos vasijas opacas; cada una de ellas contiene 100 cuentas, unas azules y otras blancas.

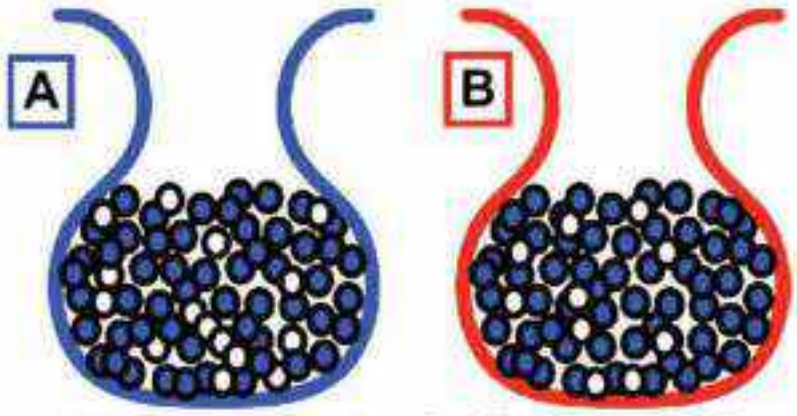


Figura 8-2. Muestras de 10 cuentas déla vasija Ay 10 cuentas de la vasija B.

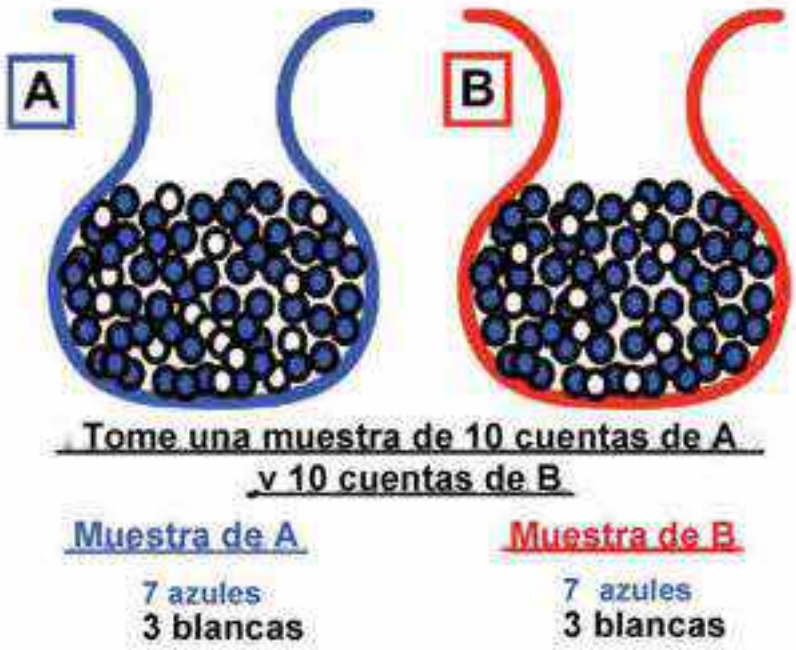


Figura 8-3. Muestras de 10 cuentas déla vasija Ay 10 cuentas de la vasija B.

De nuevo, es posible, pero no es probable. No podemos excluir esta posibilidad basándonos en nuestras muestras. Miramos a las muestras y tratamos de llegar a una conclusión respecto a todo un universo, las vasijas de las que hemos extraído las muestras.

Fijémonos ahora en el ejemplo que se muestra en la figura 8-3. De nuevo, obtenemos dos muestras. En esta ocasión, la muestra de 10 cuentas de la vasija A se compone de 7 cuentas azules y 3 blancas, y la muestra de 10 cuentas de la vasija B también se compone de 7 cuentas azules y 3 blancas. ¿Es posible que la distribución de color de las cuentas de las dos vasijas sea la misma? Claramente, podría ser. ¿Podríamos haber extraído estas dos muestras de 7 cuentas azules y 3 blancas de ambas vasijas si la distribución es en realidad de 90 cuentas blancas y 10 azules en la vasija A y 90 cuentas azules y 10 blancas en la vasija B? Sí, posiblemente, pero muy poco probable.

Cuando realizamos un estudio, sólo nos fijamos en la muestra de sujetos de nuestro estudio, como una muestra de pacientes con cierta enfermedad que están siendo tratados con el tratamiento A o con el tratamiento B. A partir de los resultados del estudio queremos extraer una conclusión de aplicación más allá de la población del estudio: ¿el tratamiento A es más eficaz que el tratamiento B en el universo total de todos los pacientes que tienen esta enfermedad que podrían tratarse con el tratamiento A o con el tratamiento B? El mismo problema que surgió con las muestras de 10 cuentas surge cuando queremos obtener una conclusión para todos los pacientes a partir de la muestra de los pacientes de nuestro estudio. Raramente, si es que ocurre alguna vez, un estudio se realiza en todos los pacientes que tienen una enfermedad o en todos los pacientes que podrían ser tratados con el fármaco en cuestión.

Con estos antecedentes, consideremos ahora un ensayo en el que se comparan grupos que reciben un tratamiento dentro de dos posibles, A y B. (Recuérdese el muestreo de cuentas que acabamos de describir.) Antes de comenzar nuestro estudio, podemos enumerar los cuatro posibles resultados del estudio (tabla 8-1):

1. Es posible que en realidad no existan diferencias en la eficacia entre el tratamiento A y el tratamiento B. En otras palabras, el tratamiento A no es ni mejor ni peor que el tratamiento B. Cuando realizamos nuestro estudio, concluimos correctamente en función de nuestras muestras que los dos grupos no se diferencian.

TABLA 8-1. Cuatro posibles conclusiones cuando se prueba si los tratamientos se diferencian o no

- **Cuando en realidad los tratamientos no difieren:**
  1. Podemos concluir correctamente que no se diferencian

     o

  2. Por error, podemos concluir que se diferencian
- **Cuando en realidad los tratamientos M difieren:**
  3. Por error, podemos concluir que no se diferencian

     o

  4. Podemos concluir correctamente que sí se diferencian

2. Es posible que en realidad no existan diferencias en la eficacia entre el tratamiento A y el tratamiento B, pero en nuestro estudio encontramos una diferencia entre los grupos y, por tanto, concluimos, basándonos en nuestras muestras, que existe una diferencia entre los tratamientos. Esta conclusión, basada en nuestras muestras, es errónea.
3. Es posible que en realidad existan diferencias en la eficacia entre el tratamiento A y el tratamiento B, pero, cuando examinamos los grupos en nuestro estudio, no encontramos diferencias entre ellos. Por tanto, basándonos en nuestras muestras, concluimos que no existen diferencias entre el tratamiento A y el tratamiento B. Esta conclusión es errónea.
4. Es posible que en realidad existan diferencias en la eficacia entre el tratamiento A y el tratamiento B, y, cuando examinamos los grupos de nuestro estudio, observamos que existen diferencias. Basándonos en estas muestras, concluimos correctamente que el tratamiento A difiere del tratamiento B.

Estas cuatro posibilidades forman el universo de resultados tras completar nuestro estudio. Fijémonos en estas cuatro posibilidades, que se presentan en una tabla de 2 X 2 (fig. 8-4): dos columnas representan la realidad (o el tratamiento A se diferencia del tratamiento B o el tratamiento A no se diferencia del tratamiento B). Las dos filas representan nuestra decisión: concluimos que se diferencian o que no se diferencian. En esta figura, las cuatro posibilidades que acabamos de enumerar se representan en cuatro celdillas en la tabla de 2 X 2. Si no existen diferencias y, basándonos en las muestras incluidas en nuestro estudio, concluimos que no existen diferencias, se trata de una decisión correcta *(celdilla a)*. Si existen diferencias y, basándonos en nuestro estudio, concluimos que existen diferencias *(celdilla d)*, también se trata de una decisión correcta. En el mejor de los casos, todas las posibilidades caerían en una de estas dos celdillas. Desafortunadamente, raramente se produce este hecho, si es que se produce alguna vez. Existen ocasiones en las que no hay diferencias entre los tratamientos, pero, basándonos en las muestras de los sujetos incluidos en nuestro estudio, concluimos erróneamente que sí son diferentes *(celdilla c)*. Esta posibilidad se denomina *error de tipo I*. También es posible que realmente existan diferencias entre los tratamientos, pero, basándonos en las muestras de nuestro estudio, concluimos erróneamente que no existe tal diferencia *(celdilla b)*; es el denominado *error de tipo II*. (En este caso, los tratamientos son diferentes, pero no hemos sido capaces de detectar la diferencia en las muestras de nuestro estudio.)

La *probabilidad* de cometer un error de tipo I se denomina a y la *probabilidad* de cometer un error de tipo II se denomina (3 (como se muestra en la fig. 8-5).

a es el denominado valor *P*, que vemos en muchas publicaciones y ha sido consolidado por muchos años de uso. Cuando leemos «*P* <0,05», se hace referencia a a. ¿Qué quiere decir que *P* <0,05? Nos indica que, basándonos en la muestra de sujetos incluidos en nuestro estudio, hemos concluido que el tratamiento A se diferencia del tratamiento B, porque hemos observado diferencias. La probabilidad de que dicha diferencia pudiera deberse al azar únicamente, y que dicha diferencia entre nuestros grupos no refleje una diferencia real entre el tratamiento A y el B, es de tan sólo 0,05 (o 1 de 20).

Prestemos atención ahora a la mitad derecha de la tabla de 2 X 2, que muestra las dos posibilidades cuando existe una diferencia real entre el tratamiento A y el B, como se muestra en la figura 8-6. Si, como vemos aquí, la realidad es que existen diferencias entre los tratamientos, sólo existen dos posibilidades. Podríamos concluir, erróneamente, que los tratamientos no se diferencian (error de tipo II). La probabilidad de cometer un error de tipo II viene designada por (3. O podríamos concluir, correctamente, que los tratamientos se diferencian. Como el total de todas las probabilidades

Figura 8-4. Posibles resultados de un ensayo aleatorizado: errores de tipo I y de tipo II.

Figura 8-5. Posibles resultados de un ensayo aleatorizado: *a* y (3.

Figura 8-6. Posibles resultados de un ensayo aleatorizado cuando los tratamientos difieren.

Figura 8-7. Posibles resultados de un ensayo aleatorizado: resumen.

debe ser igual a 1 y la probabilidad de un error de tipo II es = (3, la probabilidad de decidir correctamente basándonos en nuestro estudio que los tratamientos son diferentes, cuando existen diferencias, será igual a 1 − p. Esta probabilidad, 1 − (3, se denomina *potencia* del estudio. Nos dice cómo de bueno es nuestro estudio para identificar correctamente una diferencia entre los tratamientos cuando realmente son diferentes. ¿Cuál es la probabilidad de que nuestro estudio no pase por alto una diferencia si en realidad existe?

La tabla 2X2 completa de la figura 8-7 incluye todos los términos que hemos expuesto. En la tabla 8-2 se proporcionan múltiples definiciones para estos términos.

¿Cómo nos ayudan estos conceptos a estimar el tamaño de la muestra que necesitamos? Si nos planteamos la cuestión de cuántas personas tenemos que estudiar en un ensayo clínico, debemos ser capaces de especificar una serie de parámetros (se exponen en la tabla 8-3).

En primer lugar, debemos especificar la diferencia esperada en la tasa de respuesta. Supongamos que el tratamiento existente cura al 40% de los pacientes y vamos a probar un tratamiento nuevo. Debemos ser capaces de decir si esperamos que el tratamiento nuevo cure al 50%, al 60% o a otro porcentaje de los pacientes tratados. Es decir, ¿el nuevo tratamiento será un 10% mejor que el tratamiento habitual y curará al 50% de los pacientes, o un 20% mejor que el tratamiento habitual y curará a un 60%, o un porcentaje diferente? ¿Qué tamaño de diferencia entre el tratamiento habitual y el tratamiento nuevo queremos ser capaces de detectar con nuestro estudio?

¿Cómo llegamos normalmente a dicha cifra? ¿Qué pasa si no tenemos información sobre la que basar la estimación de la mejoría de la eficacia que podría anticiparse? Quizá estemos estudiando un nuevo tratamiento del que no existe experiencia previa. Un abordaje es buscar datos en poblaciones humanas sobre enfermedades y tratamientos similares. También podemos buscar datos relevantes en estudios en animales. En ocasiones, simplemente no podemos establecer estimaciones. En estos casos, podemos hacer una conjetura (p. ej., una

TABLA 8-2. Resumen de términos

Término	Definiciones
a	Probabilidad de cometer un error de tipo I Probabilidad de concluir que los tratamientos se diferencian cuando en realidad no difieren
P	Probabilidad de cometer un error de tipo II Probabilidad de concluir que los tratamientos no se diferencian cuando en realidad sí difieren
Potencia	1 - probabilidad de cometer un error de tipo II 1 - P Probabilidad de concluir correctamente que los tratamientos son diferentes Probabilidad de detectar una diferencia entre los tratamientos si los tratamientos en realidad son diferentes

TABLA 8-3. ¿Qué se debe especificar para estimar el tamaño de la muestra necesario en un ensayo aleatorizado?

1. La diferencia en las tasas de respuesta que se quiere detectar
2. Una estimación de la tasa de respuesta en uno de los grupos
3. El nivel de significación estadística (a)
4. El valor de la potencia deseada (1 − (3)
5. Si la prueba es unilateral o bilateral

mejoría del 30%) pero limitando la estimación: es decir, calcular el tamaño de la muestra necesario basándonos en una mejoría del 40% en la tasa de respuesta y calcular también el tamaño de la muestra necesario basándonos en una mejoría del 20% en la tasa de respuesta.

En segundo lugar, debemos contar con una estimación de la tasa de respuesta (tasa de curación, tasa de mejoría) en uno de los grupos. En el ejemplo que acabamos de exponer, dijimos que la tasa de curación actual (o la tasa de respuesta) es del 40%. Ésta es la estimación de la tasa de respuesta para el grupo que recibe el tratamiento habitual basándonos en la experiencia clínica actual.

En tercer lugar, debemos especificar el nivel de a con el que estaremos satisfechos. La elección depende del investigador; no existe nada sagrado en ningún valor específico, pero generalmente se utilizan valores de 0,05 o 0,01. En cuarto lugar, debemos especificar la potencia del estudio. De nuevo, no existe ningún valor sagrado, pero habitualmente se utilizan potencias del 80% o del 90%.

Por último, debemos especificar si la prueba va a ser unilateral o bilateral. ¿Qué significa esto? Nuestra tasa de curación actual es del 40% y vamos a estudiar un nuevo tratamiento que creemos que tendrá una tasa de curación más elevada, quizás del 50% o del 60%. Con el tratamiento nuevo queremos detectar una diferencia que sea en la dirección de la mejoría, un aumento de la tasa de curación. Por tanto, podríamos decir que sólo estudiaremos en busca de una diferencia en esa dirección, porque ésa es la dirección en la que estamos interesados; es decir, es una prueba unilateral.

El problema es que en la historia de la medicina y de la salud pública a veces nos hemos sorprendido y hemos encontrado que tratamientos nuevos que pensábamos que serían beneficiosos, realmente han sido dañinos. Si esta posibilidad es real, en nuestro estudio querríamos encontrar una diferencia en la tasa de curación en *cualquier dirección* respecto de la tasa actual, es decir, utilizaríamos una prueba bilateral, que estudiaría no sólo una diferencia que sea mejor que la tasa de curación actual, sino también una que sea peor que la tasa de curación actual. Los médicos clínicos y otros investigadores a menudo prefieren utilizar una prueba unilateral en sus estudios porque dichas pruebas requieren muestras de menor tamaño que las pruebas bilaterales. Como el número de pacientes disponibles para estudios a menudo es limitado, las pruebas unilaterales son atractivas. En ocasiones, los investigadores pueden tomar la decisión práctica de emplear una prueba unilateral incluso aunque no haya justificación conceptual para esta decisión.

Sobre este tema existen opiniones divergentes. Hay quien cree que si el investigador sólo está interesado en una dirección (mejoría) está justificado emplear una prueba unilateral. Otros creen que, siempre que la diferencia pueda ir en cualquiera de las direcciones, es necesario emplear una prueba bilateral. En una situación en la que una enfermedad concreta es mortal en el 100% de los casos, cualquier diferencia con un tratamiento nuevo sólo podría dirigirse en la dirección de la mejoría, por lo que sería apropiado utilizar una prueba unilateral.

Prestemos atención ahora a la aplicación de estos cinco factores para estimar el tamaño de muestra necesario a partir de una tabla de tamaño muestral. Las tablas 8-4 y 8-5 son selecciones de tablas de tamaños muéstrales publicadas por Gehan en 1979[1]. (En muchos libros de estadística estándar existen tablas similares.) Ambas tablas proporcionan el número de pacientes necesarios *en cada grupo* para detectar diversas diferencias en las tasas de curación con un a de 0,05 y una potencia (1 − (3) de 0,80. La tabla 8-4 está concebida para ser utilizada en una prueba bilateral y la tabla 8-5 para una prueba unilateral.

Supongamos que estamos realizando un ensayo clínico sobre dos tratamientos: uno que se utiliza habitualmente y uno nuevo. El tratamiento habitual tiene una tasa de curación del 40% y creemos que el tratamiento nuevo puede tener una tasa de curación del 60%, es decir, queremos detectar una mejoría en la tasa de curación del 20%. ¿Cuántos sujetos tenemos que estudiar? Supongamos que utilizaremos un a de 0,05, una potencia del 80% y una prueba bilateral. Por tanto, emplearemos la tabla 8-4. La primera columna de esta tabla indica la menor de las dos tasas de curación. Como la tasa de curación actual es del 40% y con nuestro nuevo tratamiento esperamos una tasa de curación del 60%, la menor de las dos tasas es el 40%, por lo que nos fijamos en esa fila de la tabla. Esperamos que el tratamiento nuevo tenga una tasa de curación del 60%, por lo que la diferencia entre las tasas de curación es del 20%. Nos desplazamos hacia abajo en la columna del 20% (la diferencia en las tasas de curación) hasta el punto de intersección con la fila del 40% (la menor de las tasas de curación) y encontramos el valor 97. Por tanto, *en cada uno de los grupos de nuestro estudio* necesitamos 97 sujetos.

TABLA 8-4. **Número de pacientes necesarios en cada grupo para detectar varias diferencias en las tasas de curación; a = 0,05; potencia (1 - p) = 0,80 (prueba bilateral)**

La menor de las dos tasas de curación	DIFERENCIAS EN LAS TASAS DE CURACIÓN ENTRE LOS DOS GRUPOS DE TRATAMIENTO													
	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70
0,05	420	130	69	44	36	31	23	20	17	14	13	11	10	8
0,10	680	195	96	59	41	35	29	23	19	17	13	12	11	8
0,15	910	250	120	71	48	39	31	25	20	17	15	12	11	9
0,20	1.090	290	135	80	53	42	33	26	22	18	16	12	11	9
0,25	1.250	330	150	88	57	44	35	28	22	18	16	12	11	-
0,30	1.380	360	160	93	60	44	36	29	22	18	15	12	-	-
0,35	1.470	370	170	96	61	44	36	28	22	17	13	-	-	-
0,40	1.530	390	175	97	61	44	35	26	20	17	-	-	-	-
0,45	1.560	390	175	96	60	42	33	25	19	—	—	—	—	—
0,50	1.560	390	170	93	57	40	31	23	—	—	—	—	—	—

Adaptada de Gehan E. Clinical trials in cancer research. Environ Health Perspect 32:31,1979.

Otro método es utilizar la tabla en una dirección inversa. Por ejemplo, consideremos una clínica para pacientes que sufren una cierta enfermedad rara. Cada año la clínica trata a 30 pacientes con la enfermedad y quiere probar un tratamiento nuevo. Dado que el número máximo de pacientes es 30, podríamos preguntarnos: «¿Qué diferencia de tamaño podríamos esperar detectar en las tasas de curación?». Podemos encontrar una diferencia de un cierto tamaño que puede ser aceptable o podemos encontrar que el número de sujetos disponibles para el estudio es simplemente demasiado pequeño. Si el número de pacientes es demasiado pequeño, tenemos varias opciones: podemos decidir no realizar el estudio, y dicha decisión debería adoptarse pronto, antes de invertir un gran esfuerzo; o podríamos decidir prolongar el estudio en el tiempo para acumular más sujetos. Por último, podríamos decidir colaborar con investigadores de otros centros para aumentar el número total de sujetos disponibles para el estudio. En un estudio que se realiza en un solo sitio, puede ser difícil identificar sesgos en la selección de participantes, pero en un estudio multicéntrico, la presencia de algún sesgo en uno de los centros sería detectable más fácilmente.

Esta sección ha demostrado el uso de una tabla de tamaño muestral. También existen disponibles fórmulas y programas informáticos para calcular el tamaño de las muestras. Los tamaños muéstrales pueden calcularse no sólo para ensayos aleatorizados, sino también para estudios de cohortes o de casos-controles (se exponen en los caps. 9 y 10).

TABLA 8-5. **Número de pacientes necesarios en cada grupo para detectar varias diferencias en las tasas de curación; a = 0,05; potencia (1 — (3) = 0,80 (prueba unilateral)**

La menor de las dos tasas de curación	DIFERENCIAS EN LAS TASAS DE CURACIÓN ENTRE LOS DOS GRUPOS DE TRATAMIENTO													
	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70
0,05	330	105	55	40	33	24	20	17	13	12	10	9	9	8
0,10	540	155	76	47	37	30	23	19	16	13	11	11	9	8
0,15	710	200	94	56	43	32	26	22	17	15	11	10	9	8
0,20	860	230	110	63	42	36	27	23	17	15	12	10	9	8
0,25	980	260	120	69	45	37	31	23	17	15	12	10	9	-
0,30	1.080	280	130	73	47	37	31	23	17	15	11	10	-	-
0,35	1.160	300	135	75	48	37	31	23	17	15	11	-	-	-
0,40	1.210	310	135	76	48	37	30	23	17	13	-	-	-	-
0,45	1.230	310	135	75	47	36	26	22	16	—	—	—	—	—
0,50	1.230	310	135	73	45	36	26	19	—	—	—	—	—	—

Adaptada de Gehan E. Clinical trials in cancer research. Environ Health Perspect 32:31,1979.

RECLUTAMIENTO Y RETENCIÓN DE PARTICIPANTES DEL ESTUDIO

Un desafío importante en la realización de los ensayos aleatorizados es el reclutamiento de un número suficiente de voluntarios elegibles y dispuestos. El fallo en el reclutamiento de un número suficiente de voluntarios puede dejar a un ensayo bien diseñado sin un número suficiente de participantes para lograr resultados estadísticamente válidos. Los participantes potenciales también deben estar dispuestos a ser aleatorizados para el ensayo. Los ensayos pueden retrasarse mucho por este problema del reclutamiento limitado y los costes para completar dichos ensayos pueden aumentar. Sin embargo, debido a las presiones para reclutar a un número suficiente de participantes, se necesita un alto nivel de vigilancia para asegurase de que los investigadores del estudio no han empleado la coacción, manifiesta o encubierta, consciente o inconscientemente, para convencer a posibles participantes para que se incorporen a un estudio. Dentro de los límites de un ensayo aleatorizado, los participantes deben ser completamente informados de los riesgos y los acuerdos adoptados con fines de compensación si se produjesen efectos adversos. También se deben adoptar los acuerdos adecuados para retribuir los gastos de los participantes, como transporte, alojamiento si es necesario, y el tiempo de los mismos, en especial si la participación se asocia con pérdida de salario. Sin embargo, el pago de incentivos en efectivo a participantes potenciales supone riesgo de coacción manifiesta o sutil y puede dar lugar a sesgos y distorsión de los resultados del estudio, en especial si los incentivos pagados son cuantiosos.

En ocasiones, la incorporación como participante en un estudio ha sido publicitada a voluntarios potenciales con el argumento de que sólo a través de la participación el participante tendrá la oportunidad de ser tratado con los tratamientos disponibles más novedosos. Sin embargo, la justificación para llevar a cabo un ensayo aleatorizado es que no sabemos qué tratamiento es mejor. Por tanto, resulta fundamental que las personas que realizan el ensayo eviten ser muy entusiastas prometiendo a los participantes beneficios que aún no han sido demostrados de modo concluyente que estén asociados con el tratamiento que se está estudiando.

Un problema relacionado es el de retener a los voluntarios durante toda la duración del estudio. Las pérdidas de seguimiento y otras formas de falta de cumplimiento pueden convertir este aspecto en un problema importante. Los participantes pueden perder interés en el estudio con el paso del tiempo o considerar la participación demasiado inadecuada, especialmente a largo plazo. Los investigadores deben valorar por qué los participantes a menudo abandonan los estudios y adoptar las medidas adecuadas para evitar pérdidas de seguimiento.

FORMAS DE EXPRESAR LOS RESULTADOS DE LOS ENSAYOS ALEATORIZADOS

Los resultados de los ensayos aleatorizados pueden expresarse de diversas formas. Pueden calcularse los riesgos de morir o de desarrollar una enfermedad o una complicación en cada grupo, y posteriormente puede calcularse la *reducción del riesgo* (eficacia). La *eficacia* del agente que se está estudiando, como una vacuna, puede expresarse mediante las tasas de desarrollar la enfermedad en el grupo vacunado y en el grupo al que se administra placebo:

$$\text{Eficacia} = \frac{\left(\begin{array}{c}\text{Tasa en los que}\\ \text{recibieron el placebo}\end{array}\right) - \left(\begin{array}{c}\text{Tasa en los que}\\ \text{recibieron la vacuna}\end{array}\right)}{\text{Tasa en los que recibieron el placebo}}$$

Esta fórmula nos informa de la cuantía de la disminución de la enfermedad gracias al uso de la vacuna. Los riesgos a menudo se calculan por *personas-años* de observación.

La eficacia, o cómo de bien funciona un tratamiento bajo condiciones «ideales», puede diferenciarse de la efectividad, o cómo de bien funciona el tratamiento en situaciones «reales». Aunque los ensayos aleatorizados evalúan con mayor frecuencia la eficacia del tratamiento, los dos términos (eficacia y efectividad) a menudo se emplean indistintamente. La eficacia y la efectividad se analizan con mayor detalle en el capítulo 17.

Otra forma de comunicar los resultados de los ensayos aleatorizados es calculando la *razón de los riesgos* entre los dos grupos de tratamiento (el riesgo relativo), que se analizará en el capítulo 11. Además, con frecuencia determinamos las *curvas de supervivencia* en cada grupo y las comparamos (v. cap. 6) para ver si existen diferencias.

Un objetivo importante de los ensayos aleatorizados es producir un efecto en la forma de ejercer la medicina clínica y en el ámbito de la salud pública. No obstante, en ocasiones los médicos pueden tener dificultades para situar los hallazgos de dichos ensayos en una perspectiva que sea relevante para su práctica. Por tanto, otro método para expresar los resultados de los ensayos aleatorizados es estimar el *número de pacientes que sería necesario tratar* (NNT) para prevenir un resultado adverso, como una muerte. Esto puede calcularse del siguiente modo:

$$\text{NNT} := \frac{1}{\left(\begin{array}{c}\text{Tasa en el}\\ \text{grupo no tratado}\end{array}\right) - \left(\begin{array}{c}\text{Tasa en el}\\ \text{grupo tratado}\end{array}\right)}$$

Así, por ejemplo, si la tasa de mortalidad en el grupo no tratado es del 17% y la tasa de mortalidad en el grupo tratado es del 12%, necesitaríamos tratar:

$$\frac{1}{17\%-12\%} = \frac{1}{0{,}05} = 20$$

personas para evitar una muerte. Las estimaciones del NNT suelen redondearse hacia arriba hasta el siguiente número entero. Este método puede emplearse en estudios de varias intervenciones, tanto para tratamientos como para medidas preventivas. El mismo abordaje también puede utilizarse para valorar el riesgo de efectos secundarios calculando el *número necesario para dañar* (NND) para causar daño en una persona más. Estas estimaciones están sujetas a un error considerable y suelen presentarse con un intervalo de confianza del 95% para poder interpretarlas correctamente. Además, tienen otras limitaciones: no tienen en cuenta la calidad de vida y tienen un valor limitado para los pacientes. Estas estimaciones, no obstante, pueden ayudar a los médicos a estimar la cuantía del efecto que podrían esperar observar utilizando el nuevo tratamiento o la nueva medida preventiva en sus prácticas.

INTERPRETACIÓN DE LOS RESULTADOS DE LOS ENSAYOS ALEATORIZADOS

Generalización de los resultados más allá de la población del estudio

Cuando realizamos un ensayo, el objetivo último es generalizar los resultados más allá de la población del estudio. Consideremos un ejemplo. Supongamos que queremos evaluar un nuevo fármaco para el lupus eritematoso sistémico (una enfermedad del tejido conjuntivo) utilizando un ensayo aleatorizado. Los diagramas de la figura 8-8 representan un ensayo aleatorizado en el que una población definida es identificada en el total de la población, y un subgrupo de esa población definida será la población del estudio. Por ejemplo, la *población total* podrían ser todos los pacientes con lupus eritematoso, la *población definida* podrían ser todos los pacientes con lupus eritematoso en nuestra comunidad

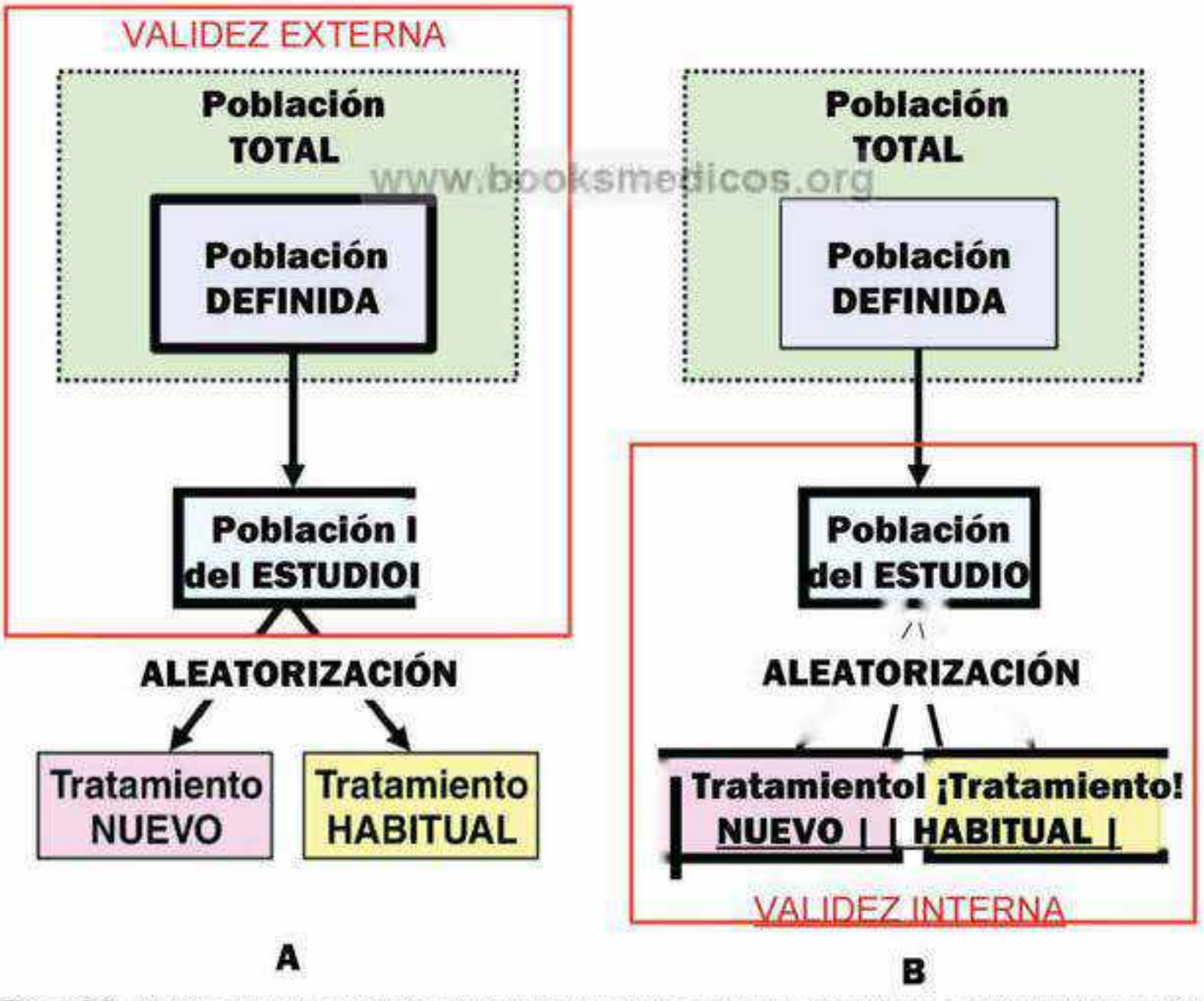


Figura 8-8. A, Validez externa (generalización) en un ensayo aleatorizado. Los hallazgos del estudio son generalizables de la población del estudio a la población definida y, presumiblemente, al total de la población. B, Validez interna en un ensayo aleatorizado. El estudio se realizó correctamente y los hallazgos del estudio son, por tanto, válidos en la población del estudio.

y la *población del estudio* podrían ser los pacientes con la enfermedad que reciben asistencia médica en alguna de las distintas clínicas de nuestra comunidad.

Si realizamos un estudio en los pacientes reclutados de varias clínicas de nuestra comunidad y observamos que un tratamiento nuevo es mejor que el tratamiento empleado habitualmente, querríamos poder afirmar que el tratamiento nuevo es mejor para la enfermedad con independencia de dónde recibe el tratamiento el paciente, y no únicamente para los pacientes de esas clínicas. Nuestra capacidad para aplicar los resultados obtenidos en nuestra población de estudio a una población más general se denomina *generalización* o *validez externa* del estudio. Queremos ser capaces de *generalizar* a partir de los hallazgos del estudio a todos los pacientes con la enfermedad en nuestra comunidad. Para ello, debemos conocer hasta qué grado los pacientes que hemos estudiado son representativos de la población definida, es decir, de todos los pacientes con la enfermedad en cuestión en nuestra comunidad (v. fig. 8-8A). Debemos caracterizar a los que no participaron en el estudio e identificar características de los pacientes del estudio que pudieran ser diferentes de las de los pacientes que no participaron en el estudio. Dichas diferencias pueden descartar que podamos generalizar los resultados del estudio a otros pacientes que no fueron incluidos en el estudio. También podemos querer generalizar nuestros resultados, no sólo a todos los pacientes con la enfermedad en nuestra comunidad, sino a todos los pacientes con la enfermedad, con independencia de dónde vivan, es decir, a la totalidad de pacientes con la enfermedad. Sin embargo, en raras ocasiones se tiene en cuenta a la población total en un ensayo aleatorizado. Aunque se espera que la población definida sea representativa de la población total, esta suposición raramente se verifica, si es que alguna vez se hace.

La validez externa debe diferenciarse de la *validez interna* (v. fig. 8-8B). Un ensayo aleatorizado tiene validez interna si la aleatorización se ha realizado correctamente y el estudio no sufre otros sesgos ni ninguno de los principales problemas metodológicos que hemos analizado. Los ensayos aleatorizados pueden considerarse el método de diseño de estudios de referencia, porque la aleatorización, si se realiza correctamente, evita que cualquier sesgo por parte de los investigadores del estudio pueda influir en la asignación del tratamiento para cada paciente. Si el estudio es lo suficientemente amplio, la aleatorización probablemente logrará la comparabilidad entre los grupos de tratamiento en cuanto a factores que pueden ser importantes para el resultado, como la edad, el sexo, la raza, etc., así como para factores que no hemos medido o de cuya importancia no somos conscientes. Los aspectos de la validez interna y de la validez externa (generalización) son puntos básicos a la hora de realizar cualquier ensayo aleatorizado y en otros tipos de diseños de estudios, que se analizarán en siguientes capítulos.

¿Qué información pueden proporcionar a un médico en ejercicio los resultados de un ensayo aleatorizado acerca de un paciente concreto?

Consideremos un escenario hipotético sencillo. Un médico está a punto de recetar un tratamiento a uno de sus pacientes. Conoce un ensayo aleatorizado de gran calidad publicado recientemente que comparaba el tratamiento A y el tratamiento B para la patología que presenta su paciente (fig. 8-9A). Como se observa en el diagrama, en el ensayo la proporción de pacientes que lograron un buen resultado (partes azules de las barras) tras recibir el tratamiento A fue mucho mayor que la proporción de pacientes que lograron un buen resultado tras recibir el tratamiento B. Los resultados del ensayo se comunicaron, por tanto, mostrando que el tratamiento A es superior al tratamiento B para esta enfermedad.

El médico conoce bien los resultados publicados del ensayo aleatorizado. Sin embargo, antes de recetar el tratamiento para su paciente basándose en los resultados del ensayo, tiene algunas preguntas cuya respuesta le podría proporcionar una guía valiosa para elegir el mejor tratamiento *para este paciente*. A continuación se exponen tres de sus preguntas a modo de ejemplo:

1. «Si mi paciente hubiese participado en el ensayo aleatorizado y hubiese sido asignado al grupo aleatorizado a recibir el tratamiento A (fig. 8-9B), ¿habría sido uno de los que mejoraron (se muestran en azul) o habría sido uno de los que no respondieron al tratamiento A (la parte blanca superior de la barra)?».
2. «Si mi paciente hubiese participado en el ensayo aleatorizado y hubiese sido asignado al grupo que recibió el tratamiento A (fig. 8-9C), ¿habría sido uno de los que sufrieron efectos secundarios (se muestran en rojo) o habría sido uno de los que no presentaron efectos secundarios con el tratamiento A (la parte blanca superior de la barra)?».
3. «Si mi paciente hubiese participado en el ensayo aleatorizado y hubiese sido asignado al grupo que recibió el tratamiento B (fig. 8-9D), ¿se habría encontrado en el grupo que mejoró tras recibir el tratamiento B (se muestra en azul) o se habría encontrado entre los que no respondieron al tratamiento B (la parte blanca superior de la barra)?»

Desafortunadamente, la mayoría de los ensayos aleatorizados no proporcionan la información que el médico necesitaría para caracterizar a un paciente concreto lo suficiente como para predecir qué respuesta podría tener ese paciente a los tratamientos disponibles.

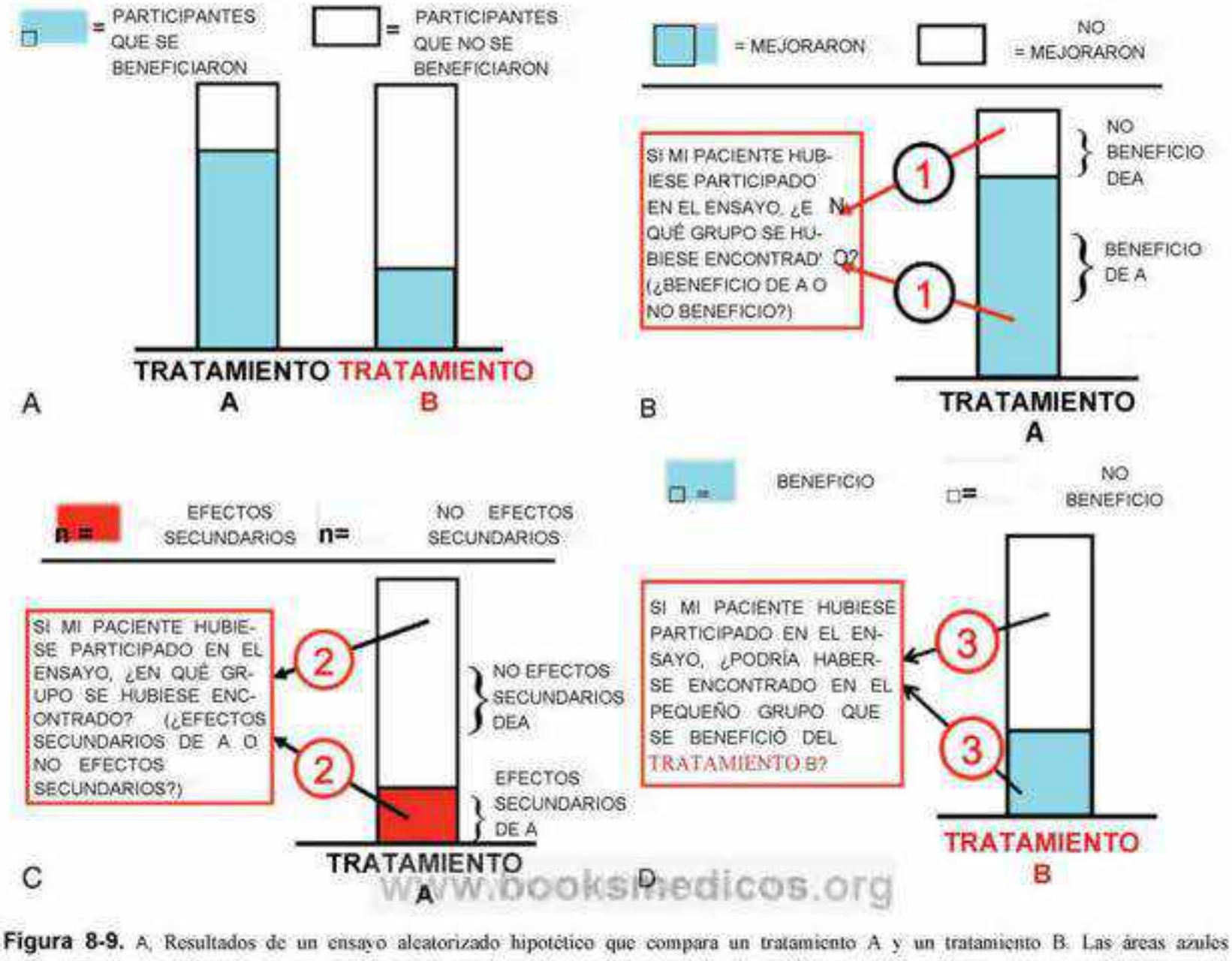


Figura 8-9. A, Resultados de un ensayo aleatorizado hipotético que compara un tratamiento A y un tratamiento B. Las áreas azules indican el número de pacientes que se beneficiaron de cada tratamiento y las áreas blancas indican los que no respondieron a cada tratamiento. B, Primera pregunta del médico. C, Segunda pregunta del médico. D, Tercera pregunta del médico.

El médico por lo general no posee la suficiente información que le ayude a decidir si sería razonable generalizar a partir de los resultados del ensayo aleatorizado a un paciente específico antes de elegir e iniciar el tratamiento. Si generaliza a su paciente, ¿a partir de qué subgrupo de participantes en el ensayo debería generalizar?

Otro factor limitante en muchos ensayos aleatorizados es que, aunque asumamos que los abandonos del ensayo fueron mínimos y que todos los participantes aceptaron ser aleatorizados, quedan preguntas por contestar: ¿podemos asumir que en el mundo «real» no aleatorizado un paciente determinado respondería del mismo modo que un paciente aleatorizado podría responder en un ensayo? ¿Qué sabemos de la personalidad y las preferencias de los participantes en los ensayos aleatorizados que nos indicarían si un paciente específico que debe ser tratado posee características similares, como los mismos valores, personalidad y preocupaciones? ¿Una persona que acepta ser aleatorizada es parecida a la población general de la que un paciente específico puede proceder para recibir tratamiento? Como destacó David Mant, los participantes de los ensayos aleatorizados generalmente no son representativos de la población general[2]. Los participantes de los ensayos son por lo general más sanos, más jóvenes y están mejor informados que los pacientes que acuden a ser tratados. Una última cuestión que se debe abordar es si hemos perdido nuestra preocupación sobre los individuos cuando reducimos a todo el mundo en el estudio a ser parte de un grupo de estudio y a menudo sólo examinamos los resultados para el grupo como un todo, perdiendo de vista las diferencias y preferencias individuales.

Investigación comparativa de eficacia (ICE)

Algunos ensayos aleatorizados están diseñados para comparar un tratamiento nuevo con un placebo. Otros ensayos aleatorizados se ocupan de la comparación de un tratamiento nuevo con un tratamiento más antiguo aceptado con el fin de determinar si el nuevo tratamiento es superior al tratamiento establecido. Más adelante en este capítulo (págs. 169-172) estudiaremos dos ejemplos de ensayos utilizados para evaluar intervenciones ampliamente aceptadas. En los últimos años también ha surgido interés en lo que se ha denominado investigación comparativa de eficacia (ICE), en la que dos o más intervenciones existentes son comparadas

con el fin de «determinar qué intervención sería más útil en una población dada o en un paciente determinado». En este tipo de abordaje, los resultados de otros tipos de diseños de estudios, que se analizan en capítulos posteriores, pueden utilizarse conjuntamente con los hallazgos de ensayos aleatorizados para intentar responder a estas preguntas.

Otro aspecto es el coste de las intervenciones. Por ejemplo, muchos tratamientos de las infecciones por VIH son muy caros y dichos tratamientos pueden ser asequibles en países desarrollados, pero puede que no lo sean en muchos países en vías de desarrollo. A medida que aparecen medicaciones más novedosas y más baratas, a menudo se realizan estudios para determinar si las alternativas más nuevas y baratas son igual de efectivas que las intervenciones más caras, cuya eficacia ya ha sido documentada. Estos estudios a menudo se denominan *estudios de equivalencia* y están diseñados para determinar si las intervenciones más baratas son igual de eficaces que los tratamientos más caros. Para estos estudios también se utiliza el término de *estudios de no inferioridad*. Estos estudios deben distinguirse de los *estudios de superioridad*, en los que fármacos de nueva aparición son evaluados para determinar si son más eficaces (superiores) que intervenciones disponibles en la actualidad.

LAS CUATRO FASES PARA PROBAR NUEVOS FÁRMACOS EN ESTADOS UNIDOS

A medida que aparecen nuevos fármacos, la Food and Drug Administration estadounidense sigue una secuencia estándar de cuatro fases para probar y evaluar estos nuevos agentes:

Ensayos de fase I. Estos ensayos son estudios farmacológicos clínicos: estudios pequeños de 20-80 pacientes que se ocupan de aspectos de seguridad del nuevo fármaco o de otros tratamientos. Examinan efectos tóxicos farmacológicos, como la seguridad, los márgenes de seguridad de las dosis en el ser humano y los efectos secundarios observados con el nuevo tratamiento. Si el fármaco pasa estas pruebas, a continuación se realizan estudios de *fase II*.

Ensayos de fase II. Los estudios de fase II consisten en investigaciones clínicas de 100-300 pacientes con el fin de evaluar la eficacia del nuevo fármaco o tratamiento y estudiar aún más su seguridad relativa. Si el fármaco pasa los estudios de fase II, a continuación pasa a ensayos de *fase III*.

Ensayos de fase III. Estos estudios son ensayos controlados aleatorizados a gran escala diseñados para valorar la eficacia y la seguridad relativa. Estos estudios a menudo se realizan con 1.000-3.000 o más participantes. El reclutamiento de esta gran cantidad de participantes puede ser muy difícil y a menudo necesita la participación de más de un centro de estudio. Cuando desde el comienzo se anticipan dificultades en el reclutamiento, el estudio puede diseñarse en la fase de planificación como un ensayo multicéntrico. Si el fármaco pasa la fase III puede ser aprobado y recibir la licencia para su comercialización.

Estudios de fase IV. Cada vez es un hecho más reconocido que ciertos efectos adversos de los medicamentos, como la carcinogénesis (cáncer) y la teratogénesis (malformaciones congénitas), pueden no manifestarse durante muchos años. También es posible que dichos efectos adversos de los nuevos fármacos puedan ser tan infrecuentes que puede que no se detecten incluso en ensayos clínicos aleatorizados relativamente extensos, o puedan volverse evidentes únicamente cuando el fármaco es usado por una gran cantidad de pacientes, una vez comercializado. Por este motivo, los estudios de *fase IV*, que también se conocen como de *vigilancia tras la comercialización*, son importantes para controlar nuevos fármacos que ya son utilizados por la población. Los estudios de fase IV no son estudios aleatorizados y no son ensayos, a diferencia de los ensayos de fase I, II y III. Como los estudios de fase IV estudian los efectos secundarios de tratamientos nuevos una vez que el tratamiento ya se comercializa, carecen de aleatorización. Con el fin de que los hallazgos de los estudios de vigilancia tras la comercialización sean válidos, resulta fundamental contar con un sistema de comunicación de efectos adversos de gran calidad. Aunque el objetivo de los estudios de fase IV a menudo es el número de efectos secundarios comunicados y el número de pacientes que recibieron el nuevo tratamiento y sufrieron efectos secundarios, los estudios de fase IV a menudo son muy valiosos para aportar pruebas adicionales sobre los beneficios y ayudan a optimizar el uso del nuevo agente.

La secuencia rigurosa que acabamos de describir ha protegido a la población estadounidense de muchos tratamientos peligrosos. En los últimos años, sin embargo, la presión para acelerar el procesamiento de nuevos fármacos para tratar la infección por VIH y el SIDA ha dado lugar a un replanteamiento de este proceso de aprobación. Parece probable que las modificaciones que terminen haciéndose en el proceso de aprobación no se limitarán a los fármacos utilizados para tratar el SIDA, sino que de hecho tendrán ramificaciones extensas en el proceso general de aprobación. Los cambios que se realicen en el futuro tendrán, por tanto, implicaciones importantes para la salud de los pacientes de Estados Unidos y de todo el mundo.

TRES ENSAYOS ALEATORIZADOS IMPORTANTES EN ESTADOS UNIDOS

Hypertension Detection and Follow-up Program

Hace muchos años, un estudio de la Veterans Administration demostró que el tratamiento de personas que sufren elevaciones importantes de la presión arterial puede reducir significativamente su mortalidad[4]. La cuestión de si el tratamiento antihipertensivo beneficia a personas con elevaciones leves de la presión arterial (presión arterial diastólica de 90-104 mmHg) no fue resuelta. Aunque podríamos ser capaces de reducir la presión arterial en dichas personas, debemos tener en cuenta el problema de los efectos secundarios de los fármacos antihipertensivos. A menos que pueda demostrarse algún beneficio para la salud de los pacientes, el uso de estos fármacos antihipertensivos no estaría justificado en las personas cuya presión arterial se encuentra mínimamente elevada.

El estudio multicéntrico Hypertension Detection and Follow-up Program (HDFP) fue diseñado para investigar los beneficios del tratamiento de la hipertensión leve a moderada. En este estudio, de 22.994 sujetos que fueron elegibles porque sufrían elevación de la presión arterial diastólica, 10.940 fueron aleatorizados a tratamiento escalonado o al grupo de tratamiento referido (fig. 8-10).

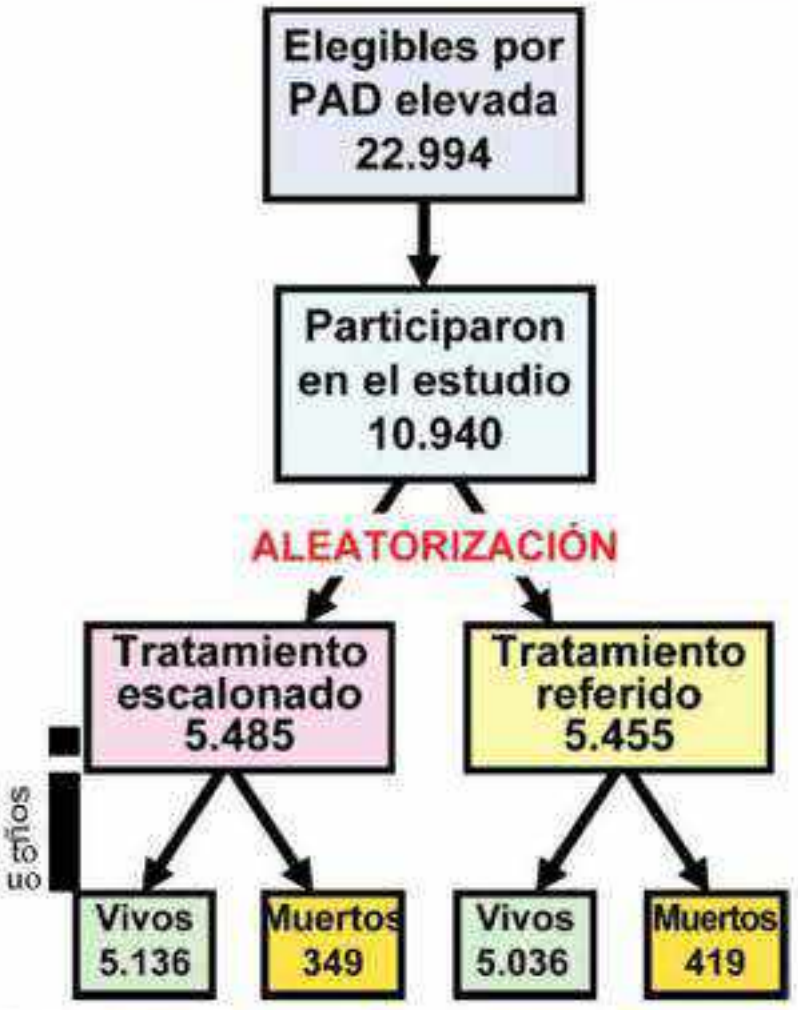


Figura 8-10. Diseño del Hypertension Detection and Follow-up Program (HDFP). PAD, presión arterial diastólica.

El *tratamiento escalonado* hacía referencia al tratamiento siguiendo un protocolo definido con precisión en el que el tratamiento se cambiaba cuando no se había obtenido una disminución especificada de la presión arterial durante un cierto período. El grupo de comparación suponía un problema: desde el punto de vista del diseño del estudio, hubiese sido deseable un grupo que no recibiese tratamiento para la hipertensión. Sin embargo, los investigadores creyeron que no sería éticamente justificable dejar sin tratamiento antihipertensivo a pacientes hipertensos. De modo que los pacientes del grupo de comparación fueron remitidos de vuelta a sus médicos, y este grupo se denominó *grupo de tratamiento referido*. A continuación se estudió la mortalidad en ambos grupos a lo largo de un período de 5 años[5].

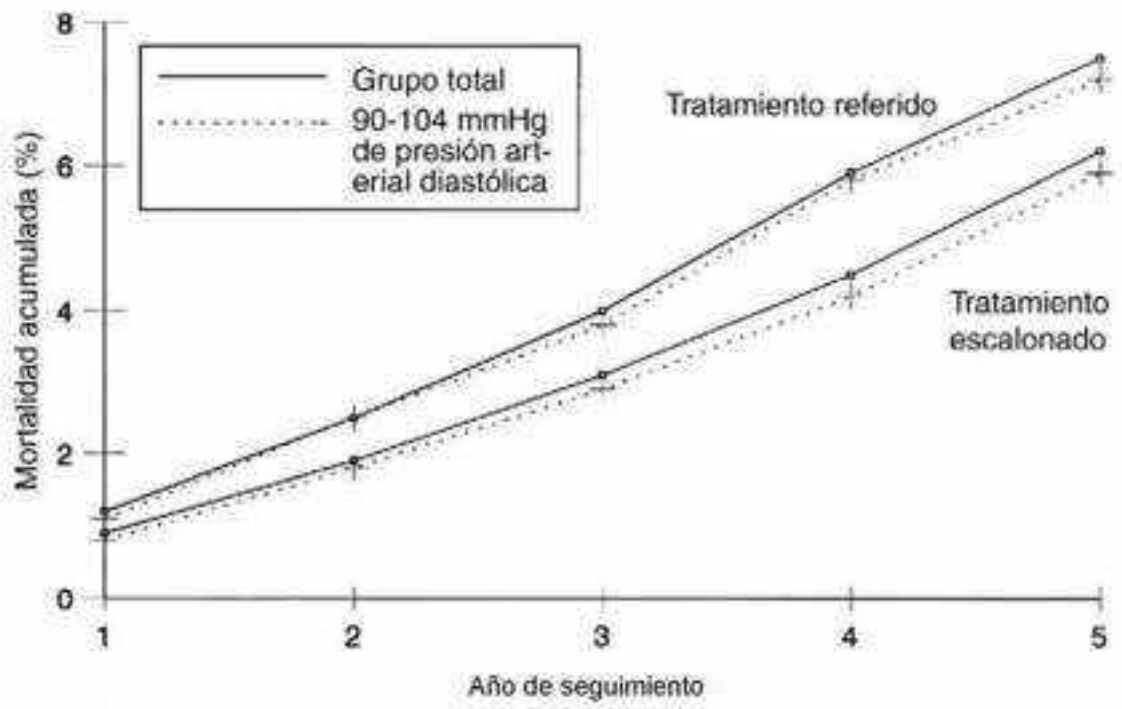


Figura 8-11. Mortalidad acumulada por todas las causas en función del nivel de presión arterial y el tipo de tratamiento recibido en el HDFP. (Adaptado de Hypertension Detection and Follow-up Program Cooperative Group: Five-year findings of the Hypertension Detection and Follow-up Program: I. Reduction in mortality of persons with high blood pressure, including mild hypertension. JAMA 242:2562-2571, 1979.)

En la figura 8-11 observamos que, en cada intervalo tras la incorporación al estudio, los pacientes del grupo de tratamiento escalonado presentaban una mortalidad inferior que los del grupo de tratamiento referido. En dicha figura vemos que se mantuvo el mismo patrón en aquellos que únicamente presentaban elevaciones leves de la presión arterial.

Los resultados se exponen con mayor detalle en la tabla 8-6, en la que se presentan los datos en función de la presión arterial diastólica al incorporarse al estudio. La columna de la derecha muestra el porcentaje de reducción de la mortalidad en el grupo de tratamiento escalonado: la mayor reducción se produjo en los pacientes con una elevación leve de la presión diastólica.

Este estudio ha tenido un gran impacto y ha logrado que los médicos traten elevaciones incluso leves de la presión arterial. Sin embargo, ha sido criticado porque carecía de un grupo no tratado para comparación. No sólo fueron remitidos estos pacientes de vuelta a sus médicos, sino que no hubo control del tratamiento que les fue proporcionado por sus médicos. Por tanto, la interpretación de estos datos es algo problemática. Incluso hoy, existe controversia acerca de si de hecho existió una objeción ética legítima a incluir un grupo no tratado en este estudio o si existió un problema ético a la hora de diseñar un estudio caro que fue difícil de organizar y que dejó tanta incertidumbre y dificultad en su interpretación.

Multiple Risk Factor Intervention Trial

Un problema grave de los ensayos a gran escala que requieren la inversión de gran cantidad de recursos, económicos y de otro tipo, y que se tardan años en completar es que su interpretación a menudo se ve empañada por un problema en el diseño o en la metodología que puede no haber sido apreciado en una fase inicial del estudio. El Multiple Risk Factor Intervention Trial (MRFIT) fue un estudio aleatorizado diseñado para determinar si la mortalidad por infarto de miocardio podría disminuir por cambios del estilo de vida y otras medidas. En este estudio, un grupo recibió una intervención especial (IE) que consistía en el tratamiento escalonado de la hipertensión y educación y formación intensiva sobre cambios del estilo de vida. El grupo de comparación recibió su tratamiento habitual (TH) en la comunidad. A lo largo de un periodo de seguimiento medio de 7 años, los factores de riesgo de cardiopatía coronaria (CC) disminuyeron más en los varones que recibieron la IE que en los que recibieron el TH (fig. 8-12).

Sin embargo, al finalizar el estudio, no se observaron diferencias estadísticamente significativas entre los grupos, ni en la mortalidad por CC ni en la mortalidad por todas las causas (fig. 8-13).

La interpretación de estos resultados se vio complicada por problemas serios. En primer lugar, el estudio fue realizado en una época en la que la mortalidad por enfermedad coronaria estaba disminuyendo en Estados Unidos. Además, no quedó claro si la falta de diferencias encontrada en este estudio se debió a que los cambios del estilo de vida no eran un factor importante o porque el grupo control, por su cuenta, adoptó los mismos cambios de estilo de vida que adoptaron muchas otras personas en Estados Unidos en ese periodo. Gran parte de la población adoptó cambios como modificaciones generalizadas de la dieta, aumento del ejercicio e interrupción del tabaquismo, por lo que el grupo control pudo haberse «contaminado» con algunos de los cambios de conducta que habían sido recomendados en el grupo de estudio de modo formal y estructurado.

Este estudio también muestra el problema de usar medidas intermedias como puntos finales de la eficacia en los ensayos aleatorizados. Como todo efecto sobre la mortalidad puede tardar años en manifestarse, resulta

TABLA 8-6. Mortalidad por todas las causas durante el Hypertension Detection and Follow-up Program

Presión arterial diastólica al comienzo (mmHg)	Tratamiento escalonado (TE)	Tratamiento referido (TR)	TASA DE MORTALIDAD A 5 AÑOS		Reducción de la mortalidad en el grupo de TE (%)
			TE	TR	
90-104	3.903	3.922	5,9	7,4	20,3
105-114	1.048	1.004	6,7	7,7	13,0
≥115	534	529	9,0	9,7	7,2
Total	5.485	5.455	6,4	7,7	16,9

De Hypertension Detection and Follow-up Program Cooperative Group: Five-year findings of the Hypertension Detection and Follow-up Program: I. Reduction in mortality of persons with high blood pressure, including mild hypertension. JAMA 242:2562-2571, 1979.

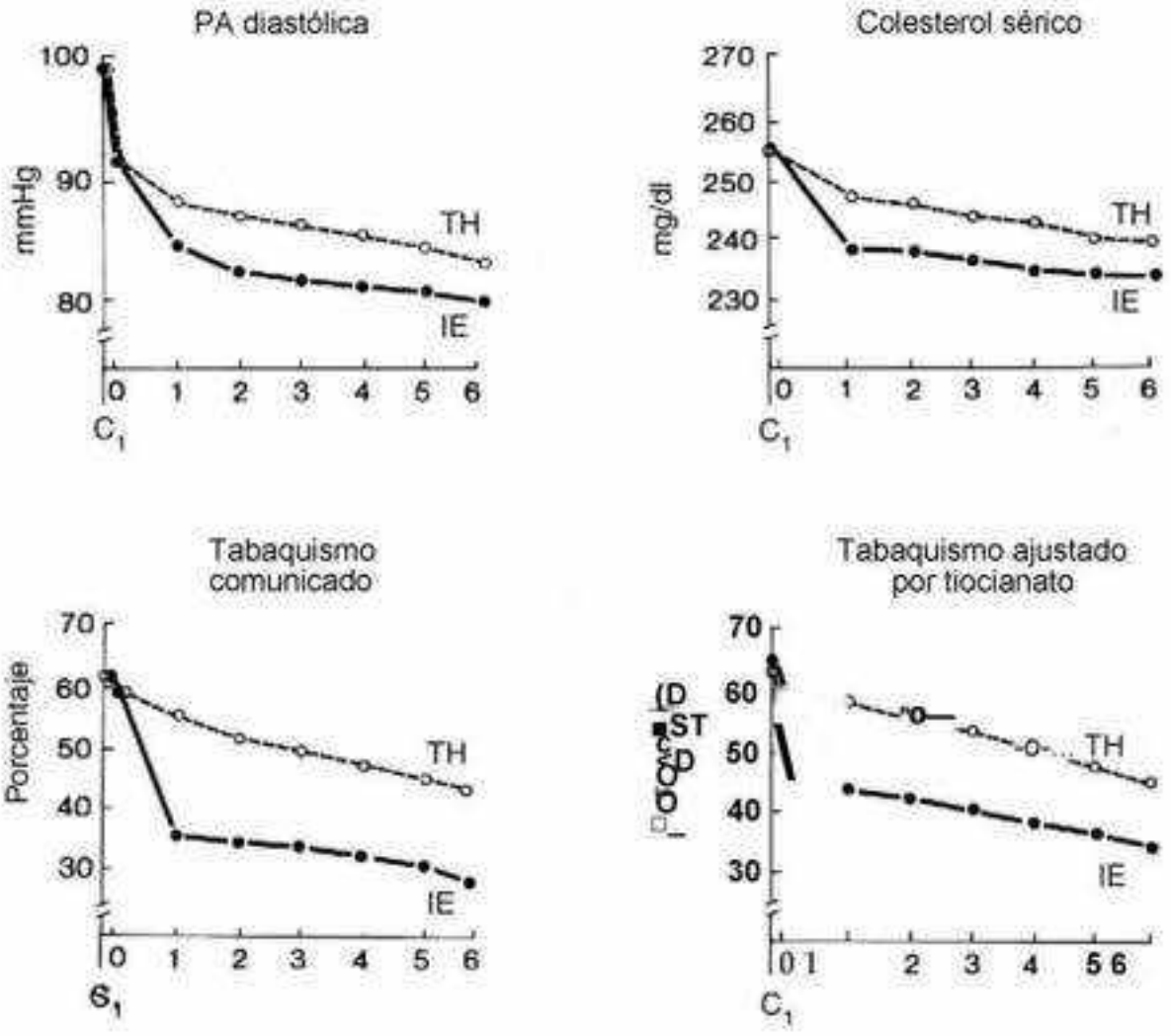


Figura 8-12. Niveles medios de factores de riesgo por año de seguimiento en los participantes del Multiple Risk Factor Intervention Trial Research Group. C_1, primera visita de cribado; IE, intervención especial; PA, presión arterial; TH, tratamiento habitual. (De Multiple Risk Factor Intervention Trial Research Group: Multiple Risk Factor Intervention Trial: Risk factor changes and mortality results. JAMA 248:1465-1477,1982.)

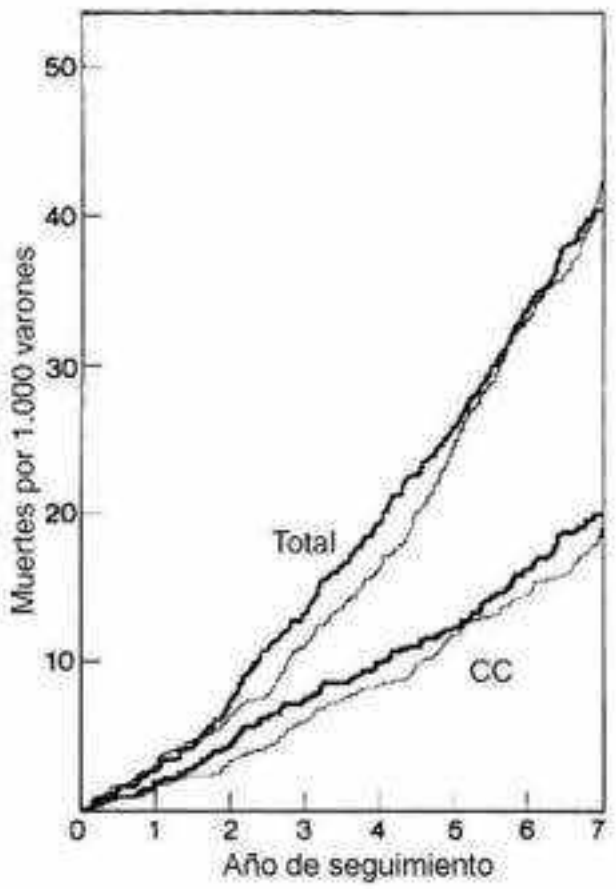


Figura 8-13. Tasas de mortalidad acumuladas por cardiopatía coronaria (CC) y totales de los participantes del Multiple Risk Factor Intervention Trial Research Group. La *línea gruesa* indica los varones que reciben el tratamiento habitual; la *línea fina* indica los varones que reciben una intervención especial. (De Multiple Risk Factor Intervention Trial Research Group: Multiple Risk Factor Intervention Trial: Risk factor changes and mortality results. JAMA 248:1465-1477, 1982.)

tentador utilizar medidas que podrían verse afectadas antes por la intervención. Sin embargo, como se observa aquí, aunque la intervención fue exitosa para reducir el tabaquismo, los niveles de colesterol y la presión arterial diastólica, no podemos concluir basándonos en estos cambios que la intervención fue efectiva, porque el objetivo del estudio era determinar si la intervención podría reducir la mortalidad por CC, lo que no ocurrió.

Debido a estos problemas, que a menudo dan lugar a dificultades en la interpretación de los hallazgos de los estudios muy amplios y caros, hay quien piensa que la inversión de los mismos fondos en diversos estudios más pequeños realizados por diferentes investigadores en diferentes poblaciones podría ser una elección más inteligente: si los resultados fueran consistentes, serían más creíbles, a pesar de los problemas de las muestras de menor tamaño que serían introducidos en las series individuales.

Estudio sobre prevención del cáncer de mama utilizando tamoxifeno

La observación de que las mujeres con cáncer de mama tratadas con tamoxifeno presentaban una menor incidencia de cáncer en la otra mama sugirió que el tamoxifeno podría ser útil para la prevención del cáncer de mama. Para estudiar esta hipótesis, se inició un ensayo aleatorizado en 1992. En septiembre de 1997,

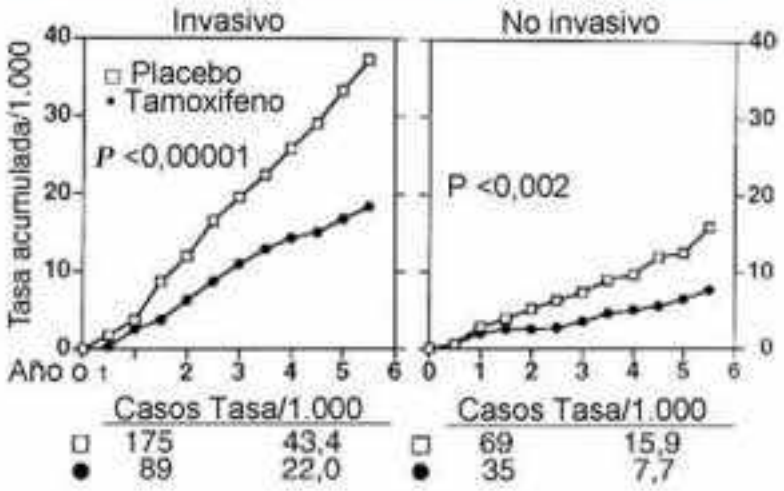


Figura 8-14. Tasas acumuladas de cáncer de mama invasivo y no invasivo en participantes que reciben placebo o tamoxifeno. (De Fisher B, Costantino JP, Wickerham DL, et al: Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-I Study. J Natl Cancer Inst 90:1371-1388,1998.)

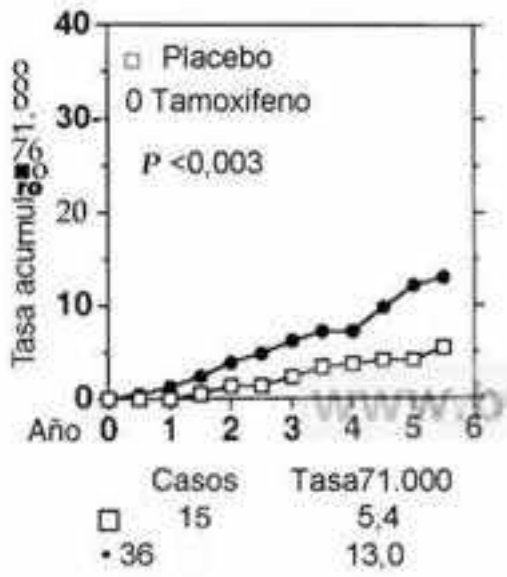


Figura 8-15. Tasas acumuladas de cáncer de endometrio invasivo en participantes que reciben placebo o tamoxifeno. (De Fisher B, Costantino JP, Wickerham DL, et al: Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-I Study. J Natl Cancer Inst 90:1371-1388,1998.)

13.388 mujeres de 35 o más años de edad se habían reclutado para el ensayo y habían sido asignadas aleatoriamente a recibir placebo o 20 mg de tamoxifeno al día durante 5 años. En marzo de 1998, un comité independiente que controlaba los datos decidió que las pruebas sobre la reducción del riesgo de cáncer de mama eran lo suficientemente sólidas como para justificar la interrupción del estudio. Como se observa en la figura 8-14, las tasas acumuladas del cáncer de mama invasivo y no invasivo se redujeron de modo importante en las mujeres tratadas con tamoxifeno. Al mismo tiempo, como se observa en la figura 8-15, las tasas de cáncer endometrial invasivo aumentaron en el grupo tratado con tamoxifeno. Por tanto, cuando se toma la decisión de utilizar tamoxifeno para la prevención del cáncer de mama, los beneficios potenciales del tamoxifeno deben sopesarse frente a la mayor incidencia de cáncer endometrial. El cuadro se ve aún más complicado por el hecho de que en la época en la que se publicaron los resultados de este estudio, dos estudios europeos más pequeños no observaron la reducción comunicada en el estudio americano. Así pues, nos encontramos ante el dilema del beneficio frente al daño; además, surge la duda de por qué otros estudios no han demostrado el mismo efecto destacado sobre la incidencia del cáncer de mama y cómo se deben tener en cuenta los resultados de dichos estudios a la hora de elaborar políticas públicas sobre esta materia.

ENSAYOS ALEATORIZADOS PARA EVALUAR INTERVENCIONES AMPLIAMENTE ACEPTADAS

Los ensayos controlados aleatorizados pueden usarse con dos propósitos fundamentales: 1) para evaluar nuevas formas de intervención antes de que sean aprobadas y recomendadas para su uso general, y 2) para evaluar intervenciones que son muy controvertidas o que han sido ampliamente usadas o recomendadas sin haber sido evaluadas adecuadamente. Para evaluar el impacto que ejercen los ensayos controlados aleatorizados sobre la práctica médica, el segundo uso demuestra el desafío de cambiar los abordajes empleados en la práctica médica habitual que pueden no haber sido evaluados adecuadamente. En esta sección presentamos dos ejemplos sobre esta materia.

Un ensayo sobre la cirugía artroscópica de la rodilla por artrosis

Alrededor del 6% de los adultos de más de 30 años de edad y el 12% de los adultos mayores de 65 años sufren dolor de rodilla intenso como resultado de la artrosis. En Estados Unidos, una intervención realizada con frecuencia en los pacientes con dolor de rodilla y signos de artrosis ha sido la cirugía artroscópica con lavado o desbridamiento de la articulación de la rodilla utilizando un artroscopio. Se ha estimado que la intervención se ha realizado cada año en más de 225.000 adultos de mediana edad y de edad avanzada, con un coste anual de más de 1.000 millones de dólares.

En diversos ensayos controlados aleatorizados se compararon pacientes que fueron sometidos a un desbridamiento o a un lavado de la rodilla con controles que no recibieron tratamiento. Los pacientes tratados comunicaron más mejoría de su dolor de rodilla que los que no fueron tratados. Otros estudios, sin embargo, en los que sólo se inyectó solución salina en la rodilla, también comunicaron mejoría de los síntomas. Así pues, resultó claro que los beneficios percibidos podrían relacionarse más con las expectativas del paciente que con la eficacia real, porque la mejoría subjetiva comunicada por los pacientes era más probable cuando a los

Figura 8-16. Diseño de un ensayo controlado sobre cirugía artroscópica para la artrosis de la rodilla. (Basado en Moseley JB, O'Malley K, Petersen NJ, et al: A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med 347:81-88,2002.)

pacientes no se les ocultaba si recibían o no tratamiento quirúrgico. Para resolver el asunto de si el lavado o el desbridamiento artroscópico reduce los síntomas de dolor de rodilla en los pacientes con artrosis, se necesitaba un ensayo controlado aleatorizado en el que los controles fuesen sometidos a un tratamiento placebo. En julio de 2002, Moseley y cois.[6] publicaron los resultados de un ensayo aleatorizado muy bien realizado sobre esta intervención, utilizando una artroscopia placebo en los controles.

El diseño de este estudio se muestra en la figura 8-16. Ciento ochenta veteranos fueron aleatorizados a grupos sometidos a desbridamiento artroscópico (59), lavado artroscópico (61) o artroscopia placebo (60). La intervención placebo consistió en una incisión cutánea y en un desbridamiento simulado sin introducción de un artroscopio. Los parámetros medidos fueron el grado de dolor de la rodilla, determinado por encuestas, y el grado de función física, determinado por encuestas y observación directa. Los pacientes fueron observados durante un período de 2 años. Los encargados de valorar el dolor y el estado funcional en los participantes, así como los mismos participantes, desconocían a qué grupo de tratamiento habían sido asignados.

Los resultados se muestran en las figuras 8-17 y 8-18. En ninguno de los grupos sometidos a intervención artroscópica se logró un mayor alivio del dolor que en el grupo sometido a la intervención placebo (v. fig. 8-17). Además, en ninguno de los grupos sometidos a alguna de las dos intervenciones artroscópicas se logró una mayor mejoría de la función física que en el grupo sometido a la intervención placebo (v. fig. 8-18).

El investigador principal del estudio, el Dr. Nelda Wray, del Houston Veterans Affairs Medical Center,

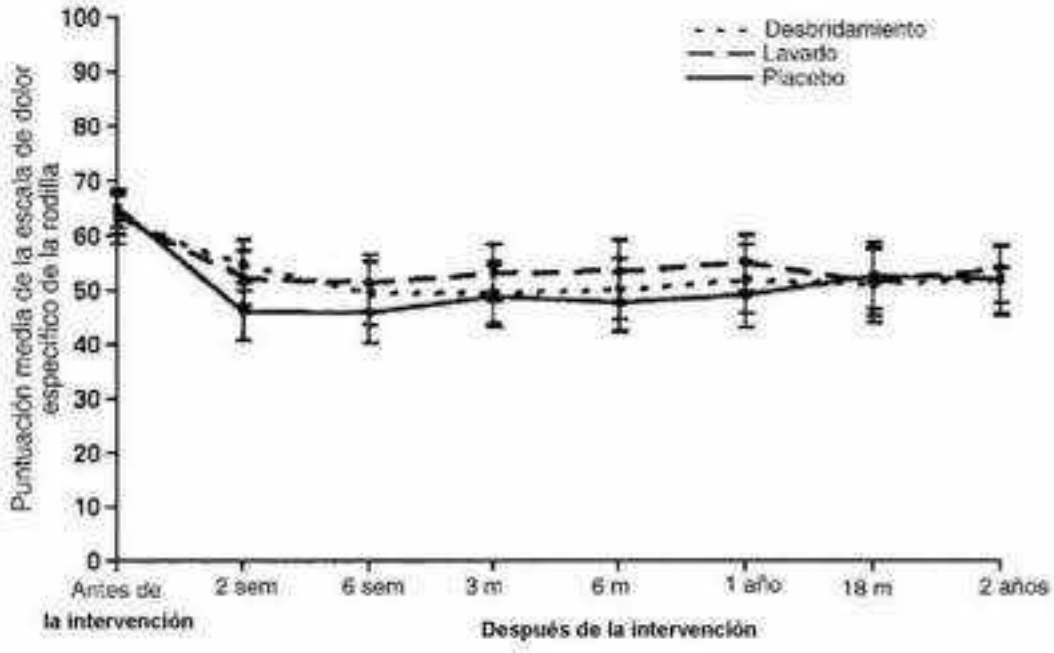


Figura 8-17. Valores medios (e intervalos de confianza del 95%) de la escala de dolor específico de la rodilla. Las determinaciones se realizaron antes de la intervención y 2 semanas, 6 semanas, 3 meses, 6 meses, 12 meses, 18 meses y 24 meses después de la intervención. Las puntuaciones más altas indican un dolor más intenso.

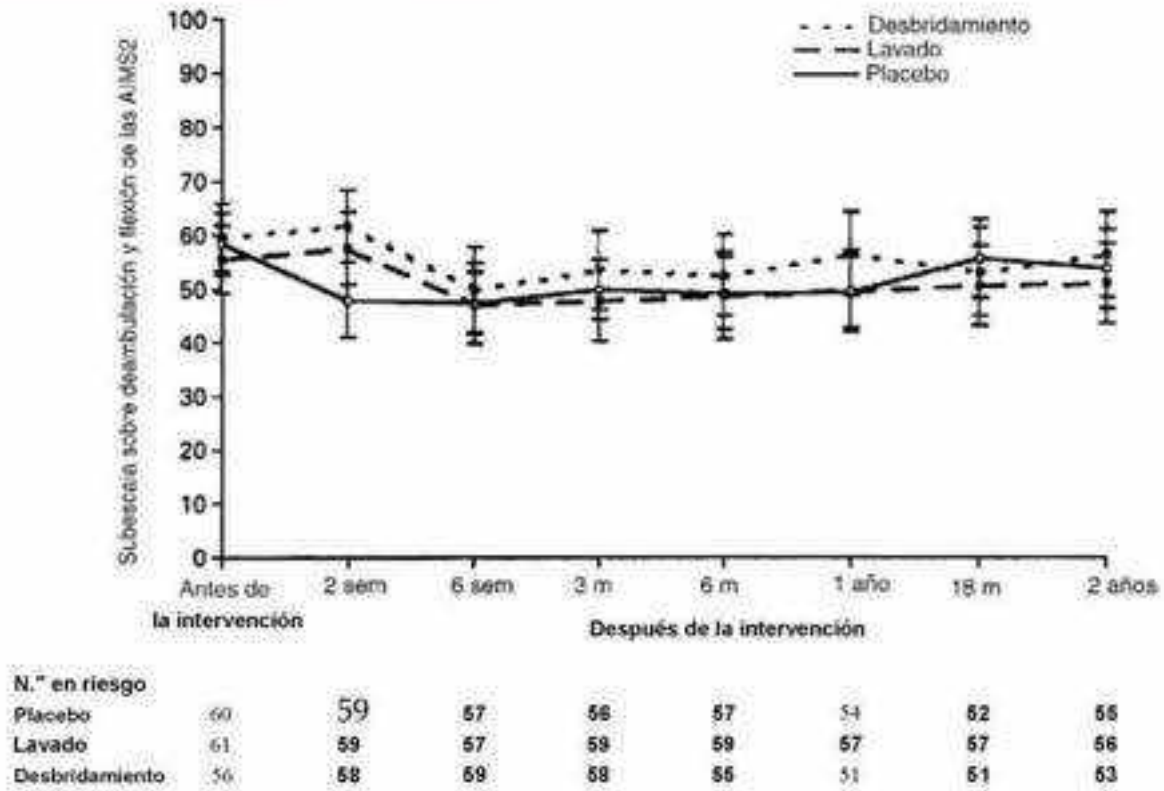


Figura 8-18. Valores medios (e intervalos de confianza del 95%) de la subescala sobre deambulación y flexión de las escalas de medición del impacto de la artrosis (AIMS2, por sus siglas en inglés). Las determinaciones se realizaron antes de la intervención y 2 semanas, 6 semanas, 3 meses, 6 meses, 12 meses, 18 meses y 24 meses después de la intervención. Las puntuaciones más altas indican una función peor.

donde se realizó el estudio, resumió los resultados diciendo: «Nuestro estudio demuestra que la cirugía no es mejor que el placebo; la intervención por sí misma no es útil.» Un mes después de la publicación de este estudio, el Department of Veterans Affairs emitió una nota de recomendación a sus médicos exponiendo que no debería realizarse la intervención quirúrgica hasta que se publicaran estudios adicionales. Según la nota de recomendación, el dolor de rodilla no era un dato suficiente para indicar la cirugía a menos que también existieran signos de «alteraciones anatómicas o mecánicas» que presumiblemente pudieran mejorar con dicha intervención.

Efecto de los grupos de apoyo psicosocial en la supervivencia de las pacientes con cáncer de mama metastásico

En 1989 se realizó un estudio en el que las mujeres con cáncer de mama metastásico eran asignadas aleatoriamente a terapia de grupo de apoyo-expresiva o a un grupo control. La terapia de apoyo-expresiva es un tratamiento estandarizado para pacientes con enfermedades potencialmente mortales que fomenta que un grupo de participantes, dirigidos por un terapeuta, expresen sus sentimientos y preocupaciones acerca de la enfermedad y su impacto. Este estudio demostró un beneficio en la supervivencia, aunque en él no se había planeado inicialmente un análisis de supervivencia. Otros ensayos de otras intervenciones psicosociales no han demostrado un beneficio en la supervivencia.

Para aclarar este aspecto, Goodwin y cols. realizaron un ensayo aleatorizado multicéntrico en el que 235 mujeres con cáncer de mama metastásico fueron aleatorizadas a un grupo que recibió terapia de apoyo-expresiva o a un grupo control que no recibió esta intervención (fig. 8-19). De las 235 mujeres, 158 fueron asignadas al grupo de intervención y 77 al grupo control.

A lo largo del período del estudio, la supervivencia no se vio prolongada en las pacientes que recibieron terapia de apoyo-expresiva (fig. 8-20). Sin embargo, el estado de ánimo y la percepción del dolor sí mejoraron, especialmente en las mujeres que estaban más

Figura 8-19. Diseño de un ensayo controlado aleatorizado sobre apoyo psicosocial en grupo en la supervivencia de pacientes con cáncer de mama metastásico. (Basado en Huston P, Peterson R: Withholding proven treatment in clinical research. N Engl J Med 345:912-914, 2001.)

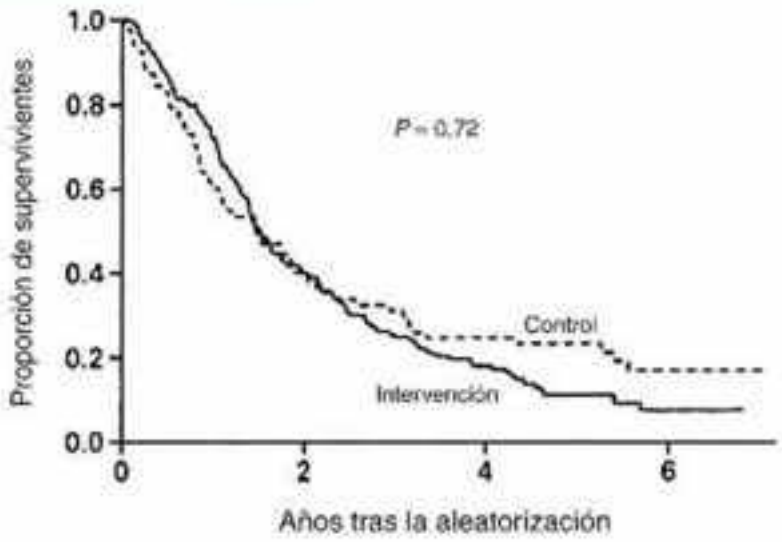


Figura 8-20. Curvas de supervivencia de Kaplan-Meier de mujeres asignadas al grupo de intervención y al grupo control. No se observaron diferencias significativas en la supervivencia entre los dos grupos.

angustiadas. Aunque los hallazgos en las publicaciones médicas son contradictorios acerca de la supervivencia y se siguen realizando estudios adicionales, los resultados de este estudio sugieren que esta intervención no se acompaña de beneficios en la supervivencia. Por tanto, los deseos de las mujeres que deciden hacer frente a su enfermedad de diferente forma, como no compartir sus sentimientos en un grupo, deben ser respetados. Además, no se debe sugerir a las mujeres que prefieren no participar en dicha terapia de grupo en ese momento difícil de sus vidas que su negativa puede estar acelerando su propia muerte.

REGISTRO DE ENSAYOS CLÍNICOS

Es un hecho conocido desde hace mucho tiempo que no se publican todos los resultados de los ensayos clínicos. Esto puede suponer un problema serio cuando se revisan los resultados de todos los ensayos clínicos publicados. Por ejemplo, si se revisan los ensayos clínicos de un nuevo fármaco pero sólo se han publicado los que muestran resultados beneficiosos y no los que muestran resultados negativos (por algún motivo), de los estudios publicados podría llegarse a la conclusión errónea de que *todos* los estudios sobre el nuevo fármaco han mostrado un beneficio claro. Este tipo de problema frecuente se denomina *sesgo de publicación* o sesgo de no publicación. Por ejemplo, Liebeskind y cois.[8] identificaron 178 ensayos clínicos controlados sobre ictus isquémico agudo publicados en inglés a lo largo de un período de 45 años desde 1955 a 1999 por medio de una búsqueda sistemática de varias bases de datos de gran envergadura. Estos ensayos reclutaron a un total de 73.949 sujetos y evaluaron 75 fármacos u otros tipos de intervención. Encontraron que el problema del sesgo de publicación era un factor importante a la hora de revisar las publicaciones sobre los ictus isquémicos agudos. Era mucho más probable que *no* fuesen publicados los ensayos que demostraban que el fármaco estudiado era dañino que los ensayos en los que los resultados indicaban que el fármaco estudiado fue neutro o beneficioso.

Varios factores explican el problema de los sesgos de publicación. Las revistas médicas tienen más interés en publicar resultados de estudios que muestran efectos espectaculares que resultados de estudios que no encontraron beneficios con un fármaco nuevo. Tanto los investigadores como las publicaciones están menos interesados en los estudios que concluyen que un tratamiento nuevo es inferior al tratamiento habitual o en los que los hallazgos no apuntan claramente hacia una u otra dirección. Existe un factor más importante que está contribuyendo a este problema: las empresas que lanzan nuevos fármacos y financian los estudios sobre dichos fármacos con frecuencia prefieren no publicar los resultados cuando el fármaco estudiado es menos eficaz que los tratamientos ya disponibles. Las empresas están claramente preocupadas por si los resultados de dichos estudios pudieran influir negativamente en las ventas del producto y afectar a los grandes ingresos potenciales que habían calculado tener con el nuevo fármaco. El resultado neto, sin embargo, es la ocultación de los datos, lo que proporciona una visión del fármaco (incluyendo su eficacia y seguridad) que no es completa, por lo que los reguladores, los médicos y la población no pueden adoptar decisiones basadas en evidencias, es decir, decisiones basadas en la información total generada por los ensayos clínicos.

La importancia del riesgo para la salud pública por el hecho de comunicar selectivamente los resultados de los ensayos clínicos y la frecuencia con la que esta comunicación selectiva tiene lugar condujo al Committee of Medical Journal Editors a adoptar una política, que entró en vigor en 2005, que obligaba a registrar *todos* los ensayos clínicos sobre intervenciones médicas en un registro público de ensayos antes de reclutar a participantes en el estudio[9]. Se consideran intervenciones médicas a los fármacos, las intervenciones quirúrgicas, los dispositivos, los tratamientos conductuales y los procesos de asistencia sanitaria. Antes de considerar la publicación de un ensayo clínico en alguna de las principales revistas médicas que adoptaron esta política, es obligatoria la inscripción del mismo en un registro accesible al público sin coste alguno.

CONSIDERACIONES ÉTICAS

En el contexto de los ensayos clínicos surgen muchos aspectos éticos. Una pregunta que se plantea con frecuencia es si la aleatorización es ética. ¿Cómo podemos

dejar a pacientes sin un tratamiento farmacológico, especialmente cuando tienen enfermedades graves y potencialmente mortales? La aleatorización es ética sólo cuando desconocemos si el fármaco A es mejor que el fármaco B. Podemos tener alguna indicación de que un tratamiento es mejor que el otro (a menudo éste es el motivo para realizar un ensayo en primer lugar), pero no estamos seguros. A menudo, sin embargo, no está claro en qué momento «descubrimos» que el fármaco A es mejor que el fármaco B. Un mejor planteamiento de la pregunta es: ¿cuándo tenemos pruebas adecuadas que apoyen la conclusión de que el fármaco A es mejor que el fármaco B? Una cuestión que ha recibido mucha atención en los últimos años es si es ético utilizar un placebo[10]. Este aspecto conlleva implícito el tema de si es ético no administrar un tratamiento cuya eficacia ha sido demostrada[11].

La pregunta también puede plantearse de modo inverso: «¿Es ético no aleatorizar?». Cuando consideramos fármacos, medidas preventivas o sistemas de asistencia sanitaria de aplicación en gran cantidad de personas, tanto en Estados Unidos como en otros países, la norma puede ser llevar a cabo un ensayo aleatorizado que aclare los aspectos del beneficio y el daño y no seguir sometiendo a las personas a efectos tóxicos innecesarios y crear falsas esperanzas, a menudo con un gran coste. Por tanto, las dudas sobre la ética de la aleatorización deben plantearse en ambas direcciones: aleatorización o no aleatorización.

Otro punto importante es si puede obtenerse realmente un consentimiento informado. Muchos protocolos de ensayos clínicos multicéntricos requieren la incorporación de los pacientes en el estudio inmediatamente tras el diagnóstico. El paciente puede que no sea capaz de otorgar su consentimiento y los familiares pueden estar tan angustiados por el diagnóstico que acaban de recibir y por sus implicaciones que tienen gran dificultad en comprender la noción de la aleatorización y dar el visto bueno para que el paciente sea aleatorizado. Por ejemplo, gran parte del progreso alcanzado en las últimas décadas en el tratamiento de la leucemia infantil ha sido resultado de protocolos multicéntricos rigurosos que han precisado la incorporación del niño al estudio inmediatamente después de establecer el diagnóstico de la leucemia. Claramente, en unos momentos en los que los padres están tan angustiados, nos podemos cuestionar si son capaces de otorgar realmente su consentimiento informado. Sin embargo, el progreso ha tenido lugar gracias únicamente a dichos ensayos rigurosos, que han salvado las vidas de tantos niños con leucemia aguda.

Por último, ¿bajo qué circunstancias debería interrumpirse un ensayo antes de lo que se había planeado inicialmente? Este tema también es complicado y puede surgir porque desde el principio se observan o efectos beneficiosos o efectos nocivos del agente estudiado, antes de haber reclutado a toda la muestra de participantes o antes de que los sujetos hayan sido estudiados durante el período de seguimiento completo. Muchos estudios cuentan con un comité externo examinador de datos que controla los datos a medida que son recibidos; el comité toma la decisión, como se ve, por ejemplo, en el Physicians' Health Study expuesto en el capítulo 7, en el que se estudiaban simultáneamente dos medicaciones en un diseño factorial: la aspirina se estudiaba para la prevención primaria de las enfermedades cardiovasculares y el beta-caroteno se estudiaba para la prevención primaria del cáncer. El comité externo examinador de datos decidió que los hallazgos sobre la aspirina eran lo suficientemente claros como para finalizar el estudio sobre la aspirina, pero que el estudio sobre el beta-caroteno debía continuar (v. págs. 151-152).

CONCLUSIÓN

Los ensayos aleatorizados son el método de referencia para evaluar la eficacia de las medidas terapéuticas, preventivas y de otro tipo tanto en la medicina clínica como en el ámbito de la salud pública. En los capítulos 7 y 8 se ha proporcionado una visión global de los métodos de diseño de estudio en los ensayos aleatorizados y las medidas empleadas para minimizar o evitar sesgos de selección y de otro tipo. Desde un punto de vista social, la generalización y los aspectos éticos son consideraciones importantes, y estos aspectos también se han analizado.

EPÍLOGO

Concluiremos esta exposición sobre los ensayos aleatorizados citando un artículo de Caroline y Schwartz que fue publicado en la revista *Chest* en 1975. El artículo se titulaba «Chicken soup rebound and relapse of pneumonia: Report of a case»[12].

En la introducción los autores expusieron lo siguiente:

> *Desde hace mucho tiempo se sabe que el caldo de pollo posee una potencia terapéutica inusual frente a una gran variedad de agentes víricos y bacterianos. De hecho, ya en el siglo xii, el teólogo, filósofo y médico Moses Maimónides escribió: «El caldo de pollo [...] se recomienda como alimento excelente, así como medicación.» Estudios anecdóticos previos sobre la eficacia terapéutica de este agente, sin embargo, no han logrado proporcionar detalles sobre la duración adecuada del tratamiento. A continuación exponemos un caso clínico en el que la interrupción abrupta del caldo de pollo dio lugar a una recidiva grave de una neumonía*[10].

A continuación, los autores presentan el caso clínico de un médico de 47 años que fue tratado de neumonía con caldo de pollo. La administración de caldo de pollo se interrumpió prematuramente y el paciente sufrió una recidiva. Al no poder contar con más caldo de pollo, la recidiva fue tratada con penicilina intravenosa.

La exposición de los autores es de especial interés. A continuación se muestra una parte:

> *La eficacia terapéutica del caldo de pollo fue descubierta por primera vez hace varios miles de años cuando una epidemia de gran mortalidad para los varones jóvenes de Egipto parecía no afectar a una minoría étnica que residía en la misma zona. La investigación epidemiológica contemporánea reveló que la dieta del grupo no afectado por la epidemia contenía grandes cantidades de una preparación cocinada hirviendo pollo con varias verduras e hierbas. Se debe destacar a este respecto que las órdenes relativas a la dieta dadas a Moisés en el monte Sinaí, aunque restringían el consumo de no menos de 19 tipos de aves, no incluían el pollo en la prohibición. Algunos eruditos creen que la receta del caldo de pollo fue transmitida a Moisés en la misma ocasión, pero fue relegada a la tradición oral cuando las escrituras fueron canonizadas. [...] Aunque el caldo de pollo se utiliza ampliamente en la actualidad frente a una variedad de trastornos orgánicos y funcionales, su elaboración está en gran medida en manos de particulares y la estandarización es casi imposible. Las investigaciones preliminares de la farmacología del caldo de pollo han demostrado que se absorbe inmediatamente tras su administración oral. [...] No se recomienda la administración parenteral.*

Este trabajo suscitó el envío de varias cartas al editor. En una, el Dr. Laurence E Greene, catedrático de Urología de la Clínica Mayo, escribió:

> *Puede que esté interesado en saber que hemos tratado exitosamente la impotencia masculina con otro compuesto derivado del pollo, la hexametilacetil lututria tetrazolamina citarabina sódica (Schmaltz [Upjohn]). Este compuesto, cuando se aplica en pomada en el pene, no sólo cura la impotencia, sino que también aumenta la libido y evita la eyaculación precoz. [...] Los estudios preliminares indican que sus efectos dependen de la dosis en vista de que la relación sexual dura 5 minutos cuando se aplica la pomada al 5%, 15 minutos cuando se aplica la pomada al 15%, / así sucesivamente.*
>
> *Hemos recibido una beca de 650.000 dólares de la National Scientific Foundation para llevar a cabo un estudio prospectivo controlado aleatorizado doble ciego. Desafortunadamente, somos incapaces de obtener un número adecuado de sujetos debido a que todos los voluntarios se niegan a participar a menos que se les asegure que serán sujetos y no controles*[13].

BIBLIOGRAFÍA

1. Gehan E: Clinical trials in cancer research, Environ Health Perspect 32:31,1979.
2. Mant D: Can randomized trials inform clinical decisions about individual patients? Lancet 353:743-746,1999.
3. IOM (Institute of Medicine) (2009): Initial National Priorities for Comparative Effectiveness Research. Washington, DC, National Academy Press, http://www.nap.edu/catalog.php?record_id = 12648. Accessed June 28,2013.
4. Veterans Administration Cooperative Study Group on Hypertensive Agents: Effects of treatment on morbidity in hypertension: Results in patients with diastolic blood pressure averaging 115 through 129 mm Hg. JAMA 213:1028-1034, 1967.
5. Hypertension Detection and Follow-up Program Cooperative Group: Five year findings of the Hypertension Detection and Follow-up Program: I. Reduction of mortality of persons with high blood pressure, including mild hypertension. JAMA 242:2562,1979.
6. Moseley JB, O'Malley K, Petersen NJ, et al: A controlled trial of arthroscopic surgery for osteoarthritis of the knee, N Engl J Med 347:81-88,2002.
7. Goodwin PJ, Leszcz M, Ennis M, et al: The effect of group psychosocial support on survival in metastatic breast cancer, N Engl J Med 345:1719-1726, 2001.
8. Liebeskind DS, Kidwell CS, Sayre JW, et al: Evidence of publication bias in reporting acute stroke clinical trials, Neurology 67:973-979, 2006.
9. DeAngelis CD, Drazen JM, Frizelle FA: Clinical trial registration: A statement from the International Committee of Medical Journal, JAMA 292:1363-1364,2004.
10. Emanuel EJ, Miller FG: The ethics of placebo-controlled trials: A middle ground, N Engl J Med 345:915-919, 2001.
11. Huston P, Peterson R: Withholding proven treatment in clinical research, N Engl J Med 345:912-914,2001.
12. Caroline NL, Schwartz H: Chicken soup rebound and relapse of pneumonia: Report of a case, Chest 67:215-216,1975.
13. Greene LF: The chicken soup controversy [letter], Chest 68:605, 1975.

PREGUNTAS DE REPASO DE LOS CAPÍTULOS 7 Y 8

1. El principal objetivo de la asignación aleatoria en un ensayo clínico es:
   a. Ayudar a asegurar que los sujetos del estudio son representativos de la población general.
   b. Facilitar el doble ciego (enmascaramiento).
   c. Facilitar la medición de las variables del resultado.
   d. Asegurar que los grupos del estudio poseen características basales comparables.
   e. Reducir el sesgo de selección en la asignación del tratamiento.

2. Un anuncio en una revista médica decía: «Dos mil pacientes con faringitis fueron tratados con nuestra nueva medicina. En 4 días, el 94% se encontraban asintomáticos.» El anuncio asegura que la medicina es eficaz. Basado en lo expuesto anteriormente, la afirmación:
   a. Es correcta.
   b. Puede ser incorrecta porque la conclusión no está basada en una tasa.
   c. Puede ser incorrecta porque no reconoce el fenómeno de cohorte a largo plazo.
   d. Puede ser incorrecta porque no se utilizó ninguna prueba estadísticamente significativa.
   e. Puede ser incorrecta porque no se utilizó un grupo control o de comparación.

3. El objetivo de un estudio *doble ciego* o *con doble enmascaramiento* es:
   a. Lograr comparabilidad entre los sujetos tratados y no tratados.
   b. Reducir los efectos de la variación del muestreo.
   c. Evitar sesgos del observador y del sujeto.
   d. Evitar sesgos del observador y variación del muestreo.
   e. Evitar sesgos del sujeto y variación del muestreo.

4. En muchos estudios que examinaban la asociación entre los estrógenos y el cáncer uterino endometrial se utilizó una prueba significativa unilateral. La suposición subyacente que justificó una prueba unilateral en vez de bilateral fue que:
   a. La distribución de la proporción expuesta seguía un patrón «normal».
   b. Antes de realizar el estudio se pensaba que los estrógenos causaban cáncer uterino endometrial.
   c. El patrón de asociación podría expresarse mediante una función en línea recta.
   d. El error de tipo II era el error potencial más importante que se debía evitar.
   e. Sólo se utilizó un grupo control.

5. En un ensayo aleatorizado, un diseño cruzado planeado:
   a. Elimina el problema de un posible efecto de orden.
   b. Debe tener en cuenta el problema de posibles efectos residuales del primer tratamiento.
   c. Necesita aleatorización estratificada.
   d. Elimina la necesidad de controlar el cumplimiento o la falta de cumplimiento.
   e. Mejora la generalización de los resultados del estudio.

6. Un ensayo aleatorizado que comparaba la eficacia de dos fármacos demostró una diferencia entre los dos (con un valor de $P < 0{,}05$). Suponga que, sin embargo, los dos fármacos en realidad no se diferencian. Éste es, por tanto, un ejemplo de:
   a. Error de tipo I (error *a*).
   b. Error de tipo II (error (3).
   c. 1 — a.
   d. 1 - p.
   e. Ninguno de los anteriores.

7. Todos los siguientes son beneficios potenciales de un ensayo clínico aleatorizado *excepto:*
   a. La probabilidad de que los grupos del estudio sean comparables es mayor.
   b. Se elimina la autoselección para un tratamiento particular.
   c. La validez externa del estudio es mayor.
   d. La asignación del siguiente sujeto no puede predecirse.
   e. El tratamiento que recibe un sujeto no está influido por sesgos conscientes o subconscientes del investigador.

Pregunta de repaso adicional en la siguiente página.

Número de pacientes necesarios en un grupo control y en un grupo experimental para una probabilidad dada de obtener un resultado significativo (prueba bilateral)

	DIFERENCIAS EN LAS TASAS DE CURACIÓN ENTRE LOS DOS GRUPOS DE TRATAMIENTO					
La menor de las dos tasas de curación	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**
0,05	420	130	69	44	36	31
0,10	680	195	96	59	41	35
0,15	910	250	120	71	48	39
0,20	1.090	290	135	80	53	42
0,25	1.250	330	150	88	57	44
0,30	1.380	360	160	93	60	44
0,35	1.470	370	170	96	61	44
0,40	1.530	390	175	97	61	44

a = 0,05; potencia (1 — (3) = 0,80.
Datos de Gehan E: Clinical trials in cancer research. Environ Health Perspect 32:31,1979.

La pregunta 8 se basa en la tabla precedente:

8. Una compañía farmacéutica sostiene que un nuevo fármaco G para una cierta enfermedad posee una tasa de curación del 50% en comparación con el fármaco H, cuya tasa de curación es sólo del 25%. Se le encarga a usted que diseñe un ensayo clínico para comparar los fármacos G y H. Utilizando la tabla precedente, calcule el número de pacientes necesarios en cada grupo de tratamiento para detectar dicha diferencia con unos valores de a = 0,05, bilateral, y (3 = 0,20.

El número de pacientes necesarios en cada grupo de tratamiento es_____.

Capítulo 9

Estudios de cohortes

Objetivos de aprendizaje

- **Describir el diseño de un estudio de cohortes y distinguirlo de un ensayo clínico aleatorizado.**
- **Ilustrar el diseño de un estudio de cohortes con dos ejemplos destacados.**
- **Comentar algunos sesgos que pueden producirse en los estudios de cohortes.**

En este capítulo y en los siguientes de la sección 2, la atención se centra en el uso de la epidemiología para dilucidar las relaciones etiológicas o causales. Los dos pasos que subyacen a los diseños de los estudios que se comentan en los capítulos 9 y 10 se muestran esquemáticamente en la figura 9-1.

1. En primer lugar, hay que determinar si existe una asociación entre un factor o una característica y el desarrollo de una enfermedad. Esto se puede lograr mediante el estudio de las características de los grupos, de las características de los individuos o de ambos factores (v. caps. 9 a 12).
2. En segundo lugar, se deducen inferencias apropiadas respecto a una posible relación causal a partir de los patrones de asociación que se han encontrado (v. caps. 14 y 15).

Figura 9-1. Si se observa una asociación entre una exposición y una enfermedad u otro resultado (1), surge la pregunta de si la asociación es causal (2).

En los capítulos 9 y 10 se describen los diseños de los estudios utilizados para el paso 1. En este capítulo se analizan los estudios de cohortes; los estudios de casos y controles y otros diseños se comentan en el capítulo 10. Estos estudios, a diferencia de los ensayos aleatorizados, se denominan colectivamente estudios *observacionales*.

DISEÑO DE UN ESTUDIO DE COHORTES

En un estudio de cohortes, el investigador selecciona un grupo de individuos expuestos y un grupo de individuos no expuestos y realiza un seguimiento de ambos para comparar la incidencia de la enfermedad (o la mortalidad por enfermedad) en ambos grupos (fig. 9-2). El diseño puede incluir más de dos grupos, aunque sólo se muestran dos grupos para fines esquemáticos.

Si existe una asociación positiva entre la exposición y la enfermedad, sería de esperar que la proporción de personas del grupo expuesto que desarrollan la enfermedad (incidencia en el grupo expuesto) fuese mayor que la proporción de personas del grupo no expuesto que desarrollan la enfermedad (incidencia en el grupo no expuesto).

Los cálculos correspondientes se muestran en la tabla 9-1. Se comienza con un grupo expuesto y un grupo no expuesto. De las $(a + b)$ personas expuestas, la enfermedad se desarrolla en a pero no en b. Por tanto, la incidencia de la enfermedad entre las personas expuestas es $\left(\frac{a}{a+b}\right)$. De forma similar, en las $(c+ d)$ personas no expuestas del estudio, la enfermedad se desarrolla en c pero no en d, por lo que la incidencia de la enfermedad entre los no expuestos es $\left(\frac{c}{c+ d}\right)$.

Figura 9-2. Diseño de un estudio de cohortes.

TABLA 9-1. Diseño de un estudio de cohortes

		Después, seguimiento para ver si			
		Se *desarrolla la enfermedad*	*No se desarrolla la enfermedad*	Total de enfermedad	Tasas de incidencia
Primero seleccionar	Expuesto	a	b	$a + b$	$\frac{a}{a+b}$
	No expuesto	c	d	$c + d$	$\frac{c}{c+d}$

El uso de estos cálculos se aprecia en un ejemplo hipotético de un estudio de cohortes que se muestra en la tabla 9-2. En este estudio de cohortes, la asociación del tabaquismo con la arteriopatía coronaria (AC) se investiga seleccionando para el estudio un grupo de 3.000 fumadores (expuestos) y un grupo de 5.000 no fumadores (no expuestos) que no presentan cardiopatía al inicio del estudio. En ambos grupos se realiza el seguimiento de la aparición de AC y se compara la *incidencia* de AC en ambos. La AC se desarrolla en 84 de los fumadores y en 87 de los no fumadores. El resultado es una incidencia de AC de 28,0/1.000 en los fumadores y 17,4/1.000 en los no fumadores.

Se debe tener en cuenta que, debido a que se están identificando casos *nuevos* (incidentes) de la enfermedad a medida que ocurren, se puede determinar si existe una relación temporal entre la exposición y la enfermedad, es decir, si la exposición precedió a la aparición de la enfermedad. Es evidente que esta relación temporal se debe establecer si hay que considerar que la exposición es una posible causa de la enfermedad en cuestión.

COMPARACIÓN DE LOS ESTUDIOS DE COHORTES CON ENSAYOS CLÍNICOS ALEATORIZADOS

Llegados a este punto, es útil comparar el estudio de cohortes observacional que se acaba de describir con el diseño de ensayo aleatorizado (cohorte experimental) descrito previamente, en los capítulos 7 y 8 (fig. 9-3).

Ambos tipos de estudios comparan el grupo expuesto con el no expuesto (o un grupo con una cierta exposición frente a un grupo con otra exposición). Debido a que, por razones éticas y de otro tipo, no se puede distribuir aleatoriamente a las personas para recibir una sustancia supuestamente perjudicial, como un posible carcinógeno, la «exposición» en la mayoría de los ensayos aleatorizados es un tratamiento o una medida preventiva. En los estudios de cohortes que investigan la etiología, la «exposición» es a menudo un agente posiblemente tóxico o carcinógeno. En ambos tipos de diseño, sin embargo, un grupo expuesto se compara con un grupo no expuesto o con un grupo sometido a otra exposición.

La diferencia entre estos dos diseños (la presencia o ausencia de asignación aleatoria) es fundamental a la hora de interpretar los hallazgos del estudio. Las ventajas de la asignación aleatoria se comentaron en los capítulos 7 y 8. En un estudio no aleatorizado, cuando se observa una asociación de una exposición con una enfermedad, queda la incertidumbre de si la asociación puede deberse a que las personas no se asignaron al azar a la exposición; tal vez no sea la exposición, sino más bien los factores que llevaron a la gente a presentar dicha exposición, lo que se asocia con la enfermedad. Por ejemplo, si se observa un mayor riesgo de una enfermedad en los trabajadores de una fábrica determinada, y si la mayoría de los trabajadores de esa fábrica viven

TABLA 9-2. Resultados de un hipotético estudio de cohortes sobre el tabaquismo y la arteriopatía coronaria (AC)

		Después, seguimiento para ver si			
		Se *desarrolla AC*	*No se desarrolla AC*	Total	Incidencia por 1.000 por año
Primero seleccionar	Fumadores de cigarrillos	84	2.916	3.000	28,0
	No fumadores de cigarrillos	87	4.913	5.000	17,4

Figura 9-3. Selección de los grupos de estudio en los estudios epidemiológicos experimentales y observacionales.

en un área concreta, el mayor riesgo de la enfermedad podría deberse a una exposición asociada a su lugar de residencia en lugar de a su ocupación o lugar de trabajo. Este tema se comenta en los capítulos 13 y 14.

SELECCIÓN DE LAS POBLACIONES DE ESTUDIO

La característica esencial en el diseño de los estudios de cohortes es la comparación de los resultados en un grupo expuesto y en un grupo no expuesto (o en un grupo con una característica determinada y en otro sin esa característica). Hay dos formas básicas para generar tales grupos:

1. Se puede crear una población de estudio mediante la selección de grupos para la inclusión en el mismo basándose en si han sido expuestos o no (p. ej., las cohortes que presentan una exposición laboral) (fig. 9-4).
2. O bien se puede seleccionar una población definida antes de que cualquiera de sus miembros se exponga o antes de identificar sus exposiciones. Se podría seleccionar una población basándose en algún factor no relacionado con la exposición (p. ej., la comunidad de residencia) (fig. 9-5) y realizar la anamnesis, o llevar a cabo análisis de sangre o de otro tipo, en toda la población. Utilizando los resultados de la anamnesis o de las pruebas analíticas, se puede separar a la población en los grupos *expuesto* y *no expuesto* (o en aquellos que tienen ciertas características biológicas y los que no), como se hizo en el estudio de Framingham, que se describe más adelante en este capítulo.

Los estudios de cohortes, en los que se espera que se produzca un resultado en una población, a menudo requieren un período de seguimiento prolongado, que dura hasta se han producido bastantes fenómenos (resultados). Cuando se emplea la segunda estrategia (en la que se identifica una población para el estudio basándose en alguna característica no relacionada con la exposición en cuestión), la exposición de interés puede que no tenga lugar durante un cierto tiempo, incluso durante muchos años después de que la

Figura 9-4. Diseño de un estudio de cohortes comenzando con los grupos expuesto y no expuesto.

Figura 9-5. Diseño de un estudio de cohortes comenzando con una población definida.

Figura 9-6. Cronología para un hipotético estudio de cohortes prospectivo iniciado en 2012.

Figura 9-7. Cronología para un hipotético estudio de cohortes retrospectivo iniciado en 2012.

población se haya definido. En consecuencia, la duración del seguimiento requerido es aún mayor con la segunda estrategia que con la primera. Hay que tener en cuenta que, con cualquiera de las estrategias, el diseño del estudio de cohortes es fundamentalmente el mismo: *se comparan personas expuestas y no expuestas*. Esta comparación es el sello distintivo del diseño de cohortes.

TIPOS DE ESTUDIOS DE COHORTES

Un problema fundamental con el diseño de cohortes que se acaba de describir es que la población de estudio a menudo debe seguirse durante un periodo prolongado para determinar si se ha producido el resultado de interés. Tomemos como ejemplo un estudio hipotético de la relación del tabaquismo con el cáncer de pulmón. Se identifica una población de estudiantes de primaria y se siguen; diez años después, cuando son adolescentes, se identifican los que fuman y los que no lo hacen. Después, se continúa el seguimiento de ambos grupos (fumadores y no fumadores) para ver quién desarrolla cáncer de pulmón y quién no. Pongamos por caso que el estudio comienza en 2012 (fig. 9-6) y supongamos que muchos niños que se convertirán en fumadores lo harán en el plazo de 10 años. Por tanto, el estatus de la exposición (fumador o no fumador) se determinará 10 años más tarde, en 2022. Para los fines de este ejemplo, se supondrá que el periodo de latencia desde que se empieza a fumar hasta que se desarrolla un cáncer de pulmón es de 10 años. Por tanto, el desarrollo de cáncer de pulmón se determinará 10 años después, en 2032.

Este tipo de diseño del estudio se denomina *estudio de cohortes prospectivo* (o también *estudio de cohortes concurrente o longitudinal*). Es *concurrente* porque el investigador identifica la población original al comienzo del estudio y, en efecto, acompaña a los sujetos al mismo tiempo a lo largo del tiempo hasta el punto en el que la enfermedad se desarrolla o no se desarrolla.

Esta estrategia tiene una serie de problemas. Una dificultad es que, tal y como se acaba de describir, el estudio requerirá al menos 20 años para completarse, lo que puede conllevar varios problemas. Si se tiene la suerte de obtener una beca de investigación, la financiación suele limitarse a un máximo de tan sólo 3-5 años. Además, con un estudio de esta duración, existe el riesgo de que los sujetos de estudio sobrevivan al investigador o de que el investigador muera antes del final del estudio. Teniendo en cuenta estas cuestiones, el estudio de cohortes prospectivo a menudo resulta poco atractivo para los investigadores que están pensando en nuevos temas que evaluar.

Hay que dilucidar si estos problemas significan que el diseño de cohortes no es práctico y si hay alguna manera de acortar el periodo de tiempo necesario para llevar a cabo un estudio de cohortes. A continuación se considerará una estrategia alternativa usando el diseño de cohortes (fig. 9-7). Supongamos que de nuevo se comienza el estudio en 2012, pero ahora se dispone en la comunidad de una antigua lista de los escolares de primaria elaborada en 1992, y además se les había encuestado con respecto a su hábito de fumar en 2002. Gracias al uso de estos recursos de datos en 2012, se puede empezar a identificar qué personas de esta población han

Figura 9-8. Cronología para un hipotético estudio de cohortes prospectivo y un hipotético estudio de cohortes retrospectivo iniciados en 2012.

desarrollado el cáncer de pulmón y cuáles no. Este tipo de estudio se denomina *estudio de cohortes retrospectivo* o *estudio de cohortes histórico* (o también *estudio prospectivo no concurrente*). Sin embargo, se debe tener en cuenta que el diseño del estudio no difiere del diseño de cohortes prospectivo (todavía se está comparando un grupo expuesto con uno no expuesto); lo que se ha hecho en el diseño de cohortes retrospectivo ha sido utilizar datos históricos del pasado para poder acortar el marco temporal para el estudio y obtener los resultados antes. Ya no es un diseño prospectivo, porque se está comenzando el estudio con una población preexistente para reducir la duración del mismo. Pero, como se muestra en la figura 9-8, los *diseños, tanto para el estudio de cohortes prospectivo como para el estudio de cohortes retrospectivo* o *histórico, son idénticos: se comparan las poblaciones expuesta y no expuesta.* La única diferencia entre ellos es el tiempo. En un *diseño de cohortes prospectivo,* la exposición y la no exposición se determinan a medida que ocurren durante el estudio, y los grupos se siguen a continuación durante varios años en el futuro y se mide la incidencia. En un *diseño de cohortes retrospectivo,* la exposición se determina a partir de los registros anteriores y el resultado (desarrollo o no desarrollo de la enfermedad) se determina en el momento de iniciar el estudio.

También es posible llevar a cabo un estudio que sea una combinación de un diseño de cohortes prospectivo y un diseño de cohortes retrospectivo. Con esta estrategia, la exposición se determina a partir de registros objetivos en el pasado (como en un estudio de cohortes histórico), y el seguimiento y la medición de resultados continúan en el futuro.

EJEMPLOS DE ESTUDIOS DE COHORTES

Ejemplo 1: estudio Framingham

Uno de los estudios de cohortes más importantes y mejor conocidos es el estudio Framingham de enfermedades cardiovasculares, que comenzó en 1948[1]. Framingham es una ciudad de Massachusetts, a unos 32 kilómetros de Boston. Se pensó que las características de su población (algo menos de 30.000 habitantes) serían apropiadas para un estudio de este tipo y facilitarían el seguimiento de los participantes.

Los residentes se consideraron elegibles si tenían entre 30 y 62 años de edad. La justificación para usar este rango de edad fue que es poco probable que las personas menores de 30 años manifiesten los criterios de valoración cardiovasculares que se evalúan durante el período de seguimiento propuesto de 20 años. Muchas personas mayores de 62 años ya tienen una enfermedad coronaria establecida, por lo que no merecería la pena estudiar la incidencia de enfermedad coronaria en las personas de este grupo de edad.

TABLA 9-3. Constitución de la población del estudio Framingham

	Número de varones	Número de mujeres	Total
Muestra aleatoria	3.074	3.433	6.507
Personas que respondieron	2.024	2.445	4.469
Voluntarios	312	428	740
Personas que respondieron sin AC	1.975	2.418	4.393
Voluntarios sin AC	307	427	734
Total de personas sin AC: grupo del estudio Framingham	2.282	2.845	5.127

AC, arteriopatía coronaria.
De Dawber TR, Kannel WB, Lyell LP: An approach to longitudinal studies in a community: The Framingham Study. Ann NY Acad Sci 107:539-556,1993.

Los investigadores buscaron un tamaño muestral de 5.000 personas. En la tabla 9-3 se muestra cómo se obtuvo la población final del estudio. Constaba de 5.127 varones y mujeres de entre 30 y 62 años de edad en el momento de la inclusión en el estudio, sin enfermedad cardiovascular en dicho momento. En este estudio se definieron muchas «exposiciones», como el tabaquismo, la obesidad, la hipertensión arterial, la hipercolesterolemia, los niveles bajos de actividad física y otros factores.

Los nuevos episodios coronarios se identificaron evaluando a la población del estudio cada 2 años y controlando a diario los ingresos en el único hospital de Framingham.

El estudio fue diseñado para comprobar las siguientes hipótesis:

- La incidencia de AC aumenta con la edad. Se produce antes y con más frecuencia en los varones.
- Las personas con hipertensión desarrollan AC a un ritmo mayor que las que son normotensas.
- La hipercolesterolemia se asocia con un riesgo mayor de AC.
- El tabaquismo y el consumo habitual de alcohol se asocian con una mayor incidencia de AC.
- El aumento de la actividad física se asocia con una disminución del desarrollo de AC.
- El aumento del peso corporal predispone a una persona a desarrollar AC.
- Los pacientes con diabetes mellitus tienen una mayor incidencia de AC.

Cuando se analiza esta lista en la actualidad, es posible preguntarse por qué este tipo de relaciones tan obvias y bien conocidas deberían haberse evaluado en un estudio tan extenso. No debe olvidarse el peligro de este enfoque «retrospectivo»; es sobre todo *gracias* al estudio Framingham, un estudio de cohortes clásico que hizo contribuciones fundamentales a nuestra comprensión de la epidemiología de las enfermedades cardiovasculares, por lo que estas relaciones son bien conocidas en la actualidad.

En este estudio se utilizó el segundo método descrito anteriormente en este capítulo para seleccionar una población para un estudio de cohortes: se seleccionó una población definida en función de la ubicación de la residencia o de otros factores no relacionados con la exposición o exposiciones en cuestión. Después, la población se observó a lo largo del tiempo para determinar qué personas desarrollaron o ya tenían la «exposición o exposiciones» de interés y, más adelante, para determinar cuáles desarrollaron el resultado o los resultados cardiovasculares de interés. Esta estrategia proporcionó una ventaja importante: permitió a los investigadores estudiar múltiples «exposiciones», como la hipertensión, el tabaquismo, la obesidad, los niveles de colesterol y otros factores, así como las complejas interacciones entre las exposiciones, mediante el uso de técnicas multifactoriales. Por tanto, mientras que un estudio de cohortes que comienza con un grupo expuesto y otro no expuesto se centra en la exposición específica, un estudio de cohortes que se inicia con una población definida puede explorar los papeles de muchas exposiciones.

Ejemplo 2: incidencia de cáncer de mama y deficiencia de progesterona

Se sabe desde hace mucho tiempo que el cáncer de mama es más frecuente en mujeres que son mayores en el momento de su primer embarazo. Esta observación suscita una pregunta difícil de responder: ¿la asociación entre la edad avanzada en el momento del primer embarazo y el mayor riesgo de cáncer de mama se relacionan con el hallazgo de que un primer embarazo precoz protege contra el cáncer de mama (y, por tanto, esa protección no existe en las mujeres que tienen un embarazo más tardío o ningún embarazo), o tanto un primer embarazo tardío como el mayor riesgo de cáncer de mama se deben a un tercer factor, como una anomalía hormonal subyacente?

Es difícil disociar estas dos interpretaciones. Sin embargo, en 1978, Cowan y cois.[2] realizaron un estudio diseñado para determinar cuál de estas dos explicaciones era probable que fuese la correcta (fig. 9-9). Los investigadores identificaron una población de mujeres que eran pacientes de la clínica de infertilidad del Johns Hopkins Hospital en Baltimore, Maryland, de 1945 a

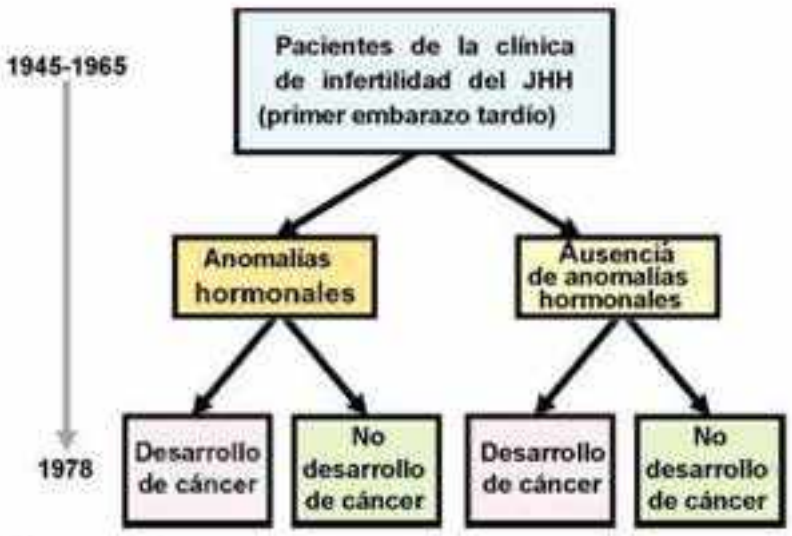


Figura 9-9. Diseño del estudio de cohortes retrospectivo de Cowan del cáncer de mama. (Datos de Cowan LD, Gordis L, Tonascia JA, et al: Breast cancer incidence in women with progesterone deficiency. Am J Epidemiol 114:209-217, 1981.)

1965. Debido a que eran pacientes de esta clínica, todas las mujeres, por definición, tenían una edad tardía en el momento del primer embarazo. En el transcurso de sus evaluaciones diagnósticas se elaboraron unos perfiles hormonales detallados para cada mujer. Por tanto, los investigadores fueron capaces de separar a las mujeres que tenían una anomalía hormonal subyacente, como la deficiencia de progesterona (expuestas), de las que no tenían dicha anomalía hormonal (no expuestas) que presentaban otra causa de infertilidad, como un problema de permeabilidad tubárica o un recuento de espermatozoides bajo del marido. A continuación, los dos grupos de mujeres se sometieron a seguimiento para detectar el desarrollo de cáncer de mama con posterioridad.

¿Cómo podrían los resultados del diseño de este estudio aclarar la relación entre la edad tardía en el momento del primer embarazo y el mayor riesgo de cáncer de mama? Si la explicación para la asociación de una edad tardía en el primer embarazo y el mayor riesgo de cáncer de mama fuese que un primer embarazo precoz protege contra el cáncer de mama, no sería de esperar que existiese ninguna diferencia en cuanto a la incidencia de cáncer de mama entre las mujeres que tienen una anomalía hormonal y las que no la tienen. Sin embargo, si la explicación del mayor riesgo de cáncer de mama es que la anomalía hormonal subyacente predispone a estas mujeres a desarrollar un cáncer de mama, sería de esperar encontrar una mayor incidencia de cáncer de mama en las mujeres con la anomalía hormonal que en aquellas sin dicha anomalía.

En el estudio se observó que, cuando se consideraba la aparición de cáncer de mama para todo el grupo, la incidencia era 1,8 veces mayor en las mujeres con anomalías hormonales que en aquellas sin tales anomalías, pero el hallazgo no era significativo desde