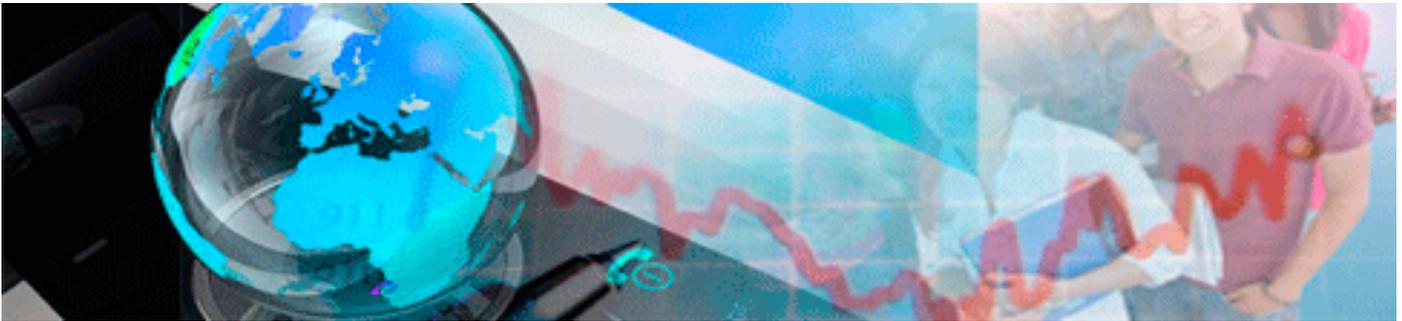


3.3 Epidemiología clínica: metodología y aplicaciones



La evaluación de la validez y la fiabilidad de las pruebas diagnósticas y de la eficacia de las intervenciones terapéuticas mediante el diseño de estudios adecuados constituye la principal tarea de la epidemiología clínica, cuyo fin último es la mejora de la calidad de la práctica clínica.

Autor: Miguel Ángel Royo Bordonada

Jefe de Área de Estudios de la Escuela Nacional de Sanidad

Se recomienda imprimir 2 páginas por hoja

Citación recomendada:

Royo-Bordonada M.A. Epidemiología clínica: metodología y aplicaciones [Internet]. Madrid: Escuela Nacional de Sanidad; 2012 [Actualizado febrero 2015]. Tema 3.3. Disponible en: dirección url del pdf.



TEXTOS DE ADMINISTRACIÓN SANITARIA Y GESTIÓN CLÍNICA
by UNED Y ESCUELA NACIONAL DE SANIDAD
is licensed under a Creative Commons
Reconocimiento- No comercial-Sin obra Derivada
3.0 Unported License.



Introducción

La epidemiología clínica surge ante la necesidad de los profesionales de la salud de disponer de evidencia científica contrastada con la que enfrentarse al contexto de incertidumbre en el que se desarrolla con frecuencia la práctica clínica. La aplicación del método científico epidemiológico al entorno clínico da respuesta a esta demanda, facilitando al profesional el uso más apropiado del arsenal diagnóstico y terapéutico disponible en cada momento.

Entre los principales impulsores de la epidemiología clínica, se han propuesto las siguientes definiciones para esta disciplina:

- Estudio de las determinantes y los efectos en las decisiones clínicas¹.
- Aplicación de los métodos biométricos epidemiológicos al estudio del diagnóstico y tratamiento por un clínico que brinda atención médica directa a pacientes².

1. Introducción

2. Evaluación de las pruebas diagnósticas

2.1. Fiabilidad

a. Índice de Concordancia Global (IC)

b. Índice Kappa (IK)

c. Coeficiente de correlación intraclass (CCI)

2.2. Validez

3. Utilización de las pruebas diagnósticas

4. Diagnóstico precoz

5. Evaluación de las intervenciones terapéuticas: Ensayo clínico

Referencias bibliográficas

- Utilización de la herramienta epidemiológica en subpoblaciones hospitalarias, con el objetivo fundamental de evaluar estrategias de diagnóstico y tratamiento³.

- Práctica de la epidemiología en la solución de problemas por los clínicos⁴.

Independientemente de la definición que escojamos, todos sus autores coinciden en que la epidemiología clínica pretende dar respuesta a las preguntas que surgen en la práctica clínica relacionadas con el diagnóstico, tratamiento, pronóstico y prevención de los pacientes, mediante el uso de un método científico contrastado que garantiza la validez y fiabilidad de las evidencias obtenidas⁵.

De esta forma, la epidemiología

clínica ha puesto los cimientos sobre los que se ha desarrollado durante las últimas décadas el fenómeno conocido como Medicina Basada en la Evidencia⁶. El reto para el profesional de la salud, en la práctica clínica diaria, consiste en enlazar de la mejor forma posible la evidencia científica obtenida a partir del planteamiento de grupo (poblacional) de la epidemiología clínica con cada paciente individual, y ahí es donde reside lo que Sackett tan acertadamente denominó la ciencia del arte de la medicina clínica⁷.

La epidemiología clínica pretende dar respuesta a las preguntas que surgen en la práctica clínica, mediante el uso de un método científico contrastado que garantiza la validez y fiabilidad de las evidencias obtenidas.

2. EVALUACIÓN DE PRUEBAS DIAGNÓSTICAS

Las pruebas diagnósticas incluyen desde la anamnesis y exploración física hasta las técnicas más modernas de diagnóstico por imagen (TAC, RMN), pasando por las pruebas de laboratorio y otro tipo de procedimientos como el electrocardiograma. A priori, podría parecer que las pruebas diagnósticas que miden parámetros biológicos son más seguras y fiables que aquellas basadas en la anamnesis, ya que estas últimas estarían sujetas a una valoración personal con un cierto componente, por así decirlo, subjetivo, mientras que las primeras tendrían un carácter más objetivo. Sin embargo, esto es pura ilusión, ya que ambos tipos de pruebas están sujetas tanto a errores aleatorios como sistemáticos. Además, con frecuencia los valores de los parámetros biológicos se obtienen a partir de una imagen o un sonido que deberán ser interpretados por la persona encargada de llevar a cabo la prueba diagnóstica.

La evaluación de una prueba diagnóstica requiere que nos fijemos en dos aspectos de la misma. Por un lado, la fiabilidad o capacidad para ofrecer resultados similares al repetir su aplicación en idénticas condiciones. Por otro lado, la validez o grado en que detecta la enfermedad o fenómeno objeto de su interés o, dicho de otra forma, la capacidad que posee de clasificar correctamente a los individuos en función de su estatus de salud-enfermedad. En la Figura 1 se observan cuatro dianas sobre las que se han realizado una serie de disparos consecutivos con el mismo arma. Asumiendo que el tirador no cometiese ningún tipo de error al apuntar y disparar, cada diana representa una situación diferente: la superior izquierda representa una medida muy válida (los disparos están en el centro de la diana) y fiable (los disparos están muy agrupados); la inferior izquierda es bastante válida (los disparos están en torno al centro de la diana) pero poco fiable (los disparos están dispersos); la superior derecha es muy fiable pero poco válida (los disparos están muy agrupados, pero alejados del centro de la diana) y la inferior derecha ni es válida ni fiable.

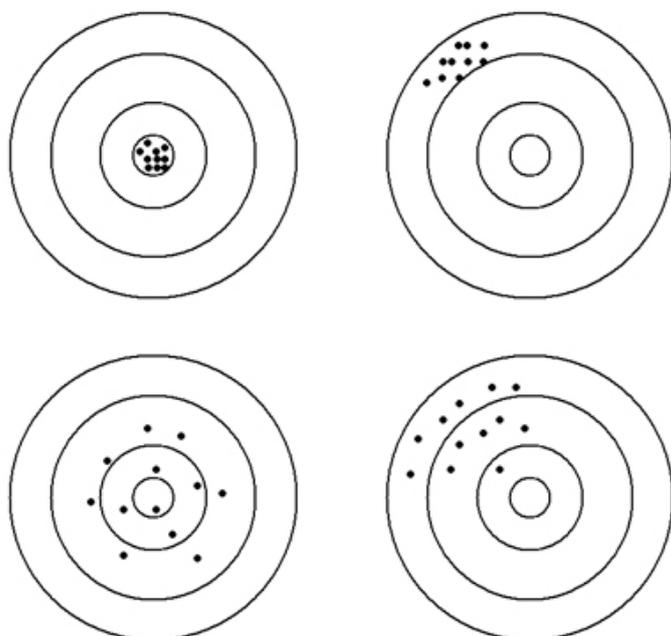


Figura 1. Evaluación de una prueba diagnóstica.

A la hora de utilizar una prueba diagnóstica en la práctica clínica tan importante es la validez como la fiabilidad de la misma, ya que la ausencia de cualquiera de ellas conlleva un error de medición que, si es lo suficientemente importante, puede alterar el resultado de la misma de forma clínicamente significativa.

2.1. Fiabilidad

La fiabilidad de una prueba diagnóstica representa su capacidad para proporcionar resultados similares al repetir su aplicación en idénticas condiciones; es decir, el grado de precisión o estabilidad conseguido en sucesivas aplicaciones de la misma. Las diversas técnicas diagnósticas de uso habitual proporcionan resultados dispares cuando las aplicamos de forma repetida en el mismo individuo. Por ejemplo, un individuo concreto que tuviese un valor real de colesterol plasmático de 200 mg/dl podría, con las técnicas diagnósticas habituales, presentar en mediciones consecutivas valores de 170 mg/dl y 230 mg/dl, cifras muy dispares y con implicaciones opuestas en términos de tratamiento. Esto es así, porque cualquier variable (biológica o no) está sometida a las siguientes fuentes de variación: la variabilidad que depende del aparato de medición, por el diferente grado de calibración

La fiabilidad de una prueba diagnóstica refleja el grado de concordancia que es capaz de proporcionar al repetir su aplicación en condiciones similares.

del mismo o por el efecto de factores ambientales como la temperatura o iluminación; la variabilidad intraobservador, al modificarse el grado de atención del profesional encargado de tomar las mediciones a causa de cansancio o condiciones ambientales, tanto físicas como psicológicas (presión ante una situación de necesidad de diagnóstico urgente); la variabilidad entre observadores, por un grado de formación previa diferente, mayor o menor agudeza visual (para discriminar una imagen radiológica) o auditiva (para auscultar un soplo), más o menos experiencia o capacidad para integrar el resto de signos y síntomas con el resultado de la prueba que se trata de medir, o un diferente nivel de motivación, entre otros posibles factores; la variabilidad biológica inherente a la característica que se desea medir (el colesterol plasmático y la tensión arterial varían a lo largo del día, el primero de forma particular en respuesta a la última comida realizada y la segunda en función de la postura, la postura -sentado, tumbado- y el nivel de estrés al que esté sujeto el individuo en el momento de la toma, como en la hipertensión de bata blanca, provocada en respuesta al simple hecho de que un profesional sanitario esté explorando a un paciente); y, por último, la variabilidad aleatoria, que representa el grado de variación que no puede atribuirse a ninguno de los factores previos y que, en consecuencia, atribuimos al azar. Dentro de este cajón desastre, resulta particularmente interesante el fenómeno de regresión a la media, que se produce siempre que utilizamos una variable diagnóstica cuantitativa para establecer el diagnóstico. En estos casos, es preciso definir un punto de corte, de forma que catalogaremos como enfermos a aquellos individuos que presenten un valor de la variable más extremo que el establecido en ese punto. Pues bien, dentro de ese subgrupo por azar habrán sido incluidos algunos individuos cuyo valor real está dentro de los límites de la normalidad y habrán quedado excluidos algunos individuos cuyo verdadero valor está fuera de estos límites. Si volvemos a medir los valores de la variable diagnóstica en el subgrupo de individuos previamente catalogados como enfermos obtendremos resultados menos extremos que en la primera ocasión o, lo que es lo mismo, más cercanos al valor medio de la variable en el grupo de partida.

La única forma de aumentar la fiabilidad de las mediciones consiste en reducir al máximo posible la variabilidad inherente a las mismas, actuando sobre cada una de las fuentes de variación mencionadas. Para ello es necesario estandarizar el

procedimiento diagnóstico y las condiciones de aplicación del mismo y entrenar por igual a los observadores para homogeneizar los criterios diagnósticos. Siempre que sea posible, todas las observaciones deberán ser recogidas por un único observador, al objeto de eliminar la variabilidad entre observadores. Por último, una forma de reducir la variabilidad aleatoria consiste en repetir la medición en varias ocasiones y tomar la media de las mismas como verdadero valor del sujeto estudiado. El "Seventh Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure"⁸ recoge todas estas recomendaciones a la hora de establecer de forma meticulosa el método a seguir para determinar la tensión arterial auscultatoria indirecta con esfigomanómetro (tabla 1).

Tabla 1. Accurate Blood Pressure Measurement in the Office

The accurate measurement of BP is the sine qua non for successful management. The equipment—whether aneroid, mercury, or electronic—should be regularly inspected and validated. The operator should be trained and regularly retrained in the standardized technique, and the patient must be properly prepared and positioned. The auscultatory method of BP measurement should be used. Persons should be seated quietly for at least 5 minutes in a chair (rather than on an exam table), with feet on the floor, and arm supported at heart level. Caffeine, exercise, and smoking should be avoided for at least 30 minutes prior to measurement. Measurement of BP in the standing position is indicated periodically, especially in those at risk for postural hypotension, prior to necessary drug dose or adding a drug, and in those who report symptoms consistent with reduced BP upon standing. An appropriately sized cuff (cuff bladder encircling at least 80 percent of the arm) should be used to ensure accuracy. At least two measurements should be made and the average recorded. For manual determinations, palpated radial pulse obliteration pressure should be used to estimate SBP—the cuff should then be inflated 20–30 mmHg above this level for the auscultatory determinations; the cuff deflation rate for auscultatory readings should be 2 mmHg per second. SBP is the point at which the first of two or more Korotkoff sounds is heard (onset of phase 1), and the disappearance of Korotkoff sound (onset of phase 5) is used to define DBP. Clinicians should provide to patients, verbally and in writing, their specific BP numbers and the BP goal of their treatment.

Tomado de National Heart, Lung, and Blood Institute (US); 2004. Report No.:04-5230⁸.

Para evaluar la fiabilidad de un procedimiento diagnóstico es preciso medir la variable en cuestión en dos o más ocasiones en cada individuo y posteriormente comparar cuán diferentes son los resultados de esas mediciones repetidas. Si lo que se pretende medir es el grado de estabilidad o consistencia del aparato de medición (precisión), intentaremos mantener constantes o sometidos al mínimo grado de variación posible el resto de factores mencionados previamente que influyen en la variabilidad del resultado. El mismo principio resulta de aplicación si lo que se pretende medir es la fiabilidad entre diferentes observadores o entre diferentes instrumentos de medida; es decir, cuando se pretende medir el grado de concordancia o equivalencia de los resultados obtenidos por los mismos. Los índices o estadísticos utilizados para medir la fiabilidad dependen del carácter de la variable que se está midiendo. Cuando se trate de variables cualitativas (enfermo-sano, leve-moderado-grave, positivo-negativo) utilizaremos el Índice de Concordancia Global y el Índice Kappa; mientras que en el caso de variables cuantitativas aplicaremos el Coeficiente de Correlación Intraclase.

a. Índice de Concordancia Global (IC)

El IC representa el porcentaje de acuerdo entre observadores (o entre mediciones repetidas por un único observador) sobre el total de observaciones realizadas. A modo de ejemplo, un panel de expertos fue invitado a revisar las historias médicas y el resto de información relevante para diagnosticar la causa de la muerte de los participantes en un ensayo clínico con estatinas y, en particular, el número de muertes debidas a enfermedad isquémica cardiaca (EIC). Los resultados de comparar los diagnósticos del panel de expertos con los del investigador del estudio se muestran en la tabla 2 (los datos entre paréntesis son los valores esperados por azar, calculados según se describe en el apartado siguiente sobre el Índice Kappa).

Tabla 2. Análisis de la concordancia entre observadores.

		Panel de Expertos		
		EIC	Otros	Total
Investigador	EIC	184 (108)	36 (112)	220
	Otros	23 (99)	180 (104)	203
	Total	207	216	423

De las 220 muertes atribuidas a la EIC por el investigador, en 36 casos el panel de expertos opinó que la causa de la muerte era diferente, mientras que de las 203 muertes atribuidas por el investigador a otras causas, 23 fueron causadas por EIC según el panel de expertos. Por tanto, $IC = (184 + 180) / 423 = 0,86$; lo que representa una concordancia del 86%.

b. Índice Kappa (IK)

En el ejemplo anterior, si en lugar de un panel de expertos hubiésemos lanzado una moneda al aire para establecer el diagnóstico de la causa de la muerte (cara=EIC y cruz=otras), en algunas ocasiones el resultado hubiese coincidido con el del investigador por puro azar. El IK permite corregir o ajustar el porcentaje de acuerdo observado (IC) por el porcentaje de acuerdo que cabría esperar exclusivamente por el efecto del azar (Ce), mediante la siguiente fórmula:

$$IK = (IC - Ce) / (1 - Ce),$$

que representa, tal como puede apreciarse en la figura 2, la concordancia existente más allá de la que se esperaría únicamente por azar. Como regla general, la fiabilidad se considera escasa si el IK es menor de 0,4; buena si se sitúa entre 0,4 y 0,75 y excelente si es superior a 0,75.

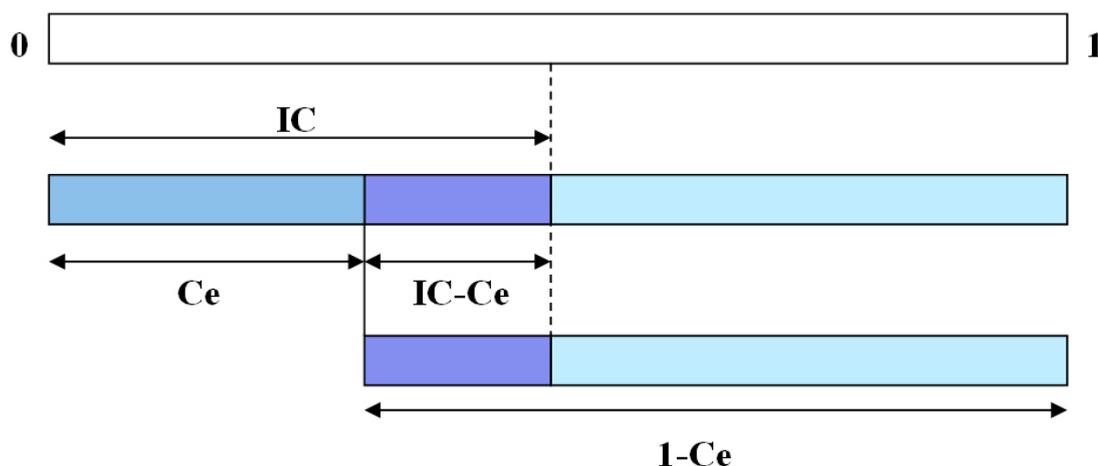


Figura 2. Índice Kappa.

Para calcular los valores esperados por azar en cada celda se multiplican los marginales de la fila y columna correspondientes

a la célula en cuestión y se divide por el total de observaciones. En nuestro ejemplo los resultados esperados son los siguientes: $220 \times 207 / 423 = 108$; $220 \times 216 / 423 = 112$; $203 \times 207 / 423 = 99$ y $203 \times 216 / 423 = 104$, aunque es suficiente con calcular el valor esperado de una de las celdas, ya que el resto pueden obtenerse a partir de las diferencias con los valores totales (marginales) de la fila o columna correspondiente. Una vez obtenidos los valores esperados, el porcentaje de concordancia debida al azar (C_e) se obtiene de forma idéntica al IC, en nuestro ejemplo $(104 + 108) / 423 = 0.5012$; es decir el 50,12%. Puesto que el IC era del 86%, el Índice Kappa resulta:

$$IK = (0.86 - 0.5012) / (1 - 0.5012) = 0.72;$$

es decir, que una vez eliminado el componente del azar, el porcentaje de acuerdo en el diagnóstico de la causa de la muerte entre el investigador del estudio y el panel de expertos fue del 72%.

c. Coeficiente de correlación intraclase (CCI)

Para cuantificar la fiabilidad de las mediciones asociadas a variables cuantitativas continuas, el índice estadístico que se deber utilizar es el CCI. Éste no debe confundirse con el coeficiente de correlación de Pearson (CCP), que refleja únicamente el grado de asociación lineal existente entre dos variables. Supongamos que multiplicamos por 10 los valores de tensión arterial obtenidos en una muestra de individuos, los nuevos valores resultantes serían absurdos ya que, por ejemplo, nadie puede tener una cifra de tensión arterial sistólica de 1.400 mmHg. El grado de equivalencia o consistencia entre los dos valores es prácticamente nulo y, sin embargo, la asociación lineal entre las dos variables es perfecta ya que con un modelo de regresión lineal podríamos predecir los resultados de la segunda variable a partir de los de la primera variable y viceversa, simplemente multiplicando o dividiendo sus valores por 10. Es decir, que mientras el CCP sería igual a 1 (el máximo posible), el CCI sería prácticamente igual a 0. El CCP ha ignorado el único aspecto que tiene importancia al evaluar la fiabilidad, la diferencia absoluta existente entre los dos valores, ya que cuanto menor es ésta mayor es el grado de concordancia entre observadores o consistencia entre mediciones repetidas.

En el ejemplo de la tabla 3, dos médicos han tomado la

tensión arterial sistólica de forma consecutiva y con el mismo esfigmomanómetro a 10 pacientes. Como puede apreciarse, las mediciones de los dos médicos son algo diferentes. Además, esta diferencia es sistemática, ya que los valores registrados por el médico B son en todos los casos superiores a los registrados por el médico A y además la diferencia es de igual magnitud casi siempre (en 8 de las 10 ocasiones la diferencia es de 5 mmHg., mientras que en las dos restantes es de 10 mmHg.). El hecho de que las diferencias vayan siempre en la misma dirección y sean de similar magnitud hace que el valor del CCP sea muy alto, concretamente de 0,95; sin embargo, esas mismas diferencias hacen que el valor del CCI sea mucho menor, concretamente de 0,64.

Tabla 3. Medición de la tensión arterial sistólica.

Paciente	Médico A (mmHg)	Médico B (mmHg)	Diferencia (B – A)
1	135	140	5
2	140	145	5
3	130	135	5
4	145	150	5
5	140	145	5
6	150	160	10
7	140	145	5
8	135	140	5
9	140	145	5
10	135	145	10
Media	139	145	6
DE	5,68	6,67	2,11

El cálculo del CCI se basa en un modelo de análisis de la varianza (de medidas repetidas). De forma simplificada, la idea consiste en descomponer la variación total observada (varianza) en los

siguientes componentes: la variabilidad debida a las diferencias entre pacientes o individuos observados (Varianza entre-individuos = σ_E); la variabilidad debida a las diferencias entre diferentes aparatos de medición o entre mediciones repetidas realizadas por el mismo observador o, como en este caso, entre diferentes observadores (Varianza intra-individuos = σ_I) y, por último la variabilidad residual, que es la parte de la variabilidad inexplicable y que atribuimos al azar (Varianza residual = σ_R). El CCI se define como el cociente entre la variabilidad entre pacientes y la variabilidad total,

$$\text{CCI} = \sigma_E / (\sigma_E + \sigma_I + \sigma_R)$$

Cuando un aparato de medición sea muy fiable o cuando la concordancia entre dos observadores sea muy alta, los resultados de las mediciones repetidas sobre un mismo individuo serán muy parecidas y, en consecuencia, la varianza intra-individuos será muy baja y el valor del cociente anterior aumentará. Cuanto más se aproxime a 0 el valor de la varianza intra-individuos, más se acercará a 1 el valor del CCI y viceversa. Puesto que es una proporción, los valores del CCI pueden variar de 0 a 1, donde como ya hemos mostrado 0 indica ausencia de concordancia y 1 concordancia o fiabilidad absoluta. Como norma general, valores inferiores a 0,4 indican escasa fiabilidad, entre 0,4 y 0,75 la fiabilidad es buena y valores superiores a 0,75 reflejan un grado excelente de fiabilidad.

2.2. Validez

La validez de una prueba diagnóstica es el grado en que mide lo que realmente trata de medir; es decir, la capacidad de clasificar correctamente a los individuos en función de su estatus de salud-enfermedad. Una prueba válida es aquella que carece de errores sistemáticos o sesgos (ej: un esfigmomanómetro que tuviese un defecto de fabricación por el que siempre midiese una tensión arterial 10 mmHg por debajo del valor real). Para estudiar la validez comparamos los resultados de la prueba bajo estudio con el verdadero estatus del individuo, o en su defecto, con los resultados del mejor instrumento disponible de medida del fenómeno bajo estudio. Esta prueba de referencia se denomina estándar de oro. Así, una prueba será válida cuando clasifique adecuadamente tanto a enfermos como a sanos o cuando sus

resultados coincidan o tengan un alto grado de concordancia con los del estándar de oro.

La validez tiene dos componentes principales, la sensibilidad (capacidad de diagnosticar correctamente a las personas que padecen la enfermedad bajo estudio) y la especificidad (capacidad de diagnosticar correctamente a las personas que no padecen la enfermedad bajo estudio). Vamos a ilustrar de estos conceptos con un ejemplo. El servicio médico de una empresa de transporte de sustancias peligrosas está valorando la implantación de una nueva prueba para el diagnóstico rápido en orina del posible consumo de cocaína entre sus conductores. Para ello, han llevado a cabo un estudio de validez de esta prueba diagnóstica sobre una muestra de 1500 individuos, 45 de los cuales eran consumidores de cocaína, con los resultados que se presentan a continuación:

La validez de una prueba diagnóstica es el grado en que clasifica correctamente a los individuos en función de su estatus de salud-enfermedad.

Tabla 4. Tabla dos por dos para la evaluación de la validez de pruebas diagnósticas

	Consumo de cocaína		Total
	SI	NO	
Resultado de la prueba			
Positivo	23 (VP)	44 (FP)	67
Negativo	22 (FN)	1411 (VN)	1433
Total	45	1455	1500

Como se aprecia en la tabla, nos podemos encontrar con cuatro posibles situaciones: Verdadero Positivo (VP), cuando el individuo padece la condición bajo estudio (consumidor de cocaína) y la prueba da positivo; Falso Negativo (FN), cuando el individuo padece la condición y la prueba da negativo; Verdadero Negativo (VN), cuando el individuo no padece la condición y la prueba da negativo; y Falso Positivo (FP), cuando el individuo no padece la condición y la prueba da positivo.

a. Sensibilidad (S)

La sensibilidad es la probabilidad de que un sujeto enfermo sea

clasificado como tal y se calcula como el porcentaje de sujetos enfermos que dan positivo al aplicarles la prueba diagnóstica. En nuestro ejemplo, $S = 23 / 45 = 0,511$; es decir, el 51% de los consumidores son detectados por el nuevo test.

b. Especificidad (E)

La especificidad es la probabilidad de que un sujeto no enfermo sea clasificado como tal y se calcula como el porcentaje de sujetos no enfermos que dan negativo al aplicarles la prueba diagnóstica. En nuestro ejemplo, $E = 1411 / 1455 = 0,9698$; es decir, el 97% de los que no consumen cocaína dan negativo en el nuevo test de detección.

c. Razón de Verosimilitud (RV)

Una prueba válida es aquella que posee una sensibilidad y especificidad elevadas. Pero la validez también puede expresarse mediante un único indicador, la RV. La RV de un resultado positivo (RV+) es la probabilidad de un resultado positivo en los enfermos dividido por la probabilidad de un resultado positivo en los no enfermos: $\text{Sensibilidad} / (1 - \text{Especificidad})$. La RV+ indica cuanto más probable es un resultado positivo en un individuo enfermo que en uno no enfermo. Si la prueba no tiene capacidad de discriminación alguna, la probabilidad del resultado positivo es igual tanto si el enfermo padece la enfermedad como si no y el valor del RV+ es igual a 1. Cuanto mayor de 1 sea la RV+ mejor discrimina la prueba.

La RV de un resultado negativo (RV-) es la probabilidad de un resultado negativo en los enfermos dividido por la probabilidad de un resultado negativo en los no enfermos: $(1 - \text{Sensibilidad}) / \text{Especificidad}$. La RV- indica cuanto menos probable es un resultado negativo en un individuo enfermo que en uno no enfermo. Cuanto menor de 1 sea la RV- mejor discrimina la prueba.

d. Curvas ROC

La determinación de la sensibilidad y la especificidad de una prueba diagnóstica requiere que trabajemos con variables dicotómicas (enfermo-no enfermo y positivo-negativo). Cuando

trabajamos con variables cuantitativas (ej.: mmHg) es preciso definir un punto de corte a partir del cual vamos a considerar que estamos ante una situación patológica y por lo tanto vamos a otorgar un resultado positivo a la prueba diagnóstica y considerar ese individuo como enfermo. Sin embargo, habitualmente no existe una cifra mágica que permita separar con precisión y exactitud los individuos enfermos de los no enfermos y, aunque existiese, los errores de medición (sistemáticos y aleatorios) provocarían una asignación errónea (falsos positivos o negativos) en un porcentaje determinado de casos. Por ej., supongamos que utilizamos como punto de corte para el diagnóstico de diabetes una cifra de 100 mg/dl de glucosa en sangre. En este caso, la sensibilidad será muy alta: todos los diabéticos serán diagnosticados ya que es muy poco probable que tengan un nivel de glucosa inferior a este límite. Sin embargo, la especificidad será muy baja: muchos individuos no diabéticos pueden tener glucemias superiores a los 100 mg/dl, con lo que aparecerán muchos falsos positivos. La situación opuesta nos encontraríamos en caso de elegir como punto de corte 140 mg/dl. Puesto que no existe un punto de corte con sensibilidad y especificidad del 100%, para seleccionar el mejor punto de corte posible se ha desarrollado un método denominado análisis de curvas de rendimiento diagnóstico (curvas ROC). Tal como se aprecia en la Figura 3, para representar una curva ROC utilizamos un eje de coordenadas, con la sensibilidad en el eje Y y el complementario de la especificidad en el eje X.

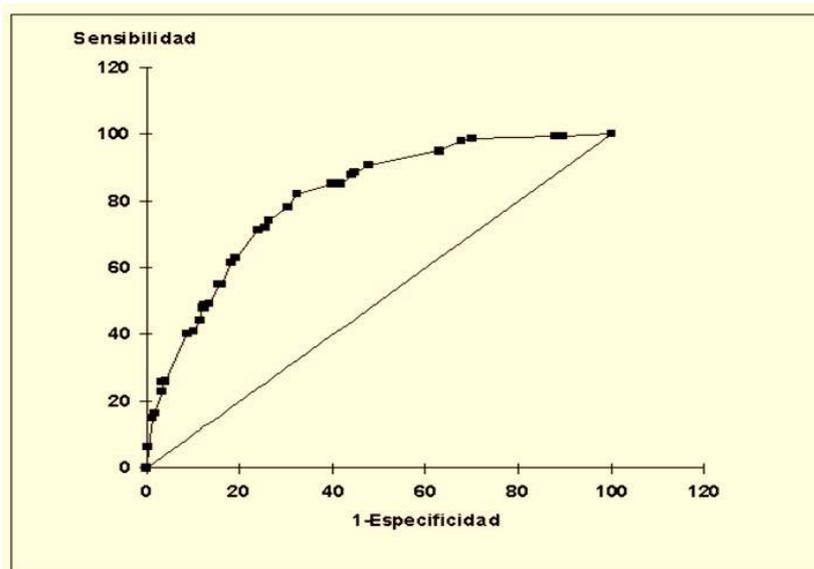


Figura 6. Curva ROC

Una prueba diagnóstica no válida es aquella cuyo trazado coincide con la diagonal. La $RV+$ y la $RV-$ en dichos puntos es igual a 1: es tan probable un resultado positivo en enfermos como no enfermos y lo mismo puede decirse de un resultado negativo. En el área por encima de la diagonal, la $RV+$ es mayor de 1 y la $RV-$ es menor de 1. La validez es mayor cuanto mayor es el área bajo la curva; es decir, cuanto más se aproxima la curva a una L invertida (ángulo superior izquierdo) mejor discrimina la prueba.

La elección del punto de corte más apropiado para decidir a partir de donde se considera positivo el resultado de la prueba depende de las consecuencias que se deriven del diagnóstico de falsos negativos y positivos. Si los costes de diagnosticar un falso positivo son claramente mayores que los de diagnosticar un falso negativo (tratamiento muy agresivo y poco beneficioso para los enfermos) interesa una mayor especificidad y el punto de corte se situará en la parte baja de la curva. Por el contrario, si el tratamiento es muy beneficioso y con escasos efectos adversos, el coste de dejar a un enfermo sin tratar (falso negativo) es muy elevado, por lo que buscaremos máxima sensibilidad y colocaremos el punto de corte en la parte alta de la curva. En esta decisión también hay que tener en cuenta que el número absoluto de falsos positivos y falsos negativos también depende de la prevalencia de la enfermedad. Si la prevalencia es baja, en la población predominan los individuos no enfermos, por lo que el número absoluto de falsos negativos será necesariamente bajo, mientras que el número de falsos positivos podrá ser muy alto. Para tratar de minimizar este posible problema buscaremos que la prueba diagnóstica tenga una elevada especificidad y por lo tanto elegiremos un punto de corte en la parte baja de la curva. Por el contrario, si la prevalencia de la enfermedad es alta interesará una elevada sensibilidad y situaremos el punto de corte en la parte alta de la curva.

2.3 Probabilidad posterior

En la práctica clínica diaria, las pruebas diagnósticas se utilizan como una herramienta de ayuda en la decisión terapéutica. Cuando aplicamos una prueba diagnóstica a un paciente concreto estamos tratando de confirmar o descartar una sospecha

diagnóstica o, mejor dicho, intentamos aumentar o disminuir esa sospecha en un grado que nos permita tomar una decisión terapéutica o, en su caso, optar por realizar nuevas pruebas diagnósticas cuyos resultados ayuden a decir si aplicamos o no un tratamiento determinado. Por ejemplo, en un individuo con hipertensión arterial (HTA) de grados 2 y 3 está indicado el tratamiento farmacológico; pero en la HTA de grado 1, las guías de prevención cardiovascular contemplan considerar la intervención farmacológica únicamente cuando el riesgo cardiovascular global es alto ($>5\%$), por lo que en este caso sería necesario realizar otras pruebas, como medición del colesterol plasmático, para poder calcular el riesgo cardiovascular y tomar una decisión terapéutica en función del mismo⁹.

La sospecha diagnóstica se basa en la estimación del riesgo de padecer la enfermedad que presenta un sujeto que acude a la consulta en función de su historial clínico, los signos y síntomas que refiere y el resultado de cuantas pruebas diagnósticas se le hayan hecho hasta ese momento en relación con la posible enfermedad bajo estudio. Este riesgo es lo que se denomina probabilidad previa (al resultado de la/s prueba/s diagnóstica/s que se esté pensando aplicar a continuación). Si la probabilidad previa es muy alta (como en la HTA de grado 3), el médico tendrá la suficiente certeza (escasa incertidumbre) acerca del riesgo de padecer la enfermedad y, en consecuencia, decidirá aplicar el tratamiento sin necesidad de acudir a nuevos procedimientos diagnósticos. En caso contrario (HTA de grado 1), aplicará nuevas pruebas diagnósticas cuyo resultado permita modificar la estimación del riesgo de padecer la enfermedad en grado suficiente para poder tomar una decisión terapéutica. Este riesgo es lo que denominamos probabilidad posterior (al resultado de la/s prueba/s diagnóstica/s que acabamos de aplicar). En función del resultado de la prueba diagnóstica hablaremos de la probabilidad posterior a un resultado positivo, el Valor Predictivo Positivo, o de la probabilidad posterior a un resultado negativo, el Valor Predictivo Negativo. Por tanto, los valores predictivos, aunque en parte dependen de la sensibilidad y la especificidad del procedimiento diagnóstico en cuestión, no son características inherentes al mismo, ya que también dependen de la probabilidad previa que, en caso de no tener ningún dato adicional sobre el individuo bajo estudio, se corresponde con la prevalencia de la misma en población general.

a. Valor Predictivo Positivo (VPP)

El VPP de una prueba diagnóstica es la tasa de verdaderos positivos o, dicho de otro modo, la probabilidad de que un sujeto positivo esté realmente enfermo, y se calcula como el porcentaje de sujetos positivos que padecen la enfermedad. En el ejemplo del test de análisis de metabolitos de cocaína en orina, $VPP = 23 / 67 = 0,343$; es decir, el 34,3% de los que han dado positivo a la prueba diagnóstica consumen cocaína.

b. Valor Predictivo Negativo (VPN)

El VPN de una prueba diagnóstica es la tasa de verdaderos negativos o, dicho de otro modo, la probabilidad de que un sujeto negativo esté realmente libre de la enfermedad bajo estudio, y se calcula como el porcentaje de sujetos negativos que no sufren la enfermedad. En el ejemplo del test de análisis de metabolitos de cocaína en orina, $VPN = 1411 / 1433 = 0,985$; es decir, el 98,5% de los que han dado negativo a la prueba diagnóstica no consumen cocaína.

En un sujeto con una probabilidad previa alta de padecer la enfermedad, un resultado negativo de la prueba diagnóstica puede hacer modificar la actitud terapéutica hacia la decisión de no tratar; pero el médico necesita tener seguridad de que ese resultado no es un falso negativo, y esta certeza sólo puede tenerse si la sensibilidad de la prueba es muy alta (tabla 4). De forma similar, en un paciente con una probabilidad previa baja de padecer la enfermedad, un resultado positivo de la prueba diagnóstica puede hacer modificar la actitud terapéutica hacia la decisión de tratar; pero el médico deseará tener la mayor seguridad posible de que ese resultado no es un falso positivo, y en este caso esa certeza sólo puede obtenerse si la especificidad de la prueba es muy alta. De todo lo anterior se deduce que el VPN depende sobre todo de la especificidad de la prueba diagnóstica y que está influenciada de mayor magnitud y relevancia cuando la probabilidad previa es baja. Asimismo, el VPP depende sobre todo de la sensibilidad de la prueba diagnóstica y está influenciada de mayor magnitud y relevancia cuando la probabilidad previa es elevada.

3. UTILIZACIÓN DE LAS PRUEBAS DIAGNÓSTICAS

Una prueba diagnóstica es tanto más eficaz cuanto mayor cambio provoca en la probabilidad de padecer la enfermedad, dado que la magnitud de este cambio es la que determina el posible cambio en nuestra actitud terapéutica (tratar o no). Dado que una prueba diagnóstica siempre conlleva costes (económicos, efectos secundarios, molestias para el paciente y su entorno, etc.), sólo se debe aplicar si la información que aporta puede cambiar la actitud terapéutica ante el paciente. Por otro lado, en la práctica clínica no existe ninguna prueba con sensibilidad y/o especificidad del 100%, por lo que el médico siempre se ve obligado a tomar decisiones con cierto grado de incertidumbre. Un error en el diagnóstico se sigue de un error en el tratamiento, bien por tratar a un individuo no enfermo (FP), bien por dejar de administrar el tratamiento a un individuo enfermo (FN). Por ello, a la hora de decidir si aplicamos una prueba diagnóstica determinada o al seleccionar un procedimiento diagnóstico entre una batería posible de los mismos, hemos de tener en cuenta la gravedad de las consecuencias derivadas tanto de diagnosticar y tratar a un individuo sin la enfermedad como de no tratar a un individuo que padece la enfermedad.

En este contexto y como regla general, las pruebas de alta sensibilidad están especialmente indicadas en las siguientes situaciones:

- El precio que hay que pagar por no detectar la enfermedad es muy alto. Esto se produce ante la sospecha de procesos peligrosos pero para los que existe un tratamiento eficaz, como la tuberculosis, la sífilis o el linfoma de Hodking.
- Para descartar hipótesis diagnósticas en las primeras etapas de un estudio. Una prueba muy sensible rara vez es negativa en presencia de la enfermedad (da pocos FN), por lo que un resultado negativo prácticamente elimina la posibilidad de que se padezca la enfermedad.
- Para detectar los casos cuando la probabilidad de la enfermedad es relativamente baja. Este es el caso de los programas de cribado poblacional, donde el propósito de la prueba es descubrir la enfermedad en personas sin síntomas de la misma.

De forma paralela, es conveniente utilizar pruebas de alta especificidad en las siguientes situaciones:

- Cuando se precisa verificar la presencia de una enfermedad (confirmación de un diagnóstico), que ha sido sugerida por otros datos. Esto se debe a que una prueba muy específica rara vez es positiva en ausencia de la enfermedad (da pocos FP), por lo que un resultado positivo prácticamente nos garantiza que el paciente padece la enfermedad.
- Cuando el diagnóstico y/o tratamiento incorrectos (FP) tienen graves consecuencias para el paciente, bien porque se trata de enfermedades estigmatizantes, como el SIDA, y/o de mal pronóstico, bien porque el tratamiento tiene importantes efectos secundarios sobre la salud o sobre el bienestar económico, emocional y/o social del paciente, como en el caso de la quimioterapia para el cáncer.

Lo ideal sería contar con una prueba diagnóstica con sensibilidad y especificidad elevadas, pero esto no siempre es posible. La elección entre una prueba u otra debe basarse en un análisis de riesgos y beneficios derivados de diagnosticar y tratar a un porcentaje de individuos sin enfermedad y viceversa; teniendo en cuenta que para una sensibilidad y especificidad dadas, la probabilidad previa de la enfermedad será la que determine el mayor o menor número de FP y FN que vamos a obtener.

4. DIAGNÓSTICO PRECOZ

El diagnóstico precoz consiste en el uso de pruebas diagnósticas para la detección de enfermedades en pacientes asintomáticos; es decir, en pacientes que todavía no han desarrollado la fase clínica de la enfermedad. El criterio básico para aplicar una prueba diagnóstica en este caso es el mismo que cuando nos enfrentamos ante un paciente sintomático: sólo se debe aplicar si la información que aporta puede cambiar la actitud terapéutica ante el paciente. Además, para llevar a cabo un programa de diagnóstico precoz es preciso que se cumplan una serie de condiciones que afectan a la enfermedad, a la prueba diagnóstica y al propio programa. Existen dos estrategias principales de diagnóstico precoz:

- a. Cribado poblacional: En este tipo de programas se invita a toda la población susceptible a someterse a la prueba diagnóstica en cuestión. Ej.: cribado de cáncer de mama a través de mamografía en mujeres mayores de 50 años.
- b. Búsqueda activa de casos: Aprovechando la visita al médico por otras causas, se invita al paciente a someterse a determinadas pruebas diagnósticas destinadas a detectar patologías no sospechadas por la clínica. Ej.: búsqueda activa de individuos con hipertensión arterial o diabetes no diagnosticadas.

El objetivo de un programa de diagnóstico precoz consiste en reducir la morbi-mortalidad y aumentar la calidad de vida de los pacientes a través de un diagnóstico y tratamiento precoces de la enfermedad. Para que un programa de diagnóstico precoz pueda cumplir su objetivo, la enfermedad debe ser susceptible de ser abordada de forma efectiva, para lo cual el programa debe cumplir una serie de requisitos. En primer lugar, es preciso que exista una fase preclínica durante la cual la enfermedad es detectable. En segundo lugar, es necesario que exista un tratamiento eficaz para la misma. Y en tercer lugar, es imprescindible que exista un punto crítico en el curso de la enfermedad, situado entre el diagnóstico precoz posible y el diagnóstico clínico habitual; de modo que el tratamiento antes de alcanzar este punto crítico afecte de forma favorable el curso del proceso, por ser más fácil de aplicar y/o más eficaz que durante la fase clínica de la enfermedad.

Además, cuanto mayor sean la gravedad y la prevalencia de la enfermedad en fase preclínica detectable mayor será el rendimiento de la prueba diagnóstica y más eficiente el programa de diagnóstico precoz. Por lo que se refiere a la prueba diagnóstica debe ser válida, fiable, fácil de aplicar, barata, con pocos efectos secundarios, aceptable por la población a la que va dirigida y con una sensibilidad muy alta (próxima o igual al 100%), de forma que permita detectar a prácticamente todos los casos existentes en fase preclínica. Una estrategia utilizada con frecuencia en los programas de diagnóstico precoz consiste en la combinación de pruebas diagnósticas. Al combinar dos pruebas en serie, primero aplicaremos una prueba muy sensible que nos permitirá detectar casi todos los casos presentes en la población, aunque a costa

Los programas de diagnóstico precoz son diseñados para tratar de reducir la morbi-mortalidad de la población mediante el diagnóstico y tratamiento precoces de aquellas enfermedades que cumplan los requisitos de aplicación.

de muchos falsos positivos. Los que obtengan un resultado negativo con esta primera prueba los declaramos sanos. A continuación, aplicaremos una prueba muy específica con la que conseguiremos descartar los falsos positivos, que ahora darán mayoritariamente un resultado negativo y serán declarados sanos, quedándonos con los verdaderos enfermos a los que aplicaremos el tratamiento, aquéllos que también han dado positivo en esta segunda prueba. Por último, el propio programa de diagnóstico precoz también tiene que cumplir una serie de requisitos: debe ser factible desde todos los puntos de vista, tanto económicamente (coste-efectivo) como por cuestiones de accesibilidad de la población diana o cualesquiera otros aspectos que puedan poner en peligro su adecuado desarrollo. Resulta particularmente importante que exista disposición y recursos para atender a todos los pacientes que sean diagnosticados por medido del programa. Entre los programas de cribado poblacional que actualmente se aplican en nuestro entorno cabe destacar el programa de detección de la fenilcetonuria y el de cáncer de cervix.

4.1. Cribado de cáncer de mama mediante mamografía

La mejor evidencia para apoyar el cribado la proporcionan los ensayos clínicos aleatorizados en los que la intervención es el cribado y las variables de resultado son la disminución de la morbimortalidad atribuida a la condición que se quiere cribar. En el caso del cáncer de mama, diversos ensayos clínicos han mostrado una reducción del 20 al 30% de la mortalidad por cáncer de mama en mujeres de 50 a 64 años de edad. No obstante, dos de estos ensayos, seleccionados en función de criterios de calidad, observaron un efecto nulo de la mamografía. En cuanto a las evidencias sobre la efectividad del cribado de cáncer de mamografía en los países donde se han instaurado estos programas, éstas son en algunos casos prometedoras, pero todavía muy limitadas; aunque se reconoce que los programas de cribado han tenido un impacto positivo en la organización global de la asistencia al cáncer de mama. Por otro lado, la mamografía implica ciertas molestias y riesgos por irradiación. Además, solo entre un 5 y un 20% de las mamografías positivas son confirmadas como cáncer de mama, por lo que entre el 80 y el 95% de las mujeres con mamografía positiva se ven sometidas a pruebas diagnósticas adicionales

(mamografías adicionales, ecografía, citología, punción, biopsia) que, exponiéndolas a nuevos riesgos, finalmente no conducen al diagnóstico de cáncer de mama. También se ha detectado sobre-diagnóstico y tratamiento con cirugía y/o radioterapia innecesario⁵. En resumen, el conjunto de la evidencia científica es con frecuencia incapaz de dar una respuesta consistente a todos los interrogantes planteados cuando se valora la introducción de un programa de cribado poblacional y, en el caso del cribado del cáncer de mama con mamografía, ampliamente aceptado en mujeres > 50 años e implantado en muchos países de nuestro entorno, la controversia continua¹⁰.

5. EVALUACIÓN DE LAS INTERVENCIONES TERAPÉUTICAS: ENSAYO CLÍNICO

Un ensayo clínico es un experimento planificado en el que, de forma prospectiva, se comparan dos o más intervenciones (o intervención y placebo) preventivas, curativas o rehabilitadoras, que son asignadas de forma individualizada y aleatoria a un grupo de pacientes para determinar su eficacia¹¹. Además, tanto la selección de los sujetos como los períodos de tratamiento y seguimiento han de tener lugar simultáneamente en todos los grupos. Dado que los participantes en un ensayo clínico son, por definición, sujetos enfermos (pacientes), las intervenciones preventivas a las que nos referimos son secundarias, dirigidas a evitar las secuelas de la enfermedad, incluyendo la muerte. Por ejemplo, una dieta rica en fibra para prevenir la recurrencia de pólipos adenomatosos de colón o el uso de estatinas para reducir la mortalidad post-infarto de miocardio. No serán objeto de nuestro estudio en este tema los ensayos cuya finalidad consiste en evaluar la eficacia de intervenciones preventivas en sujetos sanos (ensayos de prevención primaria), denominados ensayos de campo, cuando la intervención se aplica a nivel individual, o ensayos de intervención comunitaria, cuando la intervención se aplica a nivel grupal (escuela, empresa, familia, etc.).

Los ensayos clínicos evalúan las intervenciones en condiciones muy diferentes a las de su aplicación habitual, ya que los pacientes están sometidos a un estrecho seguimiento y vigilancia médica y no suelen ser representativos de todo el espectro de posibles sujetos que padecen la enfermedad en cuestión. Por lo tanto, un ensayo está evaluando el efecto de la intervención en condiciones

El ensayo clínico controlado y aleatorizado constituye la herramienta estándar para evaluar la eficacia de las intervenciones terapéuticas.

ideales (eficacia). Sin embargo, no puede evaluar como se comportará esa misma intervención en condiciones reales de uso (efectividad) ni su relación coste-beneficio (eficiencia), aspectos que se contemplan en el marco de la investigación de resultados. Los pacientes seleccionados para realizar un ensayo representan habitualmente al grupo más accesible al equipo investigador, con menos patologías asociadas y con mayores probabilidades de cumplir con el protocolo del estudio. Además, los pacientes que aceptan participar son en general sujetos motivados y más concienciados con su estado de salud, por lo que es probable que el grado de adhesión a una terapia concreta o a una modificación dietética sea mayor del que observaríamos en el resto de pacientes. Por otro lado, la estrecha vigilancia a la que están sometidos estos pacientes mientras dura el estudio hace que los incumplimientos o abandonos del tratamiento sean menos probables que en condiciones habituales.

5.1. Ensayo clínico de grupos paralelos

El ensayo de grupos paralelos constituye, por su sencillez, el diseño básico para realizar un ensayo clínico. En este modelo, cada paciente es asignado al mismo tiempo y de forma aleatoria a una intervención diferente y tras un período de seguimiento variable, se procede a medir la respuesta (Figura 7).

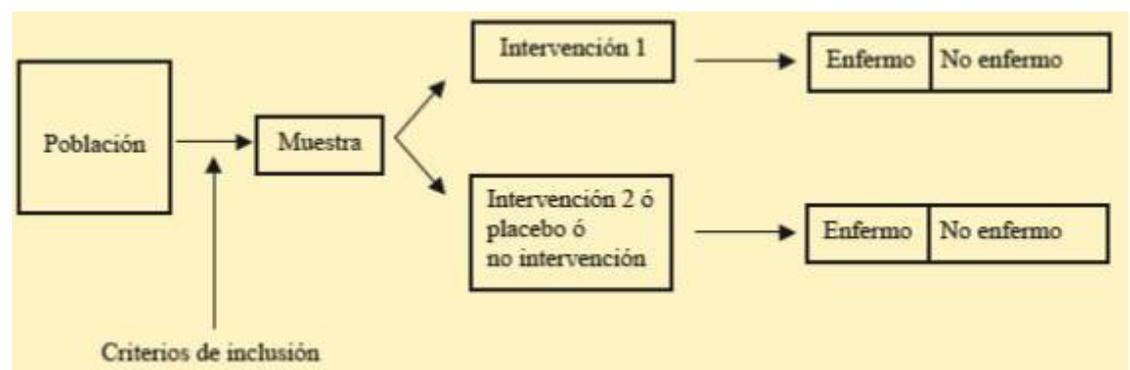


Figura 7. Ensayo clínico de grupos paralelos.

En general, se incluyen dos grupos de comparación, dada la dificultad de reclutar suficiente número de pacientes por grupo cuando se aumenta el número de tratamientos a comparar. A veces, este diseño puede contemplar un período de preinclusión previo a la aleatorización, que tiene por objeto disminuir las

tasas de no cumplimiento de la intervención o de abandono del estudio y, en ocasiones, estabilizar los valores de las variables a través de las cuales se va a medir la respuesta. Durante el período de preinclusión, todos los individuos son sometidos a la misma intervención, habitualmente la del grupo control, excluyendo del estudio a los sujetos no cumplidores. Mensink y cols. realizaron un ensayo clínico de grupos paralelos para comparar el efecto de los ácidos grasos monoinsaturados y poliinsaturados sobre el perfil lipídico plasmático. Para ello, establecieron un periodo de preinclusión durante el que los sujetos fueron asignados a una dieta rica en ácidos grasos saturados, permitiendo valorar el grado de cumplimiento de la dieta y estabilizar los lípidos sanguíneos. Una vez finalizado este periodo los sujetos cumplidores fueron asignados de forma aleatoria a una dieta rica en ácidos grasos mono o poliinsaturados respectivamente¹².

5.2. Aleatorización

La aleatorización consiste en la asignación de las unidades experimentales (sujetos participantes) a dos o más intervenciones utilizando un mecanismo aleatorio de forma que ni el investigador ni el sujeto investigado conozcan la intervención que va a ser asignada en el momento de la inclusión de éste último en el estudio, una vez obtenido su consentimiento informado para participar en el estudio¹³. Aunque un estudio aleatorizado es necesariamente controlado, el grupo control podrá recibir una intervención activa, que habitualmente será la utilizada en condiciones normales para la condición bajo estudio, inactiva e inocua (placebo) o simplemente no recibir intervención alguna. La aleatorización elimina el sesgo del investigador en la asignación de los sujetos, garantiza la validez de los niveles de significación de las pruebas estadísticas que se aplicarán en el análisis de los datos y tiende a producir grupos comparables respecto de características conocidas y desconocidas que pudieran afectar al desenlace de interés (factores pronósticos). Frente a otros métodos de control de factores de confusión utilizados en las fases de diseño (muestreo estratificado, emparejamiento) o análisis (ajuste de tasas, regresión múltiple), la aleatorización presenta la gran ventaja de que controla tanto factores conocidos como desconocidos.

La eficacia de la aleatorización depende del tamaño muestral del estudio. Cuando éste no es lo suficientemente grande -a modo

La aleatorización garantiza que el efecto observado es atribuible a la intervención bajo estudio

de orientación estableceremos el punto de corte en 100 sujetos aunque esto también depende del número de grupos analizados-no se puede garantizar el equilibrio en el tamaño de los grupos ni la distribución homogénea de los factores de confusión. Por ello, a la hora de presentar los resultados de un ensayo el primer paso debe consistir en mostrar una tabla con la distribución de los factores pronósticos conocidos en los grupos de intervención y control. Una distribución homogénea del número de sujetos y de los factores conocidos en ambos grupos constituye una prueba de que la aleatorización ha sido eficaz y nos indica que, con toda probabilidad, los grupos también serán comparables respecto de posibles factores pronósticos que por ser desconocidos no se hayan podido medir, permitiendo atribuir con seguridad el efecto observado a la intervención bajo estudio.

Un buen método de aleatorización debe de cumplir los siguientes requisitos para garantizar la calidad del mismo y la eficacia del proceso:

- a. La asignación tiene que ser ciega, es decir, desconocida por el investigador hasta el momento en que el sujeto da su autorización para participar en el estudio.
- b. La secuencia de asignación debe resultar impredecible, es decir, el investigador debería ser incapaz de adivinar con antelación la intervención a la que va a ser asignado el próximo sujeto.
- c. El proceso debe estar basado en propiedades matemáticas conocidas y ser reproducible.

5.3 Selección de los participantes en un ensayo clínico

La selección de los sujetos que participarán en el estudio debe llevarse a cabo de forma cuidadosa. Para ello, hay que empezar por definir de forma precisa el perfil de los sujetos susceptibles de participar en el estudio, por ej.: sujetos de ambos sexos, de 25 a 40 años de edad, con sobrepeso (IMC entre 25 y 30) y residentes en Madrid. A continuación se deben definir los criterios de exclusión, en función de aspectos que pudieran comprometer la participación en el estudio (ej.: enfermedades incapacitantes física o mentalmente) o que pudieran interaccionar con la

intervención bajo estudio y afectar la variable de interés. En cualquier caso, hemos de intentar ser lo menos restrictivos posible a la hora de establecer los criterios de inclusión y exclusión, al objeto de evitar que tengamos problemas para reclutar un número apropiado de sujetos.

5.4. Enmascaramiento

El enmascaramiento, o carácter ciego de un ensayo clínico, es el conjunto de medidas o precauciones que se toman con el fin de que los intervinientes en el mismo (paciente, médico, evaluador, ..) desconozcan el tratamiento al que el paciente ha sido asignado, ya que este conocimiento podría introducir sesgos en el desarrollo, análisis y evaluación de un ensayo clínico. La presentación de las intervenciones enmascaradas deberá ser idéntica, diferenciándose únicamente por la sustancia activa que contienen. Hay que tener en cuenta varios aspectos que tienen que ser idénticos en ambas intervenciones: vía y pauta de administración, forma, tamaño, color, sabor, textura, etc.. El enmascaramiento del responsable de evaluar la respuesta (paciente, médico o un tercero) a la intervención es particularmente importante cuando la variable resultado es subjetiva (ej: valoración de la intensidad del dolor, análisis de una radiografía, ..), ya que el conocimiento de la intervención asignada puede afectar la percepción sobre la variable respuesta. Los resultados más objetivos, como la muerte, son relativamente insensibles a este tipo de sesgo.

Lamentablemente, no siempre es posible aplicar el enmascaramiento, por ej. en el caso de que la intervención sea una dieta o una intervención quirúrgica. Por otro lado, en ocasiones un tratamiento puede tener efectos secundarios bien conocidos que pueden permitir al paciente realizar la identificación del mismo. Por ej.: la rifampicina tiñe de rojo la orina. Además, nunca deben enmascarse los tratamientos si por ello sometemos a algún tipo de riesgo a los pacientes. Por todo ello, para poder hacer una valoración lo más objetiva y realista posible de los resultados del ensayo clínico en relación con la eficacia de la intervención, el investigador debe describir no solo el método de enmascaramiento y las violaciones del mismo, sino también los aspectos comentados previamente, en relación con la imposibilidad de aplicarlo o los posibles riesgos de que no sea eficaz.

5.5. El placebo

La administración de una intervención produce un efecto terapéutico sobre el individuo, denominado efecto placebo, independientemente del principio activo que contenga la misma. Es decir, incluso la administración de un principio inactivo inocuo produce un efecto medible. El simple hecho de ser observado, de participar en un estudio, produce una modificación positiva en el estado clínico de los pacientes, denominada "efecto Hawthorne". A esto hay que añadir el poder de sugestión derivado del hecho de recibir una terapia, aunque ésta sea farmacológicamente inactiva, y el de la presencia de un médico, o investigador, que aplica la misma.

El efecto global de cualquier intervención tiene dos componentes: el efecto placebo y el efecto específico de la misma. Por ello, al estudiar la eficacia de una nueva terapia, es preciso aplicar una intervención alternativa al grupo control, que permita controlar el efecto placebo y discernir el efecto específico de la nueva terapia. Por razones éticas, cuando exista una terapia eficaz disponible para la condición bajo estudio no aplicaremos un placebo a los pacientes del grupo control, sino ese tratamiento habitual.

5.6. Evaluación de la respuesta y presentación de los resultados

En general, el análisis principal de un ensayo clínico debe basarse en el principio de intención de tratar, que incluye a cada paciente en su grupo de aleatorización independientemente de los cambios que pudieran ocurrir a posteriori: incumplimiento de la intervención, cambio de tratamiento, etc.. Aunque este tipo de análisis tiene una mayor tendencia a obtener resultados nulos, aumentando las dificultades para demostrar la eficacia de cualquier nueva intervención, existen dos poderosas razones para llevarlo a cabo. Por un lado, con este tipo de análisis se pretende preservar lo conseguido con la aleatorización, que los grupos sean comparables, lo cual constituye el fundamento para poder obtener conclusiones válidas (no sesgadas) sobre la eficacia de una intervención. Por otro lado, se trata de evaluar el efecto de la intervención en condiciones lo más parecidas posibles a las de la vida real (efectividad) y no sólo teniendo en cuenta las

condiciones ideales (eficacia) que proporcionan el subgrupo de pacientes que han seguido el protocolo de forma estricta.

Tradicionalmente, la medición del efecto de una intervención podía llevarse a cabo mediante medidas absolutas, como la reducción del riesgo atribuible, o relativas, como la reducción porcentual del riesgo relativo. De forma más reciente han surgido otras medidas de carácter complementario, que ponen el énfasis en facilitar la tarea del clínico a la hora de orientar el tratamiento de sus pacientes. Estas medidas de carácter más pragmático incluyen el porcentaje de pacientes libres de la condición bajo estudio y el número necesario de pacientes a tratar (NNT). El NNT constituye el número de pacientes que tendríamos que tratar para evitar que se produjera un caso con la enfermedad o condición estudiada y se calcula como el inverso de la reducción del riesgo atribuible. La presentación de los resultados mediante un único indicador, como la reducción porcentual del riesgo relativo, puede producir una impresión magnificada sobre el efecto del tratamiento, por lo que actualmente se recomienda que al presentar los resultados de un ensayo clínico se utilice la reducción porcentual del riesgo relativo acompañada de alguna magnitud que mida la reducción del riesgo en términos absolutos, particularmente el NNT. Para ilustrar esta última afirmación, nos basaremos en el ejemplo descrito por Cook y Sackett, a partir de una revisión de ensayos clínicos sobre tratamientos farmacológicos para la hipertensión arterial realizada por Collins y cols.¹⁴. Los estudios fueron divididos en dos grupos mutuamente exclusivos: aquéllos en los que todos los pacientes tenían una presión diastólica al inicio del ensayo menor de 110 mmHg (hipertensión ligera) y aquéllos en los que todos los pacientes tenían una presión diastólica al inicio del estudio menor de 115 mmHg (hipertensión moderada). Como puede apreciarse en la tabla 5, la reducción porcentual del riesgo relativo de padecer un ACV durante los 5 años siguientes con tratamiento farmacológico fue similar en los dos grupos, del 40%, lo que podría inducir a pensar que la estrategia terapéutica debería ser igual en ambos grupos. Sin embargo, la reducción absoluta del riesgo fue mucho menor en los sujetos con hipertensión ligera, dado que este grupo de pacientes tenía una tasa de ACV en ausencia de tratamiento mucho más baja que el grupo con hipertensión moderada. Traducido en términos de NNT, mientras en un caso se necesitaba tratar 13 pacientes para evitar un ACV, en el otro era preciso tratar a 167 pacientes

El análisis por intención de tratar proporciona una medida no sesgada de la eficacia de la intervención en condiciones reales.

para obtener el mismo resultado. En consecuencia, parece que la estrategia terapéutica debería ser diferente en esos dos grupos de pacientes.

Hipertensión	ACV en 5 años		Reducción del RR (Pc-Pt)/Pc	Reducción del RA Pc-Pt	Número necesario de pacientes a tratar (NNT) 1/(Pc-Pt)
	Grupo Control	Grupo Tratamiento			
Moderada					
Tasa ACV (P)	0,2	0,12	40%	0,08	13
N	16778	16898			
Ligera					
Tasa ACV(P)	0,015	0,009	40%	0,006	167
N	15165	15238			

Pc: Tasa grupo control
Pt: Tasa grupo tratamiento
Reducción del RR: Reducción porcentual del riesgo relativo de padecer un ACV. La Reducción Relativa del riesgo se calcula haciendo la diferencia de la tasa del grupo control menos la tasa del grupo tratamiento y esto se divide sobre la tasa del Grupo de Control. En el ejemplo es igual en los dos grupos.
Reducción del RA: Reducción absoluta del riesgo. Se calcula haciendo la diferencia de la tasa del grupo control menos la tasa del grupo tratamiento. En el ejemplo es menor en el grupo con hipertensión ligera.
NNT: Número necesario de pacientes que tendríamos que tratar para evitar un caso de enfermedad. Se calcula como el inverso de la reducción del riesgo atribuible.
Nota: Hay una errata en el encabezado de la tabla 5 del minitema, ya que la Reducción Relativa del riesgo se mide sobre el Riesgo Absoluto del Grupo Control (ver c señalada en rojo), no sobre Pt.

Por último, la importancia de una adecuada presentación de los resultados de los ensayos clínicos ha llevado a la elaboración de la "Declaración CONSORT", un informe con recomendaciones para mejorar la calidad de la presentación de los resultados de ensayos clínicos elaborado por investigadores, estadísticos, epidemiólogos y editores biomédicos y apoyado por numerosas revistas biomédicas y grupos editoriales. La declaración recoge los aspectos (22 ítems) que han de ser contemplados y la forma de presentarlos en los diferentes apartados (título y resumen, introducción, métodos, resultados y discusión) del informe sobre un ensayo clínico de dos grupos paralelos. La declaración, junto con las extensiones de la misma para los diferentes tipos de ensayos clínicos (en función de datos, intervenciones o diseños alternativos) se pueden consultar en <http://www.consort-statement.org/consort-statement> .

Bibliografía

1. Spitzer WO. *Clinical epidemiology (editorial)*. *J Chron Dis* 1986;39(6):411-415.
2. Fletcher RH, Fletcher SW, Wagner EH. *Epidemiología clínica*. Ed. Consulta SA, Madrid, 1989.
3. Depetris AR. *La tercera revolución epidemiológica y la promoción de salud*. En: *Conferencia internacional de promoción de salud*. Ministerio de Salud de Colombia y OPS/OMS, Santa Fe de Bogotá, 1992:70-9.
4. Feinstein AR. *Ciencia, medicina clínica y espectro de enfermedades*. En: Cecil-Loeb. *Tratado de Medicina Interna*. 14 ed. Ed. Importécnica, Madrid, 1978.
5. Imaz I, González J. *Epidemiología clínica*. En: Royo MA, Damián J. *Método Epidemiológico*. Madrid: ENS-ISCI III, 2009:169-86.
6. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Medicina basada en la evidencia. Cómo ejercer y enseñar la MBE*. ADS Printing SA. Madrid, 1997.
7. Sackett DL, Haynes RB, Tugweell P. *Epidemiología clínica. Una ciencia básica para la medicina clínica*. Ed. Díaz de Santos, Madrid, 1989.
8. *National High Blood Pressure Education Program. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure*. Bethesda (MD): National Heart, Lung, and Blood Institute (US); 2004 Aug. Report No.: 04-5230.
9. Lobos JM, Royo-Bordonada MA, Brotons C y cols. *Guía Europea de Prevención Cardiovascular en la Práctica Clínica. Adaptación Española del CEIPC 2008*. *Clin Invest Arterioscl* 2009; 21(3):124-50.
10. Baines CJ. *Rational and Irrational Issues in Breast Cancer Screening*. *Cancers* 2011; 3:252-66.

11. Pocock SJ. *Clinical trials. A practical approach*. Chichester, GB: John Willey & Sons, 1983.

12. Mensink RP, Katan MB. *Effect of a diet enriched with mono-unsaturated or polyunsaturated fatty acids on levels of low-density and high-density lipoprotein cholesterol in healthy women and men*. *N Engl J Med* 1989; 321:436-441.

13. Guallar E, Royo-Bordonada MA. *Directrices del "Comitee proprietary medicinal products": Estadística*. En Ergón ed. *II Reunión de Actualización de ensayos clínicos con medicamentos*. Madrid, 1996.

14. Cook RJ, Sackett DL. *The number needed to treat: a clinically useful measure of treatment effect*. *BMJ* 1995; 310:452-54.