

## ÉTICA Y ESTUDIOS EXPERIMENTALES

**Juan Manuel Paz Fernández**

*Jefe Laboratorio Central, Hospital General de Galicia*

El objetivo de los estudios experimentales es evaluar la eficacia de cualquier intervención preventiva, curativa o rehabilitadora.

Estos estudios se definen por dos características. La primera es que los investigadores tienen control sobre el factor de estudio, es decir, deciden qué tratamiento, con qué pauta y durante cuánto tiempo recibirá cada uno de los grupos de estudio. Hay que considerar los aspectos éticos, valorando los posibles riesgos que supone la intervención, que en ningún caso han de ser superiores a los posibles beneficios derivados de la misma.

La segunda característica es que la asignación de los individuos a los grupos de estudio se realiza de forma aleatoria. Si los grupos obtenidos de este modo son comparables y son estudiados con una misma pauta de seguimiento, cualquier diferencia observada entre ellos al finalizar el experimento puede ser atribuida, con un alto grado de convicción, a la diferente intervención a que han sido sometidos los participantes.

La metodología de los estudios experimentales puede aplicarse a cualquier intervención sanitaria sin necesidad de que sea farmacológica, como por ejemplo en la evaluación de distintas formas de citación para aumentar el cumplimiento de una medida preventiva, nuevas pautas organizativas en

la consulta, para evaluar si el consejo proporcionado por un profesional sanitario es eficaz en el tratamiento de una enfermedad, o estudiar la eficacia de distintas estrategias de formación médica.

Como **principales ventajas de los estudios experimentales** podemos citar las siguientes:

- \* Proporcionan un mayor control del factor de estudio

- \* La asignación aleatoria tiende a controlar los factores pronósticos que pueden influir en el resultado (incluidos aquellos que no se miden), y de este modo se aísla el efecto de la intervención

- \* Los estudios experimentales son los que proporcionan la mejor evidencia de una relación causa-efecto

Los **inconvenientes** más sobresalientes son:

- \* Restricciones éticas impiden que muchas preguntas puedan ser abordadas siguiendo la metodología de los estudios experimentales

- \* Se llevan a cabo en muestras muy seleccionadas (dificulta la generalización).

- \* Las intervenciones pueden ser muy rígidas, estar muy estandarizadas y diferir de lo que es la práctica habitual (dificulta la generalización)

- \* En los estudios experimentales se aborda habitualmente la relación entre una única intervención y su efecto sobre una enfermedad, mientras que en los estudios observacionales analíticos se pueden evaluar varios factores de riesgo.

- \* Suelen tener un coste elevado, aunque ello depende de la duración y la complejidad del protocolo.

El estudio experimental mas frecuente es el **ensayo clínico aleatorio**, en el que la asignación controlada del factor de estudio se realiza sobre los individuos, y cuya estructura básica se esquematiza en la figura 1.

A diferencia de otros estudios, en los experimentales el investigador controla la intervención, lo que plantea siempre dudas sobre la ética de su realización.

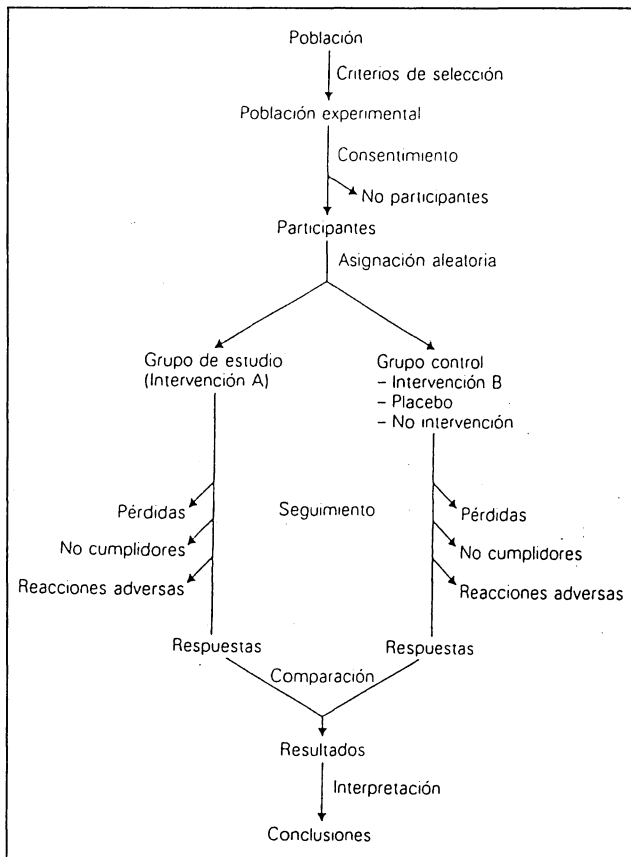


Figura 1. Estructura básica de un estudio experimental

Esquemáticamente, se pueden señalar cuatro **restricciones éticas básicas**:

1.- Los factores de estudio o exposiciones deben limitarse a los potencialmente preventivos de una enfermedad o de sus consecuencias

2.- Las exposiciones en todos los grupos de estudio deben ser igualmente aceptables según los conocimientos actuales

3.- Los sujetos incluidos en el estudio no deben ser privados en ningún momento de

las mejores medidas terapéuticas y preventivas. Por ejemplo, no es ético utilizar un placebo en una situación para la que exista un tratamiento efectivo.

4.- Los sujetos deben ser informados de su participación en un experimento y de sus posibles consecuencias.

Por principio, los procedimientos de investigación, por relevantes que pudieran ser sus conclusiones, no justifican sufrimiento o daño significativo de seres humanos. También por principio, la investigación con pacientes exige, al tiempo que se minimizan los riesgos, extremar los beneficios.

La investigación debe ser siempre realizada contando con el **consentimiento informado** de los sujetos en estudio.

La **selección** de sujetos o pacientes a incluir en el estudio debe ser justa, evitando el reclutamiento de pacientes, o

poblaciones, débiles o con niveles de autonomía disminuidos.

El material y los animales de experimentación que se utilizan para investigación deben ser tratados con extrema corrección.

Desde la perspectiva ética, **la primera premisa para considerar ética una investigación es su propia calidad.** A la hora de decidir ésta, los dos aspectos más importantes son la relevancia de los objetivos perseguidos y la metodología empleada para alcanzarlos.

La **relevancia de la investigación** debe ser considerada no sólo en cuanto a su capacidad para generar conocimiento objetivo sobre mecanismos de enfermedad, sino también en términos de utilidad para la mejora de la práctica clínica, incluyendo su evidente componente social.

La investigación para que sea ética debe ser metodológicamente impecable y que conduzca a resultados útiles y aparentemente relevantes en un contexto científico, empleando medios lícitos.

Un punto común a cualquier tipo de investigador es aceptar que debe poseer imprescindibles condiciones de honradez y veracidad.

Nosotros vamos a referirnos aquí al hecho de que "la investigación para que sea ética debe ser metodológicamente impecable" y dentro de esto a un aspecto particular, el tratamiento estadístico de los resultados de los estudios, considerando primero las pruebas terapéuticas y luego las pruebas diagnósticas.

## PRUEBAS TERAPÉUTICAS

Uno de los estudios típicos a los que se enfrenta el investigador, es al problema del

**contraste de una hipótesis** realizando la asunción de que no existe diferencia en la eficacia de dos fármacos A y B (hipótesis nula). Basándose en los resultados observados en la muestra, el investigador utiliza las pruebas de significación estadística para evaluar si existe la suficiente evidencia que le permita rechazar esa hipótesis nula y, consecuentemente, aceptar la hipótesis alternativa de que sí existen diferencias entre ambas terapias.

Se plantea la necesidad de saber cuál es el tamaño de la muestra suficiente a utilizar para detectar un efecto clínicamente importante, si de hecho existe. Esto es complicado, porque los distintos investigadores pueden tener diferentes opiniones de que tamaño de efecto es clínicamente importante. Pueden adaptarse, en aquellos casos en que sea de aplicación, por ejemplo, el máximo error clínicamente aceptable (definido por diferentes colectivos internacionales de prestigio y/o autores), o la variación achacable a la fisiológica más la analítica.

Un principio esencial en la determinación del tamaño de la muestra es que la aproximación usada se corresponda con los objetivos y diseño de la investigación y el análisis que se planifica. Por ejemplo, la aproximación adoptada dependerá de si el objetivo del estudio es evaluar una intervención o de identificar factores pronóstico, si el proyecto es una prueba clínica, un estudio caso-control, un estudio de cohorte, si hay emparejamiento o agrupamiento de sujetos, de si el análisis será una comparación de grupos, un análisis de Mantel-Haenszel, un modelo de regresión con varias covariables, etc.

El tamaño de la muestra a usar en cada estudio depende de varios factores, y por

tanto en cada caso la expresión de cálculo puede ser distinta. En el anexo pueden verse algunos ejemplos.

El **primer factor** es el tipo de resultado que se está evaluando, por ejemplo dicotómico o en escala continua.

El **segundo factor** en el cálculo del tamaño de la muestra, con respecto al tipo de variable evaluada, es el riesgo de error a tolerar, error que puede ocurrir cuando los resultados de un estudio son debidos al azar y no al tratamiento que se este evaluando. Hay dos tipos de tales errores, ambos resultantes de la variación aleatoria. El error tipo 1, conocido también como "alfa", se comete cuando el investigador rechaza la hipótesis nula siendo esta verdadera. Es equivalente a encontrar un resultado falso positivo, ya que el investigador concluye que existe una diferencia, cuando en realidad no existe. Por convención, un cambio de 1 en 20 de error tipo I o P de 0.05 se considera aceptable para la mayoría de los estudios, aunque otros valores pueden ser utilizados, como 0.01.

El error tipo II o error "beta", se comete en la situación contraria, cuando el investigador no rechaza la hipótesis nula siendo esta falsa. Es equivalente a un resultado falso negativo, ya que el investigador concluye que ha sido incapaz de encontrar una diferencia que existe en la realidad.

Lo que suele ocurrir en este caso es que de hecho en esta situación existe una diferencia clínicamente importante, pero el estudio es demasiado pequeño para detectarla. Por convención, un cambio de 2 en 10 de un error tipo II, referido como un beta de 0.2, se considera aceptable, aunque esto también puede diferir según qué circunstancias.

Si beta representa la probabilidad de un resultado falso negativo, su complementario, (1-beta), conocido como poder o potencia, representa la probabilidad de observar en la muestra una determinada diferencia o efecto (clínicamente importante), si de hecho existe en la población.

El **tercer factor** que entra en el cálculo del tamaño de la muestra es la opinión del investigador sobre qué diferencia es clínicamente importante a descubrir entre el grupo control y el problema: cuanto más pequeña (más precisión), mayor tamaño.

El **cuarto factor** es la variación inherente a la variable de interés en un grupo de población. Cuanto menor variación, menor tamaño se requerirá.

La ratio del número de sujetos problema al número de sujetos control es el factor importante final en el cálculo del tamaño de la muestra. Una ratio de 1 minimiza el tamaño de la muestra y es el usado en la mayoría de los estudios clínicos.

El concretar el tamaño de muestra suficiente a utilizar es fundamental, ya que un estudio con tamaño de muestra excesivamente alto puede ser considerado no ético a causa de la innecesaria implicación de sujetos extra y el correspondiente incremento de coste. Por otra parte, un estudio con un tamaño de muestra demasiado pequeño no será adecuado para detectar un efecto clínicamente importante. Tal estudio puede ser, de este modo, científicamente inútil y, por lo tanto, no ético en el uso de sujetos y recursos.

El tamaño calculado puede resultar mayor que el número de sujetos accesibles para el investigador, o sobrepasar los recursos disponibles. En estos casos, se puede solicitar la colaboración de otros centros.

Un objetivo frecuente en la investigación es el de **estimar un parámetro poblacional** a partir de los valores que la variable de interés adopta en los individuos de una muestra. Si la variable es continua, antes de tal estimación, es preciso chequear la posible existencia de datos aberrantes para en su caso rechazarlos utilizando algún criterio, por ejemplo el de Reed.

Así mismo, en estos estudios es necesario chequear el tipo de distribución (test de David, D'Agostino, Sesgo y Kurtosis, Shapiro-Wilk, etc.).

Como parámetros poblacionales se estimarán índices de tendencia central y de dispersión, acordes con el tipo de distribución, y sus intervalos de confianza. Ver anexo.

Otro objetivo frecuente en investigación médica es, como ya se indicó, la **comparación de la eficacia de dos tratamientos**. La conclusión se basa en gran medida en el resultado de una prueba de significación estadística que compara las respuestas obtenidas con cada una de las terapias.

Si se trata de una variable dicotómica, el contraste de significación se hará en base a un test de Student o un test ji cuadrado de Pearson (con o sin corrección de Yates según frecuencia), o un test exacto de Fisher de tratarse de porcentajes muy pequeños.

Si se trata de variable continua, habrá que contrastar resultados y sus varianzas.

Como test de homogeneidad de varianzas se utilizará el test de Fisher-Snedecor, de tratarse de distribución normal, y el de Levene en caso contrario.

Para determinar si hay diferencias estadísticamente significativas entre resultados, si se trata de grupos independientes con dis-

tribución normal y homogeneidad de varianzas, test de Student para grupos independientes; si no hay homogeneidad de varianzas, test de Behrens-Fisher.

Si son grupos relacionados y distribución normal, test de Student para datos apareados.

Si la distribución no es Gaussiana, test de Wilcoxon de rango de sumas o Mann-Whitney, si se trata de grupos independientes y test de Wilcoxon de rango de signos si se trata de grupos apareados.

También se podría ver la significación estadística determinando la diferencia de las medias o medianas, dependiendo del tipo de distribución, y sus intervalos estimativos a nivel de confianza del 95%, y ver si incluyen o no el cero. A su vez, habría que ver si las diferencias son medicamente significativas contrastándolas con criterios de significación médica.

En la significación estadística influye el nº de sujetos estudiados, de suerte que una diferencia no significativa puede hacerse significativa con tal de aumentar el nº de casos, de ahí la conveniencia de estimar la diferencia de medias o medianas y su intervalo de confianza para ver si los tratamientos son distintos o equivalentes según que el nivel inferior supere el medicamento aceptable o el superior sea menor que el medicamento aceptable.

Si son 3 o más grupos y lo que se mide es una variable continua, como test de homogeneidad de varianzas se utilizará:

\* Test C de Cochran, si las distribuciones son normales y los grupos son de igual tamaño.

\* Test de Barlett, si las distribuciones son normales y las muestras son de distinto tamaño.

\* Test de Levene, en caso de distribuciones no Gaussianas.

Como tests de contrastes de significación entre los resultados de los tratamientos:

\* Test de Tukey o de Duncan, si las distribuciones son normales, hay homocedasticidad y los grupos son de igual tamaño. Si los grupos son de distinto tamaño, test de Scheffe.

\* Si las distribuciones son gaussianas pero no hay homocedasticidad, test de Welch o de Brown-Forsythe, según que las medias extremas se correspondan con las varianzas extremas o no.

\* Test de Kruskal-Wallis, si las distribuciones no son gaussianas y se trata de grupos independientes.

\* Test de Friedman, si las distribuciones no son gaussianas y se trata de grupos relacionados.

Como test de multicomparación también se pueden utilizar los de Student o Wilcoxon, según el tipo de distribución, pero con penalización Bonferroni o Dunnett dependiendo de que se contrasten todos los grupos contra todos o todos los grupos contra un control.

En los **estudios de asociación entre variables** se calcularán los coeficientes de:

\* Coeficiente de correlación de Pearson, distribución normal bivalente.

\* Coeficiente de correlación de Spearman, distribución no gaussiana y de pocas ligaduras.

\* Coeficiente de correlación de Kendal, distribución no gaussiana y muchas ligaduras.

\* Coeficiente de correlación biserial puntal, variable continua y dicotómica.

\* Coeficiente de correlación tetracórico o

coeficiente de concordancia kappa de Cohen, si se trata de dos variables dicotómicas.

\* Coeficiente de concordancia kappa de Cohen y/o residuos ajustados de Haberman según se trate de correlacionar variables categóricas (3 o más), dependiendo de si el nº de categorías en las dos situaciones es o no el mismo.

Si se trata de buscar una **relación entre dos o más variables**, habrá que determinar la ecuación de regresión de mejor ajuste. En este tipo de relación, la ecuación tendrá o no valor como ecuación de predicción de la variable dependiente en función de la o las variables independientes, según que el error de estimación sea o no inferior al médicamente aceptable.

Para la eliminación de variables de confusión y determinación de riesgos relativos ligados a un factor, regresión logística o test de Mantel Haenszel.

Cuando se trata de datos distintos e independientes procedentes de ensayos diferentes sobre una misma experiencia, pero con tamaños muestrales no lo suficientemente altos y significaciones  $P_i$  no lo suficientemente pequeñas para poder realizar, individualmente, una afirmación conclusiva, y los datos de los distintos ensayos no puedan mezclarse en uno sólo por el distinto modo en que fueron obtenidos, o porque, aunque fueron realizados con igual protocolo (ensayo multicéntrico), la variabilidad de los datos es distinta en cada uno de ellos (si se están comparando medias), de modo que no es factible obtener una única  $P$  a través de la mezcla de los datos, pero sí por medio de la mezcla de los  $P_i$  individuales.

La técnica estadística en cuestión es el **metanálisis** que es una síntesis e integración

de los resultados de los estudios independientes que tratan del mismo tema, el proceso de combinar resultados de estudios que pueden ser usados para sacar conclusiones, por ejemplo, sobre la efectividad terapéutica, o planificar nuevos estudios. El producto final posee elementos cualitativos y cuantitativos, como tener en cuenta los resultados numéricos de los estudios individuales, así como cuestiones más subjetivas como calidad, extensión del sesgo, robustez del diseño del estudio. El metanálisis es una estrategia de revisión sistemática para dirigir cuestiones de investigación que son especialmente útiles cuando los resultados de los estudios discrepan con respecto a la magnitud o dirección del efecto, cuando los tamaños de muestra son individualmente demasiado pequeños para detectar un efecto y el nivel es estadísticamente significativo, o cuando un ensayo grande es demasiado costoso y consume mucho tiempo en su realización.

## PRUEBAS DIAGNOSTICAS

El término **prueba** significa toda fuente de información -clínica o paraclínica- cuyo resultado se utiliza en una toma de decisión clínica, lo mas frecuentemente con un objetivo diagnóstico. La función diagnóstico primordial de una prueba es evaluar, en función de su resultado, la probabilidad de una enfermedad dada, o la probabilidad de su ausencia. Esta función de la prueba, que suele ser imperfecta (se cometen errores al clasificar a una persona por su resultado), se expresa, entre otros índices, por los valores predictivos. Estos valores dependen de dos

datos constitutivos:

\* Las cualidades intrínsecas de la prueba: la sensibilidad y la especificidad

\* La probabilidad primaria o probabilidad pre-prueba de la enfermedad buscada (la prevalencia de la enfermedad).

El **valor diagnóstico de una prueba** debe ser evaluado en función del patrón oro (diagnóstico verdadero, constatación definitiva e irrefutable, mejor prueba disponible).

Su calidad no es fija y varía en función de la prevalencia de la enfermedad en la población y de la composición del método diagnóstico basado en varias pruebas.

Entendemos por **sensibilidad**,  $S$ , de un signo o de una prueba en una enfermedad, a la frecuencia del signo (o la respuesta positiva de la prueba) entre los pacientes que tienen la enfermedad. Es la probabilidad condicionada del signo o la respuesta positiva de la prueba si la enfermedad está presente, es decir, es la proporción total de enfermos que la prueba es capaz de detectar en el grupo de afectados (resultados positivos ciertos), o sea positividad en enfermedad. Se calcula dividiendo el numero de verdaderos positivos por el total de enfermos (suma de verdaderos positivos mas falsos negativos), pudiendo expresarse en tanto por ciento o tanto por uno.

La **especificidad** es la frecuencia o probabilidad condicional de la ausencia del signo o de la respuesta negativa de la prueba en los sujetos que no tienen la enfermedad, es decir, es la proporción de individuos sanos confirmados como tales por el resultado negativo de la prueba (negativos ciertos), o sea negatividad en salud. Se calcula dividiendo el numero de verdaderos negativos por el total de sanos (suma de verdaderos negativos mas

falsos positivos), pudiendo expresarse en tanto por ciento o tanto por uno.

La estimación de las cualidades diagnósticas de un signo o una prueba, implica que este signo o el resultado de la prueba no están incluidos ni en la definición de la enfermedad ni en la del grupo control (no se puede ser a la vez juez y parte).

Otros índices que tienen una importancia determinante en el proceso de decisión clínica son las **razones de verosimilitud** (de inclusión y de exclusión): son ellas, y no solamente la sensibilidad y la especificidad, quienes conjugadas con la probabilidad primaria, determinan los valores predictivos.

Se llama razón de verosimilitud del resultado R de una prueba en una enfermedad M, al cociente de la frecuencia de ese resultado en la enfermedad M y la frecuencia de ese mismo resultado en el caso de pacientes que no tienen la enfermedad.

En términos de probabilidad, la razón de verosimilitud de un evento R en favor de una situación M, es la razón de la probabilidad condicional de este evento cuando la situación (enfermedad) M está presente, sobre la probabilidad condicional de este mismo evento cuando la situación (enfermedad) no está presente.

Hay pues razones de verosimilitud distintas teniendo en cuenta que la prueba tiene respuestas diferentes. Si la respuesta es dicotómica -positiva, o patológica y negativa- hay dos razones de verosimilitud, la razón de verosimilitud positiva (L), de inclusión o razón de probabilidad positiva, y la razón de verosimilitud negativa (X), de exclusión o razón de probabilidad negativa.

La razón de verosimilitud de inclusión se

calcula dividiendo la proporción de personas que teniendo la enfermedad dan positivo, por la proporción de personas que no teniendo la enfermedad también han dado un resultado positivo, es decir dividiendo la proporción de verdaderos positivos por la proporción de falsos positivos.  $L = \text{sensibilidad} / (1 - \text{especificidad})$ .

La razón de verosimilitud de exclusión se calcula dividiendo la proporción de personas que teniendo la enfermedad dan resultado negativo, por la proporción de personas que no teniendo la enfermedad también dan negativo, es decir dividiendo la proporción de falsos negativos por la proporción de verdaderos negativos.  $X = (1 - \text{sensibilidad}) / \text{especificidad}$

La razón de verosimilitud tiene grandes consecuencias en las diversas decisiones médicas tanto por su papel sobre la eficacia diagnóstica de las pruebas como por sus implicaciones económicas y éticas (por ejemplo en el diagnóstico prenatal).

La **fiabilidad** de tal estimación depende esencialmente de la calidad metodológica pero también de los efectivos sobre los cuales reposan los cálculos. En presencia de efectivos reducidos (tamaño inferior a 20), debe utilizarse una fórmula de ajuste (Koplan J L, 1985).

La incidencia de la razón de verosimilitud sobre los valores predictivos y su utilización diagnóstica es tal que es útil analizar, al menos brevemente, el papel respectivo de sus dos componentes, la sensibilidad y la especificidad. Si la especificidad es constante, la razón de verosimilitud positiva (L) es una función lineal de la sensibilidad (S). Cuando S aumenta, L aumenta en la misma



proporción. La razón de verosimilitud negativa es igualmente una función lineal de la sensibilidad. Cuando S aumenta, X disminuye en la misma proporción.

Si la sensibilidad es constante, las dos razones de verosimilitud son funciones hiperbólicas de la especificidad, pero mientras para valores elevados de la especificidad las variaciones de este afectan mucho a la razón de verosimilitud positiva y casi nada a la razón de verosimilitud negativa, ocurre lo contrario para valores bajos de la especificidad.

La eficacia diagnóstica de una prueba en la búsqueda de una enfermedad determinada M está estrechamente ligada al crecimiento de su especificidad, calidad que precisamente, se establece en el caso de los sujetos que no tiene la enfermedad M.

Cuando las dos razones de verosimilitud son iguales a 1, la prueba no tienen ningún valor diagnóstico.

Una vez realizadas las pruebas nos planteamos lo siguiente: si la prueba da un resultado positivo, ¿cuál es la probabilidad de que el sujeto tenga la enfermedad?; si da un resultado negativo, ¿con qué seguridad podemos rechazar el diagnóstico? y, ¿cuál es el riesgo en este caso de que el sujeto este alcanzado de la enfermedad? La primera cuestión la contesta el valor predictivo positivo (VPP) de la prueba en favor de la enfermedad M; la segunda cuestión la expresa el valor predictivo negativo (VPN); la tercera cuestión puede ser formulada por  $1 - \text{VPN}$ .

El valor predictivo positivo (VPP) de una prueba T o de un signo I en favor de una enfermedad M se escribe:  $\text{VPP} = p(M/T+)$  o  $p(M/I)$ , que se lee probabilidad de la enfermedad M si la prueba T da una respuesta

positiva, o si el signo I está presente.

El valor predictivo negativo (VPN) es la probabilidad de la ausencia de la enfermedad M si la respuesta es negativa, o si el signo I está ausente, y se escribe:  $\text{VPN} = p(M/T-)$  o  $P(M/I)$ .

Los valores predictivos son probabilidades condicionadas inducidas por los resultados de la prueba o por la presencia del signo. La estimación de los valores predictivos se efectúa por aplicación del teorema de "probabilidad de las hipótesis" del Reverendo Thomas Bayes.

Las fórmulas de cálculo son las siguientes:

$$\text{VPP} = P.S / [P.E + (I-P)(1-e)]$$

$$\text{VPN} = E.(1-P) / [(1-S)P + E(1-P)]$$

donde P = Prevalencia o probabilidad pretest, probabilidad a priori.

Estas mismas fórmulas en función de las razones de verosimilitud son:

$$\text{VPP} = PL / [P(L-I) + 1]$$

$$\text{VPN} = (1-P) / [P(X-1) + 1]$$

$$1 - \text{VPN} = P1 / [PX - 1] + 1$$

Si se representa en una gráfica los valores predictivos en función de la probabilidad primaria P, se obtienen hipérbolas equiláteras, y se observa

\* Para una misma prueba, cuando P disminuye, VPP disminuye y VPN aumenta. Cuando P aumenta, VPP aumenta y VPN disminuye.

\* El valor predictivo positivo crece al principio rápidamente cuando la probabilidad crece entre 0 y 0,5, y luego mas lentamente. VPP es inferior al 50% cuando la probabilidad a priori o pretest es inferior a  $1/(1+L)$ , es decir cuando la prevalencia de la enfermedad es inferior a  $1/(1+L)$  el VPP o

probabilidad postest de la enfermedad es inferior al 50%, incluso si la prueba da un resultado positivo.

\* El valor predictivo negativo disminuye cuando aumenta la probabilidad pretest de la enfermedad es inferior a 0.5 cuando  $P = 1/(1+X)$ . Por lo tanto cuando la prevalencia es superior a  $1/(1+X)$ , incluso si la prueba da resultado negativo, la probabilidad a posteriori o post test de la enfermedad es superior al 50%. Los dos valores de probabilidad primaria, tales que  $P1 = 1/(1+L)$  y  $P2 = 1/(1+X)$ , determinan lo que se puede definir como intervalo de seguridad de la prueba. En efecto, cuando la probabilidad primaria se sitúa en este intervalo, un resultado positivo de la prueba da un VPP superior al 50% y un resultado negativo da una probabilidad secundaria inferior al 50%; lo contrario ocurriría fuera de este intervalo. Estas consecuencias matemáticas del intervalo de seguridad así definido, son de una importancia capital para la conducción del razonamiento y de la decisión diagnóstica.

Cuando las cualidades diagnosticas sensibilidad y especificidad son mejoradas, las razones de verosimilitud  $L$  aumenta y  $X$  disminuye, aumentando el intervalo de seguridad.

Para una sensibilidad constante, a valores bajos de la probabilidad pretest, la eficacia diagnóstica del resultado positivo de la prueba -medida por el VPP- es tanto mas grande cuanto la especificidad y la razón de verosimilitud  $L$  sean mas elevadas. Por el contrario, las variaciones de especificidad modifican poco la consecuencia del resultado negativo; el efecto de la probabilidad primaria es predominante.

Para valores constantes de la especificidad, las variaciones de la sensibilidad afectan poco los valores predictivos positivos; por el contrario, modifican considerablemente el efecto de un resultado negativo de la prueba, sobre todo cuando la probabilidad pretest de la enfermedad es superior a 0.5 ó 0.6.

**En resumen, cuando la probabilidad pretest  $P$  es baja, un resultado negativo de la prueba tiene poca incidencia sobre la decisión diagnóstica; el de un resultado positivo de una prueba específica es elevado. A la inversa, si la probabilidad pretest es elevada, la consecuencia diagnóstica del resultado negativo de una prueba sensible es considerable, el de un resultado positivo es débil.**

La incidencia del resultado de la prueba sobre el proceso decisional traduce la eficacia diagnóstica de la prueba. Esta eficacia se mide por el efecto de concentración y por la noción de ganancia diagnóstica.

Se llama "efecto de concentración" el aumento de la densidad -o de la probabilidad- de la enfermedad investigada que es aportada por el resultado de la prueba. El efecto de concentración se calcula por la relación: Probabilidad post test de  $M$ /probabilidad pre-test de  $M = VPP/P$ .

El efecto de concentración es muy importante en situación de despistaje de una enfermedad cuya prevalencia es baja en la población estudiada.

Se llama **ganancia diagnóstica** a la diferencia entre la probabilidad post test encontrada por el resultado de la prueba y la probabilidad pretest. Si el resultado es positivo, la ganancia diagnostica  $G(+)$  se mide por la diferencia  $p(M/T+)-p(M) = VPP - P$ . Si el

resultado es negativo, la reducción de la probabilidad de la enfermedad se mide en valor absoluto por la diferencia  $G(-) = p(M) - p(M/T-)$ .

El conjunto de ganancia diagnóstica lograda por la aplicación de la prueba, o contenido diagnóstico C de la prueba es igual a la suma de estas dos ganancias diagnósticas  $C = p(M/T+) - p(M/T-) = VPP + VPN - 1$

La ganancia diagnóstica de una prueba positiva es máxima cuando la probabilidad pretest es  $1/(1 + \sqrt{L})$ .

La ganancia diagnóstica de una prueba negativa es máxima cuando la probabilidad pretest es  $1/(1 + \sqrt{L})$ .

El contenido diagnóstico de una prueba es máximo cuando la probabilidad pretest es  $1/(1 + \sqrt{LX})$ .

Otra forma de ver la eficacia de las pruebas diagnósticas es la representación de las curvas ROC, representando en ordenadas la proporción de verdaderos positivos (sensibilidad) y en abscisas la proporción de falsos positivos (1-especificidad), para diferentes puntos de corte. Cuando mayor sea el área bajo la curva, más eficaz es la prueba.

Un problema clave para medir las cualidades diagnósticas de una prueba es determinar cuál debe ser el punto de corte. Para ello se pueden seguir distintos procedimientos; uno de ellos es maximizar la suma sensibilidad más especificidad y otro consiste en determinar el que corresponde al punto de la curva ROC cuya pendiente sea  $(1-p(M))/p(M)$  ó  $A/Bx(1-p(M))/p(M)$ , según se contemplen o no los costos, donde A represente el coste neto del tratamiento de pacientes que no tengan la enfermedad (costo de los resultados falsos positivos) y B = al beneficio neto del tratamiento de los pacientes enfermos.

Para ver si una prueba se puede utilizar o no como prueba diagnóstica es necesario verificar previamente que hay una diferencia estadísticamente significativa entre su aplicación y la prueba inútil.

Asimismo, para ver si una prueba es mejor que otra es preciso ver si hay diferencia estadísticamente significativa entre la suma sensibilidad más especificidad de una con respecto a la otra, o si el área bajo la curva ROC de una difiere significativamente de la de la otra.

Galen y Gambino han emitido cuatro postulados prácticos relativos a la **elección práctica de las pruebas diagnósticas**.

1.- Elegiremos un test sensible cuando:

- a) La enfermedad sea grave y no pueda pasar inadvertida
- b) La enfermedad sea tratable
- c) Los resultados falsamente positivos no supongan un traumatismo psicológico o económico en los individuos examinados.

Ejemplo: el feocromocitoma (fatal si es ignorado, controlable al 100% cuando se diagnostica), la fenilcetonuria, las enfermedades venéreas y otras enfermedades infecciosas curables.

2.- Utilizaremos la prueba más específica posible cuando:

- a) La enfermedad sea importante, pero difícil de curar o incurable
- b) El hecho de conocer que no se padece la enfermedad tiene una importancia sanitaria y psicológica
- c) Los resultados falsamente positivos pueden suponer un trauma psicológico y económico para el individuo examinado.

Ejemplos: esclerosis en placas, cánceres ocultos.

3.- Una prueba con alto valor predictivo del resultado positivo debe ser utilizada cuando el tratamiento de los falsos positivos pudiera tener graves consecuencias.

Ejemplo: radioterapia o lobotomía innecesarios en los pacientes que se ha sospechado erróneamente un cáncer de pulmón.

4.- Desearemos un elevado valor global de la prueba cuando:

a) La enfermedad sea importante, pero curable

b) Tanto los falsos positivos como los falsos negativos supongan un traumatismo grave y conlleven consecuencias graves.

Ejemplos: infarto de miocardio (posiblemente fatal si no es tratado, al igual que traumatizante en el caso de un falso positivo), lupus eritematoso, determinadas formas de leucemia o linfoma, diabetes mellitus.

Cuando se dispone de resultados de varias pruebas diferentes hay que recurrir a un **análisis multivariado de estas pruebas** y poder llegar a un diagnóstico diferencial, distinguiendo entre distintas situaciones gnoscológicas, así como diferenciar qué variables son discriminantes y qué variables son redundantes o que no dan información diagnóstica. Como **sistemas de discriminación** entre distintas situaciones gnoscológicas, tenemos el análisis discriminante lineal, el cuadrático o logístico, dependiendo de si sólo se manejan variables continuas con distribución normal con o sin homogeneidad de varianzas, o si se trata de distribución no gaussiana y/o variable categórica.

Otra valiosa herramienta de ayuda al diagnóstico que se puede utilizar la constituye las **redes neuronales**. Su naturaleza es radicalmente distinta de los métodos estadísticos multivariantes, en los que por diversas técnicas una variable dependiente o de respuesta puede describirse en función de un conjunto de variables independientes o predictorias. En las redes neuronales no se impone ninguna condición sobre la distribución de datos, que pueden tener una gran variabilidad o incluso ser ambiguas o presentar "ruido". Otra característica es que admiten relaciones no lineales, en las que el valor de una variable puede tener distinto efecto dependiendo de los valores de las otras variables.

Como consecuencia de su peculiar funcionamiento, estos sistemas presentan una especial habilidad para aprender a partir de una gran cantidad de ejemplos, siendo capaces de predecir patrones y tendencias y descubrir sutiles relaciones entre los datos.

En estos estudios con datos procedentes de varias pruebas, es conveniente usar los distintos sistemas matemáticos que en rigor se puedan utilizar y luego seleccionar aquel que mejor clasifique los sujetos.

En la medida que no se utilicen métodos estadísticos correctos, los estudios son científicamente rechazables, carecen de utilidad, y por lo tanto no éticos, por haber usado individuos y recursos indebidamente, ya que cualquier conclusión que se pretenda sacar en estas condiciones, carece de valor.