

**Tamizaje y Pruebas de Diagnóstico:**  
**Validez y Repetición; Sensibilidad y Especificidad**

**Objetivos:**

Familiarizarse con:

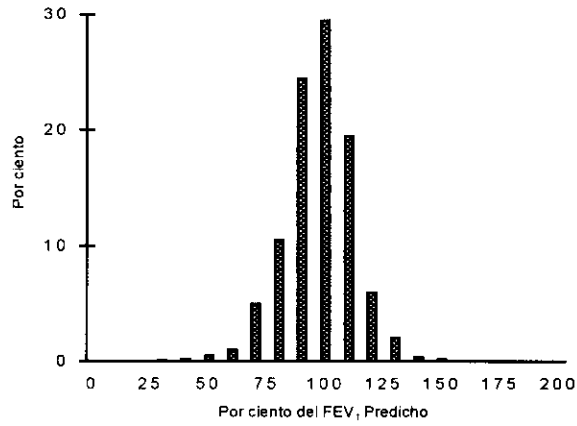
- los conceptos de validez y confiabilidad
- índices de concordancia
- la sensibilidad y especificidad como mediciones de validez
- la relación entre la sensibilidad y la especificidad y el poder predictivo negativo y positivo

**I. La Variación Biológica de las Poblaciones**

- A. Las pruebas de diagnóstico se emplean para identificar a aquellos con una enfermedad y aquellos que no la tienen. Para determinar qué es “normal” y qué es “anormal” es importante entender cómo se distribuyen en las poblaciones humanas las diferentes características. Las características continuas, como la presión arterial y el FEV<sub>1</sub> tienen distribuciones en las poblaciones, o sea, que todas las personas no tienen un mismo valor. Se puede mostrar gráficamente la distribución o se puede captar la variación con las estadísticas paramétricas como la media y la desviación estándar o los puntos percentiles.

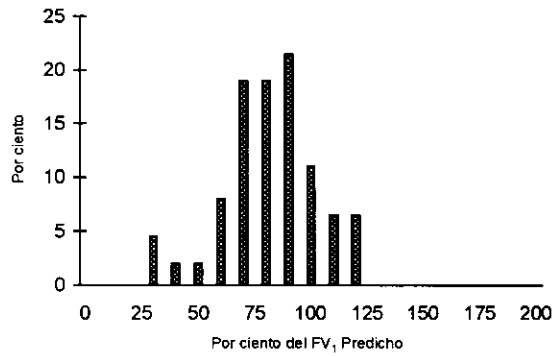
La serie de figuras siguiente ilustra la variabilidad de una medida de función pulmonar, el FEV<sub>1</sub>, en muestras de mujeres basadas en la población y que participaron en el Estudio de Harvard de las Seis Ciudades. Las figuras ilustran una casi distribución de Gauss (normal) en mujeres sin asma que nunca fumaron y un cambio hacia la izquierda en las curvas con sesgo en los grupos de asmáticas y de asmáticas que fumaron. Las distribuciones sobreponen y se podría usar únicamente el nivel de FEV<sub>1</sub> para separar las asmáticas de las no asmáticas.

**Figure 1**



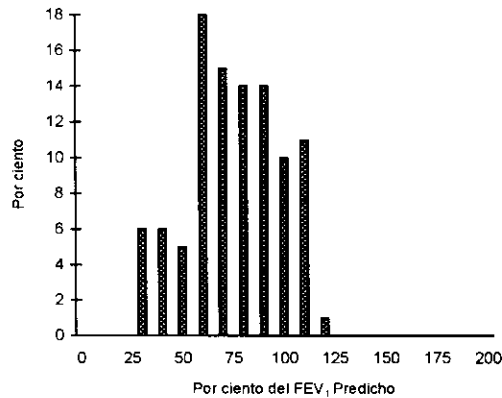
La distribución del FEV<sub>1</sub> en un estudio en seis comunidades de EUA, como el por ciento del valor predicho para los sujetos adultos femeninos que nunca fumaron y que nunca tuvieron asma.

**Figure 2**



La distribución del FV<sub>1</sub> en un estudio en seis comunidades de EUA, como el por ciento del valor predicho para los sujetos adultos femeninos que nunca fumaron y que tienen un diagnóstico de asma dado por un médico.

**Figure 3**



La distribución del FEV<sub>1</sub> en un estudio en seis comunidades de EUA, como el por ciento del valor predicho para los sujetos adultos femeninos que fumaron y que tienen un diagnóstico de asma dado por un médico.

II. **La validez de una Prueba** se define como la habilidad para identificar correctamente aquellos que tienen una enfermedad y aquellos que no tienen la enfermedad. La validez tiene dos componentes: la sensibilidad y la especificidad. Estas definiciones y conceptos se aplican tanto a las pruebas diagnósticas, ej., estableciendo la presencia o ausencia de enfermedades en pacientes, como al tamizaje, ej., identificando la enfermedad preclínica en las personas asintomáticas.

A. Sensibilidad: La habilidad de una prueba para identificar correctamente aquellos que tienen la enfermedad.

B. Especificidad: La habilidad de una prueba para identificar correctamente aquellos que no tienen la enfermedad.

C. Cálculo de la Sensibilidad y la Especificidad

1. Para calcular la sensibilidad y especificidad de una prueba necesitamos ver los componentes en el cuadro de abajo:

**Tabla 1**

		<u>Población</u>	
		Con la Enfermedad	Sin la Enfermedad
<u>Resultados de la Prueba</u>	Positivo	Tiene la Enfermedad Tiene la Prueba Positiva <b>= POSITIVOS VERDADEROS (PV)</b>	No Tiene la Enfermedad Tiene la Prueba Positiva <b>= POSITIVOS FALSOS (PF)</b>
	Negativo	Tiene la Enfermedad Tiene la Prueba Negativa <b>= FALSOS NEGATIVOS (FN)</b>	No Tiene la Enfermedad Tiene la Prueba Negativa <b>= NEGATIVOS VERDADEROS (TN)</b>

Sensibilidad =  $\frac{VP}{VP + FN}$       Especificidad =  $\frac{VN}{VN + FP}$

2. En este ejemplo estamos aplicando una prueba dicotómica (positiva o negativa) en una población de individuos, algunos de los cuales tienen la enfermedad y algunos que no la tienen. Se presentan cuatro posibilidades:

- Personas con la enfermedad tienen la prueba positiva - decisión correcta
- Personas con la enfermedad tienen la prueba negativa - decisión incorrecta
- Personas sin la enfermedad tienen la prueba negativa - decisión correcta
- Personas sin la enfermedad tienen la prueba positiva - decisión incorrecta

3. "Sabemos" que hemos tomado decisiones correctas e incorrectas porque tenemos conocimiento, de otra fuente de clasificación, sobre quién tiene verdaderamente la enfermedad y quién no la tiene. A esta otra fuente de referencia con frecuencia se le conoce como el "estándar de oro". Con frecuencia es el método corriente en uso o algún otro enfoque más costoso o invasivo para establecer el diagnóstico.
4. El ejemplo que sigue ilustrará cómo calcular la sensibilidad y la especificidad. Hay una población de 1000 individuos, de los cuales 100 tienen la enfermedad establecida por el método estándar o "estándar de oro" y 900 no tienen la enfermedad, también por el método estándar. De los que tienen la enfermedad 80 tienen una prueba positiva y 20 una prueba negativa. La sensibilidad de la prueba, o la habilidad de la prueba para identificar correctamente a aquellos con la enfermedad se calcula dividiendo a las 80 personas identificadas correctamente por las 100 personas que tienen la enfermedad, u 80%. La especificidad de la prueba, o la habilidad de la prueba para identificar correctamente a aquellos que no tienen la enfermedad se calcula dividiendo a las 800 personas identificadas correctamente por el total de 900 personas que no tienen la enfermedad, u 89%. Esto se ilustra en el cuadro de abajo:

**Cuadro 2**

**Sensibilidad y Especificidad de una Prueba**

Resultados de la Prueba	Enfermedad	No Enfermedad	Total
Positiva	80	100	180
Negativa	20	800	820
Total	100	900	1,000
Sensibilidad = $\frac{80}{100} = 80\%$		Especificidad = $\frac{800}{900} = 89\%$	

D. La relación de los cambios en la sensibilidad con los cambios en la especificidad.

1. Para las variables continuas, tales como el por ciento predicho del FEV<sub>1</sub> la decisión de dónde colocar el punto de corte para un resultado de la prueba positivo versus un resultado negativo involucra un intercambio entre la sensibilidad y la especificidad. El bajar el nivel del corte aumentará la sensibilidad pero también disminuirá la especificidad. El aumentar el nivel del corte aumentará la especificidad pero también disminuirá la sensibilidad.
2. Si es importante identificar a todos los casos de la enfermedad entonces será adecuado seleccionar el punto de corte que dará como resultado una sensibilidad alta. Por

ejemplo, la diabetes mellitus en una enfermedad seria, pero cuando se diagnostica tempranamente en cuando a su historia natural el manejo eficiente puede reducir las complicaciones. Posiblemente aprendamos que la detección y el tratamiento tempranos del asma pueden también afectar la historia natural de la enfermedad. Esta sería una situación en la que es importante para una prueba de tamizaje tener una sensibilidad alta para la detección del asma.

3. También es deseable tener una especificidad alta en el tamizaje y el diagnóstico, en particular si la condición de interés no es común. Si una enfermedad es poco común y la especificidad es baja, entonces la mayoría de los casos serán falsos positivos; en el otro extremo, si la prevalencia de una enfermedad fuera cero, entonces todos los que tuvieran una prueba positiva serían falsos positivos.

#### E. Valor predictivo

1. Otro tema importante de las pruebas diagnósticas es el valor predictivo de la prueba: o sea, si la prueba es positiva o negativa, qué probabilidad hay que el individuo tenga o no tenga la enfermedad. El valor positivo predictivo de una prueba es el por ciento de todos aquellos que tienen una prueba positiva que verdaderamente tienen la enfermedad. De manera semejante, el valor predictivo negativo es el por ciento de aquellos que tienen una prueba negativa que realmente no tienen la enfermedad.
2. El valor predictivo positivo contesta a la pregunta: "Si una persona tiene un resultado positivo a la prueba, ¿cuál es la probabilidad que él/ella tenga verdaderamente la enfermedad?". El valor predictivo negativo contesta una pregunta similar: "Si una persona tiene un resultado negativo, ¿cuál es la probabilidad de que él/ella no tengan la enfermedad?".
3. Ejemplo (vea el cuadro de abajo): Se prueban a 1000 personas en una población; 100 tienen la enfermedad y 900 no la tienen. Ochenta personas con la enfermedad y 100 personas sin la enfermedad tienen resultados positivos. Se calcularía el valor predictivo positivo dividiendo a los 80 positivos verdaderos por el número total de los que dieron resultados positivos, (180) lo que da un 44%. De aquellos que dieron un resultado negativo, 20 tenían la enfermedad y 800 no la tenían. El valor predictivo negativo se calcularía dividiendo a las 800 personas que dieron un resultado negativo y que no tenían la enfermedad por las 820 personas que tuvieron resultados negativos o 98%.

### Cuadro 3

#### Valor Predictivo de una Prueba

Resultados de la Prueba	Enfermedad	No Enfermedad	TOTAL
Positiva	80	100	180
Negativa	20	800	820
TOTAL	100	900	1,000

Valor Predictivo Positivo = $\frac{80}{180} = 44\%$	Valor Predictivo Negativo = $\frac{800}{820} = 98\%$
---	--

### III. El Valor Predictivo Positivo y su Relación con la Prevalencia de la Enfermedad (Teorema de Bayes')

Mientras mayor sea la prevalencia de una enfermedad en una población el valor predictivo positivo de una prueba usada en esa población será mayor. Esta relación se conoce como el Teorema de Bayes, un ministro que describió esta relación por primera vez. El Cuadro 4 que se muestra abajo, ejemplifica esta relación en una población hipotética de 10,000 personas, donde, ya sea 100 o 500 personas tienen la enfermedad. Con un aumento en la prevalencia de la enfermedad de 1% al 5% y con una sensibilidad y especificidad constantes, el poder predictivo aumenta de 17% a 51%.

#### Cuadro 4

##### Relación de la Prevalencia de Enfermedad con el Valor Predictivo

Ejemplo: Prueba de sensibilidad = 99%  
Prueba de especificidad = 95%

Prevalencia de la Enfermedad	Resultados de la Prueba				Valor Predictivo Positivo
	Enfermo	No Enfermo	Totales		
1%	+	99	495	594	1
	-	1	9,405	9,406	$\frac{99}{594} = 17\%$
	Totales	100	9,900	10,000	
5%	+	495	475	970	2
	-	5	9,025	9,030	$\frac{495}{970} = 51\%$
	Totales	500	9,500	10,000	

Aún teniendo una sensibilidad alta (99%) y una especificidad relativamente alta (95%), el valor predictivo positivo es bajo para una prevalencia de enfermedad del 1%.

#### IV. Valor Predictivo Positivo y Su Relación Con la Especificidad

El valor predictivo positivo también puede estar afectado por la especificidad de la prueba, como se puede ver abajo en el Cuadro 5. Si la prevalencia de la enfermedad y la sensibilidad permanecen constantes, un aumento en la especificidad del 70% al 95% aumenta el valor predictivo positivo de 27% al 69%, por el menor número de falsos positivos (450 vs. 2700). Este aumento refleja un menor número de falsos positivos conforme aumenta la especificidad.

### Cuadro 5

#### Relación entre la Especificidad y el Valor Predictivo

Ejemplo: Prevalencia = 10%

Sensibilidad = 100%

Especificidad	Resultados de la Prueba	Enfermos	No Enfermos	Totales	Valor Predictivo
70%	+	1,000	2,700	3,700	3
	-	0	6,300	6,300	$\frac{1,000}{3,700} = 27\%$
	Totales	1,000	9,000	10,000	
95%	+	1,000	450	1,450	4
	-	0	8,550	8,550	$\frac{1,000}{1,450} = 69\%$
	Totales	1,000	9,000	10,000	

#### V. Confiabilidad (Precisión, Repetición) de una Prueba

##### A. Variación Intrasujeto

Los valores registrados para medir las características humanas pueden variar a través del tiempo, aún en períodos cortos de tiempo. Por ejemplo, pueden variar las lecturas de la presión arterial tomadas en diferentes momentos del mismo día, así como las mediciones de función pulmonar.

El Cuadro 6 muestra las mediciones de flujo pico para 10 individuos que fueron examinados 30 veces, aproximadamente, en el curso de dos meses. Cada grupo de mediciones consistió de por lo menos un flujo expiratorio pico satisfactorio. Se muestran las medias, las desviaciones estándar y los valores mínimos y máximos para cada uno de los 10 individuos, mostrando los límites de la variación de las lecturas. Las desviaciones estándar cubren rangos de hasta tres veces, de cerca de 15 a 45.



### Cuadro 6

**Variación de las Mediciones de Flujo Pico Obtenidas del Mini-Wright Peak Flow Meter, de 10 Individuos Que Fueron Examinados 30 Veces en el Curso de 2 Meses**

Media	DE <sup>*</sup>	Min	Max
476.5	27.9	390	510
326.6	46.4	240	400
694.1	15.0	655	720
566.0	39.0	480	630
469.3	24.5	400	540
473.0	24.1	400	510
484.7	37.1	400	530
447.8	19.1	410	500
508.3	26.9	450	570
426.4	21.6	380	460

\* DE = Desviación Estándar

De: Lebowitz, MD y cols. Chest, 81(5):566-570, 1982.

- B. Por ciento de Concordancia Global: Es una medida de repetición. La concordancia global se calcula como se muestra abajo (Cuadro 7), como de las observaciones concordantes dividida por el total.

### Cuadro 7

**Variación del Observador o del Instrumento: Por Ciento de Concordancia**

Lectura Número Dos	Lectura Número Uno			
	Anormal	Sospechosa	Dudosa	Normal
Anormal	Ⓐ	B	C	D
Sospechosa	E	Ⓕ	G	H
Dudosa	I	J	Ⓚ	L
Normal	M	N	O	Ⓟ

$$\text{Por Ciento de Concordancia} = \frac{A + F + K + P}{\text{Total}} \times 100$$

B. Ejemplo Hipotético: Dos diferentes radiólogos leyeron las placas de Rayos X de tórax de 1000 personas. Los resultados se muestran en el Cuadro 8.

**Cuadro 8**  
**Comparación de la Interpretación de Dos Radiólogos de las Placas de Rayos X de Tórax**

Interpretación por Radiólogo #1	Interpretación por Radiólogo #2			
	Anormal	Sospechosa	Dudosa	Normal
Anormal	300	80	25	5
Sospechosa	75	50	50	60
Dudosa	50	75	75	125
Normal	15	25	50	400

$$\text{Porcentaje de Concordancia} = \frac{\text{Numero Total de Diagnosticos en los Circulos}}{\text{Numero Total de Placas}} \times 100$$

- D. Ejemplo: Se llevó a cabo un estudio para establecer la validez y la confiabilidad de los cuestionarios sobre síntomas respiratorios en un grupo de trabajadores expuestos al asbesto (Bull Eur Physiopath Resp 14:177-188, 1978).

Se encuestó en dos ocasiones a un grupo de trabajadores de los muelles que estuvieron expuestos al asbesto empleándose dos cuestionarios. Se les administró a los trabajadores un cuestionario del Consejo de Investigación Médica (estándar) y un cuestionario respiratorio clínico (clínico) en 1976 y un año después en 1977. Los resultados se presentan en el Cuadro 9.

**Cuadro 9**  
**Confiabilidad de los Cuestionarios Clínico y Estándar en 60**  
**Trabajadores con el Mismo Estado de Tabaquismo en Ambas Encuestas**

Síntoma	Clínico			Estándar		
	Prevalencia		Concordancia	Prevalencia		Concordancia
	Encuesta 1	Encuesta 2		Encuesta 1	Encuesta 2	
Disnea	1.7	6.7	.93	6.7	10.0	.93
Tos	33.3	51.7	.65*	23.3	43.3	.70**
Flema	36.7	56.7	.63*	30.0	45.0	.62
Sibilancias	3.4	13.4	.83	25.0	38.4	.73
Antecedentes de Neumonía	25.0	20.0	.82	21.7	23.3	.92
Problema de Senos	26.7	25.0	.82	28.3	30.0	.85

\* p<.05

\*\* p<.01

- E. En muchas situaciones una gran proporción de aquellos a quienes se les hace una prueba tienen resultados negativos y es frecuente que exista una concordancia considerable entre las dos pruebas respecto a estos negativos o normales. Por lo tanto, cuando se estima el por ciento de concordancia para todos los sujetos de estudio este valor puede ser alto debido al gran número de sujetos negativos. Este valor alto puede encubrir una discordancia significativa entre los observadores respecto a cuáles sujetos están clasificando como positivos.

Un modo de manejarlo es el de no considerar a aquellos sujetos a quienes ambos observadores clasificaron como negativos (celda d) y calcular el por ciento de concordancia empleando como denominador sólo aquellos a quienes al menos un observador clasificó como anormales (celdas a, b y c).

**Figura 4**

		Observador 1	
		Positivo	Negativo
Observador 2	Positivo	a	b
	Negativo	c	d

Ignorar  
←

En las observaciones pareadas en las que por lo menos una de las observaciones de cada par fue positiva, el por ciento de concordancia es igual a: 
$$a = \frac{a}{a + b + c} \times 100$$

## VI. Evaluación de las Pruebas de Tamizaje

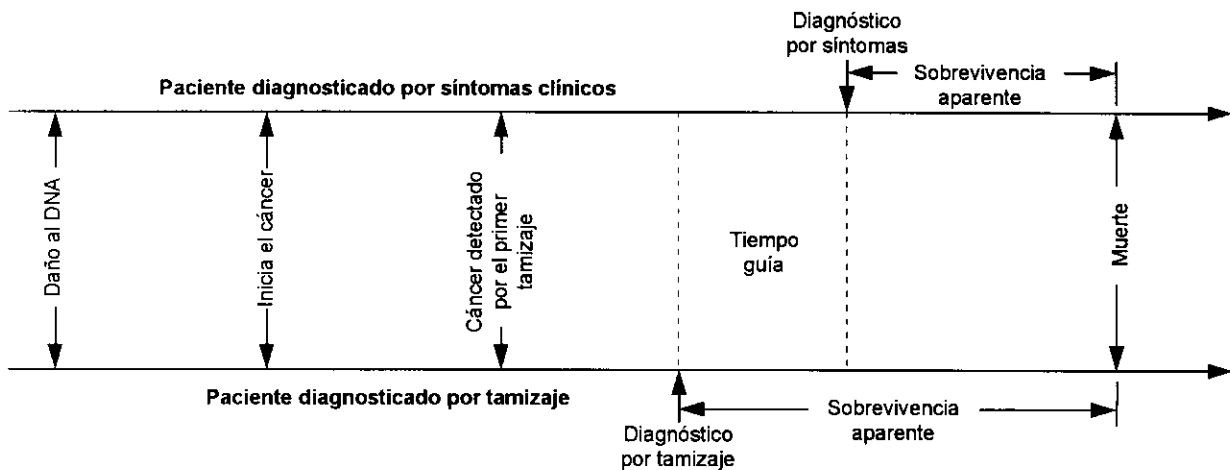
- A. **Ensayos aleatorizados:** Una prueba de tamizaje sería evaluada, de manera ideal, en una población representativa de las personas a quienes se les va aplicar la prueba. Se asignarían al azar grupos de tamaño suficiente para recibir o no recibir la prueba de tamizaje. El resultado más relevante es la reducción en la mortalidad de la enfermedad en cuestión.

Se prefiere como resultado a la mortalidad porque se sabe que el detectar tempranamente una enfermedad aparentemente aumenta la sobrevivencia, mediante el fenómeno conocido como "tiempo guía"<sup>1</sup> (Figura 5). El tiempo guía representa el intervalo en el cual se adelanta el diagnóstico; entonces, si el tamizaje produce un beneficio se debe reducir la mortalidad total, comparando a los individuos que se tamizaron con los que no se tamizaron.

<sup>1</sup> Cotéjar el término en español

**Figura 5**

Comparación de un paciente con un diagnóstico clínico de rutina de una enfermedad y un paciente con enfermedad detectada por tamizaje. El área sombreada representa el tiempo guía entre la detección por el tamizaje y el diagnóstico de rutina.



B. Datos observacionales: Aunque el ensayo clínico representa la modalidad ideal para la valoración de una prueba de tamizaje, con frecuencia las pruebas nuevas se introducen sin dicha evaluación. Los costos que se requieren para los ensayos pueden parecer prohibitivos y el razonamiento para introducir la prueba puede ser suficiente debido a su validez demostrada. Por ejemplo, la prueba de papanicolaou nunca fue evaluada en un ensayo clínico. Los datos observacionales pueden facilitar los suficientes conocimientos mediante la demostración del descenso en la mortalidad después de la introducción de una prueba, la variación geográfica consistente con la variación en el tamizaje y los cambios en el resultado intermedio (ej. estadio de la enfermedad) que es indicativo de un beneficio del tamizaje.

C. Criterios para la implementación de una prueba de tamizaje:

1. La enfermedad es del interés de la salud pública.
2. Se sabe que la intervención temprana reduce la mortalidad o la morbilidad.
3. Existe un grupo conocido de alto riesgo.
4. La prueba tiene riesgos mínimos y una validez satisfactoria.
5. El estudio de diagnóstico para obtener una prueba positiva tiene una morbilidad aceptable.

## REFERENCIAS

1. Altman DG: *Statistical Analysis of Comparison Between Laboratory Methods*. J Clin Pathol, 44(8):700, 1991.
2. Cohen J: *A Coefficient of Agreement for Nominal Scales*. Educ Psychol Meas, 20:37, 1960.
3. Fleiss JL: *Statistical Methods for Rates and Proportions*. 2nd Edition, Wiley, New York, 1981.
4. Landis JR and Koch GG: *The Measurement of Observer Agreement for Categorical Data*. Biometrics 33:159, 1977.
5. Lipman A and MacKenzie SG: *What is "Recall Bias" and Does it Exist?* Prevention of Physical and Mental Congenital Defects, Part C: Basic and Medical Science, Education, and Future Strategies, pp. 205-209, Alan R. Liss, Inc., 1985.
6. MacLure M and Willett WC: *Misinterpretation and Misuse of the Kappa Statistic*. Am J Epidemiol, 126:161, 1987.
7. MacMahon B: *Concepts of Multiple Factors*. In: Multiple Factors in the Causation of Environmentally Induced Disease, Lee DH and Kotin P (Eds). Academic Press, New York, 1972.
8. Schlesselman JJ: *Case-Control Studies: Design, Conduct, and Analysis*. Oxford University Press, New York, 1982.