UNIVERSIDADE DA CORUÑA

fic
facultade de
informática
da coruña

Evaluación de Sistemas Inteligentes

Contenido

1. Introducción
2. Tipos de evaluación
3. Características a tener en cuenta
4. Métodos cualitativos
5. Métodos cuantitativos
    1. Medidas de pares
    2. Medidas de grupo

Bibliografía

- A. Alonso Betanzos, B. Guijarro Berdiñas, A. Lozano Tello, J.T. Palma Méndez y M.J. Taboada Iglesias. Ingeniería del Conocimiento. Aspectos metodológicos. Pearson Educación 2004.

- E. Mosqueira Rey and V. Moret Bonillo. Validación de sistemas inteligentes. Tórculo Edicións 2001.

Objetivos

- Asegurar la calidad del producto desarrollado
- Asegurar el uso del SBC en dominios críticos
- Asegurar la aceptación del sistema en la rutina diaria

Problemas

- Estándar de oro
- Aproximación estándar
- Falta de:
  - Métricas de evaluación prácticas y comúnmente aceptadas
  - Falta de especificaciones concretas
  - Falta de herramientas de evaluación

Tipos de evaluación

- Verificación
- Validación
- Usabilidad
- Utilidad

Verificación

- Verificación del cumplimiento de las especificaciones
- Verificación de los mecanismos de inferencia
- Verificación de la base de conocimiento

Verificación de la Base de conocimiento

- Verificación de la consistencia
  - Reglas redundantes
    - R1: a→b y R2: a→b
    - R1: a→b, R2: b→c, R3:c→d, las reglas R4:a→d son redundantes.
  - Reglas conflictivas (inconsistencia)
    - Lógica booleana: P(x) y no P(x) deducibles a la vez
    - Lógica atributo-valor: BC U BH ---- Atr(x,a) y Atr(x,b) y a distinto de b
  - Subsunción de reglas
  - Cadenas circulares de reglas
  - Condiciones IF innecesarias

Verificación de la Base de conocimiento

- Verificación de la completitud
  - Valores no referenciados de atributos
    - Valores de atributo en ninguna premisa de regla
  - Valores ilegales de atributos
    - Valores en la premisa/conclusión que no son valores legales del atributo
  - Conclusiones inalcanzables
  - Reglas no disparables

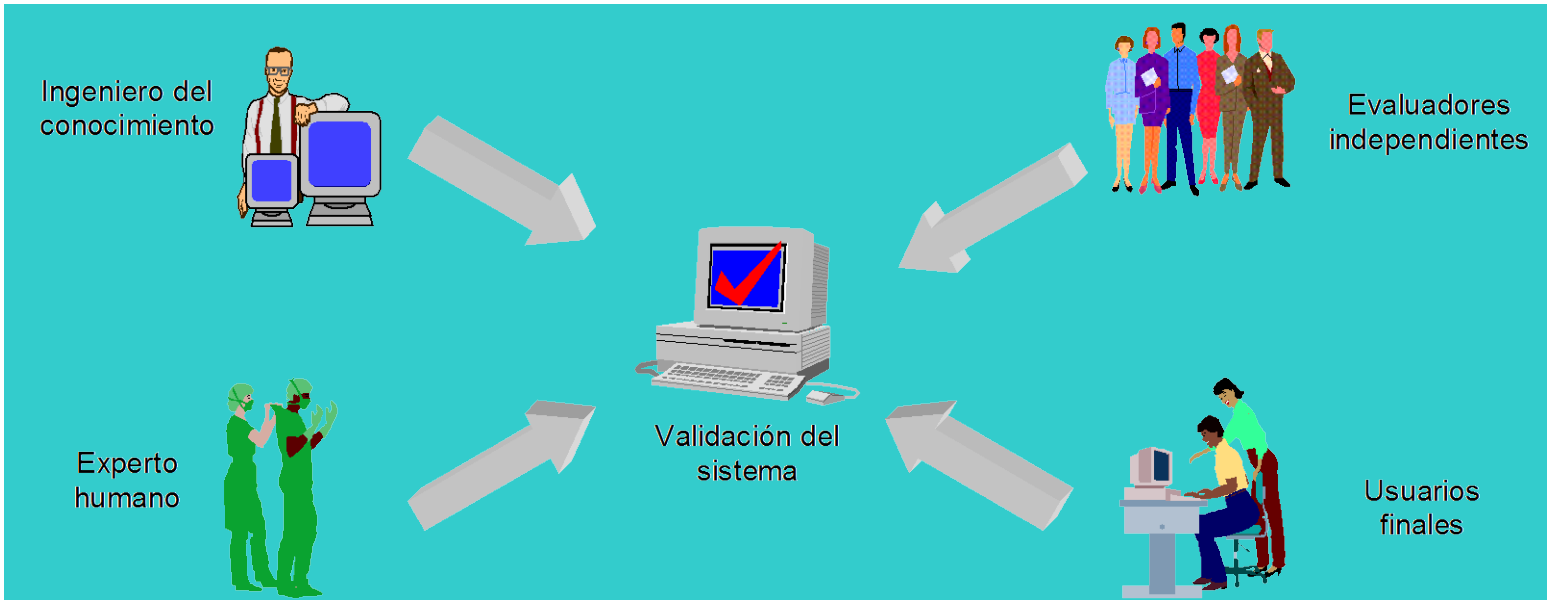
Herramientas de verificación

- Sistemas de verificación tabular
- Sistemas de propagación de restricciones
- Sistemas basados en redes de Petri
- Sistemas basados en grafos

Validación de sistemas inteligentes

- Personal involucrado en la validación
- Partes del sistema a validar
- Datos utilizados en la validación
- Referencia estándar utilizada

PERSONAL INVOLUCRADO EN LA VALIDACIÓN



- La falacia del superhombre:
  - Se le suele exigir más al sistema inteligente que al experto humano, sin tener en cuenta que el conocimiento del sistema inteligente es un modelo computacional del conocimiento de los expertos humanos

PARTES DEL SISTEMA QUE DEBEN SER VALIDADAS

- Resultados finales
  - Performance general del sistema
- Resultados intermedios
  - Descripción del funcionamiento interno del sistema
  - Permite corregir errores cometidos
- Razonamiento seguido
  - Un proceso de razonamiento incorrecto puede ser fuente de errores cuando queramos ampliar la base de conocimientos del sistema
  - Tenemos que diseñar sistemas que “piensen” como lo haría un experto humano… también en la forma

CASUÍSTICA DE LA VALIDACIÓN

- La muestra debe ser
  - Suficiente
  - Suficientemente representativa

- Proceso
  - Obtención de la casuística de validación
  - Transferencia de los datos al sistema que ha de interpretarlos
  - Resultados y criterios son la entrada del proceso de validación en el que se analiza el rendimiento del sistema

Tipos de validación

- Criterio de referencia
  - Validación contra los expertos
    - Variabilidad inter e intra expertos
    - Grupo o consenso de expertos
  - Validación contra el problema
    - Exigencias “supersistema”
    - Imposibilidad de comprobación real

VALIDACIÓN CONTRA EL EXPERTO

- Se utilizan las opiniones y las interpretaciones de los expertos humanos como criterio de validación

- Puede haber discrepancias entre expertos o sesgos en este tipo de validación
  - Factores externos: estrés,…
  - Pueden no ser independientes
  - Pueden ser ambiguos
  - Pueden pertenecer a distintas escuelas de pensamiento
  - Pueden tener sus propias ideas sobre el sistema que están validando y, por lo tanto, no ser objetivos

VALIDACIÓN CONTRA EL EXPERTO

- Hay tres procedimientos diferentes:
  - Validación contra un único experto
    - Ventajas
      - Suele haber al menos un experto disponible
    - Inconvenientes
      - La validación puede no ser fiable
  - Validación contra un grupo de expertos
    - Ventajas
      - No estamos supeditados a una única opinión
      - Permite comparar el grado de consistencia entre expertos del dominio
    - Inconvenientes
      - Los expertos no son todos iguales: ¿Cómo medir el rendimiento del sistema?
  - Validación contra un consenso de expertos
    - Ventajas
      - En teoría es el método más objetivo y fiable
    - Inconvenientes
      - Puede haber un experto especialmente influyente
      - ¿Cómo se mide el consenso?

VALIDACIÓN CONTRA EL PROBLEMA

- Nuestro sistema: ¿acierta realmente, o resuelve convenientemente, el problema planteado?

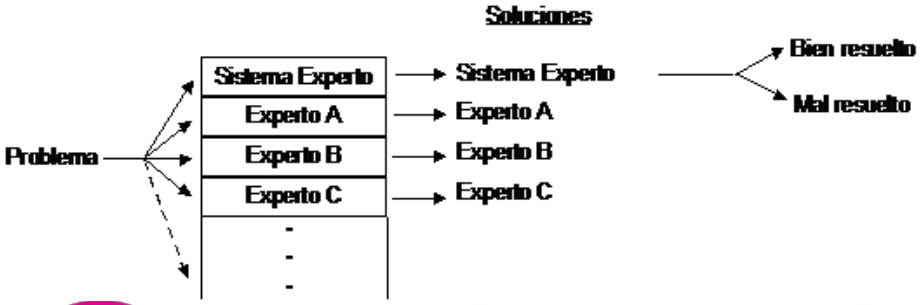
- Ventajas
  - Método completamente objetivo
  - La solución real puede verse en el problema
  - Si nuestro sistema discrepa con el experto humano, pero coincide con la respuesta del problema, la credibilidad del sistema aumenta

- Inconvenientes
  - Falacia del superhombre
  - No siempre puede realizarse una validación contra el problema

TIPOS DE CASOS DE VALIDACIÓN

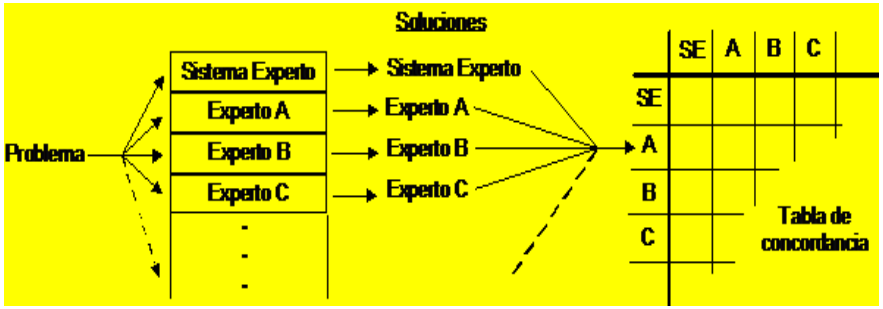
VALIDACIÓN PROSPECTIVA

- Sobre casos reales todavía no resueltos y análisis de las interpretaciones propuestas (relacionada con validación orientada al problema).
- No se utilizan casos almacenados en bases de datos sino casos que en ese momento están siendo tratados por expertos humanos.
- Se puede evaluar, no sólo la corrección de los resultados, sino aspectos referentes al uso del sistema.
- El problema surge, al igual que en la validación contra el problema, cuando el dominio de aplicación es crítico, y el sistema intenta manipular el entorno
- Se utiliza una vez que hemos validado el sistema en un entorno de desarrollo y utilizando casos históricos, y se desea realizar una nueva validación en el campo de aplicación del sistema.
- Es recomendada para sistemas que tienen capacidades de predicción.



VALIDACIÓN RETROSPECTIVA

- Sobre casos históricos ya resueltos y almacenados
- Es la más comúnmente realizada en los sistemas inteligentes, y los casos utilizados pueden incluir como referencia de validación tanto opiniones de expertos humanos, como la solución real al problema planteado (aunque generalmente suele ser la primera).
- Se suele utilizar en las etapas de desarrollo del sistema, antes de que este se instale en su entorno de trabajo habitual



Métodos de validación

- Métodos de validación cualitativos
  - Validación de superficie
  - Prueba de Turing
  - Test de campo
  - Validación de subsistemas
  - Análisis de sensibilidad
  - Grupos de control
- Métodos de validación cuantitativos
  - Medidas de pares
  - Medidas de grupo

Métodos cualitativos

Emplean técnicas subjetivas para la comparación del rendimiento.

- Validación superficial
  - Validación informal en la que el ingeniero de conocimiento, el usuario y el experto humano analizan el rendimiento del sistema.
- Test de Turing
  - Se presentan una serie de casos al sistema y a diversos expertos humanos. Posteriormente estos casos son pasados a otros expertos que deben criticarlos e identificar al sistema experto.
- Test de campo (pruebas beta)
  - Se coloca al sistema en el campo en el que se va a utilizar finalmente y se observan los errores que se producen.
- Validación de subsistemas
  - Se descompone el sistema en subsistemas y se valida cada uno de estos por separado.
- Análisis de sensitividad
  - Sirve para realizar cambios sistemáticos en algunas variables y ver que efecto tiene en la respuesta del sistema.

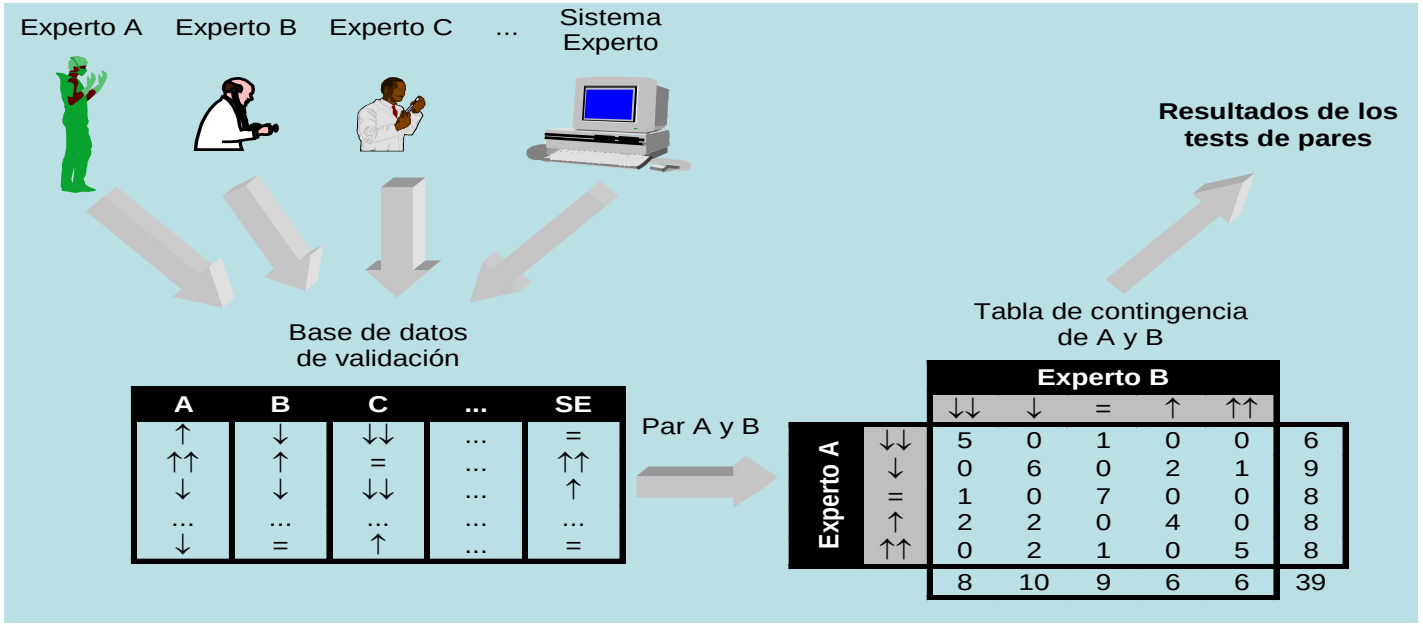
Métodos cuantitativos

Emplean medidas estadísticas para comparar el rendimiento de un sistema inteligente con un criterio de validación

- Medidas de pares:
  - Pretenden evaluar el grado de acuerdo y/o asociación entre los resultados de dos expertos.
  - Se dividen en medidas de acuerdo y medidas de asociación
- Medidas de grupo
  - Se utilizan cuando no existe una referencia estándar como criterio, por lo que es necesario validar contra un grupo de expertos
  - Utilizan como información de entrada los resultados de las medidas de pares obtenidos para cada posible par de expertos
- Ratios de acuerdo
  - Se encargan de medir el acuerdo existente entre un experto y una referencia estándar (un consenso entre expertos, o la solución real al problema).
  - Para esto mismo podríamos utilizar los tests de pares, pero la diferencia es que ahora sí sabemos cuál es la referencia correcta, y podemos establecer el error cometido.

Validación Cuantitativa
Medidas de pares

- Obtención de medidas de pares:
  1. Se desarrolla una tabla de contingencia, que relaciona los resultados de los expertos a considerar
  2. Se extrae de la tabla la medida de pares deseada. Distinguimos dos grupos: medidas de acuerdo y medidas de asociación.



A	B	C	...	SE
↑	↓	↓↓	...	=
↑↑	↑	=	...	↑↑
↓	↓	↓↓	...	↑
...	...	...	...	...
↓	=	↑	...	=

Experto A \ Experto B	↓↓	↓	=	↑	↑↑	
↓↓	5	0	1	0	0	6
↓	0	6	0	2	1	9
=	1	0	7	0	0	8
↑	2	2	0	4	0	8
↑↑	0	2	1	0	5	8
	8	10	9	6	6	39

Validación Cuantitativa
Medidas de pares (II)

- Tablas de contingencia
  - Es una tabla que relaciona de forma cruzada datos categóricos
  - La utilizaremos para relacionar los resultados de dos expertos.
  - Cada celda n_{ij} representa el número de casos en los que el experto A selecciona la categoría *i*, y el experto B la categoría *j*. Se denominan frecuencias absolutas.

		Resultados experto B				
		1	2	...	*k*	Totales
Resultados experto A	1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
	2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
	...	...	...	...	...	...
	k	n_{k1}	n_{k2}	...	n_{kk}	$n_{k.}$
	Totales	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	$n_{..} = N$

Validación Cuantitativa
Medidas de pares

- El número total de casos observados en la fila *i* ($n_{i.}$) o en la columna *j* ($n_{.j}$) se denomina como frecuencia absoluta marginal.
- El total de casos de la muestra es el término $n_{..}$, y también se representa como N

$$n_{i.} = n_{i1} + n_{i2} + \ldots + n_{ik} = \sum_{j=1}^{k} n_{ij}$$

$$n_{.j} = n_{1j} + n_{2j} + \ldots + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

$$n_{..} = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{k} n_{.j}$$

- Otra forma de generar las tablas de contingencia es en base a frecuencias relativas, o proporciones. De esta forma, la frecuencia relativa correspondiente a la celda n_{ij} se calcula así:

$$p_{ij} = \frac{n_{ij}}{N}$$

Validación Cuantitativa
Medidas de pares

- Ejemplo de base de datos de validación
  - Cuatro expertos humanos (A, B, C, D) y un sistema experto (SE)

Casos	A	B	C	D	SE
1	ALTO	ALTO	ALTO	ALTO	ALTO
2	ALTO	BAJO	ALTO	ALTO	BAJO
3	BAJO	BAJO	NORMAL	NORMAL	BAJO
4	NORMAL	BAJO	NORMAL	NORMAL	NORMAL
5	MUY ALTO	MUY ALTO	ALTO	ALTO	MUY ALTO
6	BAJO	BAJO	BAJO	BAJO	NORMAL
7	MUY BAJO	MUY BAJO	NORMAL	BAJO	BAJO
8	NORMAL	NORMAL	NORMAL	ALTO	NORMAL
9	NORMAL	NORMAL	BAJO	MUY BAJO	NORMAL
10	BAJO	BAJO	BAJO	ALTO	BAJO

		Experto SE					
		MUY BAJO	BAJO	NORMAL	ALTO	MUY ALTO	
Experto A	MUY BAJO	0	1	0	0	0	1
	BAJO	0	2	1	0	0	3
	NORMAL	0	0	3	0	0	3
	ALTO	0	1	0	1	0	2
	MUY ALTO	0	0	0	0	1	1
		0	4	4	1	1	10

		Experto SE					
		MUY BAJO	BAJO	NORMAL	ALTO	MUY ALTO	
Experto A	MUY BAJO	0.0	0.1	0.0	0.0	0.0	0.1
	BAJO	0.0	0.2	0.1	0.0	0.0	0.3
	NORMAL	0.0	0.0	0.3	0.0	0.0	0.3
	ALTO	0.0	0.1	0.0	0.1	0.0	0.2
	MUY ALTO	0.0	0.0	0.0	0.0	0.1	0.1
		0.0	0.4	0.4	0.1	0.1	1

Medidas de pares: Ejemplos

	Sistema de referencia	
Sistema a evaluar	*Positivos*	*Negativos*
Positivos	Verdaderos positivos	Falsos positivos
Negativos	Falsos negativos	Verdaderos negativos

	Criterio de referencia		
Sistema	*Tempranas*	*Tardías*	*Variables*
Tempranas	52	2	0
Tardías	2	49	3
Variables	0	2	51

Table 2
Validation results obtained for the CAFE system for each of the parameters/patterns of a CTG analysis (TP = true positives, FP = false positives)

Validated issue	%TP	%FP	Validation criteria*
Symbolic classification of the baseline	100.0	0.0	1
Abrupt baseline changes (even if the symbolic category does not change)	91.8	12.7	1
Abrupt baseline changes indicating a bradycardia	95.0	6.2	1
Abrupt baseline changes indicating a tachycardia	100.0	0.0	1
Silent rhythm	94.4	0.0	1
Saltatory rhythm	62.5	64.7	1
Accelerations	86.5	15.5	1
Decelerations	83.8	21.3	1
Classification of decelerations by type	94.4	2.8	2
Early	96.3	1.9	2
Late	92.5	4.6	2
Variable	94.4	1.9	2
Detection of artefacts	83.5	0.0	3

*1 = A′, crossed opinion of expert A.
2 = classification provided in cardiotocographic record atlases.
3 = total agreement among experts exists.

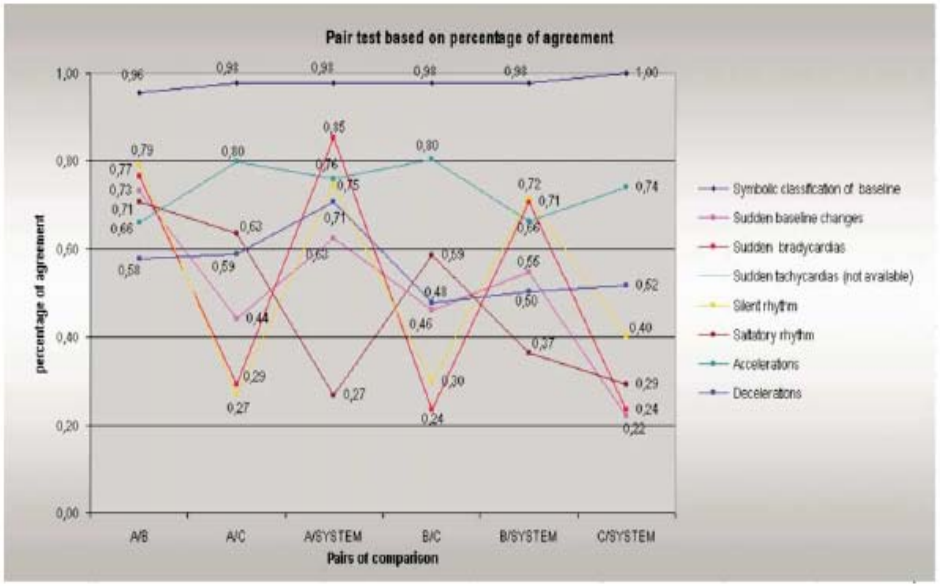


Fig. 16. First validation round. Results of the pairs' test comparing experts and system for each of the validation categories by means of the agreement index. A, B, C are the three experts participating in the study.

Medidas de pares

- Algunas de las más utilizadas
  - Índice de acuerdo
  - Índice de acuerdo dentro de uno
  - Kappa
  - Kappa ponderada.

Medidas de Acuerdo
Índice de acuerdo

- Índice de acuerdo:
  - Es el cociente entre el número de observaciones de acuerdo y el número de observaciones totales:

$$Indice = \frac{\sum_{i=1, j=1, i=j}^{k} n_{ij}}{N} = \sum_{i=1, j=1, i=j}^{k} p_{ij}$$

- Propiedades:
  - Toma valores en [0,1], donde 1 es **acuerdo completo** y 0 es **desacuerdo completo**.
  - No está afectado por el orden de las categorías
  - Un experto tiene acuerdo completo consigo mismo.
  - El acuerdo es una relación transitiva y simétrica.

Medidas de Acuerdo
Índice de acuerdo

- Ventaja:
  - Sencillez de interpretación
- Desventaja:
  - No tiene en cuenta los acuerdos debidos a la casualidad (defecto importante si el número de categorías posibles es pequeño).
  - No matiza la importancia de los errores (los trata a todos de la misma forma)
- Ejemplo:
  - Con las tablas anteriores, el valor resultante sería:

$$Porcentaje = \frac{0+2+3+1+1}{10} = \frac{7}{10} = 0.7$$

  - Y por proporciones:

$$Porcentaje = 0.0 + 0.2 + 0.3 + 0.1 + 0.1 = 0.7$$

Medidas de pares (III): Índice de acuerdo. Ejemplos

	Experto A		
Sistema	*Tempranas*	¬ *Tempranas*	$p_{i.}$
Tempranas	61	21	$(61+21)/158 = 0{,}52$
¬ *Tempranas*	28	48	$(28+48)/158 = 0{,}48$
$p_{.j}$	$(61+28)/158 = 0{,}56$	$(21+48)/(158) = 0{,}44$	

Po= 61/158 + 48/158= 0,69

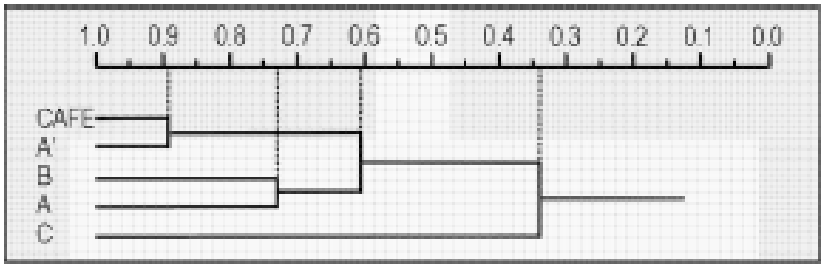


Fig. 19. Second validation round. Cluster analysis comparing experts and system overall. Groups are built on the basis of the group average taking the agreement index as the measurement distance. A, B, and C represent the clinicians in terms of experience. A′ corresponds to the second opinion of Expert A.

Medidas de Acuerdo
Índice de acuerdo dentro de uno

- En ciertas escalas semánticas, es difícil distinguir entre categorías contiguas (por ejemplo: “bajo” y “algo bajo”).

- El índice de acuerdo dentro de uno considera como acuerdos parciales aquellos casos que se distinguen en una única etiqueta lingüística consecutiva.

$$\text{Índice de acuerdo dentro de uno} = \frac{\sum_{\substack{i=1,\, j=1 \\ i=j,\, i=j\pm 1}}^{k} n_{ij}}{N} = \sum_{\substack{i=1,\, j=1 \\ i=j,\, i=j\pm 1}}^{k} p_{ij}$$

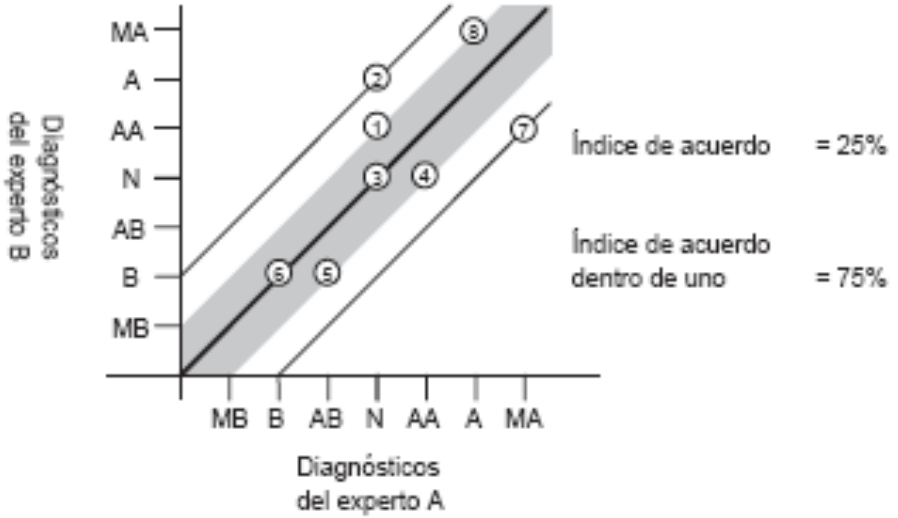
Medidas de pares (IV): Índice de acuerdo dentro de uno

Escala semántica para la clasificación simbólica de una determinada variable	Muy Bajo (MB)	Bajo (B)	Algo Bajo (AB)	Normal (N)	Algo Alto (AA)	Alto (A)	Muy Alto (MA)

(a)

Casos	Experto A	Experto B
1	Normal	Algo Alto
2	Normal	Alto
3	Normal	Normal
4	Algo Alto	Normal
5	Algo Bajo	Bajo
6	Bajo	Bajo
7	Muy Alto	Algo Alto
8	Alto	Muy Alto

(b)



(c)

Medidas de Acuerdo
Kappa

- Motivación
  - Primeros intentos de corregir acuerdos debidos a la casualidad basados en el test chi-cuadrado.
  - Pero este tipo de medidas no es adecuado, ya que mide grados de asociación: dos expertos pueden estar asociados, pero no en la dirección del acuerdo.

- Ideas generales sobre el índice kappa
  - Propuesto por Cohen (1960).
  - Se basa en dos cantidades:
    - Proporción de acuerdo observado (p_0).
    - Proporción de acuerdo esperado debido a la casualidad (p_c).
  - Por tanto, 1-p_c representa el máximo acuerdo posible eliminada la casualidad, y p_0-p_c el acuerdo obtenido eliminando casualidad.

Medidas de Acuerdo
Kappa

- Cálculo del índice kappa:
  - La expresión general es la siguiente:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

  - El término p_0 es el porcentaje de acuerdo visto anteriormente. El término p_c se calculan mediante la siguiente expresión:

$$p_c = \sum_{i=1, j=1, i=j}^{k} p_{i.} p_{.j}$$

  - Para calcular p_c, tres pasos:
    1. Proporciones marginales
    2. Acuerdo debido a la casualidad en cada celda (producto de las proporciones marginales correspondientes)
    3. Índice p_c (suma de los acuerdos debidos a casualidad en la diagonal principal).

Medidas de Acuerdo
Kappa

- Cálculo de p_c

		Experto SE					
		MUY BAJO	BAJO	NORMAL	ALTO	MUY ALTO	
Experto A	MUY BAJO	0.0 (0.0) ②	0.1	0.0	0.0	0.0	0.1 ①
	BAJO	0.0	0.2 (0.12)	0.1	0.0	0.0	0.3
	NORMAL	0.0	0.0	0.3 (0.12)	0.0	0.0	0.3
	ALTO	0.0	0.1	0.0	0.1 (0.02)	0.0	0.2
	MUY ALTO	0.0	0.0	0.0	0.0	0.1 (0.01)	0.1
		0.0 ①	0.4	0.4	0.1	0.1	1

$p_o = 0.70$
$p_c = 0.27$
$\kappa = 0.589$ ③

Medidas de Acuerdo
Kappa

- Interpretación de kappa
  - kappa < 0
    - El acuerdo observado es menor al que se esperaría debido a la casualidad.
  - kappa = 0
    - El acuerdo observado es exactamente igual al que se esperaría debido a la casualidad.
  - kappa = 1
    - El acuerdo es completo.

- Posibles valores de kappa:
  - Si $p_0=p_c$ entonces kappa es cero.
  - Si $p_0>p_c$ kappa es positiva.
    - El valor máximo de kappa es +1 (en el caso de acuerdo perfecto p_0 =1).
  - Si $p_0<p_c$ el valor es negativo.
    - Es difícil determinar el límite negativo de kappa. De todas formas, un valor negativo de kappa no indicaría error debido a la casualidad, sino un error sistemático en la interpretación de las categorías.

- Valores bajos de kappa también pueden indicar una distribución poco balanceada entre las distintas clases.

Medidas de Acuerdo
Kappa

- Interpretación básica de kappa de Landis y Koch

Kappa	*Nivel de acuerdo*
< 0.00	Nulo
0.00 – 0.20	Insuficiente
0.21 – 0.40	Ligero
0.41 – 0.60	Moderado
0.61 – 0.80	Sustancial
0.81 – 1.00	Casi perfecto o perfecto

- Inconvenientes:
  - Problema: trata todas las discordancias de la misma manera: todas las celdas que no pertenecen a la diagonal principal se penalizan por igual.
  - Por tanto la penalización por equivocarse entre “Muy Alto” y “Alto” es la misma que la de equivocarse entre “Muy Alto” y “Muy Bajo”.

- Para solucionar esto, se desarrolla el coeficiente de kappa ponderada.

Indice de acuerdo kappa: Ejemplos

	Experto A		
Sistema	*Tempranas*	¬ *Tempranas*	$p_{i.}$
Tempranas	61	21	$(61+21)/158 = 0{,}52$
¬ *Tempranas*	28	48	$(28+48)/158 = 0{,}48$
$p_{.j}$	$(61+28)/158 = 0{,}56$	$(21+48)/(158) = 0{,}44$	

$$P_0 = \sum_{i=1}^{k} N_{ii}/N = (61+48)/158 = 0{,}69$$

$$\kappa = \frac{p_0 - p_c}{1 - p_c} = \frac{0{,}69 - (0{,}52 \times 0{,}56 + 0{,}44 \times 0{,}48)}{1 - (0{,}52 \times 0{,}56 + 0{,}44 \times 0{,}48)} = 0{,}38$$

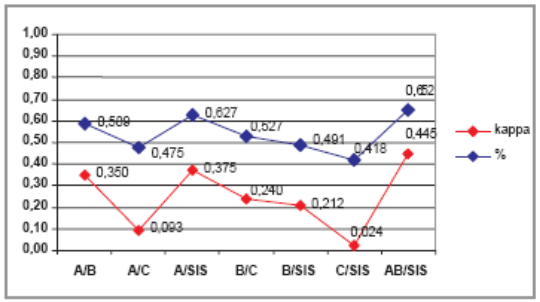
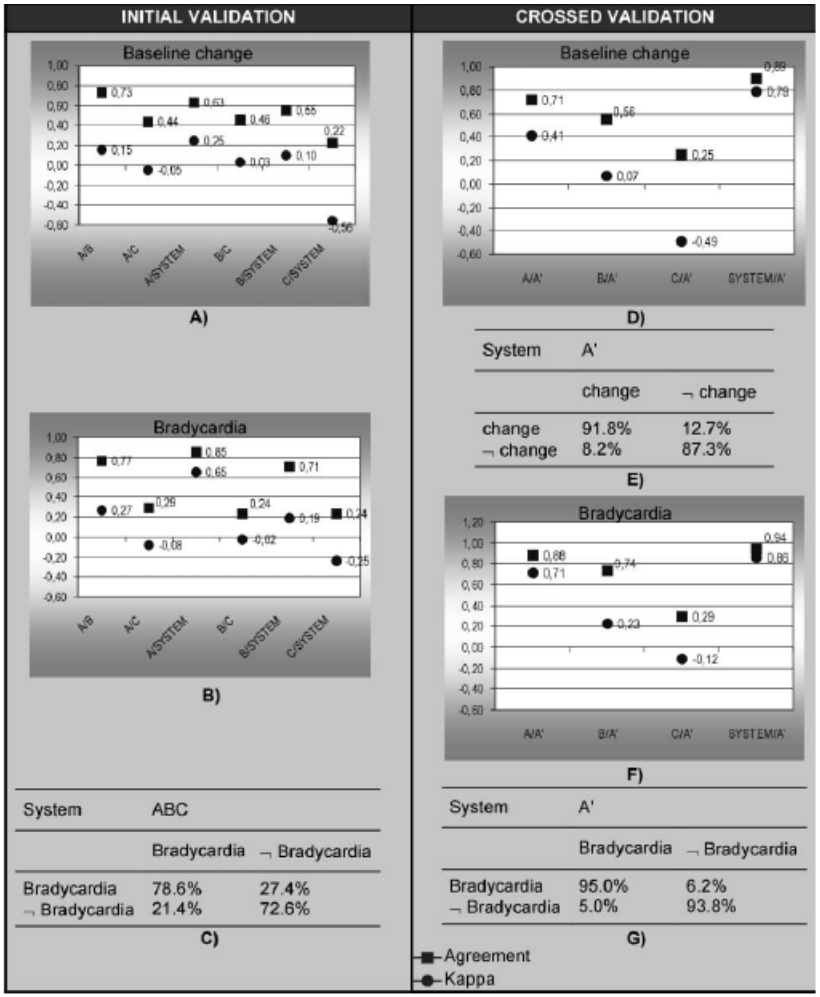


Figura 11.4: Pruebas de Pares frente a los pares de comparación. A, B, C son los expertos humanos por orden de experiencia y SIS es el sistema inteligente de diagnóstico. AB se refiere a la opinión consensuada de los expertos A y B.



Medidas de Acuerdo
Kappa ponderada

- Kappa ponderada:
  - Medida de acuerdo que corrige aquellos acuerdos debidos a la casualidad y pondera de forma distinta los desacuerdos encontrados.
  - La ponderación se realiza definiendo una matriz de pesos en la que a cada par de categorías *ij* se le asigna un peso v_{ij}. En la diagonal principal se suele asignar el peso 0.

- Propiedades:
  - Si multiplicamos todos los valores de la matriz de pesos por un valor positivo, el índice no cambia.

Medidas de Acuerdo
Kappa ponderada

- Cálculo:
  - En primer lugar, definimos las proporciones de desacuerdo q_0=1-p_0 y q_c=1-p_c. El índice kappa original queda ahora:

$$\kappa = \frac{q_c - q_0}{q_c} = 1 - \frac{q_0}{q_c}$$

  - Ahora reemplazamos estas proporciones por aquellas que tienen en cuenta los pesos:

$$q_0^{'} = \frac{\sum_{i=1,j=1}^{k} v_{ij} p_{oij}}{v_{\max}} \qquad q_c^{'} = \frac{\sum_{i=1,j=1}^{k} v_{ij} p_{cij}}{v_{\max}}$$

  - El valor de kappa ponderada es:

$$\kappa_w = 1 - \frac{q_0^{'}}{q_c^{'}}$$

Medidas de Acuerdo
Kappa ponderada

- Pesos de acuerdo:
  - También podemos definir kappa ponderada en base a pesos de acuerdo $w_{ij.}$
    - Asignamos el acuerdo máximo a la diagonal.
    - En el resto de las celdas los valores irán decreciendo hasta que no haya ningún acuerdo (convenientemente con peso 0).
  - Definimos la proporción de acuerdo ponderado observado y debido a la casualidad así:

$$p_0^{'} = \frac{\sum_{i=1,j=1}^{k} w_{ij} p_{0ij}}{w_{\max}} \qquad p_c^{'} = \frac{\sum_{i=1,j=1}^{k} w_{ij} p_{cij}}{w_{\max}}$$

  - La expresión de kappa ponderada queda:

$$\kappa_w = \frac{p_0^{'} - p_c^{'}}{1 - p_c^{'}}$$

Ratios de acuerdo y Coef. Jaccard

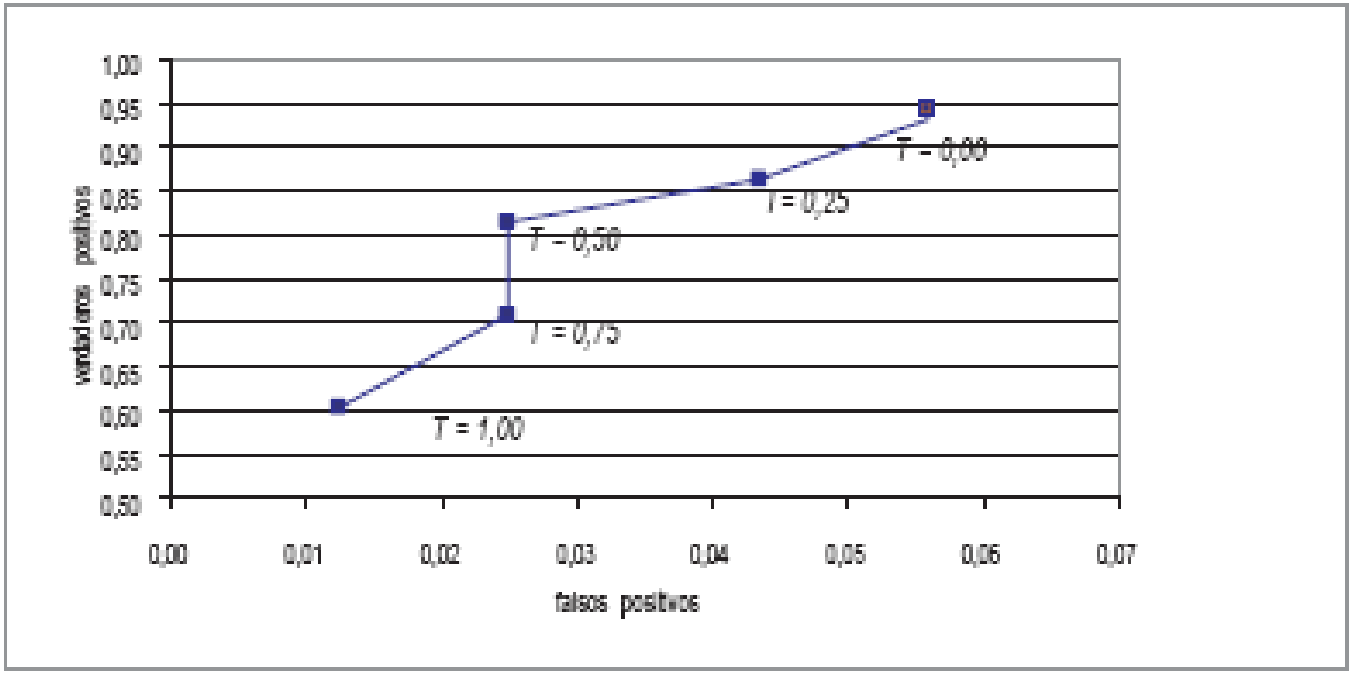
- Ratios de acuerdo
  - Comparan las interpretaciones de uno de los integrantes (sistema o experto) con una referencia estándar (consenso entre expertos, solución real, etc.)
  - Se construye una tabla de contingencia y se calcula S,E, VPP y VPN.
    - S= VP/(VP+FN)
    - E= VN/(FP+VN)
    - VPP=VP(VP+FP)
    - VPN=VN(VN+FN)

Clase	S	E	VPP	VPN
Tempranas	0,96	0,98	0,96	0,98
Tardías	0,93	0,96	0,91	0,96
Variables	0,94	0,98	0,96	0,97
Total	0,94	0,97	0,94	0,97

- Coeficiente de Jaccard
  - Sólo para casos con dos categorías de clasificación
  - Elimina del cálculo del porcentaje de acuerdo los casos en los que dos elementos del par de comparación no han diagnosticado la categoría considerada
  - Jaccard= VP(VP+FP+FN)
  - Es más adecuado para la coincidencia de los resultados positivos

Medidas de pares

- Curvas ROC



Validación

Metodos con o sin referencia estándar

- Una referencia estándar no siempre va a estar disponible
  - La opinión de los expertos es susceptible de variar.
  - Los expertos pueden no ser independientes.
  - La adquisición de la información puede influir en la opinión de los expertos
- Métodos para desarrollar una referencia estándar.
  - Método Delphi.- Se pasan las opiniones de los expertos a otro grupo de expertos, de forma anónima, para que lo someta a evaluación. El proceso sigue hasta que se llega a un consenso.
- Si no podemos desarrollar una referencia estándar
  - Medidas de Williams
  - Análisis Cluster
  - Escalamiento multidimensional.

Metodos sin referencia estándar

Medidas de Williams

Dado un conjunto de expertos y un experto aislado, ¿dicho experto está de acuerdo con el conjunto de expertos tan a menudo como lo están los miembros del grupo entre sí?

$$P_{(a,b)} = \frac{\sum_{i_a=1}^{k}\sum_{i_b=1}^{k} m_{i_a i_b}}{N} \quad i_a = i_b$$

$$P_n = 2\frac{\sum_{a=1}^{n-1}\sum_{b=a+1}^{n} P_{(a,b)}}{\left[n(n-1)\right]}$$

$$P_0 = \frac{\sum_{a=1}^{n} P_{(0,a)}}{n}$$

$$I_n = \frac{P_o}{P_n}$$

Ejemplo

Expertos: 0, 1, 2 y 3

$$P_3 = \frac{P(1,2) + P(1,3) + P(2,3)}{3}$$

$$P_0 = \frac{P(0,1) + P(0,2) + P(0,3)}{3}$$

$$I_3 = \frac{P_o}{P_3}$$

Metodos sin referencia estándar

Medidas de Williams

%	A	B	C	D	SE
A		0.8	0.6	0.4	0.7
B	0.8		0.4	0.2	0.7
C	0.6	0.4		0.6	0.4
D	0.4	0.2	0.6		0.3
SE	0.7	0.7	0.4	0.3	

Metodos sin referencia estándar

Medidas de Williams

$$In = \left(\frac{Po}{Pn} \right)$$

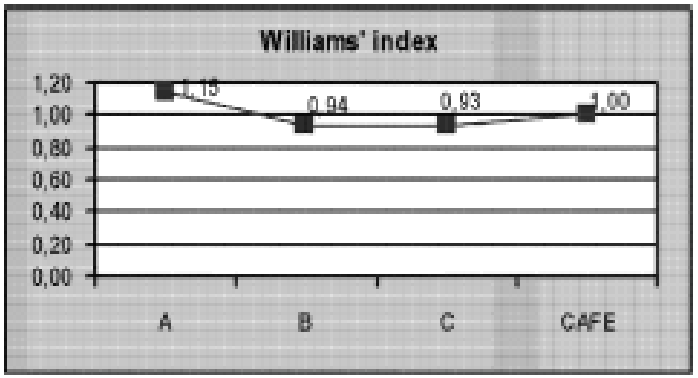
$$P0 = \left(\frac{1}{4} \right)(0.7 + 0.7 + 0.4 + 0.3) = \frac{21}{4} = 0.525$$

$$Pn = \frac{2}{4(4-1)}(0.8 + 0.6 + 0.4 + 0.4 + 0.2 + 0.6) = 0.500$$

$$In = 1.05$$

Índice Williams	**Interpretación:** El acuerdo entre el experto aislado y el resto de los expertos es ...
$0 \leq I_n < 1$	... menor que entre los expertos del grupo
$I_n = 1$	... igual al acuerdo entre los miembros del grupo
$I_n > 1$	... mayor que el acuerdo entre los miembros del grupo

Indice de Williams. Ejemplo



Value	Interpretation
> 1.00	Agreement between isolated expert and group of experts is greater than agreement among members of group
1.00	Agreement between isolated expert and group of experts is equal to agreement among members of group
[0.00-1.00)	Agreement between isolated expert and group is less than agreement among members of group

Fig. 18. First validation round. Williams' index (calculated on the basis of the agreement index) and its interpretation. A, B, C correspond to the three experts participating in the study.

Medidas de grupo

- Cluster jerárquico

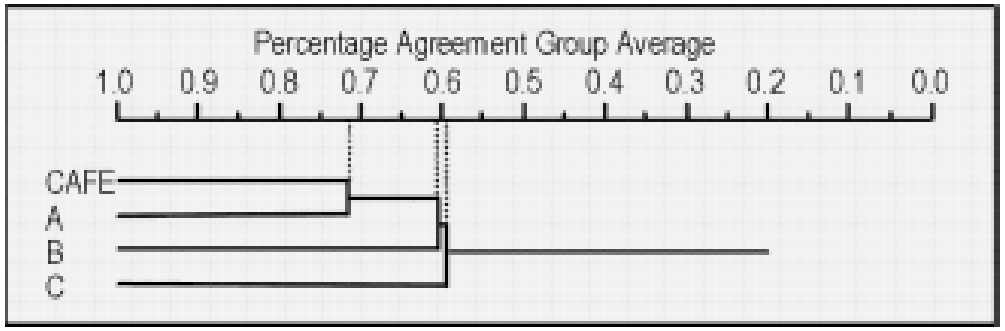


Fig. 17. First validation round. Cluster analysis comparing experts and system overall. Groups are built on the basis of the group average taking the agreement index as the measurement distance. A, B, and C represent the clinicians in terms of experience.

Medidas de grupo

- EMD

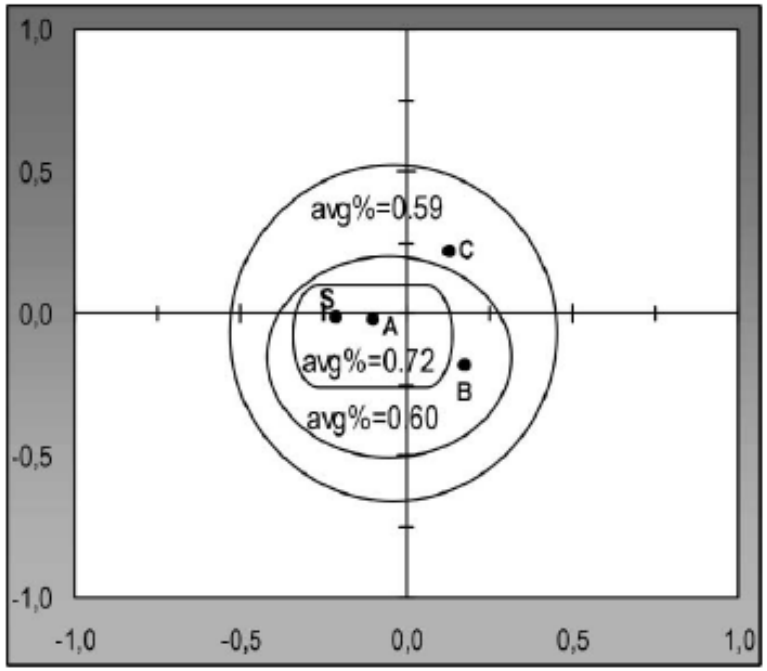


Fig. 11. Results of the MDS using the percentage agreement. The cluster analysis is superimposed in the form of bubles: A, B, C = experts; S = system; avg% = group averaged percentage agreement among the cluster's members.

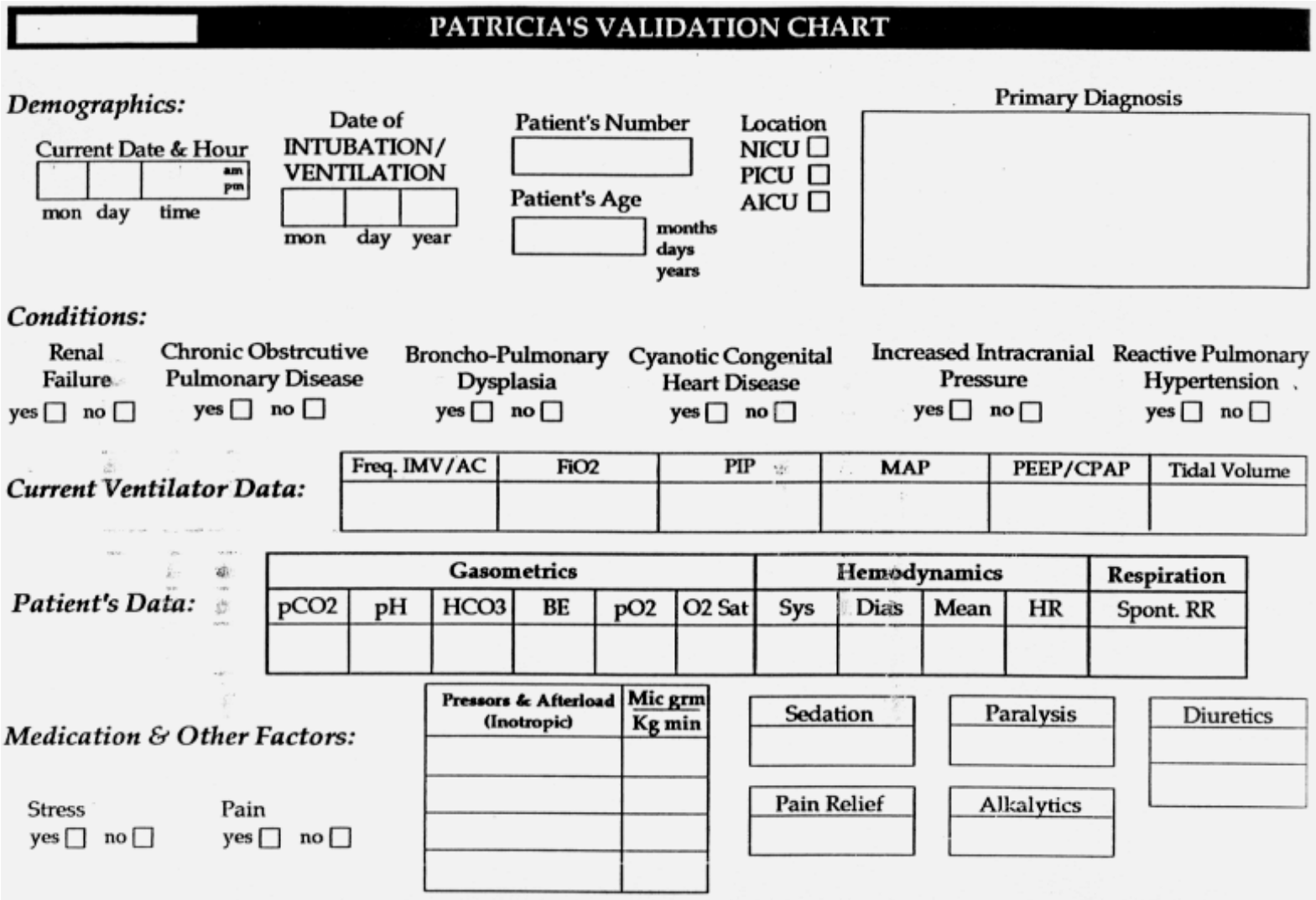
•Medidas de dispersión y tendencias

ERRORES COMETIDOS EN LA VALIDACIÓN

- Errores de comisión
- Errores por omisión

	Sistema válido	Sistema no válido
Sistema aceptado como válido	DECISIÓN CORRECTA	ERROR TIPO II Riesgo para usuario
Sistema no aceptado como válido	ERROR TIPO I Riesgo para ingeniero	DECISIÓN CORRECTA

Un ejemplo de validación



PATRICIA'S VALIDATION CHART

Demographics:

Current Date & Hour — mon day time (am pm)

Date of INTUBATION/VENTILATION — mon day year

Patient's Number

Patient's Age — months days years

Location: NICU ☐ PICU ☐ AICU ☐

Primary Diagnosis

Conditions:

Renal Failure	Chronic Obstrcutive Pulmonary Disease	Broncho-Pulmonary Dysplasia	Cyanotic Congenital Heart Disease	Increased Intracranial Pressure	Reactive Pulmonary Hypertension
yes ☐ no ☐	yes ☐ no ☐	yes ☐ no ☐	yes ☐ no ☐	yes ☐ no ☐	yes ☐ no ☐

Current Ventilator Data:

Freq. IMV/AC	FiO2	PIP	MAP	PEEP/CPAP	Tidal Volume

Patient's Data:

Gasometrics						Hemodynamics				Respiration
pCO2	pH	HCO3	BE	pO2	O2 Sat	Sys	Dias	Mean	HR	Spont. RR

Medication & Other Factors:

Pressors & Afterload (Inotropic)	Mic grm Kg min

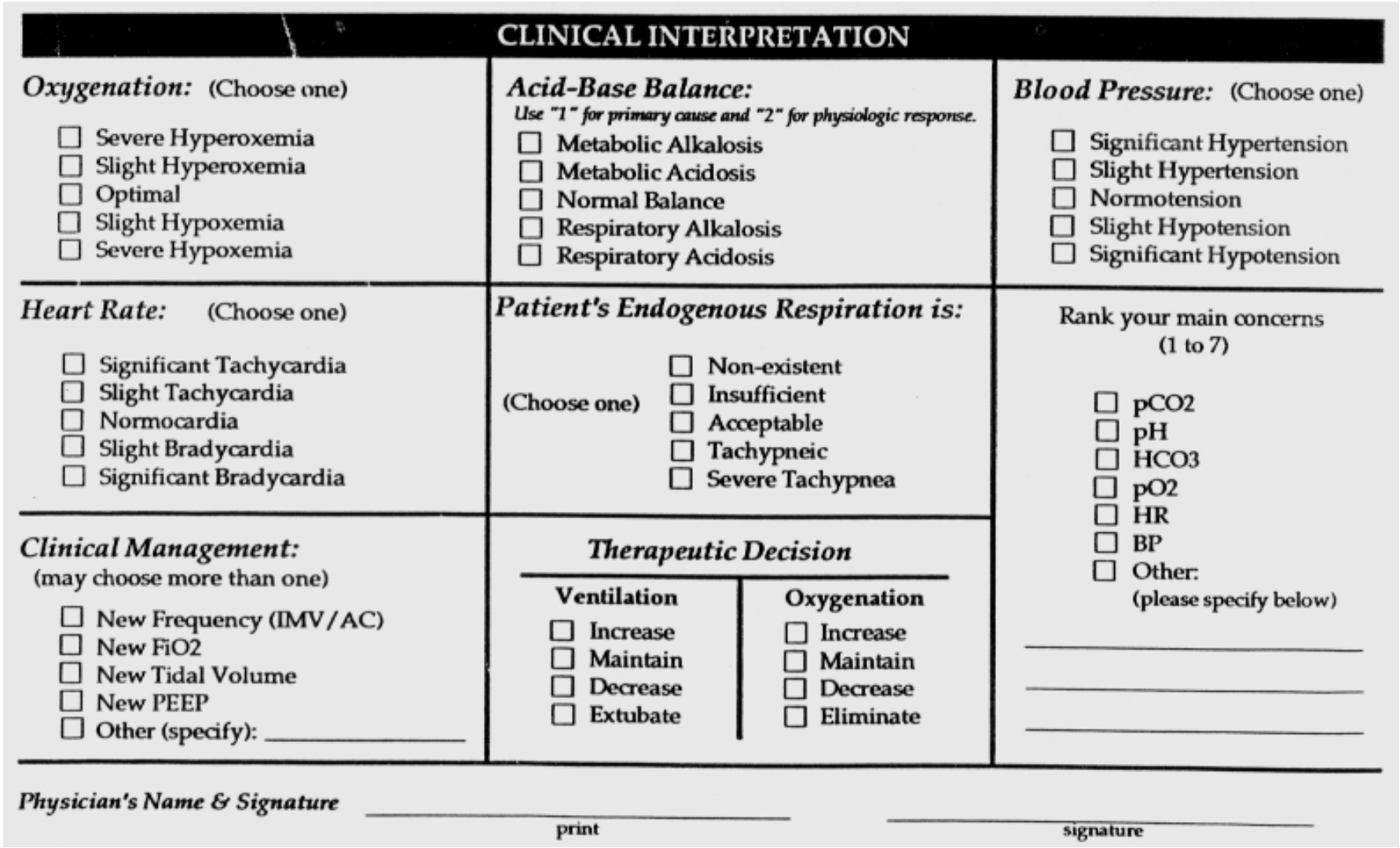
Sedation | Paralysis | Diuretics

Pain Relief | Alkalytics

Stress yes ☐ no ☐

Pain yes ☐ no ☐

Un ejemplo de validación



CLINICAL INTERPRETATION

Oxygenation: (Choose one)

- ☐ Severe Hyperoxemia
- ☐ Slight Hyperoxemia
- ☐ Optimal
- ☐ Slight Hypoxemia
- ☐ Severe Hypoxemia

Acid-Base Balance:

Use "1" for primary cause and "2" for physiologic response.

- ☐ Metabolic Alkalosis
- ☐ Metabolic Acidosis
- ☐ Normal Balance
- ☐ Respiratory Alkalosis
- ☐ Respiratory Acidosis

Blood Pressure: (Choose one)

- ☐ Significant Hypertension
- ☐ Slight Hypertension
- ☐ Normotension
- ☐ Slight Hypotension
- ☐ Significant Hypotension

Heart Rate: (Choose one)

- ☐ Significant Tachycardia
- ☐ Slight Tachycardia
- ☐ Normocardia
- ☐ Slight Bradycardia
- ☐ Significant Bradycardia

Patient's Endogenous Respiration is:

(Choose one)

- ☐ Non-existent
- ☐ Insufficient
- ☐ Acceptable
- ☐ Tachypneic
- ☐ Severe Tachypnea

Rank your main concerns (1 to 7)

- ☐ pCO2
- ☐ pH
- ☐ HCO3
- ☐ pO2
- ☐ HR
- ☐ BP
- ☐ Other: (please specify below)

_____

_____

_____

Clinical Management:

(may choose more than one)

- ☐ New Frequency (IMV/AC)
- ☐ New FiO2
- ☐ New Tidal Volume
- ☐ New PEEP
- ☐ Other (specify): _____

Therapeutic Decision

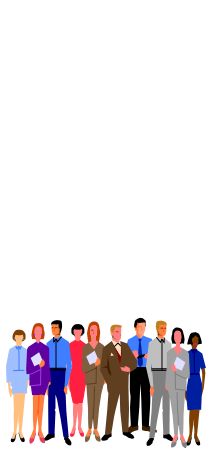
Ventilation	Oxygenation
☐ Increase	☐ Increase
☐ Maintain	☐ Maintain
☐ Decrease	☐ Decrease
☐ Extubate	☐ Eliminate

Physician's Name & Signature

_____ print

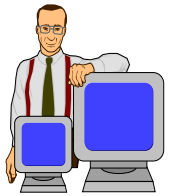
_____ signature

Un ejemplo de validación

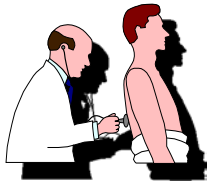


119 casos reales

PATRICIA



Experto Colaborador



Médico que atendió el caso (clínico)

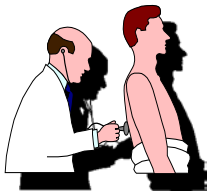
Porcentajes de acuerdo totales en todas las

Clínico vs. Experto Colaborador	79%
Clínico vs. Sistema Experto	78%
Sistema Experto vs. Experto Colaborador	92%

Un ejemplo de validación

PATRICIA

147 casos reales

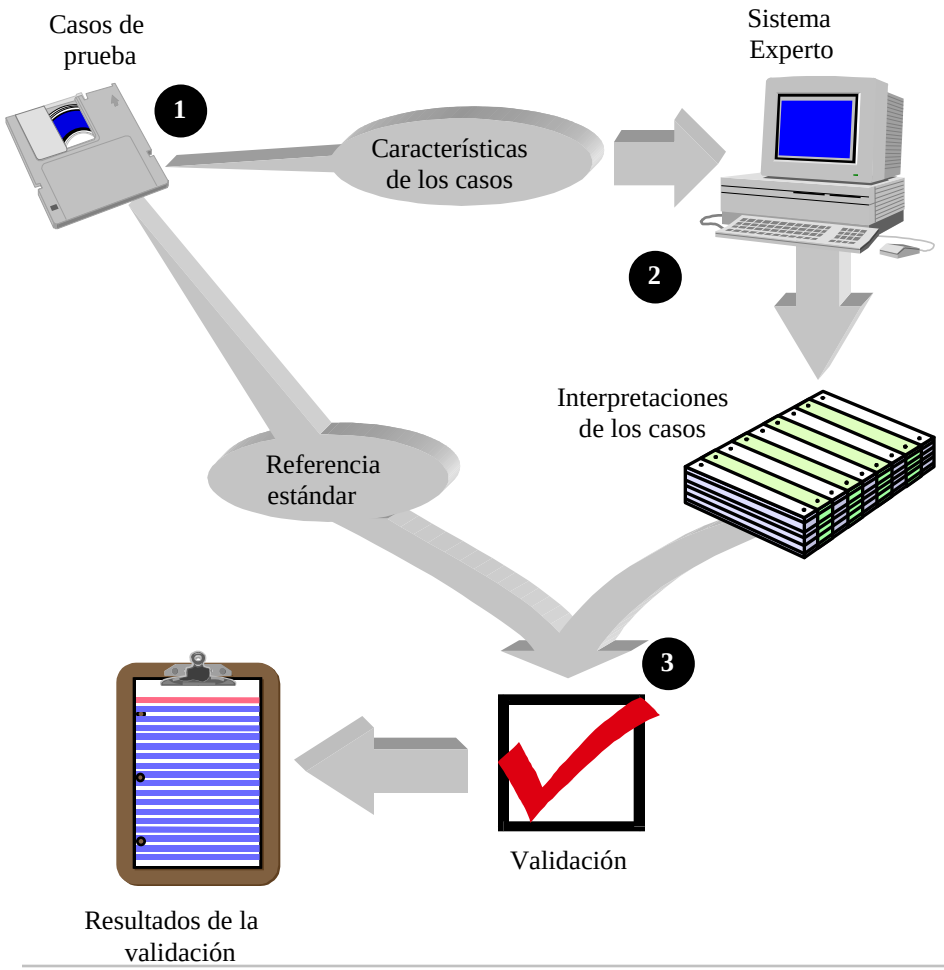


Equipo Médico

Porcentajes de acuerdo por categoría diagnóstica

Oxigenación	92%
Balance Ácido-Base	74%
Hemodinámica	87%
Terapia Ventilatoria	71%

Un ejemplo de validación



- Dominio UCI:
  - No es fácil establecer referencias estándar
  - Nunca podríamos asegurar que las interpretaciones y prescripciones de un experto sigan siempre los mismos principios
  - El estrés y el entorno contribuyen a desvirtuar comportamientos
  - Pueden aparecer soluciones equivalentes aunque no idénticas

Un ejemplo de validación

- Criterios con carácter general:

  - Si el dominio de aplicación es un dominio crítico, en el que no es posible reconsiderar decisiones una vez han sido tomadas, entonces los métodos prospectivos no son apropiados.

  - Evidentemente, si no existe una referencia estándar, o si tal referencia es muy difícil de obtener, la validación debe llevarse a cabo sin tales consideraciones.

  - Si la salida del sistema es un conjunto de interpretaciones que están lingüísticamente etiquetadas según una escala ordinal, entonces podemos considerar el uso de medidas cuantitativas, como índices de concordancia o medidas Kappa.

Un ejemplo de validación

- Esquema de la validación formal de PATRICIA
  - Contexto retrospectivo
  - Con medidas de pares y técnicas cuantitativas
  - Efectuar un análisis de grupo tratando de identificar referencias estándar, y posicionando a PATRICIA dentro del grupo de expertos colaboradores.

Un ejemplo de validación

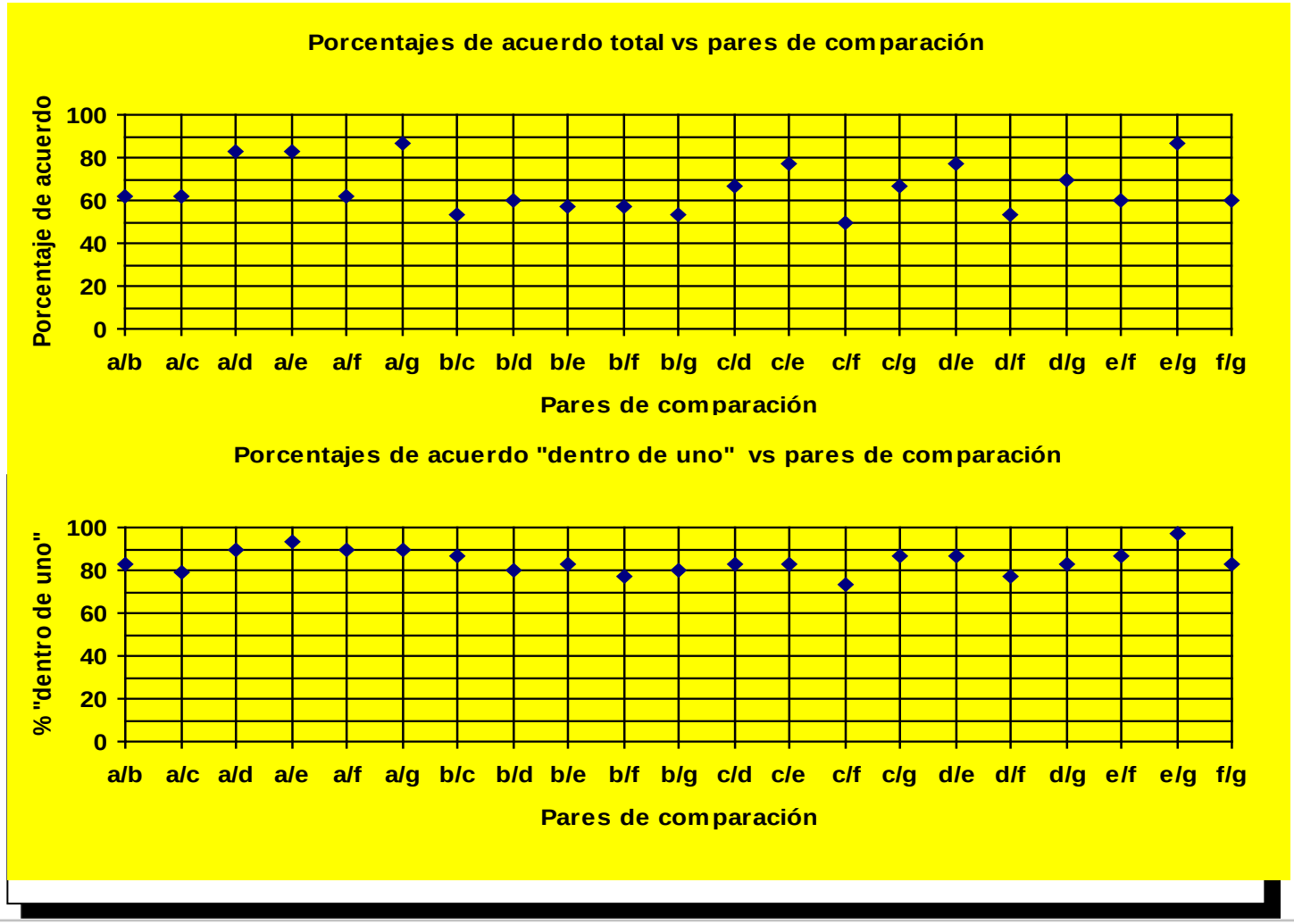
- Etapas:
  - Labores de interpretación
    - OXIGENACION
    - BALANCE ACIDO-BASE
    - RESPIRACION ENDOGENA
    - PRESION ARTERIAL
    - FRECUENCIA CARDIACA
  - Labores de sugerencias terapéuticas
    - MANEJO OXIGENATORIO
    - MANEJO VENTILATORIO

Un ejemplo de validación

- Medidas realizadas:
  - Indices de concordancia entre expertos (incluido el sistema)
  - Indices de concordancia en uno
  - Indices kappa
  - Indices kappa ponderada
  - Medidas de Williams
  - Análisis Clúster

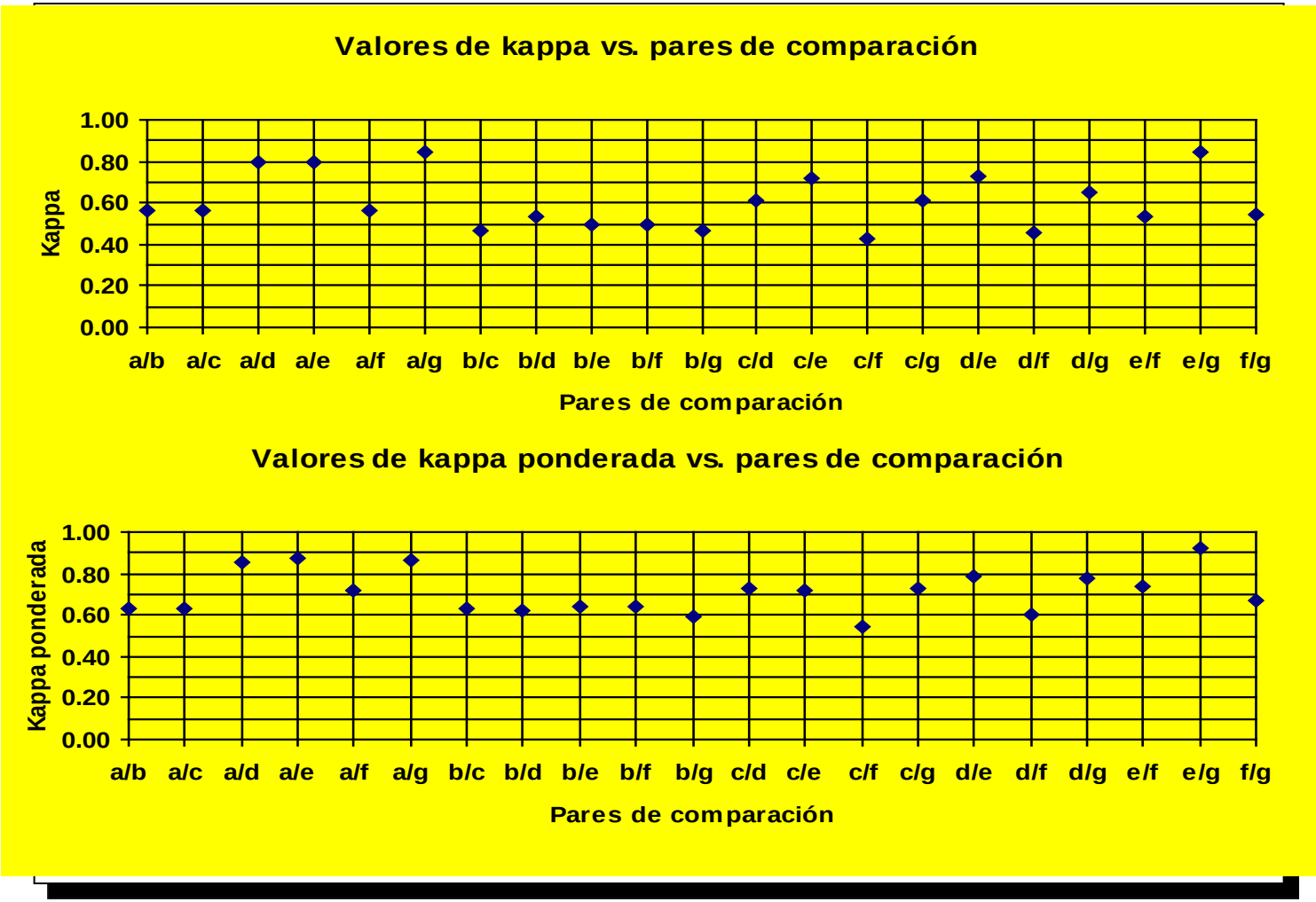
Un ejemplo de validación

Balance Ácido-Base



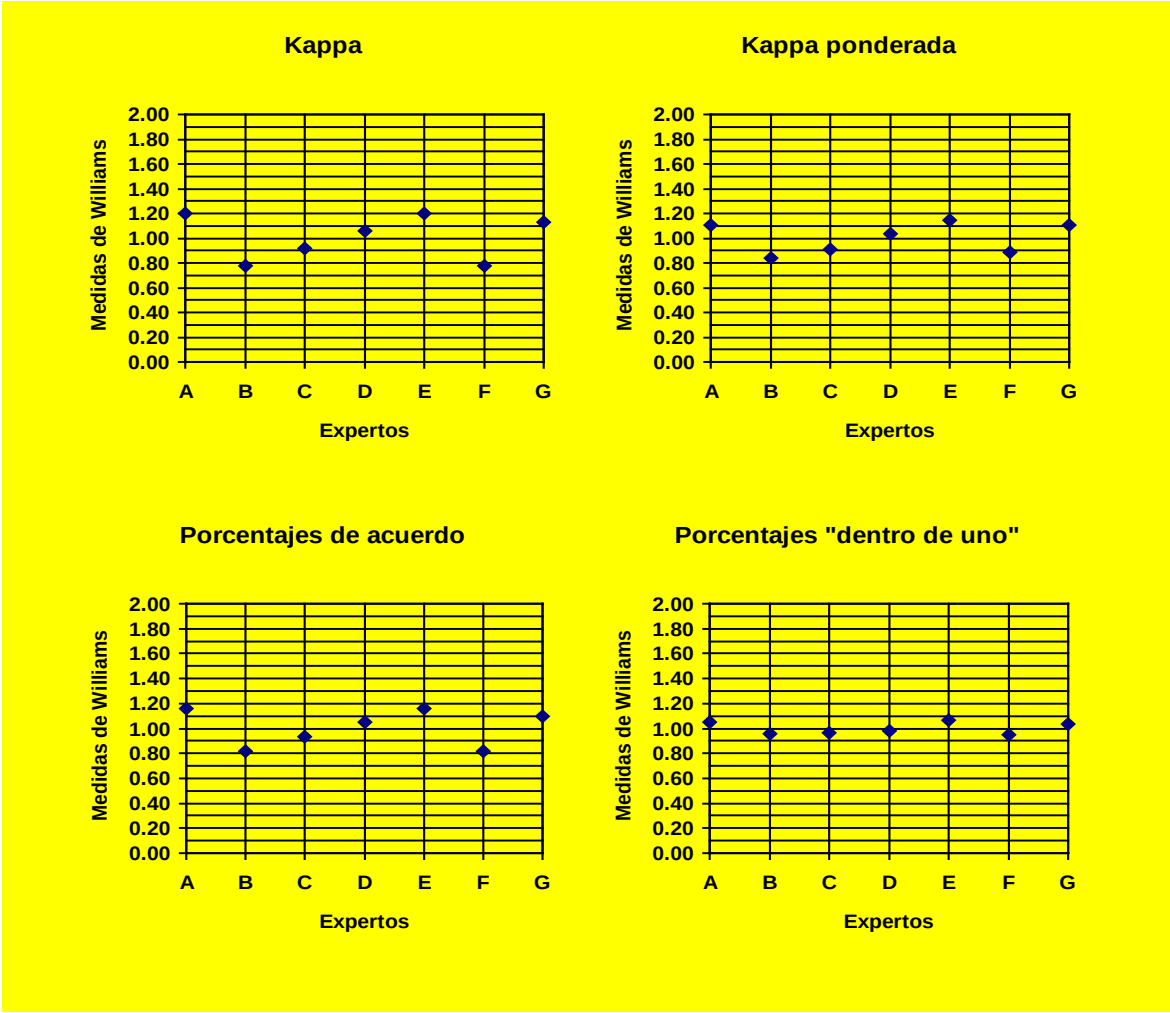
Un ejemplo de validación

Balance Ácido-Base



Un ejemplo de validación

Balance Ácido-Base



USABILIDAD DE SISTEMAS

- Métodos heurísticos
  - Técnicas heurísticas, desarrolladas por expertos, que analizan los interfaces de los módulos, evalúan su arquitectura y determinan sus puntos fuertes y débiles desde la perspectiva del usuario

- Métodos subjetivos
  - Obtienen información de los usuarios sobre prototipos operativos del prototipo en desarrollo (observación directa, cuestionarios, entrevistas, grupos de control,…)

- Métodos empíricos
  - Obtención de datos objetivos acerca de cómo los usuarios utilizan el sistema

MÉTODOS HEURÍSTICOS

- Análisis del sistema y detección de problemas de amigabilidad y calidad
  - Cuestionarios ergonómicos
  - Inspección de interfaces
  - Evaluación de la navegación
  - Análisis formales

MÉTODOS SUBJETIVOS

- Conocimiento de la opinión de los usuarios sobre la propia usabilidad del sistema
  - Pensar en alto
  - Observación
  - Cuestionarios
  - Entrevistas
  - Grupos de control
  - Retroalimentación con el usuario

EJEMPLOS DE CUESTIONARIOS CERRADOS

Escala simple ¿Puede realizarse ...?

	SI	NO	NS/NC
	☐	☐	☐

Escala multipunto ¿Está de acuerdo con ...?

Completamente en desacuerdo ☐ ☐ ☐ ☐ ☐ ☐ ☐ Completamente de acuerdo

Escala de Lickert ¿Está de acuerdo con ...?

Completamente en desacuerdo	En desacuerdo	Ligeramente en desacuerdo	Neutral	Ligeramente de acuerdo	De acuerdo	Completamente de acuerdo

Escala diferencial semántica Clasifica el módulo ... de acuerdo a los siguientes parámetros

	Extremadamente	Bastante	Ligeramente	Neutral	Ligeramente	Bastante	Extremadamente	
Fácil								Difícil
Claro								Confuso

Escala de orden Ordena los siguientes comandos según su utilidad

PEGAR ☐ DUPLICAR ☐ AGRUPAR ☐ BORRAR ☐

MÉTODOS EMPÍRICOS

- Se trata de sacar conclusiones basadas en datos objetivos obtenidos sobre cómo los usuarios utilizan el sistema
  - Exactitud
    - Número de errores provocados durante un determinado lapso de tiempo
  - Velocidad
    - Celeridad en la interacción con el sistema
  - Exactitud y velocidad son magnitudes inversamente proporcionales

MEDIDAS OBJETIVAS DE USABILIDAD

- Número de tareas diversas que pueden realizarse en un determinado periodo de tiempo
- Proporción entre interacciones correctas y errores
- Número de errores cometidos por el usuario
- Tiempo consumido en la realización de una tarea específica
- Tiempo consumido en la recuperación de errores
- Número de características del sistema que son utilizadas por los usuarios

RESUMEN

- Verificación, validación y análisis de usabilidad son fundamentales para desarrollar software de calidad
- Estas fases deben formar parte del ciclo de desarrollo del sistema
- Las metodologías de desarrollo y diseño deben incluir explícita y específicamente la ubicación idónea de las tareas de verificación, validación y usabilidad
- La realización de estas tareas requiere el dominio de técnicas específicas
- La evaluación de sistemas debe ser contemplada como un proceso global de análisis del rendimiento del sistema en cuestión

UNIVERSIDADE DA CORUÑA

fic
facultade de
informática
da coruña

Evaluación de Sistemas Inteligentes