Revista de la Asociación Mexicana de
MEDICINA CRITICA
Y TERAPIA INTENSIVA
Vol. XX, Núm. 1 / Ene.-Mar. 2006
pp 16-24

Concordancia en el juicio clínico al calificar la gravedad de los pacientes en los servicios de admisión

M en C Leticia Rodríguez-Pimentel,* Dr. Juan Garduño-Espinoza,† Dr. Noé Estrada Padilla,* Dr. Rodolfo Silva Romo*

RESUMEN

Objetivo: Determinar la eficacia de los médicos para evaluar el grado de severidad de la enfermedad en los pacientes admitidos en un Servicio de Emergencias.

Diseño: Estudio transversal y comparativo.

Lugar: Servicio de Admisión Continua de un hospital de tercer nivel de la ciudad de México.

Método: Dos médicos internistas calificaron de manera independiente y ciega la gravedad de la enfermedad de 20 pacientes que ingresaron a un Servicio de Admisión Continua. Se utilizaron seis escalas: tres escalas subjetivas (Escala de evaluación clínica que calificó a los pacientes como moderado, gravemente y muy gravemente enfermos, clasificación del estado físico de la Sociedad Americana de Anestesiología [ASA]; y una escala visual análoga) y tres escalas objetivas (Escala fisiológica aguda [SAPS II]; Sistema de intervenciones terapéuticas [TISS] e Índice de comorbilidad de Charlson).

Resultados: La escala de clasificación clínica tuvo una concordancia del coeficiente Kappa = 0.22 y la clasificación de la ASA de 0.25 (p < 0.02), la concordancia de la escala visual análoga medida a través de Ri fue de 0.59, la de TISS 0.60 (IC 95% [0.01 a 0.84]), el índice de comorbilidad de Charlson de 0.86 (IC 95% [0.71 a 0.95]) y SAPS II 0.96 (IC 95% [0.84 a 0.98]).

Conclusión: El uso de las escalas objetivas de gravedad de la enfermedad es más apropiado para evaluar a los pacientes que ingresan a los servicios de emergencia.

Palabras clave: Gravedad, acuerdo, escala análoga visual, ASA, SAPS II, TISS, comorbilidad.

SUMMARY

Objective: *To determine the efficacy of physicians to evaluate the degree of illness severity of patients admitted to an emergency room.*

Design: *Transversal and comparative study.*

Setting: *Emergency Room of a tertiary care hospital, Mexico City.*

Method: *Two internist physicians qualified independently and blindly the severity of illness of twenty patients admitted to an emergency room. Six scales were used: three subjective scales (Clinical Evaluation Scale which qualified the patients as mild, seriously and very seriously ill; Physical Status Classification of the Anaesthesiologist American Society [ASA]; and Visual Analogue Scale; and three objective scales (Simplified Acute Physiology Score [SAPS II]; Therapeutic Intervention Scoring System [TISS] and Charlson Comorbidity Index).*

Results: *Clinical Classification Scale had an agreement of Kappa coefficient = 0.22 and ASA classification = 0.25 (p < 0.02), the agreement of Visual Analogue Scale determined through Ri was 0.59, TISS 0.60 (IC 95% [0.01 to 0.84]), Charlson Comorbidity 0.86 (IC 95% [0.71 to 0.95]) and SAPS II Score 0.96 (IC 95% [0.84 to 0.98]).*

Conclusion: *The use of objective scales of illness severity is more appropriate to evaluate the patients admitted to an emergency room.*

Key words: *Severity, agreement, Visual Analogue Scale, ASA, SAPS II, TISS, comorbidity.*

Entre los factores que contribuyen a la decisión de hospitalizar o no a un enfermo, se encuentra el atribuirle un nivel de gravedad. En la práctica habitual, esto se realiza de manera subjetiva, pudiendo existir discrepancias en el examen clínico simultáneo del mismo paciente por otros médicos, además de que esta evaluación puede modificarse por otra evidencia de mayor peso, referida como “dura”, tal

* Maestro en Ciencias. Servicio de Admisión Continua, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social.

† Doctor en Ciencias. Jefatura de la División de Calidad y Desempeño, Coordinación de Atención Médica, Instituto Mexicano del Seguro Social.

como lo son los resultados de estudios de laboratorio o gabinete.

Diversos autores han indicado que la gravedad de la enfermedad en el momento de la admisión del paciente a la unidad de atención médica es el factor pronóstico de sobrevida más importante, seguida de la comorbilidad y la capacidad funcional.[1] Ya que en los servicios de admisión y urgencias se requiere tomar decisiones rápidas, con el fin de mejorar las decisiones clínicas, surge la necesidad de contar con un índice que disminuya en lo posible la subjetividad y por ende contribuya a una buena concordancia ante la evaluación por diversos médicos (distintos turnos, diversos interconsultantes) y a reducir la posibilidad de realizar ingresos innecesarios o por lo contrario, altas prematuras con el consecuente riesgo para el paciente.

VALIDEZ Y REPRODUCIBILIDAD

Las pruebas de diagnóstico y pronóstico requieren contar de validez y reproducibilidad. En el caso de las mediciones de laboratorio se debe cumplir con dos propiedades, la primera es la exactitud (accuracy) y se refiere a la coincidencia de la medición con un patrón de referencia, también se le ha llamado validez, certeza, conformidad, correctividad.

La segunda propiedad es la reproducibilidad (reproducibility) que es la obtención del mismo resultado cuando una medición se realiza en forma repetida por la misma persona o por otra diferente, y los sinónimos utilizados son: consistencia, confiabilidad, precisión, o repetibilidad.

En ciencias sociales el término fiabilidad (reliability) se utiliza como la capacidad de un instrumento para distinguir entre individuos, se calcula como el consciente de la varianza debida a los sujetos de estudio dividido entre la varianza total. Lo que cuenta para aceptar a un instrumento como fiable no es la magnitud del error sino la relación de éste y el rango en que se mueven habitualmente las mediciones; algunos autores sustituyen el término fiabilidad por el de consistencia (concistency), concordancia y reproducibilidad.[2-4]

En la práctica clínica las condiciones en las que se realizan las mediciones no suelen ser perfectamente controlables, se reconocen tres fuentes de variabilidad: la real debida a los sujetos en estudio, la debida al método de medición y la debida a la posible influencia subjetiva del observador en el proceso de medición. La calidad de las mediciones no sólo condiciona la calidad de las investigaciones, sino también la calidad de las decisiones clínicas que se apoyan en dichas mediciones. En este contexto el objetivo es identificar las causas de las discrepancias para poder corregirlas.

NIVEL DE GRAVEDAD

Para evaluar el nivel de gravedad de un enfermo se han utilizado diversas escalas e índices, entre ellas se encuentran:

Variables categóricas o subjetivas. Hasta fecha reciente, la cuantificación más común de gravedad usada para estimar el pronóstico de los pacientes ha sido una escala descriptiva de tres categorías basada en la opinión del médico: delicado, grave, muy grave; sin embargo, el error de clasificación es común en tales variables, frecuentemente nulificando su utilidad.

Variables objetivas. Debido a que las variables objetivas son confiables y estandarizadas su mayor contribución es la reducción de la variabilidad de los juicios y discriminación en los pronósticos.[5,6]

Entre los instrumentos ideados para medir el nivel de gravedad de los pacientes se encuentra la clasificación del estado físico de la Sociedad Americana de Anestesiólogos (ASA), se diseñó originalmente en 1940 para estandarizar las categorías del estado físico de los pacientes para estudios estadísticos y registros hospitalarios, para que fuera posible una interpretación uniforme), los pacientes se clasifican en cinco clases: I. Ninguna enfermedad salvo la patología quirúrgica, II. alteración general moderada, III. alteración general intensa, IV. alteración general intensa con amenaza patente para el paciente y V. paciente moribundo.[7] Owens en 1978 realizó un estudio para evaluar la consistencia interna de esta clasificación entre anestesiólogos certificados y concluyó que la clasificación del estado físico de ASA es útil pero le faltaba precisión.[8]

A nivel de las terapias intensivas existen diversos índices de gravedad, los cuales, por medio de la medición de variables fisiológicas, edad, comorbilidad, causa del ingreso, procedencia etc., establecen pronósticos para auxiliar al juicio clínico y determinar cuáles pacientes pueden beneficiarse de tales servicios. Su uso se encuentra limitado a las terapias intensivas y requiere determinar diversas variables, lo que supondría poco práctico su empleo en los servicios de urgencias, entre ellos se encuentran el modelo de probabilidad de mortalidad (Mortality Probability Models (MPM)),[9] la evaluación de salud crónica, fisiología aguda, edad (Acute Phy-

siology, Age, Chronic Health Evaluation [APACHE])[10,11] y el puntaje fisiológico agudo simplificado (Simplified Acute Physiology Score [SAPS]);[12] este último incluye 12 variables fisiológicas, edad, tipo de admisión y tres enfermedades comórbidas. El puntaje obtenido es convertido por medio de un modelo de regresión múltiple a una probabilidad de mortalidad hospitalaria en porcentaje (por ejemplo, 18 puntos indican una probabilidad de mortalidad de 3%, 59 puntos una mortalidad de 66%).

El sistema de puntaje de intervenciones terapéuticas (Therapeutic Intervention Scoring System, TISS) es un índice introducido en 1974, califica el tipo y cantidad de procedimientos realizados al enfermo, conforme al puntaje obtenido se establece una relación entre la gravedad del paciente y el número idóneo de enfermeras requerido para su atención y puede ser usado con fines administrativos o propósitos clínicos, con menos de 10 puntos corresponde a clase I, de 10 a 19 puntos clase II, de 20 a 39 puntos clase III y más de 40 puntos clase IV.[13]

Ya que las condiciones comórbidas pueden alterar el riesgo de muerte, algunos índices de gravedad lo incluyen dentro de su desarrollo, o bien puede ser calculado individualmente, como el índice de enfermedades comórbidas,[14] o el ideado por Charlson en 1986,[15] que identifica factores de confusión (tipo y número de enfermedades comórbidas) relacionados con la supervivencia de pacientes. Conforme a este puntaje las tasas de mortalidad a un año son: 0 puntos 12%, 1-2 puntos 26%, 3-4 puntos 52% y más de 5 puntos 85%.

MÉTODOS ESTADÍSTICOS PARA EVALUAR LA CONCORDANCIA

De manera paradójica, hay controversia con respecto a cuáles son los mejores métodos estadísticos para medir el acuerdo, se sugiere que sean seleccionados conforme el objetivo a alcanzar, por ejemplo si se requiere probar la validez de criterio (comparar una medida imperfecta con el verdadero estatus o "estándar de oro") o la concordancia (evaluar el acuerdo entre dos o más evaluadores o entre medidas imperfectas); según la escala de medición de la variable estudiada (nominal, ordinal, de intervalo o de razón) y según el tipo de variable (discreta o continua).

En 1975 Fleiss comentó que no existía un índice de acuerdo informativo *per se*, y dadas las múltiples mediciones de acuerdo propuestas por la literatura sugirió el uso de sólo un par de ellos[16] que evalúan este aspecto de manera general.

Sin embargo, actualmente se sugiere además tener en cuenta los diferentes componentes del desacuerdo, esto es, ¿el desacuerdo es por una diferencia en la definición de la característica medida (asociación o correlación entre evaluadores), o por una diferencia en los límites de los intervalos que los evaluadores utilizaron (sesgo del evaluador)?[17]

Al separar los diferentes componentes del desacuerdo, es posible definir qué pasos se requieren para mejorar el acuerdo.

ESCALAS BINARIAS Y NOMINALES

El estadístico Kappa (κ) ha sido uno de los métodos más utilizados para escalas binarias y nominales, sin embargo sólo verifica si los evaluadores acuerdan más que lo que el azar puede predecir, en realidad no es una medida del nivel de acuerdo, no está claro cómo el azar afecta las decisiones de los jueces o evaluadores y cómo se puede corregir.[17-21]

Para identificar los componentes del desacuerdo en este tipo de escalas una alternativa es utilizar el coeficiente de correlación policórica (PLCORR), que es una medida de asociación bivariada de variables categóricas ordenadas que resultan de dividir en varias categorías (policotomizar) una o dos variables continuas subyacentes, y en las que es difícil asignar valores numéricos que preserven sus relaciones relativas. Este método es invariante al ancho entre categorías y además se pueden comparar escalas con distintas categorías.[22]

Para determinar la presencia de sesgo entre los evaluadores (sobre o subclasificación), se puede utilizar la prueba de homogeneidad marginal (HM), la cual se refiere a la igualdad (diferencia no significativa) entre la proporción marginal de los renglones y su correspondiente proporción en las columnas de las tablas de distribución, significando que las frecuencias con las que los dos evaluadores usan las diversas categorías son las mismas.[23]

ESCALAS DE INTERVALO

Para analizar el acuerdo en general con datos en escala de intervalo frecuentemente se utiliza el coeficiente de correlación intraclase (Ri), ya que con este método existen diversas opciones, para seleccionar la adecuada se debe decidir en primer término si los datos deben ser tratados con un análisis de varianza (ANOVA) de una vía o de dos vías, también debe

elegirse entre modelos con efectos fijos, con efectos aleatorios o con efectos mixtos.[17,24-27]

Por otra parte, los procedimientos gráficos proporcionan importante información, entre ellos se encuentra el método de Bland-Altman analiza la concordancia entre 2 métodos que utilizan las mismas unidades de medida, grafica la diferencia entre las dos observaciones contra su media.[28]

OBJETIVO

El objetivo principal del estudio fue determinar el acuerdo entre evaluadores al calificar el nivel de gravedad de los pacientes en la práctica médica cotidiana en un Servicio de Admisión Continua.

El objetivo secundario fue analizar los distintos componentes del desacuerdo entre ellos.

MATERIAL (PACIENTES) Y MÉTODOS

Diseño del estudio. Se realizó un estudio transversal comparativo en el Servicio de Admisión Continua del Hospital de Especialidades, Centro Médico Nacional SXXI IMSS, que es un hospital institucional de 3er nivel de atención médica del adulto y cuenta con diversas especialidades médicas y quirúrgicas.

Evaluadores: Participaron dos médicos con especialidad en medicina interna, con experiencia clínica de 2 y 9 años en el servicio, graduados en el mismo hospital.

El tamaño de muestra requerido se calculó para dos evaluadores, utilizando las tablas propuestas por Walter SD y col.[29] para estudios de concordancia cuando se usa la correlación intraclase, teniendo un nivel de significancia de $\alpha = 0.05$, poder de 0.80 ($\beta = 0.20$) con un coeficiente de correlación intraclase (Ri) mínimo aceptable de 0.70 (ρo) y esperando fuera de 0.90 ($\rho 1$).

Pacientes: Se seleccionaron al azar 20 pacientes del área de observación del turno matutino, 10 mujeres y 10 hombres, con edad media de 56 años (límites 28-86). Siete de estos enfermos tenían urgencias quirúrgicas y trece urgencias médicas *(cuadro I)*.

Todos los pacientes contaban con exámenes de laboratorio, y se evaluaba su ingreso a hospitalización. De manera independiente y a ciegas, fueron calificados según su nivel de gravedad por los dos médicos participantes, con un lapso de 30 minutos promedio entre la evaluación efectuada por el primer médico y el segundo. Se utilizó una hoja de captación de datos que contenía las seis escalas a probar:

1. Escalas subjetivas: 1. Escala de evaluación clínica, graduada como delicado, grave y muy grave. 2. Clasificación del estado físico de la Sociedad Americana de Anestesiólogos (ASA). 3. Escala análoga visual de diferencial semántico, de nueve grados.
2. Índices objetivos. 1. Puntaje simplificado de fisiología aguda (SAPS II). 2. Sistema de puntaje de intervenciones terapéuticas (TISS). 3. Índice de comorbilidad de Charlson.

Iniciaba el evaluador A, las escalas se aplicaban en orden, empezando por las escalas subjetivas (ASA y EAV y EEC) seguidas de los índices objetivos (Comorbilidad, SAPSII y TISS). El tiempo promedio de aplicación para cada una de las escalas subjetivas fue de 1 minuto o menos, con el índice de comorbilidad fue de alrededor de 3 minutos, con el SAPS II entre 5 y 10 minutos y con el TISS de más de 15 minutos.

ANÁLISIS ESTADÍSTICO

Se describieron las variables conforme a su distribución, para ello se utilizaron medidas de tendencia central y dispersión.

Cuadro I. Descripción de pacientes.

Urgencias médicas	Edad	Sexo
Leucemia mieloide aguda M4	51	F
Leucemia mieloide aguda M3	52	F
Leucemia mieloide aguda M5	65	F
Leucemia granulocítica	28	F
Neumonía y deshidratación	29	M
Neumonía y SIDA	44	F
Tuberculosis meníngea	67	F
Angor hemodinámico	64	M
Hemorragia de tubo digestivo bajo	57	F
Pancreatitis aguda edematosa	36	F
Cáncer hepático	64	M
Enfermedad ampollosa, erisipela	75	F
Insuficiencia suprarrenal	65	M
Urgencias quirúrgicas		
Hidrocolecisto	82	M
Colangitis, cáncer	86	M
Úlcera péptica perforada	85	F
Fascitis necrotizante, sepsis	40	F
Hematoma intraabdominal	35	M
Fístula arteriovenosa	32	M
Urinoma	52	M
x	56 (28-86)	

Para determinar si el acuerdo observado entre los evaluadores era mayor que el acuerdo esperado por el azar, en el caso de las escalas ordinales con dos categorías se utilizó kappa (κ) y para más de dos categorías kappa ponderado (κp) y para identificar los distintos componentes del desacuerdo (correlación y sesgo) se utilizó la correlación policórica (PLCORR) y la prueba de homogeneidad marginal (HM). Para evaluar el acuerdo en general en los índices en escala de intervalo y continua se eligió el coeficiente de correlación intraclase (Ri) utilizando un análisis de varianza de dos vías con efectos mixtos, con los evaluadores como efectos fijos y a los pacientes como efectos aleatorios. Para evaluar los distintos componentes del desacuerdo se utilizó la correlación de Pearson (r) y para detectar sesgo la prueba t pareada (t).[30]

Se utilizaron los paquetes estadísticos SPSS versión 8 y SAS.

RESULTADOS

Escalas subjetivas: El *cuadro II* indica los valores crudos, medidas de resumen y concordancia cuando se utilizaron escalas subjetivas.

1. Escala de evaluación clínica (delicado, grave, muy grave). Ambos evaluadores calificaron a 6 pacientes (30%) como delicados, a trece (65%) como graves y a uno (5%) como muy grave. Coincidieron exactamente en 12 pacientes.

Cuadro II. Valores crudos, medidas de resumen y concordancia cuando se utilizaron escalas subjetivas.

Escala	EEC		ASA		EAV	
Evaluador	A	B	A	B	A	B
No. paciente:						
1	1	2	3	3	4	6
2	2	1	4	3	6	5
3	1	2	3	3	4	6
4	2	2	3	4	6	6
5	1	2	2	2	3	6
6	1	1	3	3	6	5
7	2	2	4	4	7	8
8	2	2	3	3	7	6
9	2	2	3	2	6	5
10	3	2	3	3	8	7
11	2	3	3	3	7	8
12	2	2	3	3	6	5
13	2	2	4	3	7	6
14	2	2	3	3	6	7
15	2	2	4	3	7	6
16	2	1	3	2	6	4
17	2	2	4	3	7	6
18	1	1	1	2	2	3
19	1	1	3	2	6	3
20	2	1	4	2	7	5
Mediana	2	2	3	3	6	6
Acuerdo	Kp = 0.22 p = n/s		Kp = 0.25 p < 0.02		Ri =0.59 p = n/s	
Asociación o correlación	PLCORR = 0.44 p = 0.055		PLCORR = 0.52 p < 0.01		r = 0.42 p = n/s	
Sesgo	HM = 14 p = n/s		HM = 14 p = n/s		t = 0.72 p = n/s	

EEC. Escala de evaluación clínica
ASA. Clasificación de la Sociedad Americana de Anestesiólogos
EEV. Escala análoga visual
κp. Kappa ponderada
Ri. Coeficiente de correlación intraclase
PLCORR. Correlación policórica
r. r Pearson
HM. Homogeneidad marginal
t. Prueba t pareada

Acuerdo por el azar: Al efectuar kappa ponderado el acuerdo no fue mayor al esperado por el azar $\kappa p = 0.22$ (p = n/s). También se hicieron cortes y se consideraron dos grupos: delicado (grupo 1) y grave-muy grave (grupo 2), se obtuvo una $\kappa = 0.29$ (p = n/s); al realizarlo como delicado-grave (grupo 1) y muy grave (grupo 2) fue de $\kappa = 0$.

Componentes del desacuerdo: El coeficiente de correlación policórica demostró una correlación de PLCORR = 0.44 con significancia limítrofe (p = 0.055) y la prueba de homogeneidad marginal no identificó sesgo.

2. Clasificación de ASA. El evaluador A clasificó en clase 1 a un paciente (5%), en clase 2 a un paciente (5%), en clase 3 a doce pacientes (60%), en clase 4 a seis pacientes (30%), y ninguno en clase 5. El evaluador B calificó en clase 2 a seis pacientes (30%), en clase 3 a doce pacientes (60%), en clase 4 a dos pacientes (10%), y ninguno en clase 5. Coincidieron exactamente en 10 pacientes.

Acuerdo por el azar: El resultado de kappa ponderado fue de $\kappa p = 0.25$ ($p < 0.02$) *(cuadro II)*. Al realizar su análisis considerando dos grupos, clase 1 a 2 y 3 a 5 se obtuvo una $\kappa = 0.41$ ($p < 0.02$), y en clase 1 a 3 y 4 a 5 el resultado fue $\kappa = 0.12$ (p = n/s).

Componentes del desacuerdo: La correlación policórica fue de PLCORR = 0.52 ($p < 0.01$) y la prueba de homogeneidad marginal no identificó sesgo.

El *cuadro III* señala la concordancia entre estas dos escalas subjetivas ordinales.

3. Escala análoga visual de diferencial semántico. El evaluador A colocó en la clasificación 2 a un paciente (5%), en la 3 a uno (5%), en la 4 a dos (10%), en la 6 a ocho (40%), en la 7 a siete (35%), y en la 8 a uno (5%). El evaluador B colocó en la clasificación 3 a dos pacientes (10%), en la 4 a uno (5%), en la 5 a cinco (25%), en la 6 a ocho (40%), en la 7 a dos (10%) y en la 8 a dos (10%). Sólo concordaron exactamente en uno.

Cuadro III. Concordancia entre escalas subjetivas ordinales con diferentes categorías correlación policórica (PLCORR).

	EECA	EECB	ASAA	ASAB
EECA		0.43 p = 0.055	0.62 p < 0.01	0.43 p = n/s
EECB			0.04 p = n/s	0.62 p < 0.01
ASAA				0.52 p < 0.01
ASAB				

EECA. Escala de evaluación clínica, observador A
EECB. Escala de evaluación clínica, observador B
ASAA. Escala de la Asociación Americana de Anestesiólogos, observador A
ASAB. Escala de la Asociación Americana de Anestesiólogos, observador B

Acuerdo en general: Al realizar el coeficiente de correlación intraclase se obtuvo un Ri = 0.59 (IC 95% -0.03, 0.84).

Componentes del desacuerdo: La correlación de Pearson fue de r 0.42 (p n/s) y la prueba t pareada no demostró sesgo.

Índices objetivos: El *cuadro IV* indica los valores crudos, medidas de resumen, acuerdo en general, correlación y sesgo cuando se utilizaron índices objetivos.

1. SAPS II. El evaluador A calificó con menos de 20 puntos a 40% de los pacientes, entre 21 y 39 puntos a 50% y de 40 a 59 puntos a 10%. El evaluador B calificó con menos de 20 puntos a 50% de los pacientes, entre 21 y 39 puntos a 40% y de 41 a 52 puntos a 10%. Hubo una coincidencia exacta en la calificación de 10 pacientes *(figura 1)*.

Acuerdo en general: El coeficiente de correlación intraclase fue de Ri = 0.96 (IC 95% 0.84 a 0.98).

Componentes del desacuerdo: La correlación de Pearson fue de r = 0.94 ($p < 0.01$) y la prueba t pareada demostró la presencia de sesgo significativo ($p < 0.01$).

2. TISS. El evaluador A calificó con 5 puntos o menos a 50% de los pacientes, entre 6 y 10 puntos a 35% y entre 11 y 15 puntos a 15%. El evaluador B calificó con 5 puntos o menos a 30% de los pacientes, entre 6 y 10 puntos a 65% y con 12 puntos a uno (5%). La coincidencia exacta entre evaluadores sólo ocurrió en 2 pacientes.

Acuerdo en general: El coeficiente de correlación intraclase fue de Ri = 0.60 (IC 95% de - 0.01 a 0.84).

Componentes del desacuerdo: La correlación de Pearson fue de r = 0.47 ($p < 0.05$) y la prueba t pareada descartó sesgo.

3. Comorbilidad. Se obtuvo una media de tres enfermedades comórbidas por paciente. El evaluador A calificó de 0 a 3 puntos a 60% de los pacientes, de 4 a 6 puntos a 30% y de 7 a 8 a 10% El evaluador B calificó de 0 a 3 puntos a 70% de los pacientes, de 4 a 6 a 15% y de 7 a 9 a 15%. Hubo una coincidencia exacta entre los evaluadores en 9 pacientes.

Acuerdo en general: Se obtuvo un coeficiente de correlación intraclase de Ri = 0.86 (IC 95% 0.71 y 0.95) *(cuadro IV)*.

Componentes del desacuerdo: La correlación de Pearson fue de r = 0.79 ($p < 0.01$) y la prueba t pareada descartó sesgo.

Cuadro IV. Valores crudos, medidas de resumen, acuerdo en general, correlación y sesgo cuando se utilizaron índices objetivos.

Índice	SAPS II		TISS		Comorbilidad	
Evaluador	A	B	A	B	A	B
Paciente						
1	24	24	6	3	4	5
2	48	49	6	7	6	7
3	29	29	5	5	3	3
4	23	23	10	6	7	6
5	22	22	3	2	4	1
6	29	26	4	5	1	3
7	25	18	5	4	3	2
8	28	24	6	6	6	5
9	18	16	15	7	2	2
10	59	52	14	7	5	2
11	34	19	5	8	1	1
12	19	19	5	6	0	0
13	18	18	8	7	3	3
14	35	30	5	7	8	8
15	19	19	11	12	3	2
16	18	18	5	6	2	2
17	19	19	4	7	2	2
18	8	8	4	5	0	0
19	15	13	6	7	1	3
20	38	30	7	8	5	9
Mediana	23.5	20.5	5.5	6.5	3	2.5
(límites)	(8-59)	(8-52)	(3-15)	(2-12)	(0-8)	(0-9)
Acuerdo general Ri	0.96		0.60		0.86	
(I.C. 95%)	(0.84 a 0.98)		(-0.01 a 0.84)		(0.71 a 0.95)	
Correlación r	0.94		0.47		0.79	
	p < 0.01		p < 0.05		p < 0.01	
Sesgo t	t = 2.87		t = .68		t = 0.0	
	p < 0.01		p = n/s		p = n/s	

SAPS II. Puntaje simplificado de fisiología aguda (Simplified Acute Physiology Score).
TISS. Sistema de puntaje de intervenciones terapéuticas (Therapeutic Intervention Scoring System).
Co. Índice de comorbilidad de Charlson.
Ri. Coeficiente de correlación intraclase
r. r Pearson
t. t pareada

DISCUSIÓN Y CONCLUSIONES

En el presente estudio, ante la evaluación de diferentes escalas se observó que el índice denominado SAPS II mostró la mayor concordancia (Ri 0.96), seguido del índice de comorbilidad de Charlson y del TISS (Ri = 0.86 y 0.60 respectivamente). Entre las escalas subjetivas la clasificación de ASA mostró el mejor puntaje (κp 0.25).

Es de hacerse notar que ambos evaluadores estaban más familiarizados con la escala de evaluación clínica y con la de ASA, y que en algunos casos, se observaron errores en la transcripción de los datos y estimaciones inexactas, como por ejemplo en la escala de Glasgow contenida en la determinación del índice SAPS II, por lo que el grado de acuciosidad y experiencia con el procedimiento seleccionado influye en los resultados (varianza debido al método de medición).[31,32]

Las características de los examinadores, tal como la experiencia laboral (varianza debido a los observadores) y el número de enfermedades comórbidas en los examinados (varianza debida a los sujetos de estudio) también contribuyeron a la variabilidad de los resultados.

No creemos que todos estos factores sean limitantes del estudio, sino por el contrario, se corrobora que diversas situaciones que suceden en la

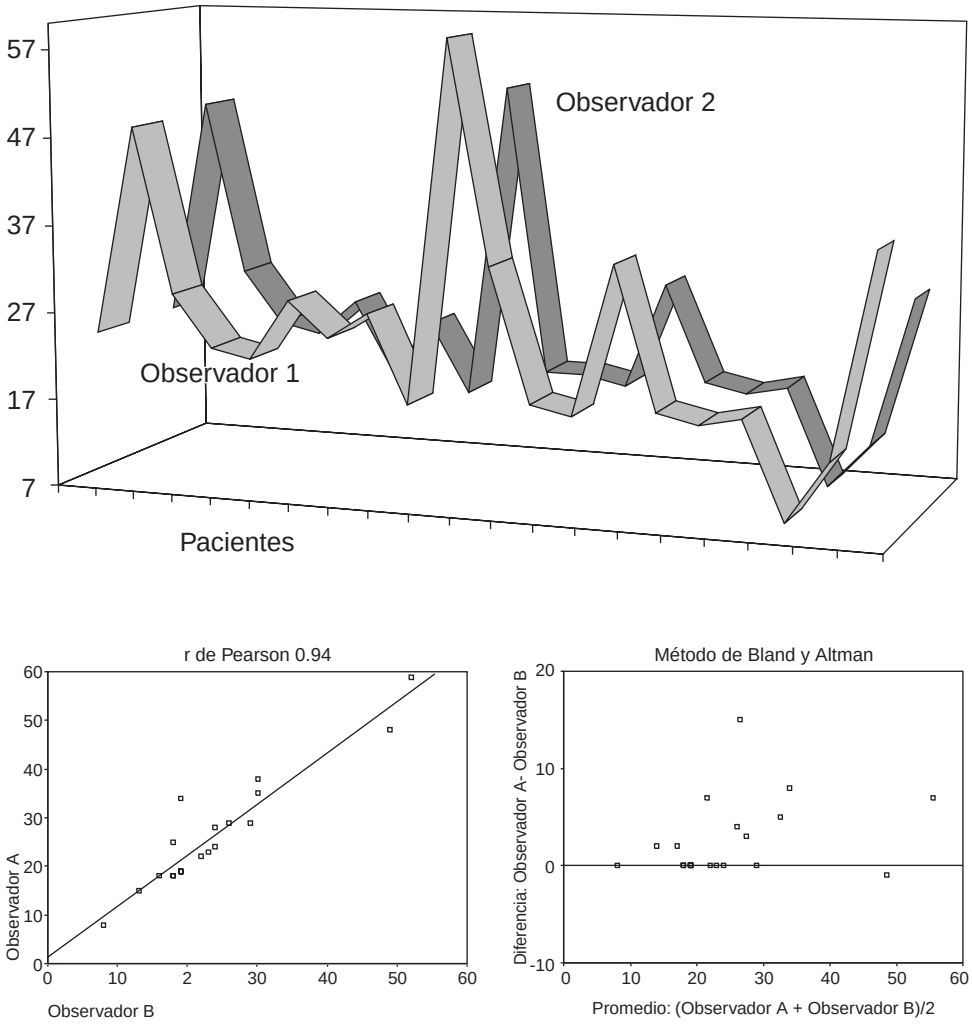

Figura 1. *Puntaje simplificado de fisiología aguda "SAPS II".*

práctica clínica habitual explican las discordancias incluso con el uso de índices puntuales, de ahí surge la necesidad de encontrar la mejor clasificación posible para calificar el nivel de gravedad de los enfermos en los servicios de admisión, en la que estos factores tengan menos influencia en este tipo de evaluaciones.

Una buena clasificación es útil para una aplicación más efectiva y racional de la terapéutica médica y probablemente disminuya ingresos hospitalarios innecesarios, reduciendo costos, morbilidad, mortalidad, iatrogenia e infecciones nosocomiales, así como para orientar al médico, paciente y familiares sobre las decisiones que deben tomarse acerca del tipo y grado de intervención médica.[33-35]

Desde el punto de vista del análisis estadístico se observaron similitudes entre los puntajes obtenidos con el Ri y con el coeficiente de correlación de Pearson, y se demostró la utilidad de desglosar los componentes del desacuerdo, al identificarse en el caso del índice SAPS II, que a pesar de mostrar un mayor acuerdo en general y una correlación alta existió un sesgo, ya que el evaluador A tendió a calificar con mayor gravedad a los pacientes.

Se concluye que aunque más elaborados y con un tiempo de aplicación discretamente mayor, en ausencia de un "estándar de oro" y dado que han mostrado mayor concordancia, es preferible el uso de índices objetivos como el SAPS II. También se pone de manifiesto la necesidad de interpretar y de utilizar adecuadamente las diversas técnicas estadísticas empleadas en las mediciones de acuerdo.

Ya se ha hecho notar la causa de la discrepancia clínica y se han sugerido recomendaciones para disminuirla.[36]

Nota: Otros autores indican que las correlaciones como la de Pearson evalúan la fuerza de la asociación linear, no el acuerdo entre dos variables, sin embargo en este trabajo se utiliza conjuntamente con una medición de sesgo para evaluar los componentes del desacuerdo.

BIBLIOGRAFÍA

1. Pompei P, Charlson ME, Douglas G. Clinical assessments as predictors of one year survival after hospitalization: implications for prognostic stratification. *J Clin Epidemiol* 1988;41:275-84.
2. Fajardo-Gutiérrez A, Yamamoto-Kimura LT, Garduño-Espinoza J, Hernández-Hernández DM, Martínez García MC. Consistencia y validez de una medición en la investigación clínica pediátrica. Definición, evaluación y su interpretación. *Bol Med Hosp Infant Mex* 1991;48: 367-81.
3. Merino CE. Observaciones y mediciones. En: Moreno Altamirano, Cano Valle, García Romero, editores. *Epidemiología clínica*. 2ª ed. México: Nueva Editorial Interamericana, 1994:69-97.
4. Latour J, Abraira V, Cabello JB, López JS. Las mediciones clínicas en cardiología: validez y errores de medición. *Rev Esp Cardiol* 1997;50:117-128.
5. Knaus WA, Wagner D, Lynn J. Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science* 1991;254:389-94.
6. Burnum JF. Medical diagnosis through semiotics. Giving meaning to the sign. *Ann Intern Med* 1993;119:939-43.
7. Saklad M. Grading of patients for surgical procedures. *Anesthesiology* 1941;2:281-84.
8. Owens WD, Felts JA, Spitznagel EL. ASA physical status classifications: A study of consistency of ratings. *Anesthesiology* 1978;49:239-43.
9. Lemeshow S, Teres D, Klar J, Spitz JA, Gehlbach SH, Rapaport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993;270:2478-86.
10. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. *Crit Care Med* 1985;13:818-29.
11. Knaus WA, Wagner DO, Draper EA, Zimmerman JE et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991;100:1619-36.
12. Bone CR, Lemeshow S, Sauler F. A new simplified acute physiology score (SAPS II) based on a European/North American Multicenter Study. *JAMA* 1993;270:2957-63.
13. Keene AR, Cullen DJ. Therapeutic intervention scoring system. Update 1983. *Critical Care Medicine* 1983;11:1-3.
14. Miskulin DC, Athienites NV, Yan G, Martin AA, Ornt DB et al. Comorbidity assessment using the Index of Coexistent Diseases in a multicenter clinical trial. *Kidney International* 2001;60:1498-1510.
15. Charlson ME, Pompei P, Ales KL, Mackenzie R. A new method of classifying prognostic co morbidity in longitudinal studies: development and validation. *J Chron Dis* 1987;40:373-83.
16. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* 1975;31: 651-659.
17. Uebersax J. Statistical methods for rater agreement. Disponible en: *http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm* (consultado en junio 2002).
18. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;41:949-958.
19. Feinstein A, Cicchetti D. High agreement but low kappa: 1. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543-549.
20. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistics. *Am J Epidemiol* 1987;126:161-169.
21. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* 1996;7:199-202.
22. Drasgow F. Polychoric and polyserial correlations. In: Kotz S, Johnson NL, editors. *Encyclopedia of Statistical Sciences*. New York: John Wiley & Sons, 1986:68-74.
23. Models for matched pairs. In: Agresti A, editor. *An introduction to categorical data analysis*. New York: Wiley, 1996:226-256.
24. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979;86:420-428.
25. Nichols DP. *Choosing an intraclass correlation coefficient*. SPSS Keywords, No. 67, 1998.
26. Yaffe RA. *Enhancement of reliability analysis: application of intraclass correlations with SPSS/Windows v.8*.
27. Intraclasss correlation coefficient. In: Kotz S, Johnson NL, editors. *Encyclopedia of Statistical Sciences*. New York: John Wiley & Sons, 1986:212-216.
28. Bland M, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-310.
29. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Statist Med* 1998;17:101-110.
30. Downie N, Heath R. *Métodos estadísticos aplicados*. 5th ed. México:. Harla, 1986:95.
31. Holt AW. Prospective evaluation of residents and nurses as severity score data collectors. *Crit Care Med* 1992; 20:1688-91.
32. Prasad K. The Glasgow coma scale: a critical appraisal of its clinimetrics properties. *J Clin Epidemiol* 1996;49:755-763.
33. Soulen JL, Duggan AK, DeAngelis CD. Identification of potentially avoidable pediatric hospital use: admitting physician judgment as a complement to utilization review. *Pediatrics* 1991;94:421-424.
34. Oye RK. A simple method for rating illness severity at admission and expected functional status at discharge: how should we use the information? *Am J Med* 2000;109: 250-251.
35. Horn SD, Sharkey PD, Bucle JM, Backofen JE, Averill RF, Horn RA. The relationship between severity of illness and hospital length of stay and mortality. *Med Care* 1991;29:305-17.
36. Sacket DL. Clinical disagreement: I. How often it occurs and why. *CMAJ* 1980;123:499-01.

Correspondencia:
Dra. Leticia Rodríguez Pimentel,
Servicio de Admisión Continua, Hospital de Especialidades, Centro Médico Nacional SXXI,
Avenida Cuauhtémoc Núm. 330, Colonia Doctores, México, D.F.
Tel: 5627 69 00 ext. 21772
E-mail: lropi@yahoo.com