Scores de Gravedad: Qué hay atrás de las siglas

Enrique Laffaire
Médico Especialista en Terapia Intensiva
Miembro Titular de la SATI
Director del Centro Cochrane de la Argentina

Desde la creación de las primeras unidades de Terapia Intensiva, se ha avanzado notablemente en técnicas de soporte y tratamiento y en una mejor comprensión de fenómenos fisiopatológicos. Esto ha permitido el reemplazo de la función de órganos afectados, a veces por períodos prolongados. Sin embargo, estos avances han ido acompañados de un incremento en el consumo de recursos. Las unidades de TI son ahora uno de los grandes consumidores de los recursos hospitalarios. Por este motivo, la evaluación de la efectividad se convierte cada vez más en una prioridad en la investigación médica.

La evaluación de la efectividad de una intervención (ya sea una droga, un procedimiento, un programa) de acuerdo a criterios metodológicos estándar debería llevarse a cabo mediante un estudio comparativo aleatorizado. En la evaluación de las UTI, esto demandaría asignar pacientes a una unidad o a una sala general. Se plantean entonces problemas éticos debido a la creencia general acerca de la utilidad de las UTI. Por lo tanto se ha recurrido a la comparación con controles históricos (la experiencia previa a la introducción de las UTI o de alguna intervención en particular). Este tipo de estudios tienen debilidades metodológicas insalvables. En la práctica, la evaluación de las UTI se limita a comparar si la evolución de los pacientes es acorde a un estándar esperado. Ese estándar puede ser los resultados de unidades de excelencia o de grupos de unidades no seleccionadas. La comparación cruda de la mortalidad en distintas terapias intensivas no es adecuada. Otros factores pronósticos, diferentes de la calidad de atención, influyen poderosamente.

Por tales motivos, se ha tratado de generar modelos de predicción que contemplen variables clínicas que influyen en el pronóstico. El primero de estos modelos fue el *Acute Physiology and Chronic Health Evaluation* (APACHE)[1], desarrollado en 1981, en la George Washington University. Luego se desarrollaron otros modelos como el *Simplified Acute Physiology Score* (SAPS)[2] y mejoras y simplificaciones de ambos (SAPS II, APACHE II y APACHE III)[3,4].

El artículo del Comité de Scores de la SATI, en el presente número de MEDICINA INTENSIVA, se ocupa de la validez de estos scores en el contexto nacional y nos da oportunidad de analizar las fortalezas, debilidades y la utilidad de estos instrumentos.

Una manera de comprender mejor estos scores es conocer su proceso de desarrollo. Existen varias etapas en la construcción de un modelo de predicción que se podrían resumir así:

1. Selección de la población y la medida de resultado: si está construyendo un modelo de predicción en terapia intensiva, surge como una cuestión obvia la selección de la población. Sin embargo, los modelos disponibles no incluyen a todos los pacientes admitidos en una terapia intensiva. Con frecuencia se excluyen grupos peculiares como pacientes ingresados por enfermedad coronaria, pacientes en el postoperatorio de cirugía cardiovascular o quemados. En un estudio europeo (Euricus) se comprobó que sólo alrededor del 40 % de los pacientes respondían a la población con las que se desarrolló el SAPS II. La elección de la mortalidad hospitalaria como medida de resultado en terapia intensiva es intuitiva. Sin embargo la mortalidad hospitalaria depende en sólo de la calidad de la atención de la UTI, sino de la evolución posterior en el hospital. Además, con frecuencia crecientes se considera que a la mortalidad como un marcador insuficiente, y señala la necesidad de incorporar medidas de calidad de vida en la evaluación de las UTIs [5]]
2. Selección e identificación de las variables predictoras: debe decidirse que variables serán medidas para valorar su influencia en la evolución de los pacientes. Por motivos estadísticos, el número de variables a probar no puede ser infinito. En general, grupos de expertos eligen un numero de variables clínicas que por motivos biológicos o por experiencia previa se supone que guardan relación con la evolución (por ejemplo, edad, presencia de hipotensión, fiebre). Estas variables además deben poseer un bajo

grado de variabilidad intra e interobservador (un mismo sujeto debería encontrar el mismo resultado puesto frente a la situación de evaluar la variable en oportunidades diferentes y dos sujetos deben encontrar el mismo resultado en el mismo sujeto).

3. Construcción del modelo. Estas variables son medidas en poblaciones grandes y se registran los resultados y la evolución del paciente. De todas las variables registradas se selecciona un número menor por técnicas estadísticas (regresión logística) y por juicio clínico. Se intenta buscar el modelo más simple (menor número de variables). En este proceso se descartan variables que "se mueven" en conjunto. Por ejemplo, si todos los pacientes con hipotensión tuvieran acidosis y los pacientes normotensos tuvieran pH normal, no se deberían incluirse ambas variables, sino sólo una de ellas. El resultado de las técnicas estadísticas son coeficientes que le adjudican un peso a las modificaciones de cada variable con respecto al valor basal. Mediante una ecuación de regresión se calcula cuales la probabilidad (en este caso de muerte), de acuerdo al valor de las variables elegidas.
4. Validación del modelo: la validación consiste en determinar si las prediccione efectuadas con ese modelo se condicen con la evolución real de los pacientes. La validación puede llevarse a cabo utilizando los datos con los que se construyó el modelo: se selecciona al azar un grupo de pacientes (la mitad o 2/3 del total) y se comprueba si en esos pacientes el modelo predice adecuadamente. Todos los modelos disponibles fueron validados al menos por este método. Sin embargo, aunque fueron elegidos al azar, los integrantes de la muestra de validación siguen perteneciendo a la población original y es mas probable que en esa población el modelo sea útil. Por ese motivo, se lleva a cabo una validación externa usando un grupo de pacientes distinto, en otros hospitales, otros países u otros años y se verifica la adecuación de la predicción. Existen dos elementos esenciales para evaluar si la predicción de un modelo es adecuada en una población determinada:
   -La calibración que compara la mortalidad observada y la mortalidad predicha. Actualmente se mide en la práctica mediante la prueba de Hosmer-Lemeshow.
   -La discriminación que mide la capacidad del modelo para distinguir entre pacientes que viven y aquellos que mueren. El instrumento de medida es la curva ROC. Un área bajo la curva ROC de 1 corresponde a una discriminación perfecta y un área de 0.5 a la ausencia de discriminación. Estos dos conceptos son difíciles de interpretar per o puede ncomprenderse mejor considerando un ejemplo. Consideremos que un modelo de predicción se aplica sobre un grupo de 1000 pacientes internados en UTI que tienen una mortalidad del 30%. Estos mil pacientes están compuestos por 500 pacientes con shock séptico que tuvieron una mortalida del 60% y 500 pacientes ingresados en postoperatorios de cirugía torácica con 0% de mortalidad. Si el modelo le asigna una mortalida del 30% a ambos grupos tendrá una excelente calibración (Muertes observadas 300, muertes predichas 300) y una pésima discriminación. Por otro lado, otro modelo aplicado a la misma población le asigna una probabilidad de muerte del 80% a los pacientes con shock séptico y del 20% a los pacientes en postoperatorios torácicos. Este modelo distingue entre los pacientes con alta y baja mortalida d (buena discriminación) pero tiene mala calibración (muertes observadas 300, muertes predichas 500). Las quinientas muertes predichas están compuestas por el 80% de los 500 pacientes con shock séptico y el 20% de los 500 pacientes en postoperatorio.

Se han propuesto numerosos usos para los scores de gravedad. Habiendo analizado los pasos necesarios para su desarrollo y las fuentes de error que acechan en cada uno de ellos, podemos considerar la utilidad de los scores de gravedad en diferentes situaciones.

El primer escollo en la utilidad de los scores de gravedad es el motivo de la publicación en el presente numero de MEDICINA INTENSIVA: ¿Tiene validez el score en el contexto en el que se lo va a utilizar? Otros estudios en poblaciones diferentes[6-11] han tenido resultados similares a los del articulo del Comité de Scores de la SATI: una discriminación relativamente buena y unacalibración deficiente

Además, con frecuencia no se aplican rigurosamente las normas de utilización de cada score. Aplicación a poblaciones para las que no fue diseñado (por ejemplo quemados), oportunidad de obtención de los datos (al ingreso, durante las primeras 24 hs) o distinta definición de las variables son algunas de los desvíos habituales.

Los scores se han propuesto para la toma de decisiones en pacientes individuales.(12). Sin embargo, los scores estiman probabilidades, y para un clínico es completamente inútil saber que su paciente tiene una probabilidad del 59% de sobrevivir.

Otra utilidad que se ha propuesto para los scores es asesorar en la asignación de recursos. De esta manera se propuesto tratar de manera diferente a grupos de pacientes de alto o bajo riesgo. Una propuesta habitual es derivar a zonas de menor complejidad a pacientes de bajo riesgo. Sin embargo, debe recordarse que el score predice bajo riesgo a ese grupo de pacientes estando internados

en UTI, *es decir sometidos a la atención y cuidados particul*ares de un área crítica. Es posible que sean de bajo riesgo porque están internados en UTI.

Con respecto a los pacientes de alto riesgo se ha desarrollado el concepto de "futilidad" o inutilidad. Se refiere a la ineficacia de la atención de pacientes con altísima probabilidad de muerte. Se ha concluido que los scores no son suficientemente poderosos para establecer predicciones con certeza[13-17]. Estudios de costo-efectividad mostraron una gran sensibilidad de los modelos a pequeñas variaciones en las estimación del riesgo (18).

La evaluación de la performance de las unidades es la aplicación mas común de los scores. Se ha propuesto hacerlo a través de la medición de la Standardised Mortality Ratio (SMR) que consiste simplemente en dividir el número de muerte predichas por el numero de muertes observadas. Una relación superior a uno implica una performance mejor que el grupo de referencia y una relación inferior a uno, un peor desempeño. Para esta utilidad valen todos los requisitos anteriormente señalados (adaptación al contexto local, adherencia rigurosa a la metodología del score, utilización en poblaciones adecuadas) pero además debe señalarse la importancia del tamaño de la población analizada. Como en cualquier prueba estadística, un pequeño tamaño de muestra puede impedir que se detecten diferencias significativas cuando éstas realmente existen.

¿Los scores de gravedad son inútiles? No, pero deben ser utilizados cuidadosamente y sobre todo deben dedicarse mayores esfuerzos de investigación, como el realizado por el Comité de Scores de la SATI, para mejorarlos y comprender más completamente sus aplicaciones.

BIBLIOGRAFIA

1. Knauss WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE – Acute Physiology and Chronic Health Evaluation: a physiologicall y based classification system Crit Care Med 1981; 9:591-7
2. Le Gall JR, Loirat P, Alperovitch A et al. A Simplified Acute Physiologic Score for ICU patients. Crit Care Med 1984;12:975-7
3. Knauss WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of diseas e classificati on system. Crit Care Med 1985;13:818-29
4. Knauss WA, Wagner DP, Draper EA et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991; 100:1619-36
5. Suter P, Armagandis A, Beaufils F, et al Predicting outcome in ICU patients: consensus conference organized by the ESICM and the SRLF. Intensive Care Med 1994,20: 390-7
6. Bastos PG, Sun X, Wagner DP, Knauss WA, Zimmerman JE. The Brazil APACHE III Study Group. Application of the APACHE III prognostic system in Brazilian intensive care units: A prospective multicenter study. Intensive Care Med 1996,22:564-570
7. Nouira S, Belghith M, Elatrous S, et al Predictive value of severity score systems: Comparison multicenter study of four models in Tunisian adult ICUs. Crit Care Med 1998; 26:852-859
8. Beck DH, Taylor GL, Millar B et al Prediction of outcome from intensive care: A prospective cohort study comparing Acute Physiology and Chronic Health Evaluation II and III prognostic systems in a Unitd Kingdom intensive care unit. Crit Care Med 1997; 25:9-15
9. Apolone G, Bertolini G, D'Amico R et al The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: Results form GiViTi. Intensive Care Med 1996;22: 1368-1378
10. Moreno R, Reis Miranda D, Fidier V. Evaluation of two outcome prediction models on an independent database. Crit Care Med 1998: 26:50-61
11. Rowan KM, Kerr JH, Major E, et al. Intensive Care Society's APACHE II study in Britain and Irelan d II. Outcome comparisons of intensive care units after adjustment for case mix by the America n APACHE II method. BMJ 1993;307:977-81
12. Knauss WA, Rauss A, Alperovitch A et al. Do objective estimates of chance for survival influence decisions to withhold or withdraw treatment. Med Decis Making 1990;10:163-171
13. Teres D, Rapoport J. Identifying patients with high risk of high cost. Chest 1991;99:530-1
14. Rapoport J, Teres D, Lemeshow S, Avruni n JS, Haber R. Explaining variability of cost using a severity of illness measure for ICU patients. Med Care 1990;28:338-48
15. Rapoport J, Teres D, Lemeshow S. Can futility be defined numerically?. Crit Care Med 1998;26:1781-1782
16. Emanuel EJ, Emanuel LL: The economics of dying. N Engl J Med 1994; 330:540-554
17. Society of Critical Care Medicine Ethics Comittee: Consensus statement of the Society of Critical Care Medicine Ethics Comitte e regardin g futile and other possibly inadvisable treatments. Crit Care Med 1997;25:887-891
18. Glance LG, Osler T, Shinozaki T. Intensive care unit prognostic scoring system to predict death: A cost effectiveness analysis. Crit Care Med 1998; 26:1842-1849