EDITORIAL

Validación y actualización de los modelos clínicos predictivos: ¿cómo y por qué?

Validation and updating of clinical prediction models: why and how?

EWOUT W. STEYERBERG

¿Los pronósticos de un modelo predictivo desarrollado previamente son válidos para mis pacientes? Esta es una pregunta difícil que se trató cuidadosamente en un estudio reciente focalizado en la validez del puntaje GRACE para predecir mortalidad hospitalaria. (1)

¿Porqué es importante este estudio de gran calidad de Mangariello y Gitelman? En la literatura médica, existe un creciente número de publicaciones sobre modelos predictivos. Se proponen nuevos métodos, como aprendizaje automático con datos clasificados, aprendizaje automático profundo, e inteligencia artificial. Independientemente del método de desarrollo, el problema central es si el modelo o algoritmo predictivo proporciona pronósticos válidos a los médicos y a sus pacientes que dependen de ellos como fuente de información y para la toma de decisiones. De hecho, los autores plantean que pueden existir muchas diferencias en distintos ámbitos. Estas incluyen diferencias en las características de los pacientes, los sistemas de salud, y el contexto socioeconómico, además de cambios en el tratamiento a lo largo del tiempo. Todas estas diferencias pueden hacer que un modelo desarrollado previamente no sea válido para el entorno en particular donde se aplica el modelo, por ejemplo el Hospital Dr. Juan A. Fernández en Buenos Aires, siendo necesaria su actualización para ese contexto específico. (2)

Validación: ¿cómo y por qué?

¿Cómo se debería realizar una validación? El primer problema es si es esperable que un modelo desarrollado con anterioridad sea aplicable a un grupo de validación. El trabajo presentado incluyó pacientes de dos centros en Argentina en el desarrollo del modelo internacional GRACE. Por lo tanto, es probable que el grupo de validación estuviera relacionado con el grupo de desarrollo. Además, el modelo incluía un conjunto clínicamente razonable de predictores, y había sido propuesto por un grupo internacional de expertos.

La guía TRIPOD proporciona más elementos para llevar a cabo una validación, especialmente en su detallado documento de Explicación y Elaboración. (3) Los elementos importantes para la validación incluyen un tamaño de muestra adecuado y métodos apropiados.

a) Tamaño de la muestra: Un tamaño de muestra adecuado implica al menos 100 eventos en la validación. (4) Por lo tanto, si la mortalidad hospitalaria es del 5%, se necesita un tamaño de muestra de al menos 2000 pacientes para obtener resultados confiables. También, los datos faltantes pueden ser imputados mediante métodos estadísticos avanzados para hacer uso completo de toda la información disponible, aún en el caso de que haya algunos registros de pacientes incompletos. Ambas condiciones se cumplieron en el presente estudio para el grupo total de pacientes (2104 pacientes, 117 eventos). (1) Sin embargo, el tamaño de la muestra fue un factor limitante en el subgrupo de SCA noST, con solo 35 eventos. En consecuencia, es imposible separar un desempeño aparentemente adecuado de la falta de potencia para detectar uno incorrecto.
b) Métodos adecuados: Los aspectos clave son la capacidad discriminatoria y la calibración. (2). Generalmente, la discriminación se evalúa por medio del área bajo la curva ROC (ABC), también llamada concordancia o estadístico c. La calibración evalúa si los riesgos estimados concuerdan con la frecuencia del evento. Los autores enfatizan correctamente la evaluación gráfica por sobre la prueba estadística. Si queremos evaluar la capacidad de un modelo de guiar la toma de decisiones, se necesitan medidas más modernas, incluyendo el "Beneficio Neto". (5) Este último acepta el número de clasificaciones de datos verdaderos positivos y penaliza las clasificaciones de falsos positivos, cuando separamos aquellos pacientes de riesgo alto versus bajo mediante un modelo predictivo. El peso relativo está definido por el contexto clínico, lo cual es mejor que utilizar un peso estadístico. (6)

Interpretación y consecuencias de la falta de validez

¿Cómo debemos interpretar los resultados? El puntaje GRACE fue bien desarrollado. El modelo tiene validez interna adecuada, sin riesgos de sobreestimación ya que el tamaño de la muestra fue muy grande (Tabla 1), pero obviamente está lejos de ser perfecto para predecir quienes van a morir y quienes no, y puede ser inapropiado para grupos específicos de pacientes. La validación externa de este trabajo confirma que la capacidad discriminatoria sigue siendo adecuada cuando el modelo se aplica en otro entorno. El ABC fue de 0.83 en el desarrollo del modelo y de 0.87 en la validación externa, (1) lo cual es atribuible a una mayor heterogeneidad [variabilidad de los valores predictores entre pacientes: "mezcla de casos" ("case-mix")]. (2)

Rev Argent Cardiol 2019;87:264-265. http://dx.doi.org/10.7775/rac.es.v87.i4.15891
VER ARTÍCULO RELACIONADO: Rev Argent Cardiol 2019;87:231-238. http://dx.doi.org/10.7775/rac.es.v87.i4.15346

Tabla. Síntesis de los principales problemas de validez interna y externa de modelos predictivos clínicos. (2)

Tipo de validez	Problemas	Descripción	Posibles soluciones
Validez interna	Sobrestimación	El modelo describe el entorno de desarrollo, pero no es válido para el entorno del cual se derivó	Mayor tamaño de la muestra, modelización cuidadosa. Cuantificación mediante validación cruzada o bootstrapping
	Subestimación	El modelo pierde algunos patrones importantes en los datos, tornándolo inválido para algunos tipos de pacientes.	Nunca se debe excluir una modelización cuidadosa. Tomar en cuenta que la elaboración de cualquier predicción está basada en la definición específica del modelo.
Validez externa	Definiciones y mediciones de predictores	El modelo se desarrolló con distintas definiciones de predictores, invalidando su desempeño en otros contextos.	Tratar de atenerse a la definición ('elementos de datos comunes'). Considerar que las diferencias potenciales pueden impactar sobre la validez del modelo.
	Predictores perdidos	Diferencias en las características de los predictores no incluidas en el modelo relacionado al entorno específico y no explicadas por diferencias en la distribución de valores de predictores que están en el modelo.	Compararar los entornos de desarrollo y validación. Considerar que cualquier modelo predictivo es limitado. Muchas características predictivamente relevantes pueden no estar incluidas en un modelo, por dificultad en su medición o por se todavía desconocidas.

La calibración de las predicciones es el aspecto de la validez más relevante para pacientes individuales y es el talón de Aquiles de los modelos predictivos. (7) En el estudio de validación presentado, los resultados fueron peores que lo esperado, lo cual puede tener dos interpretaciones muy diferentes. Uno es que el cuidado fue subóptimo en el entorno de la validación. El otro es que el modelo fue inadecuado debido a que algunos predictores que fueron diferentes en el contexto de la validación en comparación con el entorno de desarrollo no fueron incluidos en el modelo (Tabla 1). Por lo tanto, pueden existir diferencias en el riesgo basal entre los entornos de desarrollo y validación que no fueron capturados por el modelo. (1, 2) ¿Podemos aplicar entonces, el modelo GRACE en este hospital específico, en Argentina, en Sudamérica? Los resultados de validación de este estudio apoyan el concepto de un modelo global, con efectos predictores que son ampliamente válidos. (8) Sin embargo, su aplicación local requiere de actualización al entorno específico. Un enfoque simple consiste en actualizar el modelo cruzándolo con un modelo local, de modo que las predicciones sean en promedio correctas. (8, 9) Nuevos enfoques para ese tipo de actualización necesitan cuidado, junto con la disponibilidad de una recolección de datos mucho más rutinaria y el deseo de utilizar sistemas de autoaprendizaje. (10) Estas predicciones más actualizadas apoyarán la toma de decisiones y contribuirán a mejorar los resultados de pacientes individuales.

BIBLIOGRAFÍA

1. Mangariello BN, Gitelman PC. Validation of the GRACE Score (Global Registry of Acute Coronary Events) as Predictor of In-hospital Mortality in Acute Coronary Syndromes in Buenos Aires. Rev Argent Cardiol 2019;87:231-8.

2. Steyerberg EW. New York: Springer; 2009. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009. http://doi.org/dtwhd3.

3. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:1-73. http://doi.org/gfrkkz.

4. Vergouwe Y1, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475-83. http://doi.org/bj7bng

5. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38. http://doi.org/bj7bng

6. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6. http://doi.org/gcx6nq

7. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. JAMA 2018;320:27-8. http://doi.org/gd4rgm