¿Cómo utilizamos los puntajes de riesgo?

Luciano Oscar Lucas

Luciano Oscar Lucas

Médico de staff del Servicio de Cardiología del Hospital Italiano de Buenos Aires. Ciudad Autónoma de Buenos Aires, Argentina.

Acta Gastroenterol Latinoam 2021;51(3):242-244

Recibido: 06/07/2021 / Aceptado: 21/08/2021 / Publicado online: 27/09/2021 / https://doi.org/10.52787/kidl6065

Román Conroy comienza el capítulo "From epidemiological risk to clinical practice by way of statistics" del libro *Therapeutic Strategies in Cardiovascular Risk* con una frase de Tavia Gordon: "The power and elegance of the logistic function make it an attractive and elegant statistical instrument, but in the end we cannot push a button and hope that everything will come out all right. Because frequently it will not".[1]

En la medicina actual los puntajes de riesgo tienen una amplia difusión. Los utilizamos diagnosticar y estratificar el riesgo en la internación y nos orientan a la hora de tomar conductas y de evaluar el riesgo al alta y en el largo plazo. Dado que su aplicación está muy extendida y que se usan con mucha frecuencia, es crucial que sean poco complejos, fáciles de usar y que su utilización no demande mucho tiempo. Ahora bien, una vez que encontramos un puntaje de riesgo que cumple con estos atributos fundamentales, surgen otras preguntas: ¿cumple con las leyes a las que están sujetos los puntajes?, ¿funciona?, ¿siempre y en cualquier situación? ¿Por qué este y no otro?

Los puntajes funcionan de igual manera que un método de diagnóstico común (como pueden ser la troponina para el síndrome coronario agudo o el NT proBNP para la insuficiencia cardíaca) y se les aplican las mismas leyes de sensibilidad y especificidad. A su vez, tienen algunas particularidades a la hora de ser generados y de evaluar su funcionamiento.

En forma simplificada, y sin entrar en vericuetos estadísticos complejos, los dos requisitos más relevantes que deben cumplir son: en primer lugar, que el punto final que prediga sea claro, estandarizado y fácilmente replicable. Una cosa es la muerte como punto final, y otra la disnea. Este último no parece ser el mejor punto final a la hora de generar un puntaje de riesgo, puesto que el diagnóstico se presta a subjetividad en muchos casos.

En segundo lugar, debe surgir de una muestra representativa de la población sobre la cual va a aplicarse.[2] A este respecto podemos utilizar dos de ellos a modo de ejemplo. El puntaje de Framingham deriva de una cohorte de 5345 personas oriundas de esa comunidad que fueron seguidas hasta por doce años, y el punto final pautado estaba compuesto de muerte atribuida a enfermedad

Correspondencia: *Luciano Oscar Lucas*
Correo electrónico: luciano.lucas@hospitalitaliano.org.ar

coronaria, infarto de miocardio, angina o "insuficiencia" coronaria.[3] Mientras tanto, más acá en el tiempo, el grupo del Proyecto SCORE utilizó una cohorte de 200.000 individuos de once países europeos, seguidos hasta por trece años, y el punto final evaluado fue la muerte cardiovascular. En su generación se utilizaron también factores genéticos y medioambientales subrogantes de las diferentes regiones geográficas en la función.[4] A simple vista, parecería que el segundo requisito fue diseñado a partir de una muestra más representativa de la población sobre la cual iba a ser aplicado y que, además, su punto final era bien concreto. Aun considerando estas diferencias trascendentes, en su editorial sobre la publicación del puntaje SCORE, Topol y col. critican el punto final utilizado en ambos puntajes. En el primer caso por ambiguo y sensible a sesgos y en el segundo por insuficiente.[5] Esto demuestra lo difícil que puede ser encontrar los puntos finales adecuados.

Una vez generado, la forma correcta de evaluar el desempeño de los puntajes de riesgo es mediante la medición de tres características: la eficacia de discriminación, la calibración y la capacidad de reclasificación que se da ante el agregado de variables al modelo original.

La primera de las cualidades, la eficacia de discriminación, es la idoneidad de la función para separar a aquellos que poseen una alta probabilidad de presentar el punto final evaluado de aquellos que no la poseen. Esto puede evaluarse dentro de la misma población que utilizamos para generar la función, a partir de una parte de la muestra que se reserva antes de la confección del puntaje a los fines de su validación: a esto llamaremos *validez interna*. También existe una validez externa, que es la más importante y surge de aplicar la función a otras poblaciones y evaluar su capacidad discriminativa.

Para entender cuán tan trascendente es que la población sobre la cual aplicamos el puntaje sea similar a aquella a partir de la cual se generó, podemos usar de ejemplo la edad. A la hora de aplicar los puntajes de riesgo a pacientes en rangos etarios extremos –por ejemplo, mayores de 65 años–, debemos tener cuidado, porque si estas edades no están bien representadas en la muestra a partir de la cual se generó el puntaje, podría suceder que el rendimiento de la función no sea del todo bueno si lo aplicamos a una población que en su mayoría es de edad igual o mayor. A su vez, estos modelos plantean un coeficiente beta idéntico para todos los rangos etarios y puede ser que en la práctica esto no ocurra. Incluso puede suceder que otros factores afecten la función de diferente manera según la edad. Esto podría corregirse, por lo menos parcialmente, generando un factor de interacción entre la edad y los demás factores incluidos en la función.[9]

La calibración, en cambio, es una medida de cuán fidedigna es la predicción. Es decir, ¿cuántos de los que se predijo que iban a presentar el punto final realmente lo presentaron? Tanto el lugar como el momento en los que se aplica la fórmula afectarán la calibración. Por ejemplo, si la aplicamos en un lugar que presenta una mayor prevalencia del evento que el lugar en donde fue generado el puntaje, habrá una tendencia a la infraestimación. Ocurrirá lo contrario si lo aplicamos en un lugar donde la incidencia del punto final es menor: en este caso habrá una tendencia a la sobreestimación. La mejor calibración se obtiene al aplicar la fórmula en una población de características similares a aquella a partir de la cual fue generada.[2, 6] Esto es así porque la probabilidad previa, al aplicar cualquier tipo de test o puntaje, es trascendente en el rendimiento que este tendrá. ¿Será necesario, entonces, hacer algo con la fórmula para que calibre mejor en mi población? La respuesta es que muchas veces sí. En efecto, así lo han hecho con el puntaje de riesgo de Framingham países como China[7] y el Reino Unido.[8]

Por último, está la reclasificación, que es una medida del porcentaje de individuos que presentaron o no presentaron un evento y que fueron correctamente reclasificados a una nueva categoría luego de añadirse alguna variable de riesgo a la fórmula.[2] Es un concepto más nuevo que los anteriores, pero se encuentra muy en boga actualmente.

A modo de ejemplo conceptual, es cuestión de tiempo para que la genética nos muestre sustratos de riesgo para la enfermedad cardiovascular, como lo ha hecho con otras patologías. Hoy en día no encontramos ningún puntaje de riesgo que considere de manera directa variables genéticas en su fórmula. Es posible pensar que en el futuro el genotipo tendrá tanta o más relevancia que el fenotipo, en el cual nos basamos hoy para estratificar el riesgo de nuestros pacientes. Cuando la ciencia halle estas variables y se las incorpore a los modelos existentes, la cuantificación de la capacidad de reclasificación será trascendente para medir el aporte de estas variables a los diversos puntajes.

En conclusión, los puntajes de riesgo, sea cual fuere, están sujetos a las mismas leyes que los demás estudios o métodos diagnósticos. Los debemos utilizar teniendo siempre en cuenta la probabilidad previa de la población sobre la que estamos aplicándolo, puesto que el rédito que le extraeremos a la estimación depende de esa población. Incluso cuando la población posea características similares, la distribución etaria dentro de ella debe ser tenida en cuenta, puesto que, al ser un predictor de los más importantes, si el rango etario está poco representado la función no va a rendir adecuadamente. La capacidad discriminativa, la calibración y el poder de reclasificación

son las tres características que debemos evaluar en un puntaje de riesgo.

Los puntajes, si bien son una herramienta, nunca podrán reemplazar el criterio médico del profesional que lo está aplicando. La frase de Tavia Gordon con la que iniciamos el texto, quizá con algunos matices y complejidades mayores, continúa siendo cierta en nuestros días.

Cite este artículo como: *Lucas LO. ¿Cómo utilizamos los puntajes de riesgo? Acta Gastroenterol Latinoam. 2021;51(3):242-4. https://doi.org/10.52787/kidl6065*

Referencias

1. Gordon T. Editorial: Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies. J Chronic Dis. 1974;27(3):97-102.
2. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol. 2009;54(14):1209-27.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47.
4. Conroy RM, Pyörälä K, Fitzgerald AP, *et al.* Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987-1003.
5 Topol EJ, Lauer MS. The rudimentary phase of personalised medicine: coronary risk scores. Lancet. 2003;362(9398):1776-7.
6 Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. Heart. 2006;92(12):1752-9.
7. Liu J, Hong Y, D'Agostino RB Sr, *et al.* Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. JAMA. 2004;291(21):2591-9.
8. Brindle P, May M, Gill P, *et al.* Primary prevention of cardiovascular disease: a web-based risk score for seven British black and minority ethnic groups. Heart. 2006;92(11):1595-602.
9. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.

How Do We Use Risk Scores?

Luciano Oscar Lucas

Cardiologist, Staff Physician, Hospital Italiano de Buenos Aires. Ciudad Autónoma de Buenos Aires, Argentina.

Luciano Oscar Lucas

Acta Gastroenterol Latinoam 2021;51(3):245-247

Received: 06/07/2021 / Accepted: 21/08/2021 / Published online: 27/09/2021 / https://doi.org/10.52787/kidl6065

Román Conroy, in the book *Therapeutic Strategies in Cardiovascular Risk*, begins his chapter "From epidemiological risk to clinical practice by way of statistics" with a quote from Tavia Gordon: "The power and elegance of the logistic function make it an attractive and elegant statistical instrument, but in the end we cannot push a button and hope that everything will come out all right. Because frequently it will not".[1]

In today's medicine, risk scores are widely spread. We use them for diagnosis and risk stratification in hospitalization. These scores guide us when we decide therapeutic behaviors and assess the risk at discharge and in the long term. Since their application is widespread and they are frequently used, it is crucial that they must be easy to use, uncomplicated and not too time-consuming. Now, once we find one that owns these fundamental attributes, other questions arise: does it comply with the laws to which the scores are subject? Does it always work in any situation? Why this one and not another?

The scores work in the same way as a common diagnostic method (such as troponin for acute coronary syndrome or NT proBNP for heart failure), and the same laws of sensitivity and specificity apply to them. In turn, they have some peculiarities when they are generated and their performance is evaluated.

In a simplified way, and without going into complex statistical twists and turns, the two most relevant requirements that must be met are, firstly, that the endpoint it predicts be clear, standardized and easily replicable. Death as an endpoint is one thing, dyspnea is another. The latter does not seem to be the best endpoint when generating a risk score, since the diagnosis in many cases is subjective.

The second requirement is that it must arise from a representative sample of the population on which it is to be applied.[2] In this regard, we can use two of them as examples. The Framingham score is derived from a cohort of 5,345 people from that community, followed for up to twelve years, and the scheduled endpoint was composed of death attributed to coronary artery disease, myocardial infarction, angina or coronary "insufficiency".[3] Meanwhile, closer in time, the SCORE Project

Correspondence: *Luciano Oscar Lucas*
Email: luciano.lucas@hospitalitaliano.org.ar

group used a cohort of 200,000 individuals from eleven European countries, followed for up to thirteen years, and the endpoint evaluated was fatal cardiovascular disease. They also used surrogate genetic and environmental factors from the different geographical regions in its generation[4] At first glance, it would seem that the second requirement was designed from a more representative sample of the population on which it was to be applied and that, moreover, its endpoint was very specific. Even considering these important differences, in their editorial to the publication of the SCORE, Topol *et al*. criticize the endpoint used in both scores. In the first case because it is ambiguous and sensitive to biases and in the second because it is insufficient.[5] This shows how difficult it can be to find the appropriate endpoints.

The correct way to evaluate the performance of the risk scores, once generated, is by measuring three characteristics: discrimination efficiency, calibration and reclassification capacity when variables are added to the original model.

The first of the qualities, discrimination efficiency, is the suitability of the function to separate those who have a high probability of presenting the evaluated endpoint, from those who do not. We can evaluate it within the same population that we use to generate the function, using a part of the sample that is set aside, prior to the scoring, for validation purposes; we will call this internal validity. There is also an external validity, which is the most important and comes from applying the function to other and assessing its discriminatory capacity.

To understand how important it is that the population where we apply the score is similar to the one that generated it, we can use age as an example. When applying risk scores to patients in extreme age ranges, for example over 65 years of age, we must be careful, since these ages are not well represented in the sample from which the score was generated, the performance of the function may not be entirely good if we apply it to a population that is mostly of the same or older age. In turn, these models propose an identical beta coefficient for all age ranges, and this may not be the case in practice. It may happen that other factors affect function differently depending on age. This could be corrected, at least partially, by creating an interaction factor between age and the other factors included in the function.[9]

Calibration, on the other hand, is a measure of how reliable the prediction is. That is, how many of those who were predicted to submit the endpoint actually submitted it? Both where and when the formula is applied will affect the calibration. For example, if we apply it in a place that has a higher prevalence of the event than the place where it was generated, there will be a tendency to underestimate. The opposite will occur in a place where the incidence of the endpoint is lower. In this case there will be a tendency to overestimate. The best calibration is obtained by applying the formula to a population of similar characteristics to the one from which it was generated.[2, 6] This occurs because the prior probability, when applying any type of test or score, is transcendent in the performance that it will have. Will it be necessary to do something with the formula, so that it calibrates better in my population? The answer is often yes. In fact, countries such as China[7] and the United Kingdom[8] have done so with the Framingham risk score.

Finally, the reclassification: it is a measure of the percentage of individuals who did or did not present an event and who were correctly reclassified to a new category after adding some risk variable to the formula.[2] It is a newer concept than the previous ones but is widely used nowadays.

As a conceptual example, it is a matter of time for genetics to show us risk substrates for cardiovascular disease, as it has done with other pathologies. Today, we do not find any risk score that directly considers genetic variables in its formula. It is possible to think that, in the future, the genotype will be as or more relevant than the phenotype, which we use currently to stratify the risk of our patients.

When science finds these variables and incorporates them into existing models, the quantification of the reclassification capacity will be important to measure the contribution of these variables to the various scores.

In conclusion, risk scores, whatever they may be, are subject to the same laws as other studies or diagnostic methods. We must use them always taking into account the prior probability of the population on which we are applying it, since the results that we will extract from the estimate depends on that population. Even though it has similar characteristics, the age distribution within it should be taken into account, because it is one of the most important predictors: if the age range is poorly represented, the function will not perform adequately. Discrimination ability, calibration, and reclassification power are the three characteristics that we must evaluate in a risk score.

Scores, although they are a tool, can never replace the medical criteria of the professional who is applying them. Tavia Gordon's phrase, with which we began the text, perhaps with some nuances and greater complexities, continues to be true today.

Cite this article as: *Lucas LO. How Do We Use Risk Scores? Acta Gastroenterol Latinoam. 2020;51(3):245-7. https://doi.org/10.52787/kidl6065*

References

1. Gordon T. Editorial: Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies. J Chronic Dis. 1974;27(3):97-102.
2. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol. 2009;54(14):1209-27.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47.
4. Conroy RM, Pyörälä K, Fitzgerald AP, *et al.* Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987-1003.
5. Topol EJ, Lauer MS. The rudimentary phase of personalised medicine: coronary risk scores. Lancet. 2003;362(9398):1776-7.
6. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. Heart. 2006;92(12):1752-9.
7. Liu J, Hong Y, D'Agostino RB Sr, *et al.* Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. JAMA. 2004;291(21):2591-9.
8. Brindle P, May M, Gill P, *et al.* Primary prevention of cardiovascular disease: a web-based risk score for seven British black and minority ethnic groups. Heart. 2006;92(11):1595-602.
9. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.