Med Intensiva. 2017;41(6):327–329

medicina intensiva

www.elsevier.es/medintensiva

EDITORIAL

Buena ciencia

Good science

V. Modesto i Alapont*

Unidad de Cuidados Intensivos Pediátrica, Hospital Universitari i Politècnic La Fe, Valencia, España

Disponible en Internet el 23 de diciembre de 2016

En el presente número de MEDICINA INTENSIVA[1], la doctora Patricia García Soler et al. de la UCI pediátrica del Hospital Regional Universitario Carlos Haya de Málaga publican un excelente artículo que pretende validar un sistema transcutáneo de medición de la concentración de hemoglobina (Hb) en niños críticos con riesgo de sangrado. La metodología empleada es impecable, el análisis riguroso y la conclusión honesta. A mi parecer, muy buena ciencia.

En la literatura sobre pruebas diagnósticas, el objetivo clínicamente relevante suele consistir en evaluar la exactitud de una medición. La exactitud (*accuracy* en la literatura anglosajona) es una propiedad que se establece en una determinada dimensión: la comparación de dicha prueba diagnóstica con otra medición que se toma como «patrón oro» *(gold standard)*. La exactitud se mide con las razones de verosimilitud o sus logaritmos decimales (los llamados «pesos de la evidencia», inventados por Alan Turing en 1940-1941 para conseguir descifrar el código nazi de las máquinas enigma[2]).

Sin embargo, hay que saber que la concordancia o fiabilidad, una propiedad que se establece en otra dimensión distinta, es una condición previa y necesaria para poder establecer exactitud. Lo que pretendemos evaluar cuando medimos la concordancia entre 2 mediciones, o cuando estudiamos la consistencia o reproducibilidad de una misma medida repetida en el tiempo, es la «discrepancia» entre ambas mediciones (o repeticiones): el sesgo en el que incurrimos cuando tomamos una medida por la otra. Solo si 2 medidas son concordantes (es decir, no discrepantes) y una de ellas es el «patrón oro», podrá evaluarse posteriormente la exactitud de la otra como método diagnóstico.

En el análisis de concordancia entre variables con datos categóricos, se utiliza el índice kappa (ponderado) de Cohen. Pero con relativa frecuencia, en la investigación clínica se tiene la necesidad de valorar la concordancia entre mediciones cuantitativas. En este caso, no nos sirve el coeficiente de correlación R^2 de Pearson y es erróneo utilizarlo para este propósito[3]. Dos mediciones pueden tener una correlación altísima y estadísticamente significativa, y sin embargo estar en diferentes escalas. Y lo que pretendemos evaluar cuando medimos la concordancia es la «identidad» entre ambas, no solo la capacidad de variar por influencia mutua. La mejor forma de medirla parece ser mediante el coeficiente de correlación y concordancia (CCC) de Lin[4].

Una alternativa muy adecuada es la elegida por la Dra. García Soler y su equipo. Se trata del cálculo del llamado coeficiente de correlación intraclase (CCI), que estima el promedio de las correlaciones entre todas las posibles ordenaciones de los pares de observaciones disponibles. Para entender la correlación intraclase, supongamos que todas las observaciones de una variable se ordenan en m grupos (en este caso los 2 grupos: medición de Hb transcutánea y de Hb del laboratorio) que contienen *n* observaciones cada uno. Y supongamos también (hipótesis nula) que no hay motivos para esperar que haya diferencias en el nivel medio de la variable entre los *m* grupos. Si esas diferencias existieran, las observaciones del mismo grupo tenderían a estar positivamente correlacionadas entre sí y su variabilidad sería

* Autor para correspondencia.
Correo electrónico: vicent.modesto@gmail.com

http://dx.doi.org/10.1016/j.medin.2016.10.006

diferente a la de las observaciones del otro grupo. Esta correlación es lo que se conoce como correlación intraclase.

La fórmula para el cálculo se basa en un modelo de análisis de la varianza de efectos aleatorios de un factor (ANOVA). La idea es que la variabilidad total de las mediciones se puede descomponer en 2 componentes: la variabilidad debida a las diferencias entre los distintos sujetos (varianza *entre sujetos*) y la debida a las diferencias entre los métodos de medición de la variable para cada sujeto (varianza *intra sujetos*). El CCI, un coeficiente paramétrico que puede considerarse el equivalente del estadístico kappa para variables continuas, se calcula entonces como la proporción que supone la varianza *entre sujetos* sobre la variabilidad total[5]. Como es una proporción, el CCI toma valores entre 0 y 1: está próximo a 1 si la variabilidad observada se debe fundamentalmente a las diferencias entre los sujetos, y no a las diferencias entre los métodos de medición (o entre los observadores); y toma el valor 0 en caso contrario. Aunque la interpretación es subjetiva, se asume por consenso una escala para valorar el CCI como medida de reproducibilidad: valores inferiores a 0,4 indican poca reproducibilidad y valores iguales o superiores a 0,75 reproducibilidad excelente. Los valores intermedios se consideran con una fiabilidad adecuada. La principal limitación del uso del CCI, además de las que derivan del incumplimiento de las hipótesis de aplicación del modelo ANOVA (normalidad, igualdad de varianzas e independencia de los errores), es su dependencia tanto del rango de variación de la escala de medida como del número de métodos de medición (o de observadores). Así, por ejemplo, si una medición presenta una variabilidad muy reducida, puede obtenerse un CCI bajo sin que esto signifique un método poco fiable.

Los investigadores malagueños utilizan, también en su análisis, el popular, el gráfico de Bland-Altman, y con él detectan la principal debilidad del método transcutáneo de medición de la Hb. *A priori*, fijan en ± 1 g/dl el límite de tolerancia para el sesgo en la medición (la diferencia máxima en las 2 medidas de la concentración de Hb para que el nuevo método incruento se considere fiable). Puede parecer una decisión arbitraria, pero a mi juicio es muy adecuada y está basada en razones clínicas bien fundamentadas: en el ámbito de los cuidados intensivos pediátricos, en el que método transcutáneo no invasivo pretende ser usado, una diferencia de 1 g/dl puede hacernos modificar la actitud terapéutica. Sería óptimo que la posible diferencia fuese inferior a dicho margen. Sin embargo, tanto el análisis estadístico como el gráfico de Bland-Altman objetivan, tal y como los autores señalan en el texto de los resultados, que «La media de las diferencias entre los valores de laboratorio y los del pulsi-cooxímetro fue de $0{,}66 \pm 1{,}46$ g/dl, con una mediana de 0,5 g/dl (RIQ: $-0{,}2$-1,4). La mediana de las diferencias en valores absolutos fue de 0,8 g/dl (RIQ: 0,4-1,7)». Es decir: el límite de tolerancia queda claramente rebasado. De hecho el IC 95% de la media poblacional del sesgo obtenido en la UCIP de Málaga (asumiendo normalidad) es $-2{,}2$-52 g/dl: el ± 1 g/dl (límite de tolerancia) es un valor posible al 95% de confianza. Este sesgo es 10 veces superior al obtenido en un reciente trabajo sobre el tema[6], pero existen otros estudios que comunican sesgos similares e incluso superiores[7–9].

Así que la magnitud posible del sesgo en que se incurre podría invalidar la medida y afectar claramente a la relevancia clínica del método transcutáneo. De hecho, tal y como señalan los investigadores, en la tabla 1 «se aprecia que el 60,9% de las mediciones por cooximetría no invasiva presentaba una diferencia ≤ 1 g/dl respecto al valor del analizador del laboratorio central» lo que significa que casi el 50% de las mediciones se encuentran fuera del límite de tolerancia establecido por los mismos autores *a priori*. Pero es que, además, con sus modelos multivariantes correctamente ajustados según el criterio informativo de Akaike (AIC), Soler et al. detectan que precisamente el «índice de perfusión» es una covariable que afecta mucho a la concordancia. Este es un dato clínicamente muy relevante, porque si en alguna población pediátrica tiene valor medir de manera continua la Hb para detectar precozmente la aparición de anemia, es en los pacientes con riesgo de sangrado potencial (y por eso el muestreo de esta investigación se ha realizado sobre estos niños). Y constituye un hándicap muy negativo que precisamente en los pacientes con más riesgo de anemia (con inestabilidad hemodinámica por *shock* hemorrágico incipiente, y por tanto con alta probabilidad de tener alterado el índice de perfusión) la medición transcutánea pierda fiabilidad. A mi entender es en estas situaciones clínicas donde se necesita que la medición sea más fiable.

Esta importante limitación es uno de los principales asuntos comentados en la discusión y así lo han intentado reflejar los autores en sus conclusiones. Por eso, como decía al principio, me parece «buena ciencia» un estudio que utiliza el texto no solo para señalar las limitaciones del propio trabajo, sino también para intentar encontrar argumentos contrarios que refuten las hipótesis que se pretendían inicialmente proponer como ciertas. Es decir, el investigador debe intentar ser él mismo el juez más severo e implacable de lo que va a proponer como verdad científica. En palabras del Premio Nobel Richard Feynman[10] «La idea es intentar dar toda la información para contribuir a que sean los otros los que juzguen el valor de tu contribución: no solo la información que pueda conducir al juicio en una dirección concreta u otra». La integridad constituye un principio fundamental del pensamiento científico.

Bibliografía

1. García Soler P, Camacho Alonso JM, González Gómez JM, Milano Manso G. Monitorización no invasiva transcutánea de la concentración de hemoglobina en pacientes críticos pediátricos con riesgo de sangrado. Med Intensiva. 2017;41.
2. Good IJ. Studies in the history of probability and statistics XXXVII. A M Turing's statistical work in World War II. Biometrika. 1979;66:393–6.
3. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther. 1981;29:111–23.
4. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. J Am Stat Assoc. 2002;97:257–70.
5. Pita Fernández S, Pértega Díaz S. La fiabilidad de las mediciones clínicas: el análisis de concordancia para variables numéricas. Atención Primaria en la red. Fisterra.com. [actualizado 12 Ene 2004; consultado Oct 2016] Disponible en: http://www.fisterra.com/mbe/investiga/conc_numerica/conc_numerica.asp
6. Phillips MR, Khoury AL, Bortsov AV, Marzinsky A, Short KA, Cairns BA, et al. A noninvasive hemoglobin monitor in the pediatric intensive care unit. J Surg Res. 2015;195:257–62.

7. Dewhirst E, Naguib A, Winch P, Rice J, Galantowicz M, McConnell P, et al. Accuracy of noninvasive and continuous hemoglobin measurement by pulse co-oximetry during preoperative phlebotomy. J Intensive Care Med. 2014;29:238–42.
8. Amano I, Murakami A. Use of non-invasive total hemoglobin measurement as a screening tool for anemia in children. Pediatr Int. 2013;55:803–5.
9. Jung YH, Lee J, Kim HS, Shin SH, Sohn JA, Kim EK, et al. The efficacy of noninvasive hemoglobin measurement by pulse co-oximetry in neonates. Pediatr Crit Care Med. 2013;14:70–3.
10. Feynman RP. La ciencia del culto al cargamento. Discurso inaugural 1974 en Caltech.