Variabilidad de observador en procedimientos diagnósticos afines a la cirugía cardiotorácica

Guillermo Careaga Reyna*♦ Juan Garduño Espinosa**♦ Rubén Argüero Sánchez***

Resumen

La reproducibilidad de resultados es una de las condiciones principales para la validez de un escrito médico. La medición de la variabilidad de observador es una estrategia que nos permite identificar el grado de precisión de los resultados obtenidos en la investigación clínica que aborde o requiera la utilización de instrumentos de diagnóstico clínico. Se hace en el presente trabajo una presentación de los conceptos sobre el tema y posteriormente revisamos la literatura de nuestro país en el área cardiológica y neumológica publicada de enero de 1980 a julio de 1993, encontrando que en 13 años, en cuatro revistas, sólo dos artículos informan haber medido la variabilidad de observador en trabajos de ecocardiografía, electrocardiografía, microscopía y hemodinamia, lo que nos lleva a concluir que se debe hacer más énfasis en la determinación de la misma ya que puede contribuir a explicar las discrepancias observadas en una investigación, además de las que son atribuibles al azar, al paciente, a los equipos auxiliares de diagnóstico o a una muestra insuficiente.

Palabras clave: Variabilidad, cirugía cardiotorácica, intra-observador, inter-observador, concordancia.

Summary

The results reliability is a predominant condition in the validity of clinical research. The observer agreement determination is a useful procedure in the confirmation of the obtained results in clinical research. The present work is a presentation of the basis about observer agreement and a review of national publications in cardiology, neumology and cardiothoracic surgery between january 1980 and july 1993, where we found just two publications with the application of observer agreement determination in ecocardiography, electrocardiography, microscopy and hemodinamics. It was concluded that is necessary the increase in determination of observer agreement and inform this results in our publications, excluding the variability attributed to patients, diagnostics equipment or the circunstances.

Key words: *Variability, cardiothoracic surgery, intra-observer, inter-observer, agreement.*

Introducción

Una de las características más importantes en la obtención de la información científica es la confiabilidad, entendida ésta como la consistencia que muestran los datos a través del tiempo en las mismas condiciones y que es corroborada por la replicación.[1]

No toda la variabilidad que observamos en el comportamiento de los organismos o sujetos en estudio es intrínseca. Por el contrario, una gran porción de la variabilidad del comportamiento es una función de la interacción del organismo con las variables ambientales,[1] tales como el tipo de instrumentos utilizados para la medición de diversos fenómenos o eventos motivo de estudio, así como las características del observador y las modificaciones a través del tiempo en la interpretación o medición de la misma observación.

Con base en lo anterior, podemos inferir que la reproducibilidad de una prueba puede estar influida por las diferentes condiciones del individuo y/o laboratorio, por la variabilidad entre los diferentes observadores de un mismo fenómeno (cuyo ejemplo sería las diversas opiniones expresadas por un grupo de radiólogos con respecto a una misma radiografía de tórax) y, finalmente, la variabilidad intra-observador, entendida ésta como la que presentaría una persona que evalúa un mismo evento con diferente perspec-

* Cirujano cardiotorácico. Alumno de la maestría en Ciencias Médicas, sede Centro, Hospital de Cardiología del Centro Médico Nacional Siglo XXI del IMSS.
**Profesor titular del curso de Epidemiología, Clínica para maestría en Ciencias Médicas, sede Centro, Centro Médico Nacional Siglo XXI del IMSS.
*** Cirujano cardiotorácico. Tutor académico de maestría y doctorado en Ciencias Médicas; director del Hospital de Cardiología del Centro Médico Nacional Siglo XXI del IMSS. Miembro titular de la Academia Nacional de Medicina.

tiva, situación que se demostraría con un residente a quien se le pide interpretar un electrocardiograma (ECG) por la mañana y nuevamente el mismo ECG durante la noche.[2]

El problema de la variabilidad intra e inter-observador no es nuevo en medicina. En 1947 Birkelo y asociados reportaron una significativa diferencia entre la opinión diagnóstica de radiografías de tórax de pacientes tuberculosos y al comparar la impresión diagnóstica de cinco expertos radiólogos ocurrió en 20 por ciento de los casos una clara discrepancia inter-observador.[3]

Desde hace varias décadas se han estudiado aspectos muy importantes en la investigación cardiovascular, sin embargo sus resultados no han podido compararse entre sí debido a la diversidad de enfoques tanto como por la variabilidad entre los distintos observadores.[4]

Un ejemplo clásico de lo anterior es la evidencia obtenida, desde 1965, de una considerable variabilidad de observador en la aplicación individual de diversos procedimientos clínicos aceptados universalmente, incluyendo el ECG,[5] de cuyo análisis dependen en forma evidente aplicaciones terapéuticas a determinado paciente. Ello orienta a pensar que las predicciones clínicas nunca serán certeras pero sí claramente probabilísticas. Si a esto agregamos la aparición de nuevos métodos diagnósticos en el área de la cardiología pudiera ocurrir un decremento en la capacidad predictiva para esos métodos por dos razones: la variabilidad inherente al procedimiento en sí y la variabilidad consecuente en su interpretación.[6]

Quizá con el panorama antes mencionado podamos entender la importancia de conocer no sólo la sensibilidad, especificidad y valor predictivo[7] de las diversas pruebas diagnósticas y pronósticas sino también de la variabilidad entre los observadores, y dentro de ellos mismos en diversas circunstancias al interpretar la prueba. Tal vez el estudio más comentado de decisión médico-quirúrgica que nos puede permitir dejar más clara la importancia de la evaluación de la variabilidad de observadores, sea el realizado por Barkwin en 1945,[8] al someter a evaluación a 1000 niños para determinar la necesidad de efectuarles amigdalectomía; 611 fueron operados y los restantes 389 fueron sometidos a revisión por otro grupo de pediatras quienes opinaron que el 45 por ciento requería amigdalectomía. Finalmente los 215 pacientes restantes a quienes no fue indicado el procedimiento quirúrgico en los dos primeros exámenes fueron evaluados por otro grupo de pediatras quienes indicaron el procedimiento en 47 por ciento. Cabe resaltar que los pediatras sólo examinaron a los pacientes en quienes el análisis previo indicaba no efectuar amigdalectomía.

El anterior es un ejemplo claro de que el entrenamiento tiende a conferir un grado determinado de certeza y de expectativas de ciertos resultados, lo cual es producto de la experiencia la cual nunca cesa y se adquiere por la exposición continua a la información, su intercambio y la interpretación de resultados. Por otra parte, algunos estudios formales han mostrado que la variabilidad de observador es innata e independiente de la experiencia.[9]

Al momento actual podemos enunciar dos formas claras de variabilidad o desacuerdo clínico. La primera ocurre al comparar el juicio clínico de un signo, síntoma o diagnóstico con los resultados de pruebas de laboratorio, biopsia y estudio de autopsia, como cuando Day y colaboradores compararon el ECG fetal con la auscultación del foco fetal, donde los observadores tuvieron una variabilidad de 20 por ciento con diferencias de hasta 15 latidos por minuto en relación a los resultados obtenidos en el ECG.[10] El segundo tipo de desacuerdo ocurre cuando el juicio clínico de un médico se compara con el resultado obtenido en el examen clínico por otro médico o bien, cuando se realiza una segunda evaluación por los mismos observadores.[11]

Entre las causas de no reproducibilidad de resultados en un estudio, Sackett y cols,[11] atribuyen a tres posibilidades etiológicas la variabilidad de resultados. La primera causa se refiere al examen: un ambiente inadecuado, interacción inadecuada entre el examinado y el observador y/o el uso o funcionamiento incorrecto del instrumento para la medición. La segunda posibilidad es la variación del examinado: cambios biológicos, efectos de la enfermedad o de los medicamentos y de la memoria para proporcionar la misma información. Finalmente, la causa atribuida al examinador (variabilidad intra e inter-observador): por variación biológica de los sentidos, la tendencia a anotar en los expedientes inferencias en lugar de evidencias, los diferentes esquemas de clasificación de un mismo fenómeno y los sesgos de sospecha diagnóstica.

Variabilidad en las áreas de cardiología y neumología

En el área cardiovascular y torácica, entre los estudios orientados a la detección de la variabilidad de observador, uno de ellos evaluó a tres expertos angiólogos en la detección de presencia o ausencia de pulsos en extremidades pélvicas encontrando que la concordancia intra-observador fue de 73 a 87 por ciento (concordancia observada), cuando se analizaban específicamente los pulsos femoral, tibial posterior y dorsal del pie; y la variabilidad inter-observador fue menor cuando la revisión de los pacientes se hacía en condiciones similares;[12] asimismo, Godfrey y colaboradores evidenciaron una considerable variabilidad en la detección de signos clínicos de obstrucción de las vías respiratorias en pacientes con enfermedad pulmonar obstructiva crónica.[13]

Aunque se han desarrollado esfuerzos para disminuir el desacuerdo inter-observador en el ECG, Gorman y cols,[14] compararon la interpretación de dos grupos de médicos que evaluaron 561 electrocardiogramas encontrando que el pri-

mer grupo detectó 307 anormalidades que el segundo grupo no observó, haciendo notar que estas discrepancias son vitales en un estudio diagnóstico de rutina del cual dependen decisiones terapéuticas importantes. Sugirieron que el empleo de índices o códigos puede eliminar la variación de observador debido a la interpretación clínica pero no la inconsistencia intrínseca que ocurre al definir una onda del ECG.

En otro estudio de 14 cardiólogos que revisaron 38 ECG, y que recibieron un instructivo específico donde se les indicaba anotar sólo si el estudio era normal, anormal o no adecuado, encontraron una variabilidad inter-observador entre el 5 y el 58 por ciento, y en proporciones similares la variabilidad intra-observador, con lo que se hizo evidente que la participación de un solo observador como substituto de varios evaluadores no es una garantía aceptable de confiabilidad.[15]

Por otro lado, Hull y colaboradores en 1979 compararon el efecto del warfarín vía oral contra dosis bajas de heparina en el tratamiento a largo plazo de la trombosis venosa profunda y en sus conclusiones para explicar sus resultados indican sesgo de sospecha diagnóstica aparentemente controlado al aplicar medidas para cegar a los médicos encargados de examinar los estudios, pero no hacen alusión a la variabilidad de observador aunque emplearon flebografías como método diagnóstico, las cuales se conoce son susceptibles de producir gran variabilidad al momento de interpretarlas.[16]

Un ejemplo más de la trascendencia de la determinación de la variabilidad de observador es el estudio realizado por McNaulty y colaboradores donde encontraron modificaciones a través del tiempo de la función ventricular al cateterizar en dos días consecutivos a 17 pacientes, situación que se puede explicar por la variabilidad de los sujetos examinados, pero no se puede descartar la variación de los observadores, evento que - ellos reconocen- no midieron, a pesar de que los procedimientos se efectuaron en dos laboratorios de hemodinamia diferentes.[17]

Otra área de discrepancia es el ventriculograma y la angiografía coronaria, donde la contractilidad miocárdica es evaluada por el cálculo de la fracción de eyección y una evaluación subjetiva de la función ventricular por el hemodinamista. En el procedimiento de evaluación de un cardiópata la lectura adecuada de un angiocardiograma juega un rol muy importante, por lo que para determinar su grado de variabilidad se evaluaron 22 médicos, quienes interpretaron 13 angiografías en forma independiente, analizando cada estudio en dos ocasiones con un intervalo de varias semanas entre la primera y la segunda evaluación. La concordancia intra-observador varió de 63 a 92 por ciento y la correlación inter-observador fue de 0.34 ("r" de Pearson). El grupo estuvo formado por cardiólogos y cirujanos cardiovasculares y aunque los primeros lo hicieron mejor no hubo diferencias significativas entre unos y otros, sin embargo queda claro que es recomendable para estudios que involucren interpretación de angiocardiogramas anotar la medición de la variabilidad inter y/o intra-observador según sea el caso.[18]

Finalmente, esta variación en la interpretación de angiocardiogramas no es exclusiva de esa área, hay variabilidad en la interpretación de radiografías, como ya se ha ejemplificado, de ECG, en la elaboración de historias clínicas, en estudios de histopatología (como lo menciona Feinstein con respecto al diagnóstico histopatológico del cáncer pulmonar), etc.[19-20]

A pesar de ello hay aceptación general de que la angiografía coronaria es un método objetivo y adecuado para el diagnóstico de la presencia y gravedad de lesiones coronarias. El grupo de Zir, del Hospital General de Massachusetts, analizó la variabilidad inter-observador exclusivamente orientado al porcentaje de detección de lesiones arteriales (presumiblemente el aspecto más objetivo de la angiografía coronaria), encontrando un elevado grado de variabilidad al utilizar el coeficiente de correlación "r" (0.44 a 0.85), agregando además que esta variación sorprendió por ser los 4 médicos evaluados integrantes del mismo grupo y participar en la interpretación conjunta para diagnóstico y envío o no a tratamiento quirúrgico de las lesiones coronarias.[21]

¿Qué ocurre en nuestro medio?

Con la finalidad de revisar la situación en México respecto a la evaluación de la variabilidad inter e intra-observador en cardiología, neumología y cirugía cardiotorácica, se revisaron los artículos publicados en cuatro revistas médicas nacionales, de enero de 1980 a julio de 1993; el objetivo fue detectar la aparición de artículos orientados a evaluar la variabilidad de observador y además analizar los artículos donde se utilizaron medidas a través de la exploración física u otros instrumentos, y en los que hubiera sido conveniente medir la variabilidad de los observadores involucrados.

Se descartaron las publicaciones donde se presentaron resultados de casos aislados, las orientadas a técnicas quirúrgicas o a presentar resultados de medidas terapéuticas médico-quirúrgicas, dejando solamente los artículos que presentaron análisis de la aplicación de estudios diagnósticos.

Los resultados se presentan en el Cuadro 1, y llama la atención que en la gran mayoría de los estudios no se tomó en cuenta la medición de la variabilidad de observador, además de que ninguno de los artículos menciona el número de observadores involucrados en la evaluación de sus casos. Sólo uno de los artículos relacionados a la ecocardiografía la tomó en cuenta, y en microscopía de luz uno de ellos determina el índice de certeza diagnóstica; en esa misma área, otro artículo indica que los procedimientos se efectuaron por triplicado, pero no hace consideraciones en cuanto a la variabilidad presentada en cada tercia de observaciones y cómo fue considerada en el análisis final de sus resultados.

Lo anterior no necesariamente es un reflejo de que la medición de la variabilidad de observador en nuestro medio

Cuadro 1. Número de estudios en México que midieron la variabilidad de observador de acuerdo a los instrumentos evaluados.

Prueba estudiada	Total de artículos	Artículos que evaluaron la variabilidad*
Ecocardiografía	23	1
Angiografía coronaria y ventriculograma	7	0
Electrocardiograma	3	0
Microscopía de luz	3	1
Microscopía electrónica	3	0
Total	39	2 (5%)

*Número total de artículos de cada área que tomaron en cuenta la variabilidad de observador en sus resultados

se pase por alto en el desarrollo de trabajos de investigación clínica, pues probablemente los servicios clínicos y quirúrgicos del área cardiológica y neumológica ya tengan sus estándares, sin embargo es importante presentarlos para cada proyecto de investigación, con lo cual se puede clarificar más la precisión de resultados obtenidos y no sólo atribuir la variabilidad a los equipos electromédicos utilizados, técnicas aplicadas o al paciente estudiado.

Alternativas para medir y disminuir la variabilidad de observador

Veamos a continuación qué opciones se han descrito para medir y en su caso disminuir la variabilidad de observador. El grupo de Bioestadística y Epidemiología de la Universidad Mc Master de Canadá ha propuesto las siguientes condiciones para disminuir la discrepancia intra e inter-observadores: un primer punto sería efectuar las evaluaciones en un ambiente adecuado y constante, la segunda sugerencia es desarrollar claves o procedimientos estandarizados en la evaluación, los cuales deben ser repetidos a lo largo de la misma y los hallazgos encontrados deben ser corroborados con testigos y con estudios paraclínicos adecuados, a lo que finalmente se agregaría la evaluación por un colega "cegado". La tercera aportación es también trascendente pues se refiere a la forma en que efectuamos nuestros registros; habitualmente el clínico anota sus inferencias y cada vez más raramente la evidencia encontrada, y la recomendación es hacer exactamente lo opuesto, auxiliándonos, por otro lado de las técnicas propuestas por las ciencias sociales cercanas a la medicina.[22]

Una vez cumplidas las condiciones anteriores debemos aceptar algún irreductible grado de diferencia entre una observación y otra o entre uno y otros observadores, aunque esta mínima diferencia siempre debe estar dentro de límites tolerables.[9] ¿Cuáles son esos límites tolerables? Como respuesta a esa pregunta lo primero que debemos aceptar es que ese límite debe ser medible y para tal efecto se han desarrollado diversos sistemas que van desde los coeficientes de correlación utilizados en otras condiciones, hasta índices diseñados exclusivamente para medir la concordancia de observador, tales como el SDAI (*Standard Desviation Agreement Index*) que es la desviación estándar del índice de concordancia o el EAI (*Experience Agreement Index*) índice de concordancia experiencia propuestos en 1966,[23] o el índice "kappa" propuesto por Cohen en 1960,[24] adaptado por Spitzer a partir de 1967[25] y revisado por él mismo en 1974.[26]

El índice "kappa" se desarrolló por necesidad en psiquiatría, donde muchas de sus mediciones son subjetivas y por tanto sujetas a mayor variabilidad entre uno y otro examinador. Este índice es la proporción de concordancia corregida y varía de -1 que indica total desacuerdo, a +1 que indica total concordancia, pasando por el cero (nula concordancia); sus características de muestreo son conocidas y por lo tanto pueden ser sometidas a pruebas de significancia estadística mostrando la concordancia real del observador no influida por el azar dividiéndola entre la concordancia probable no ocasionada por el azar.[24] Un agregado al índice "kappa", es la "kappa" ponderada, que además de las características anteriores nos permite evaluar variables con un nivel de medición ordinal.[25]

Por último podemos mencionar lo siguiente: conociendo la existencia real de la variabilidad intra e inter-observador es necesario aceptarla, aplicar las recomendaciones arriba anotadas para disminuirla y medirla en cada ocasión que se utilice un instrumento clínico de medición. Además, debemos desarrollar estándares de "normalidad" y rangos de variabilidad de observador en nuestra práctica cotidiana,[27] con lo que podemos aspirar a que nuestras observaciones serán más certeras. Una vez hecho lo anterior es muy recomendable obtener una concordancia intra-observador siempre mayor que la inter-observador,[28] aunque las dos deben ser siempre cercanas a +1 si se aplica el índice "kappa".

El control y la medición de la variabilidad de observador debe iniciarse desde las escuelas de medicina y los hospitales que tienen estudiantes en entrenamiento de pre y postgrado, donde se deben incluir entrevistas y exámenes de pacientes con supervisión directa y frecuente o enriquecer los ya existentes. Más aún, las sociedades, consejos de especialidad e instituciones docentes deben continuar estimulando la práctica de médicos que tengan la finalidad de medir y mejorar la reproducibilidad de sus hallazgos clínicos e interpretaciones.[29,30]

Referencias

1. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics. Norwalk Conn. San Mateo Ca: Appleton & Lange, 1990: 57.
2. Riegelman RK, Hirsch RP. Cómo estudiar un estudio y probar una prueba. Bol Of Sanit Panam 1991; 111(5): 449.
3. Birkelo CC, Chamberlain WE, Phelps PS y cols. Tuberculosis case findings -a comparison of effectiveness of various roentgenographic and photofluorographic methods. JAMA 1947; 133: 359.
4. Blackburn H. The electrocardiogram in cardiovascular epidemiology. Problems in standardized application. Ann N Y Acad Sc 1965; 126: 882.
5. Astrand I. Exercise electrocardiograms recorded twice with an 8-year interval in a group of 204 women and men 48-63 years old. Acta Med Scand 1965; 178: 27.
6. Shapiro A. The evaluation of clinical predictions. N Engl J Med 1977; 296:1509.
7. Anderson RE, Hill RB, Key Ch. The sensitivity and specificity of clinical diagnostics during five decades. JAMA 1989; 261: 1610.
8. Barkwin H. Pseudodoxia pediatrica. N Engl J Med 1945; 239: 691.
9. Spodick DH. On experts and expertise: the effect of variability in observer performance. Am J Cardiol 1975; 36: 592.
10. Day E, Maddern L, Wood C. Auscultation of foetal heart rate: an assessment of its error and significance. Br Med J 1968; 4: 422.
11. Departament of Clinical Epidemiology and Biostatistics, Mc Master University. Clinical disagreement: I how often occurs and why? Can Med Assoc J 1980; 123: 499.
12. Meade TW, Gardner MJ, Cannon P y cols. Observer variability in recording peripheral pulses. Br Heart J 1968; 30: 661.
13. Godfrey S, Edwards RHT, Campbell EJM y cols. Repeatability of physical signs in airway obstruction. Thorax 1969; 24: 4.
14. Gorman PA, Calatayud J, Abraham S y cols. Observer variation in interpretation of the electrocardiogram. Med Ann D C 1964; 33: 97.
15. Blackburn H, Blomqvist G, Freiman A y cols. The exercise electrocardiogram: differences in interpretation, report of a technical group on exercise electrocardiography. Am J Cardiol 1968; 21: 871.
16. Hull R, Delmore T, Gento E y cols. Warfarin sodium versus low-dose heparin in long-term treatment of venous thrombosis. N Engl J Med 1979; 301: 855.
17. Mc Naulty JH, Kremkau EL, Rosch J y cols. Spontaneous changes in left ventricular function between sequential studies. Am J Cardiol 1974; 34: 23.
18. Detre KM, Wright E, Murphy ML y cols. Observer agreement in evaluating coronary angiograms. Circulation 1975; 52: 979.
19. Haynes RB, Sackett DL, Tugwell P. Problems in the handling of clinical and research evidence by medical practitioners. Arch Intern Med 1983; 143: 1971.
20. Feinstein AR, Gelfman NA, Yesner R: Observer variability in the histopathologic diagnosis of lung cancer. Am Rev Respir Dis 1970; 101: 671.
21. Zir LM, Miller SW, Dinsmore RE, et al: Interobserver variability in coronary angiography. Circulation 1976; 53: 627.
22. Departament of Clinical Epidemiology and Biostatistics Mc Master University. Clinical disagreement: II How to avoid and to learn from our's mistakes. Can Med Assoc J 1980; 123: 613.
23. Armitage P, Blendis LM, Smyllie HC. The measurement of observer disagreement in the recording of signs. J Roy Statist Soc 1966; 129: 98.
24. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measmt 1960; 20: 37.
25. Spitzer RL, Cohen J, Fleiss JL y cols. Quantification of agreement in psychiatric diagnoses, a new approach. Arch Gen Psychiat 1967; 17: 83.
26. Spitzer RL, Fleiss JL. A re-analysis of the reliability of psychiatric diagnosis. Br J Psychiat 1974; 125: 341.
27. Castell DO, O'brien KD, Muench H y cols. Estimation of liver by percussion in normal individuals. Ann Intern Med 1969; 70: 1183.
28. Koran LM. The reliability of clinical methods, data and judgements. N Engl J Med 1975; 293: 695.
29. Seegal D, Wertheim AR. On the failure to supervise student's performance of complete medical examinations. JAMA 1962; 180: 476.
30. Weiner SL. Teaching of physical diagnosis. Ann Intern Med 1974; 80: 772.