Como citar este artículo: Simancas M, Arévalo L. Desempeño de cuatro métodos estadísticos para evaluación de la concordancia prueba-reprueba de variables continuas en una muestra. Revista Biosalud 2017; 16 (1): 19-29 DOI: 10.17151/biosa.2017.16.1.4

DESEMPEÑO DE CUATRO MÉTODOS ESTADÍSTICOS PARA EVALUACIÓN DE LA CONCORDANCIA PRUEBA-REPRUEBA DE VARIABLES CONTINUAS EN UNA MUESTRA

Miguel Simancas Pallares[1]
Luisa Arévalo Tovar[2]

RESUMEN

Objetivo: Comparar el desempeño de cuatro pruebas estadísticas para la evaluación de la confiabilidad prueba/re-prueba de variables continuas. **Métodos:** estudio de simulación estadística desarrollado dentro en el marco de un estudio de pruebas diagnósticas *in* vitro en 120 dientes que cumplieron con los criterios de selección. Para cada diente posicionado en un dispositivo de estandarización se tomaron dos radiografías digitales (T_0 y T_1) a las cuales se evaluó la longitud dental. Los datos se analizaron con estadística descriptiva y luego la comparación estadística a través de "t" de Student pareada, coeficiente de correlación intraclase, coeficiente de correlación de Pearson y coeficiente de correlación y concordancia de Lin en el paquete Stat v.13.2 para Windows (StataCorp., TX., USA). **Resultados:** La media de longitud dental para T0 fue 21,15 mm y para T1 21,07 mm. La prueba "t" de Student reveló una diferencia de medias de 0,089 (P=0,00). El coeficiente de correlación intraclase fue 0,877 (IC 95%: 0,43 – 0,98), coeficiente de correlación de Pearson 0,93 y el coeficiente de correlación y concordancia de Lin 0,93 (IC 95%: 0,908 – 0,956). **Conclusiones:** La selección de una prueba estadística para evaluación de concordancia prueba/re-prueba debe hacerse teniendo en cuenta los objetivos del estudio en cada contexto y la posibilidad de cada método estadístico de valorar la presencia de error en los datos. Así, un método que actualmente cumple con este requerimiento esencial es el coeficiente de correlación y concordancia de Lin por lo cual se recomienda su uso en futuros estudios.

Palabras clave: reproducibilidad de resultados, estadística y datos numéricos, error, pruebas de hipótesis.

PERFORMANCE OF FOUR STATISTICAL METHODS FOR THE ASSESSMENT OF TEST RE-TEST RELIABILITY OF CONTINUOUS VARIABLES IN A SAMPLE

RESUMEN

Objective: To compare the performance of four statistical tests in continuous variables test/re-test reliability assessment. Methods: Statistical simulation study developed in the framework of an in vitro diagnostic test study including 120 teeth which met the inclusion criteria. Each tooth was positioned in a standardization device and was taken two digital x-rays (T0 and T1) in which we assessed tooth-length. Data were analyzed with descriptive statistics and

[1] M.Sc. Epidemiología Clínica. Departamento de Investigación. Facultad de Odontología. Universidad de Cartagena. E-mail: msimancasp@unicartagena.edu.co
[2] Esp. Periodoncia. Departamento de Medicina Oral y Cirugía. Facultad de Odontología. Universidad de Cartagena. E-mail: larevalot@unicartagena.edu.co

ISSN 1657-9550 (Impreso) ISSN 2462-960X (En línea)
DOI: 10.17151/biosa.2017.16.1.4
Recibido: abril 14 de 2016 - Aceptado: octubre 24 de 2016
Biosalud, Volumen 16 No. 1, enero - junio, 2017. págs. 19 - 29

then a statistical comparison was done with paired Student's "t" test, intraclass correlation coefficient, Pearson correlation coefficient and Lin's concordance correlation coefficient in Stata v.13.2 for Windows (StataCorp., TX., USA). Results: The average dental length for T0 was 21.15 mm and for T1 21.07 mm. Student's "t" test revealed an average difference of 0.089 (P=0.00). The intraclass correlation coefficient 0.877 (95% CI: 0.43 – 0.98), Pearson's product-moment correlation coefficient 0.93, and Lin's concordance correlation coefficient 0.93 (95% CI: 0.908 – 0.956). Conclusions: Selection of a statistical test for test/re-test reliability assessment should be made having in mind the research objectives in any context and the possibility of each method for error assessment. Thus, a method that currently complies with this essential requirement is Lin's concordance correlation coefficient, which is recommended for future test re-test research studies.

Key words: reproducibility of results, statistics, and numerical data, error, hypothesis-testing.

INTRODUCCIÓN

Los servicios de salud constantemente introducen al mercado novedosos dispositivos para la evaluación de una variable de interés (1, 2). En odontología por ejemplo, las sondas periodontales electrónicas (3), la radiovisiografía (4), la tomografía axial computarizada (TAC) (5) y los métodos digitales para la realización de cefalometrías (6), suponen mejoras en el desempeño de los procesos operativos. No obstante, en la mayoría de las ocasiones los resultados de las investigaciones sobre estos auxiliares diagnósticos no incluyen evaluaciones de su precisión y exactitud que permitan hacer inferencias sobre el uso generalizado de las mismas, lo que se traduce en poca validez externa y paradójicamente, limita su aplicación en otros contextos clínicos.

En este sentido, los instrumentos que se utilicen para cuantificar el nivel de daño o certeza diagnóstica deben basarse en estudios de pruebas diagnósticas (7). Estos estudios permiten hacer inferencias sobre el grado de acuerdo -concordancia-, validez o incluso características receptivas al operador de cualquier instrumento. La concordancia hace referencia al grado en que dos o más observadores, métodos, técnicas u observaciones están de acuerdo sobre el mismo fenómeno observado (8). Recobra importancia cuando se desea conocer si con un método o instrumento nuevo, diferente al habitual, se obtienen resultados equivalentes de tal manera que eventualmente uno y otro puedan ser intercambiados (7). Está enmarcada en los análisis derivados de los estudios de pruebas diagnósticas que a saber, puede evaluar dos aspectos: **consistencia** (cuando uno de los métodos evaluados incluye el *gold-standard* o patrón de oro) o **conformidad** (cuando ninguno de los dos métodos o técnicas evaluadas se asume como patrón de oro) (9).

Respecto de la validez de una prueba diagnóstica, esta debe incluir adecuadas representaciones de **exactitud** que corresponde al grado en que una medición refleja la realidad de un fenómeno, capacidad de medición o clasificación de un método o instrumento para aquello que originalmente fue propuesto, y **precisión** que corresponde al grado en que los puntajes de una medición se encuentran libres de error de medida en diferentes momentos del tiempo (10). Teniendo en cuenta esto, los abordajes estadísticos para evaluación de la concordancia deben incluir en sus fórmulas, medidas de precisión y exactitud para hacer que una prueba diagnóstica pueda ser considerada -en su contexto- válida (11).

Un aspecto importante en la evaluación de las pruebas diagnósticas es que el instrumento debe demostrar que las mediciones realizadas en dos (o más) momentos del tiempo son estables, es decir, que tenga la misma capacidad de medición a lo largo del tiempo (12). A esto se le conoce como estabilidad prueba/re-prueba o concordancia prueba/re-prueba (CPRP).

Para tratar de sobrevenir todas estas dificultades, diversos métodos estadísticos se han propuesto y empleado en la literatura científica para evaluación de la concordancia prueba/re-prueba de una variable continua, entre estos: la prueba "t" de Student pareada, el coeficiente de correlación intraclase (CCI), el coeficiente de correlación de Pearson y el coeficiente de correlación y concordancia de Lin. También, Bland & Altman han propuesto un método para evaluación de límites de acuerdo entre dos sistemas de medición (8, 13). La prueba **"t" de Student pareada** se emplea para comparar promedios de una variable cuantitativa antes y después o en dos grupos distintos (14, 15). El **CCI**, una formulación especial del coeficiente de Pearson, es uno de los métodos más empleados para evaluación de concordancia de una variable continua y se define como la evaluación de la proporción de variabilidad total que es debida a la variabilidad de los sujetos, basada en un modelo de análisis de varianza (ANOVA) con medidas repetidas (16). Por su parte, el coeficiente de correlación de **Pearson**, mide la probabilidad de establecer una ecuación lineal entre dos variables, en la que por cada cambio de unidad en una de ellas, se espera un cambio de unidad (correlativo) en la otra (17). Finalmente, el **coeficiente de correlación y concordancia de Lin**, evalúa qué tan lejos se desvían los datos observados por dos métodos u observadores de una línea a partir del origen y a 45° en un plano cartesiano, que corresponde a la línea de perfecta concordancia. Este estadístico combina una medida de precisión, representada por el coeficiente de correlación de Pearson y una medida de exactitud, representada por el coeficiente de corrección del sesgo (7, 13).

Tal y como se evidenció anteriormente, estos métodos suponen indicaciones precisas de acuerdo con su definición estadística y no todos pueden ser empleados en las mismas condiciones o situaciones de estudio (8). Así, dada la amplia variabilidad de abordajes estadísticos disponibles, el uso indiscriminado de estas pruebas estadísticas en la literatura y la falta de unidad en los criterios para su uso en la evaluación de la concordancia (18-20), el objetivo del presente estudio fue comparar el empleo de cuatro pruebas estadísticas (t de Student pareada, coeficiente de correlación de Pearson, coeficiente de correlación intraclase y coeficiente de correlación y concordancia de Lin) para la evaluación de la CPRP de variables continuas.

MATERIALES Y MÉTODOS

Tipo de estudio: este estudio originalmente pretendió determinar la concordancia prueba/re-prueba *in vitro* de un dispositivo de estandarización radiográfica. Sin embargo, *ad hoc* con los datos obtenidos de ese estudio, se desarrolló la presente simulación, como un ejercicio de ejemplificación de los resultados de cuatro técnicas en una misma situación de cumplimiento de supuestos (una muestra).

Población: molares maxilares o mandibulares.

Muestra y Muestreo: 120 molares maxilares o mandibulares que cumplieron con los criterios de selección. El tamaño de la muestra se determinó teniendo en cuenta una correlación esperada de 0,90, valor en el cambio de las medias de 0,20 mm, coeficiente de correlación y concordancia esperado de 0,90 de acuerdo con los criterios propuestos por McBride para la evaluación de concordancia de variables continuas (21). El valor en el cambio de las medias se realizó de acuerdo con la significancia de esta magnitud en el contexto clínico. El cálculo del tamaño se realizó en el paquete GenStat v.12 [VSN International., UK] para Windows.

Criterios de selección
Criterios de inclusión: molares superiores e inferiores, derechos e izquierdos con formación radicular completa y ápice cerrado.
Criterios de exclusión: dientes con caries coronal o radicular, dientes con destrucción coronal por trauma dento-alveolar.

Variables
Variable de medida: longitud dental. Esta variable se clasifica como una variable cuantitativa, continua de razón. Se definió como la distancia en milímetros (mm) desde el punto más apical radicular visible radiográficamente hasta el punto coronal más visible radiográficamente (22).

Protocolo del estudio: inicialmente se procedió a realizar un riguroso proceso de limpieza de los órganos dentarios objeto de estudio. Estos dientes fueron sumergidos en una solución de hipoclorito de sodio al 5,25% durante 24 horas y fueron esterilizados en autoclave modelo EA-600[a] (Tuttnauer USA Co., LTD., NY., USA) con presión 1.2Kg/m^2 durante 45 minutos hasta lograr su completa esterilización. Una vez estériles los órganos dentarios, se procedió a diseñar con cera tropical una plantilla de estabilización del órgano dentario en un 30% de profundidad en sentido meso-distal, lo cual logró fijar la posición del espécimen. Una vez lograda la estabilización, se procedió a ubicar el espécimen en un dispositivo de estandarización radiográfica.

Una vez definida y registrada la posición del espécimen, se procedió a la toma de la radiografía inicial (T_0) empleando el radiovisiógrafo Dr. Suni Plus (Suni Medical Imaging., San Jose del Oro., CA., USA) con equipo de rayos X de pared (RAIOS X TIMEX 70C PAREDE GELO 127V +4%., Rod Abrao Assed. Km53 +450m – Ribeirao Preto – Sao Paulo – Brasil) pre-especificado a 7 miliamperios y 70 kilovoltios. Todas las radiografías fueron tomadas con una duración de 0,20 segundos de exposición por el mismo operador.

Las imágenes obtenidas fueron almacenadas en el software del radiovisiógrafo empleando un sistema de codificación alfanumérico compuesto por dos consonantes, las cuales identificaban el proyecto de investigación, tres dígitos en numeración arábiga y dos dígitos para indicar el tiempo de toma de la imagen. Así, la radiografía AB00101 indicaba el espécimen que pertenecía al proyecto de investigación AB, primer espécimen y primer tiempo de toma. Este sistema de codificación no fue de conocimiento público, lo cual permitió una identificación secuencial y objetiva de las mediciones. Pasados 15 días, se obtuvo la segunda radiografía (T_1) la cual se archivó en la misma carpeta electrónica de la primera toma pero identificada con el código final 02. La segunda radiografía se obtuvo empleando los mismos parámetros de exposición y posición (empleando el dispositivo) que la primera toma, a fin de lograr imágenes completamente reproducibles.

Un odontólogo entrenado, calibrado (CCI: 0,87: IC 95%: 0,437 – 0,981) y con experiencia >5 años en el sistema de medición radiográfica, realizó las mediciones de longitud dental de cada una de las unidades muestrales empleando la herramienta de medición del software y sin aplicar filtros de brillo, contraste o magnificación. La longitud dental (en mm), se registró en un formato para recolección de la información, diseñado para tal fin por el equipo de investigadores. Obtenidas las segundas imágenes, el mismo odontólogo realizó las mediciones y registró en un segundo formato que no contenía la primera medición con la finalidad de evitar sesgos de medición. Adicionalmente y con la intención de evitar sesgos, un auxiliar de investigación presentó de forma aleatoria (con respecto a la primera serie de radiografías) las imágenes de forma cegada en un ordenador portátil DELL Latitude E6510 (Dell INC, USA) de monitor 15.6".

Recolección y procesamiento de la información: obtenidas las radiografías en T_0 y T_1, la información de los registros de recolección se digitó en una tabla matriz diseñada en Microsoft

Excel (Redmond, WA, USA). Esta digitación contó con verificación periódica a fin de minimizar errores. Asimismo, la tabla matriz contó con copias de seguridad periódicas para evitar la pérdida de información.

Análisis estadístico: digitada y depurada la información, se procedió a realizar el análisis estadístico. Inicialmente se realizó análisis de normalidad de la distribución de los datos empleando la prueba Shapiro-Wilks. Posteriormente, se procedió con análisis descriptivo empleando medidas de tendencia central y dispersión. Dado que los datos siguieron una distribución normal, se reportó media y desviación estándar para cada variable. Asimismo, al evaluar la homocedasticidad con el test de Levene se encontró que los grupos provenían de una distribución con varianzas iguales. Se aplicaron entonces cuatro pruebas estadísticas: "t" de Student pareada, coeficiente de correlación intraclase (CCI), coeficiente de correlación de Pearson y coeficiente de correlación y concordancia de Lin. Todos los análisis se realizaron empleando el paquete Stata v.13.2 para Windows [StataCorp., College Station., TX., USA].

Consideraciones éticas: el presente estudio se clasificó de acuerdo con la resolución 008430 de 1993 como un estudio sin riesgo debido a que no se realizaron intervenciones sobre los pacientes. Los órganos dentarios provenían de un banco de dientes derivado de otras investigaciones sobre humanos con aprobación del Comité de Ética en Investigaciones, y para los cuales se obtuvo consentimiento informado escrito, detallando que esos órganos pudieran ser utilizados en procesos de investigación dentro de la Universidad de Cartagena y garantizando la confidencialidad de la información suministrada inicialmente. La presente investigación se acogió también a lo expresado en la declaración de Helsinki, modificada en Edimburgo 2008.

RESULTADOS

Estadística descriptiva

La estadística descriptiva se muestra en la Tabla 1.

Tabla 1. Estadística descriptiva para cada uno de los momentos de medición radiográfica. Desv. Est: desviación estándar.

	Momento	
Estadístico	**$T_{0\ (mm)}$**	**$T_{1\ (mm)}$**
Media	21,15	21,07
Desv. Est.	2,10	2,03
Mediana	21,22	21,14
Máximo	26,28	26,55
Mínimo	16,22	16,58

"t" de Student pareada

Los estadísticos derivados de esta prueba de hipótesis se muestran en la Tabla 2.

Tabla 2. Estadísticos derivados del análisis "t" de Student. Desv. Est: desviación estándar.

	n	Media	Desv. Est	IC 95%	P-valor
Momento 1	117	21,15	2,10	20,77 - 21,54	0,20
Momento 2	117	21,07	2,03	20.69 - 21,44	
Diferencia		0,088	0,7583	-0,05 - 0,22	

Coeficiente de correlación y concordancia de Lin

Se obtuvo estimador de 0,93 (IC 95%: 0,908 – 0,956), coeficiente de correlación de Pearson 0,93, pendiente 1,03, intercepto -0,54, factor de corrección del sesgo 0,99. Los límites de acuerdo (gráfico de Bland & Altman) se muestran en la figura 1.

Coeficiente de correlación de Pearson (r)
Se obtuvo estimador 0.933.

Coeficiente de correlación intraclase (CCI)
Se obtuvo estimador 0,877 (IC 95 %: 0,437 – 0,981).

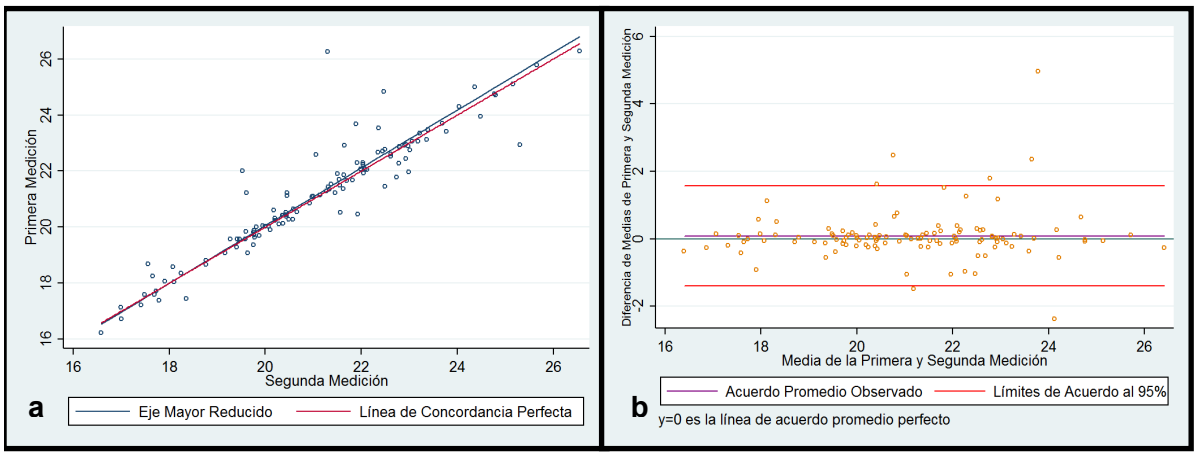

Figura 1. Gráfico de Bland & Altman que muestra los límites de acuerdo.

DISCUSIÓN

Diversas investigaciones han empleado un sinnúmero de pruebas estadísticas en investigación biomédica para la evaluación de problemas de medición, sin llegar a un consenso universal sobre sus indicaciones (8). La mayoría de estos métodos incluyen pruebas de hipótesis (pruebas "t" de Student pareadas, análisis de varianzas-ANOVA) o coeficientes de correlación (Pearson y coeficiente de correlación intraclase) (18, 22). Otros métodos incluyen análisis de regresión y el coeficiente de variación (23, 24). Es de notar que entre los métodos menos considerados está la técnica propuesta por Bland & Altman (límites de acuerdo) y la descrita por Lin (coeficiente de correlación y concordancia de Lin)(25).

El análisis de dos pruebas diagnósticas o una prueba en diferentes momentos del tiempo, depende en gran medida del nivel de error presente en los datos. En términos generales, puede ser de dos tipos: aleatorio y sistemático (26). El error aleatorio se refiere a características o situaciones (fuentes de error) que son debidas a factores "no controlables" tales como la suerte, la atención prestada por el observador y la variabilidad biológica normal y que de una u otra forma, afectan un puntaje particular. Tales fuentes de error, de una forma aleatoria, aumentan o disminuyen los puntajes de la prueba en diferentes momentos del tiempo. Por su parte, el error sistemático o sesgo, se refiere a factores (fuentes) constantes que afectan todos los puntajes de la misma forma (sistemáticamente) (27).

La prueba "t" de Student ha sido empleada frecuentemente para comparar promedios de una prueba (test) o incluso en estudios prueba re-prueba para saber si existe algún sesgo estadísticamente significativo entre las pruebas (22, 28). Aunque es útil, no debe ser empleada

como tal para la evaluación de la concordancia dado que el estadístico t no provee información sobre el nivel de precisión y exactitud entre las evaluaciones de las pruebas (o momentos de la prueba). Además, la detección de diferencias estadísticamente significativas depende, en gran medida, de la magnitud de error aleatorio entre las pruebas (29). Específicamente, debido a la naturaleza de la fórmula matemática empleada para el cálculo del estadístico de prueba, los errores sistemáticos son menos detectables si existe alto error aleatorio entre estas pruebas (o momentos de la prueba).

En consonancia con lo anterior, al haber empleado esta prueba de hipótesis para comparar la estabilidad del dispositivo antes y después y no obtener diferencias estadísticamente significativas entre los promedios (p=0,20), esto puede interpretarse en que el promedio de longitud dental antes y después es el mismo, y que por tanto, no existe variación media entre las mediciones. Así, el dispositivo produce (antes y después) una imagen geométricamente estable sin diferencias estadísticamente significativas, y que por tanto puede sugerirse para estandarización de la posición de las unidades muestrales. No obstante, debe hacerse especial énfasis en que esta prueba no es adecuada para tal fin puesto que como se ha mencionado anteriormente, no es capaz de detectar la presencia de errores sistemáticos entre los dos momentos; requerimiento vital de los métodos estadísticos cuando se evalúa concordancia (8).

Otro de los métodos estadísticos probados en el presente estudio incluye el coeficiente de correlación de Pearson que ha sido la técnica más empleada para la evaluación de la concordancia. Usualmente obtener un estadístico >0,80 con p-valor <0,05 indica que existe concordancia entre los dos métodos o un mismo método es estable en el tiempo (8). No obstante, emplear este método para estos fines se considera estadísticamente incorrecto puesto que no puede valorar la presencia de error sistemático como se mencionó anteriormente para la prueba "t" de Student (29, 30). Además, el rango de valores observado en el análisis incrementa el coeficiente si ésta incluye valores extremos, sobreestimando entonces la correlación obtenida entre las variables (31). Así, debe hacerse énfasis en que este coeficiente mide la intensidad de asociación lineal entre dos mediciones (correlación) pero no proporciona información sobre el acuerdo observado (concordancia)(7). De acuerdo al valor de este coeficiente obtenido en la presente investigación (ρ=0,93), el lector puede concluir que el método es estable en el tiempo. No obstante, al presentar las limitaciones teóricas anteriormente mencionadas, su uso se ve restringido en estadística.

De otro lado, el hecho de haber obtenido un CCI=0,83 en la presente investigación, puede interpretarse también que el método es estable en el tiempo y que puede ofrecer imágenes geométricamente estables ya que un estimador >0,75 puede calificarse como excelente. Aunque este estadístico ha sido ampliamente usado (32), tiene la gran ventaja sobre ρ de ser univariado en vez de bivariado y que puede ser empleado cuando más de un re-test (más de un momento, T_α) se compara con un test (momento T_0)(33); presenta al igual que los métodos anteriores, una serie de limitaciones que restringen su aplicabilidad en el estudio de la concordancia de variables continuas: primero, no estima o discrimina la variabilidad entre los métodos de medición, y segundo, tiene varios supuestos difíciles de cumplir: a. los métodos provienen de una muestra al azar de una población de métodos (o en este caso momentos del tiempo), b. el error de medición es similar para cada uno de los métodos, y c. al igual que el coeficiente de Pearson, depende de la magnitud de los valores de la muestra de estudio (34). Adicionalmente, tiene la dificultad de la interpretación del estimador y su traducción a la relevancia desde el punto de vista clínico, tal como sucede con el coeficiente de Kappa y que los datos necesitan tener una distribución simétrica (normal)(7) lo cual en el presente no

representó un obstáculo, pero en la mayoría de los estudios es difícil encontrar datos que sigan una distribución normal. Respecto de su reporte en investigaciones, algunos autores sugieren informar el ICC acompañado de otros estimadores, basados en el cumplimiento de los objetivos trazados en la investigación (8).

En consonancia con lo anterior, existe la necesidad de un abordaje estadístico que combine medidas de precisión y exactitud (identificación de la magnitud del error sistemático presente en las mediciones). Así, un abordaje ideal es el coeficiente de correlación y concordancia de Lin (ρ_c). Este coeficiente se definió reescalando la desviación cuadrática media entre los métodos de medida de forma tal que puede adoptar valores entre -1 (perfecta discordancia) y +1 (concordancia perfecta)(13). Aumenta de valor en función de: a)la cercanía del eje principal o la pendiente de la curva de regresión de las parejas de datos obtenidos en la línea de perfecta concordancia (coeficiente de corrección de sesgo: Cb) que permite evaluar la exactitud de los datos obtenidos que para el caso del presente estudio fue 0,99; y b)en función de la dispersión alrededor de la línea de mejor ajuste o línea de regresión de las parejas de los datos obtenidos, siendo éste el reflejo de la precisión de las mediciones obtenidas y corresponde al coeficiente de correlación de Pearson (0,93)(35). Así, con estos estimadores, los resultados del presente estudio con ese abordaje estadístico la CPRP pueden evaluarse como moderada (21).

Altman & Bland reconocieron diversas limitaciones de los tres anteriores métodos presentados ("t" de Student pareada, CCI y Pearson) e incluso métodos como el error estándar del método y el coeficiente de variación y propusieron un método complementario al coeficiente de correlación y concordancia de Lin para evaluación de la concordancia a través de una crítica evaluación de la presencia (y magnitud) de errores aleatorios o sistemáticos en la muestra estudiada (31). La principal diferencia entre estos estadísticos es que generaliza hallazgos individuales a nivel poblacional de las diferencias entre los diferentes momentos del tiempo (prueba/re-prueba)(8). En este sentido, uno de los primeros pasos en el análisis de límites de acuerdo (método de Bland & Altman) es presentar y explorar los datos de prueba/re-prueba con un gráfico de Bland & Altman, el cual grafica las diferencias individuales entre los métodos o momentos versus las respectivas medias individuales (36). El siguiente paso es entonces la interpretación de estos límites de acuerdo; estos deben ser interpretados que para un nuevo individuo de la población estudiada, se esperaría (con una probabilidad aproximada de 95%) que la diferencia entre alguno de los dos momentos debería estar entre los límites de acuerdo. Así, en el contexto del presente estudio, para una nueva unidad muestral con un 95% de probabilidad la diferencia entre alguno de los dos momentos de medición está entre -1,398 y 1,575mm±0,0174.

Una de las limitaciones del análisis de la concordancia a través del coeficiente de correlación y concordancia de Lin y el método de Bland & Altman para el análisis de los límites de acuerdo es que los datos deben seguir una distribución normal (13); sin embargo, existe evidencia que soporta la utilidad de este estadístico aun en datos que no siguen una distribución normal (37). Otra limitación es el tamaño de muestra; diversos estudios sugieren n>40 para este tipo de estudios (38).

En términos generales, el método de Bland & Altman permite conocer si las diferencias entre los dos métodos son sistemáticas o al azar. Se espera que la diferencia promedio entre los dos métodos sea cero y que el 95% de las diferencias se encuentren dentro de 1,96 de las desviaciones estándar de dicho promedio (39). Un paso muy importante para determinar la significancia clínica y estadística de estos estimadores es la correlación existente entre la diferencia y la media (0,082; P-valor: 0,31) y la inspección visual de los patrones de orientación de los datos en el gráfico. Siempre que esta

correlación (en caso que sea significativamente >0), indica heterocedasticidad de los datos y que por tanto antes de informar los límites de acuerdo, debería realizarse una transformación de los datos (8); además, gráficamente en caso que exista tal heterocedasticidad existe un patrón claro de dispersión de la nube de puntos. Así, para el presente estudio no se encontró heterocedasticidad (no existe evidencia estadística y tampoco gráfica), por tanto no existe necesidad de transformar los datos y los límites de acuerdo pueden ser informados de forma natural y como se describieron previamente. Con los anteriores hallazgos puede confirmarse, la ausencia de errores sistemáticos entre las dos mediciones lo cual permite hacer inferencias sobre la precisión y exactitud al evaluar la CPRP.

CONCLUSIONES

Poniendo de manifiesto toda la evidencia científica, la escogencia de los métodos estadísticos para evaluar la CPRP debe hacerse evaluando el contexto de cada estudio particular y los objetivos de la investigación. Teniendo en cuenta las limitaciones del presente estudio y su marcada aplicabilidad a una muestra, el CCC de Lin se propone como el método ideal para este tipo de situaciones ya que permite una efectiva evaluación de la presencia de error sistemático y del aleatorio y por tanto permite llegar a conclusiones sobre la precisión y exactitud de los métodos o mediciones en distintos momentos del tiempo y así, sobre la capacidad de generalización de resultados.

RECOMENDACIONES

Comparar el rendimiento de las pruebas por medio de la simulación de datos, por ejemplo por el procedimiento Montecarlo; manipulando condiciones críticas para el desempeño de las pruebas como tamaño de la muestra, tamaño del efecto y cumplimiento de supuestos. Adicionalmente, calcular error tipo II o la potencia de los contrastes y comparar la tendencia central de la variable dependiente a través de procedimientos como MANOVA.

REFERENCIAS

1. Bahrololoomi Z, Ezoddini F, Halvani N. Comparison of Radiography, Laser Fluorescence and Visual Examination for Diagnosing Incipient Occlusal Caries of Permanent First Molars. Journal of Dentistry. 2015; 12(5):324-32.
2. Gomez J. Detection and diagnosis of the early caries lesion. BMC oral health. 2015; 15 Suppl 1:S3.
3. Niederman R. Manual and electronic probes have similar reliability in the measurement of untreated periodontitis. Evidence-based dentistry. 2009; 10(2):39.
4. Kumar LV, Sreelakshmi N, Reddy ER, Manjula M, Rani ST, Rajesh A. Clinical Evaluation of Conventional Radiography, Radiovisiography, and an Electronic Apex Locator in Determining the Working Length in Primary Teeth. Pediatric Dentistry. 2016; 38(1):37-41.
5. Leonardi Dutra K, Haas L, Porporatti AL, Flores-Mir C, Nascimento Santos J, Mezzomo LA, et al. Diagnostic Accuracy of Cone-beam Computed Tomography and Conventional Radiography on Apical Periodontitis: A Systematic Review and Meta-analysis. Journal of Endodontics. 2016;42(3):356-64.
6. Tadinada A, Mahdian M, Sheth S, Chandhoke TK, Gopalakrishna A, Potluri A, et al. The reliability of tablet computers in depicting maxillofacial radiographic landmarks. Imaging science in dentistry. 2015; 45(3):175-80.
7. Cortés-Reyes E, Rubio-Romero JA, Gaitán-Duarte H. Statistical methods for evaluating diagnostic test agreement and reproducibility. Revista Colombiana de Obstetricia y Ginecología. 2010; 61(3):247-55.
8. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Medicine. 1998; 26(4):217-38.
9. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clinical Pharmacology & Therapeutics. 1981; 29(1):111-23.
10. Aravena PC, Moraga J, Cartes-Velásquez R, Manterola R. Validity and Reliability in Dental Research. Int J Odontostomat. 2014; 8(1):69-75.
11. Alarcón A, Muñoz S. Medición en salud: algunas consideraciones metodológicas. Rev Med Chile. 2008;1 36(1):125-30.
12. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley and Sons; 1986.
13. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45(1):255-68.
14. Gómez-Gómez M, Danglot-Banck C, Vega-Franco L. Choosing a statistical test. Second part. Revista Mexicana de Pediatría. 2013; 80(2):81-5.
15. Student. The probable error of a mean. Biometrika. 1908; 6(1):1-25.
16. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin. 1979; 86(2):420-8.
17. Pita-Fernández S, Pértega-Díaz S. Relación entre variables cuantitativas. Cad Aten Primaria. 1997; 4:141-4.
18. Chuan A, Thillainathan S, Graham P, Jolly B, Wong D, Smith N, et al. Reliability of numerical scales used for direct observation of procedural skills. Anaesthesia and intensive care. 2016; 44(2):201-8.
19. Flores-Mir C, Rosenblatt MR, Major PW, Carey JP, Heo G. Measurement accuracy and reliability of tooth length on conventional and CBCT reconstructed panoramic radiographs. Dental press journal of orthodontics. 2014; 19(5):45-53.
20. Oznurhan F, Tuzuner T, Baygin O, Unal M, Kapdan A, Ozturk C. Accuracy of three different apex locators and visual exam in primary teeth with and without root resorption in vitro. European journal of paediatric dentistry: Official Journal of European Academy of Paediatric Dentistry. 2014;15(4):381-4.

21. McBride GB. A proposal for strenght-of-agreement criteria for Lin's concordance correlation coefficient. National Institution of Water & Atmospheric Research Ltd, 2005 HAM2005-062.

22. Oliveira ML, Vieira ML, Cruz AD, Boscolo FN, De Almeida SM. Gray scale inversion in digital image for measurement of tooth length. Brazilian Dental Journal. 2012; 23(6):703-6.

23. Scaf G, Morihisa O, Loffredo L de C. Comparison between inverted and unprocessed digitized radiographic imaging in periodontal bone loss measurements. Journal of applied oral science: revista FOB. 2007; 15(6):492-4.

24. Feltz CJ, Miller GE. An asymptotic test for the equality of coefficients of variation from k populations. Statistics in Medicine. 1996; 15(6):646-58.

25. Carrasco J, Jover L. Métodos estadísticos para evaluar la concordancia. Medicina Clínica (Barcelona). 2004; 122(Supl 1):28-34.

26. Olds T. Five errors about error. Journal of Science and Medicine in Sport / Sports Medicine Australia. 2002; 5(4):336-40.

27. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. Journal of Strength and Conditioning Research / National Strength & Conditioning Association. 2005; 19(1):231-40.

28. Bodur H, Odabas M, Tulunoglu O, Tinaz AC. Accuracy of two different apex locators in primary teeth with and without root resorption. Clinical Oral Investigations. 2008; 12(2):137-41.

29. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. International Journal of Epidemiology. 1995; 24 Suppl 1:S7-14.

30. Bates BT, Zhang S, Dufek JS, Chen FC. The effects of sample size and variability on the correlation coefficient. Medicine and Science in Sports and Exercise. 1996; 28(3):386-91.

31. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476):307-10.

32. Sánchez-Ayala A, Farias-Neto A, Vilanova LS, Costa MA, Paiva AC, Carreiro AD, et al. Reproducibility, Reliability, and Validity of Fuchsin-Based Beads for the Evaluation of Masticatory Performance. Journal of Prosthodontics: official journal of the American College of Prosthodontists. 2015.

33. Baumgartner TA. Norm-referenced measurement: reliability. Measurement concepts in physical education and exercise science. Champaing, IL: Human Kinetics; 1989. p. 45-72.

34. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Computers in Biology and Medicine. 1990; 20(5):337-40.

35. Cepeda MS, Africano JM, Polo R, Alcala R, Carr DB. Agreement between percentage pain reductions calculated from numeric rating scores of pain intensity and those reported by patients with acute or cancer pain. Pain. 2003; 106(3):439-42.

36. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995; 346(8982):1085-7.

37. Carrasco JL, Jover L, King TS, Chinchilli VM. Comparison of concordance correlation coefficient estimating approaches with skewed data. Journal of Biopharmaceutical Statistics. 2007; 17(4):673-84.

38. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

39. Bishop D. Reliability of a 1-h endurance performance test in trained female cyclists. Medicine and science in sports and exercise. 1997; 29(4):554-9.