Vergleich von Messmethoden

– Artikel Nr. 24 der Statistik-Serie in der DMW –

Comparing methods of measurement

Autoren U. Grouven [1] R. Bender [1] A. Ziegler [2] S. Lange [1]

Institut
[1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln
[2] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck

Übereinstimmung von Bewertungen

In Artikel Nr. 23 [1] der DMW-Statistik-Serie haben wir die Übereinstimmung zwischen Bewertern bei kategoriellen Variablen behandelt. In diesem Fall ist der Kappa-Koeffizient nach Cohen das Verfahren der Wahl [1]. Die vorliegende Arbeit befasst sich mit der Bewertung der Übereinstimmung bei Daten mit kontinuierlichem Messniveau.

Klinische Messwerte weisen immer eine mehr oder weniger große Ungenauigkeit auf. Der Messfehler resultiert dabei zum Einen aus der Ungenauigkeit (dem „Auflösungsvermögen“) des Messverfahrens selbst, die ihrerseits wiederum verschiedene Komponenten haben kann, und zum Anderen aus intraindividuellen Schwankungen der Ausprägung der beim Patienten zu messenden Variable. In solchen Fällen werden häufig Messungen wiederholt oder unterschiedliche Messverfahren eingesetzt. Dabei sind zwei Aspekte von Bedeutung: 1) Wie gut ist die Übereinstimmung von wiederholten Messwerten einer bestimmten Messmethode („repeatability“, „reliability“), d. h. die Bestimmung des Messfehlers der Methode ist von Interesse, und 2) wie gut ist die Übereinstimmung unterschiedlicher Messmethoden untereinander („agreement“). Beispielsweise möchte man einen teuren oder zeitaufwändigen Labortest durch ein günstigeres bzw. schnelleres Verfahren ersetzen. Voraussetzung hierfür ist jedoch eine hinreichende Übereinstimmung der erzielten Messergebnisse.

In der Praxis werden derartige Daten häufig mit nicht adäquaten Analyseverfahren ausgewertet. Anhand des folgenden Beispiels diskutieren wir zunächst die Mängel und Nachteile dieser Verfahren und beschreiben dann geeignete und effiziente Verfahren zur Auswertung von kontinuierlichen Daten beim Vergleich von Messmethoden.

Ein einführendes Beispiel

Tab. 1 zeigt manuelle (SBD1) sowie maschinelle Messwerte (SBD2) des systolischen Blutdrucks an 30 Personen (Beispiel 1, Teildatensatz aus [2]). Von Interesse ist es, wie gut die beiden Messmethoden übereinstimmen („agreement“).

Ein erster sinnvoller Schritt besteht darin, die Daten grafisch aufzutragen (**Abb. 1**a). Es zeigt sich ein positiver linearer Zusammenhang der Messwerte.

Korrelation

Ein häufig beobachtetes Vorgehen ist die Berechnung des Korrelationskoeffizienten als Maß für den Grad der Übereinstimmung zwischen den Messungen. Die Korrelation hat im vorliegenden Beispiel einen hohen Wert von $r = 0{,}90$ mit einem zugehörigen p-Wert von $p < 0{,}0001$. Eine hohe Korrelation ist jedoch nicht gleichbedeutend mit einer hohen Übereinstimmung. Dies lässt sich folgendermaßen veranschaulichen: Erhöht man die Werte der manuellen Messung um 20% (SBD1*) und verringert gleichzeitig die Werte der maschinellen Messung um 20% (SBD2*) (**Tab. 1**), so erhöht sich offensichtlich der Unterschied zwischen den Messwerten. Das heißt, der Grad der Übereinstimmung wird deutlich geringer. Die Korrelation zwischen SBD1* und SBD2* bleibt jedoch unverändert und hat weiterhin einen Wert von $r = 0{,}90$.

Der p-Wert von $p < 0{,}0001$ bezieht sich darüber hinaus auf den Test der Hypothese, dass die Korrelation gleich Null ist, d. h. die Messwerte völlig unabhängig voneinander sind. Diese Hypothese ist bei der Frage nach Übereinstimmung zwischen den Messwerten jedoch völlig irrelevant. Bei zwei Messwerten, die dasselbe messen sollen, wird man natürlich eine gewisse Korrelation der Daten erwarten. Hinzu kommt die Tatsache, dass die Korrelation abhängig ist vom betrachte-

Schlüsselwörter
- Messmethoden
- Bland-Altman-Methode
- Intraklass-Korrelationskoeffizient
- Wiederholbarkeit

Key words
- Methods of measurement
- Bland-Altman method
- Intraclass correlation coefficient (ICC)
- Repeatability

Bibliografie
DOI 10.1055/s-2007-959047
Dtsch Med Wochenschr 2007; 132: e69–e73 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. rer. biol. hum. Ralf Bender
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail Ralf.Bender@iqwig.de

Tab. 1 Systolische Blutdruckmessungen (SBD1 = manuell, SBD2 = maschinell) an 30 Personen (Teildatensatz aus [2]) in mm Hg.

Nr	SBD1	SBD2	SBD1–SBD2	SBD1*	SBD2*	SBD1**	SBD2**	SBD1**–SBD2**
1	107	124	-17	128,4	99,2	99	132	-33
2	108	128	-20	129,6	102,4	100	136	-36
3	82	98	-16	98,4	78,4	74	106	-32
4	104	135	-31	124,8	108,0	96	143	-47
5	112	124	-12	134,4	99,2	104	132	-28
6	124	136	-12	148,8	108,8	116	144	-28
7	102	112	-10	122,4	89,6	94	120	-26
8	112	135	-23	134,4	108,0	104	143	-39
9	112	122	-10	134,4	97,6	104	130	-26
10	100	111	-11	120,0	88,8	92	119	-27
11	104	111	-7	124,8	88,8	96	119	-23
12	122	125	-3	146,4	100,0	114	133	-19
13	110	122	-12	132,0	97,6	102	130	-28
14	104	114	-10	124,8	91,2	96	122	-26
15	102	126	-24	122,4	100,8	94	134	-40
16	114	137	-23	136,8	109,6	122	129	-7
17	102	115	-13	122,4	92,0	110	107	3
18	120	112	8	144,0	89,6	128	104	24
19	138	113	25	165,6	90,4	146	105	41
20	144	133	11	172,8	106,4	152	125	27
21	154	166	-12	184,8	132,8	162	158	4
22	134	140	-6	160,8	112,0	142	132	10
23	166	154	12	199,2	123,2	174	146	28
24	150	170	-20	180,0	136,0	158	162	-4
25	144	154	-10	172,8	123,2	152	146	6
26	130	141	-11	156,0	112,8	138	133	5
27	140	154	-14	168,0	123,2	148	146	2
28	148	131	17	177,6	104,8	156	123	33
29	220	226	-6	264,0	180,8	228	218	10
30	192	184	8	230,4	147,2	200	176	24

*Bei den mit * und * * bezeichneten Werten handelt es sich um zu Demonstrationszwecken erzeugte künstliche Werte, nähere Details siehe Text.*

ten Wertebereich. Die Korrelation ist höher bei einer großen Spannweite der Stichprobenwerte. Beschränkt man die Stichprobe im Beispiel 1 auf Messungen aus dem Bereich 90–150 mm Hg, so verringert sich der Korrelationskoeffizient auf einen Wert von r = 0,43. Lin [3] hat eine Modifikation des Korrelationskoeffizienten, den ***Konkordanz-Korrelationskoeffizienten***, vorgeschlagen, der die Streuung der Messwerte um die Winkelhalbierende misst und somit die Verzerrung („Bias"), d. h. die durchschnittliche Differenz der Messwerte, mit berücksichtigt. Der Nachteil des Konkordanz-Koeffizienten ist, dass bei einem niedrigen Wert unklar ist, ob dieser durch eine systematische Verzerrung oder durch eine hohe Streuung der Werte verursacht wurde. Auch das Problem der Abhängigkeit vom Wertebereich der Messwerte bleibt bestehen [4].

Regression

Ähnlich ungeeignet für einen Methodenvergleich ist die gewöhnliche Regressionsanalyse. Hierbei wird eine Regressionsgerade an die Datenpaare angepasst, wie in **Abb. 1**b für den Beispieldatensatz dargestellt. Die Regressionsgerade zeigt eine gute Anpassung an die Daten, erlaubt jedoch keine Aussagen über die Übereinstimmung der einzelnen Messwerte. Ein Test, ob der Steigungsparameter den Wert Null hat, ist gleichbedeutend mit einem Test auf Korrelation gleich Null. Dies ist, wie oben ausgeführt, ein ungeeignetes Vorgehen. Informativer ist ein Vergleich der Punktewolke mit der Winkelhalbierenden, welche der Gleichheit (d. h. der völligen Übereinstimmung) der Messmethoden entspricht (d. h. Achsenabschnitt = 0 und Steigungsparameter = 1). Sind jedoch – was beim Vergleich von Messmethoden in aller Regel der Fall ist – beide Messverfahren mit Zufallsfehlern behaftet, so lässt sich zeigen, dass auch bei Übereinstimmung der Methoden der erwartete Steigungsparameter kleiner als 1 und der Achsenabschnitt größer als 0 ist [5]. Ein entsprechender Test ist daher nicht aussagekräftig. Zudem besteht, wie für die Korrelation beschrieben, auch bei der Regression eine Abhängigkeit vom Wertebereich der erhobenen Daten.

Intraklass-Korrelationskoeffizient

Eine weitere statistische Maßzahl, die zur Analyse von Methodenvergleichen eingesetzt wird, ist der Intraklass-Korrelationskoeffizient (ICC) [6]. Der ICC wurde entwickelt, um die Abhängigkeit zwischen Paaren von Messwerten X_1 und X_2 zu quantifizieren, wenn die Reihenfolge der Messwerte keine Rolle spielt und beide Messwerte als Zufallsstichprobe aus einer Population möglicher Messwerte angesehen werden können [7]. Später wurden unterschiedliche Varianten des ICC für den Vergleich zufällig ausgewählter Messungen (z. B. wiederholte Messungen mittels einer Messmethode) sowie für fest vorgegebene Messungen (z. B. Vergleich zweier konkreter Messmethoden) vorgeschlagen. Der ICC lässt sich mit Hilfe von Varianzkomponenten aus geeigneten Varianzanalysemodellen mit festen und zufälligen Effekten (mixed models) berechnen [8]. Der ICC kann jedoch auch bei hoher Übereinstimmung der Messungen kleine Werte annehmen, wenn die Streuung zwischen den Messmethoden klein ist im Verhältnis zur Streuung der Messungen

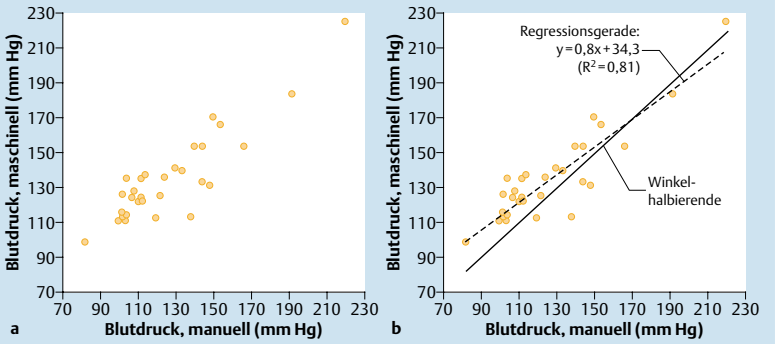

Abb. 1 (**a**) Streudiagramm der Daten aus Beispiel 1 (Tab. 1). (**b**) Streudiagramm der Daten mit Regressionsgerade und Winkelhalbierender.

innerhalb einer Messmethode [8]. Das heißt, ebenso wie der normale Korrelationskoeffizient ist auch der ICC abhängig vom betrachteten Messwert-Bereich. Aus diesen Gründen ist der ICC zur Beurteilung der Übereinstimmung von Messmethoden nur eingeschränkt einsetzbar [7]. Eine detaillierte Beschreibung und kritische Diskussion der Eigenschaften des ICC und dessen Anwendung in Methodenvergleichsuntersuchungen geben Müller & Büttner [6]. Es sei noch darauf hingewiesen, dass ein enger Zusammenhang besteht zwischen dem ICC und dem gewichteten Kappa [1, 9].

Mittelwert-Vergleich (t-Test)

Ist man an der Übereinstimmung zweier Messmethoden interessiert, so ist es sinnvoll, die Differenz der Messwerte X_1–X_2 zu betrachten. Ein häufig beobachtetes Vorgehen ist, eine formale Testprozedur mit Hilfe des gepaarten ***t***-Tests auf eine nicht vorhandene Verzerrung durchzuführen. Ein solches Vorgehen gibt jedoch keinerlei Aufschluss über die Übereinstimmung der Methoden und führt gar zu paradoxen Ergebnissen, was anhand der Daten aus Beispiel 1 demonstriert werden soll. Die geschätzte Verzerrung, d. h. die mittlere Differenz d der Blutdruckwerte, ist gleich -8,4 mit einer Standardabweichung von 12,9. Ein gepaarter ***t***-Test hat den Wert -3,56 mit einem zugehörigen ***p***-Wert von 0,001. Man würde hier also auf eine Nicht-Übereinstimmung der Methoden schließen.

Wir wollen nun die Blutdruckdaten SBD1 und SBD2 künstlich wie folgt verändern: bei den ersten 15 Messwerten wird SBD1 um 8 Einheiten reduziert und SBD2 um 8 Einheiten erhöht, bei den letzten 15 Messwerten ist es genau umgekehrt (**Tab. 1**, Variablen SBD1** und SBD2** mit zugehöriger Differenz SBD1**–SBD2**). Dies führt dazu, dass die durchschnittliche Differenz d** der modifizierten Werte unverändert bleibt. Allerdings variiert die Größenordnung der Abweichungen der Einzelmessungen viel stärker. Dieses schlägt sich in einer höheren Standardabweichung von 25,2 nieder. Die zugehörige Teststatistik hat einen Wert von -1,83, entsprechend einem nicht signifikanten ***p***-Wert von 0,08. Obwohl sich also die modifizierten Datenpaare deutlich stärker voneinander unterscheiden als die Original-Werte, würde man hier anhand des ***t***-Tests eine Übereinstimmung der Messmethoden zum üblichen 5%-Niveau nicht ablehnen können.

Bland-Altman-Methode

Für die Beurteilung der Übereinstimmung von Messmethoden ist eine alleinige Berücksichtigung der ***durchschnittlichen*** Differenz der Messwerte, also der Verzerrung, nicht ausreichend. Von entscheidender Bedeutung ist die Betrachtung der ***Streuung*** der Differenzen der ***einzelnen*** Messwertepaare.

Bland & Altman [10] haben ein einfaches grafisches Verfahren vorgeschlagen, das die Verzerrung und die Streuung der Daten berücksichtigt und mit dessen Hilfe systematische Abweichungen, Ausreißer sowie Abhängigkeiten der Varianz von der Größe der Messwerte beurteilt werden können. Hierbei wird die Differenz der anhand der verschiedenen Methoden ermittelten Blutdruckwerte für jeden Patienten berechnet (d. h. SBD1–SBD2) und gegen den Mittelwert der beiden Messungen (d. h. (SBD1+SBD2)/2) grafisch aufgetragen. Der Mittelwert der beiden Messergebnisse stellt dabei die bestmögliche Schätzung des unbekannten wahren Wertes dar. Ein Plot der Differenz gegen eine der beiden Messungen ist bei mit Messfehlern behafteten Werten nicht geeignet, da in diesem Fall Differenz und Einzelmesswerte auch bei Unabhängigkeit der beiden Messwerte korreliert sind, und somit zu irreführenden Ergebnissen führen können [11]. Der „Bland-Altman-Plot" ist in **Abb. 2** dargestellt. Anhand dieses Plots lassen sich Größenordnungen und Muster der individuellen Abweichungen zwischen den Messmethoden deutlich besser ablesen als bei einer einfachen Punktewolke der Messwerte gegeneinander (**Abb. 1**). Bei einer hinreichend symmetrischen Verteilung der Differenzen liegen 95% der Werte im Bereich $d \pm 2 \times s$, wobei s die Standardabweichung der Differenzen bezeichnet. Diese Grenzen werden als ***Übereinstimmungsgrenzen*** („limits of agreement") bezeichnet und zusammen mit der Verzerrung d in die Grafik eingezeichnet. Für das Datenbeispiel 1 ergeben sich die Übereinstimmungsgrenzen wie folgt:

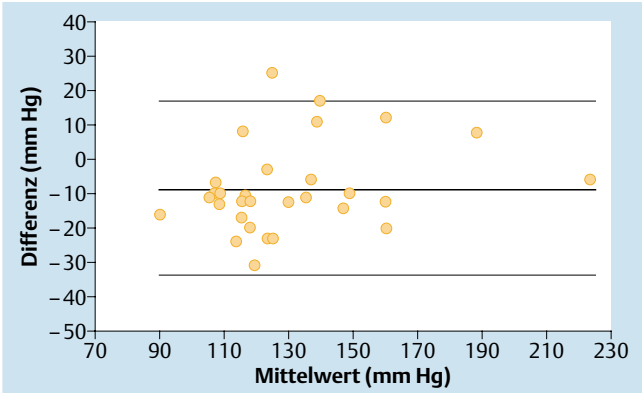

Abb. 2 Bland-Altman-Plot für die Beispieldaten mit Verzerrung d=Mittelwert der Blutdruckdifferenzen (SBD1-SBD2) und Übereinstimmungsgrenzen $d \pm 2 \times s$, wobei s die Standardabweichung der Differenzen bezeichnet.

$$d - 2 \times s = -8{,}4 - 2 \times 12{,}9 = -34{,}2 \text{ und } d + 2 \times s = -8{,}4 + 2 \times 12{,}9 = 17{,}4$$

Das heißt, dass die manuelle Messung des Blutdrucks (SBD1) in 95% der Fälle einen Wert liefert, der bis zu 34,2 mm Hg kleiner bzw. bis zu 17,4 mm Hg größer ist als der maschinell erhobene Messwert (SBD2). Eine klinische Beurteilung dieser Werte ermöglicht nun eine Einschätzung der Übereinstimmungsgüte zwischen manuell und maschinell gemessenen Blutdruckwerten.

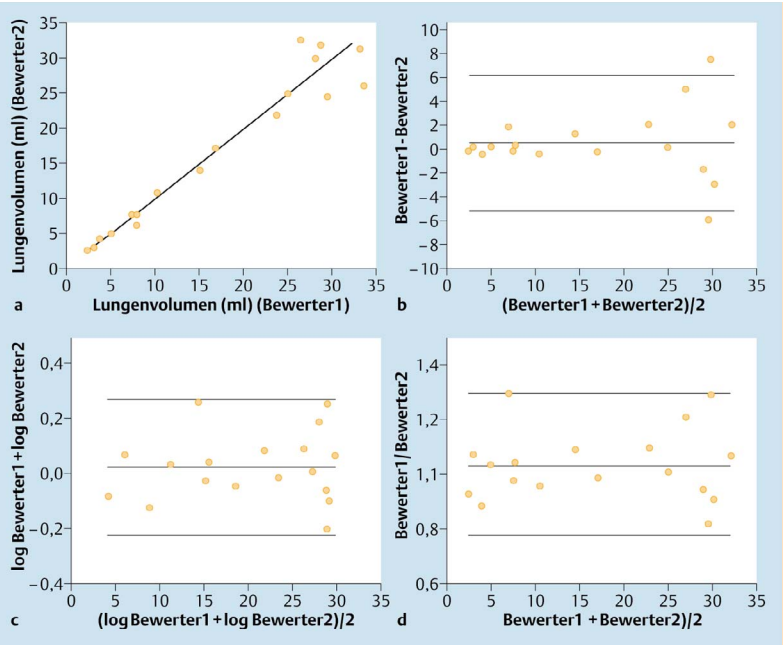

Abb. 3 (**a**) Streudiagramm mit Winkelhalbierender für die Daten aus Beispiel 2. (**b**) Bland-Altman-Plot für Originaldaten. (**c**) Bland-Altman-Plot für logarithmierte Daten. (**d**) Bland-Altman-Plot für Quotient der Daten.

Erweiterungen und Anmerkungen zum Bland-Altman-Verfahren

Transformation der Messwerte

Um sicherzustellen, dass die errechneten Vertrauensgrenzen über den gesamten Messbereich Gültigkeit haben, ist eine Voraussetzung der Bland-Altman-Auswertung, dass die Differenzen zwischen den Messmethoden keine systematischen Veränderungen aufweisen. Dies ist in der Realität aber nicht immer der Fall. So lässt sich häufig eine erhöhte Variabilität der Differenzen mit steigender Größenordnung der Messwerte beobachten.

Dies soll an folgendem Beispiel 2 illustriert werden (modifizierte Daten nach [12]). Bei 18 schwangeren Frauen wurde das fötale Lungenvolumen mit Hilfe eines Ultraschall-basierten, 3-dimensionalen Bildgebungsverfahrens von zwei unabhängigen Bewertern bestimmt. **Abb. 3**a zeigt das Streudiagramm der von den beiden Bewertern ermittelten Messwerte, einschließlich der Winkelhalbierenden. Es deutet sich eine höhere Variabiltität der Messwertdifferenzen bei höheren Lungenvolumenwerten an. Dies ist noch deutlicher im zughörigen Bland-Altman-Plot (**Abb. 3**b) zu erkennen. Bei einer solchen Struktur der Daten lässt sich mit Hilfe einer logarithmischen Transformation der Messwerte eine gleichförmigere Variabilität über den gesamten Messwertbereich erreichen. Der Bland-Altman-Plot der logarithmisch transformierten Werte ist in **Abb. 3**c dargestellt. Durch die Transformation wurde eine gleichförmigere Streuung der Daten erreicht. Für die Verzerrung der logarithmierten Daten ergibt sich ein Wert von 0,02, die Übereinstimmungsgrenzen sind -0,23 und 0,27. Zur Interpretation der Werte müssen diese auf die Originalskala rücktransformiert werden. Das geschieht durch die Anwendung der Exponentialfunktion. Entsprechend ergibt sich eine Verzerrung von 1,02 mit Übereinstimmungsgrenzen von 0,79 und 1,31 auf der Originalskala. Dabei ist jedoch zu beachten, dass die Transformation der Differenz zweier Werte auf der logarithmischen Skala einen dimensionslosen Quotienten liefert. Der rücktransformierte Bias-Wert von 1,02 auf der Originalskala besagt somit, dass der von Bewerter 1 ermittelte Wert um durchschnittlich 2% größer ist als die Messung von Bewerter 2. Die Übereinstimmungsgrenzen besagen, dass für 95% der Fälle der Messwert von Bewerter 1 zwischen 21% kleiner und 31% größer ist als die Messung von Bewerter 2. Alternativ zur log-Transformation der Daten kann man auch direkt den Quotienten der Messwerte betrachten, was zu ähnlichen Ergebnissen führt (**Abb. 3**d).

Führt eine logarithmische Transformation bei komplizierteren Zusammenhängen zwischen Mittelwert und Differenzen nicht zum Ziel, so lassen sich mit Hilfe regressionsanalytischer Ansätze geeignete Übereinstimmungsgrenzen herleiten. Hierauf soll allerdings nicht näher eingegangen werden; nähere Details finden sich in [2].

Konfidenzintervalle für die Übereinstimmungsgrenzen

Bei der Berechnung der Verzerrung und der Übereinstimmungsgrenzen aus einer konkreten Stichprobe handelt es sich um ***Schätzungen*** des wahren Wertes in der zugrunde liegenden Population. Die Ergebnisse sind somit einer Zufallsschwankung unterworfen. Unter der Annahme, dass die Differenzen der Messwerte einer Normalverteilung folgen, lassen sich Standardfehler und Konfidenzintervalle für Verzerrung und Übereinstimmungsgrenzen berechnen [2, 10]. Für die Daten aus Beispiel 1 reicht das 95%-Konfidenzintervall für den geschätzten Bias-Wert d=-8,4 von 8,1 bis 17,7. Die Konfidenzintervalle für untere und obere Übereinstimmungsgrenze sind -42,6 bis -25,8 und 9,0 bis 25,8. Aufgrund des eher kleinen Stichprobenumfangs sind die Konfidenzintervalle recht breit und die Ergebnisse somit mit einer entsprechend großen Unsicherheit behaftet.

Wiederholbarkeit von Messungen

Ein wichtiger Aspekt bei Methodenvergleichen ist die Wiederholbarkeit der Messwerte der einzelnen Methoden. Eine hohe Variabilität zwischen wiederholten Messungen beeinträchtigt die Güte der Übereinstimmung mit einer Vergleichsmethode. Eine schlechte Wiederholbarkeit einer oder beider Methoden hat eine schlechte Übereinstimmung zwischen den Messmethoden zur Folge [11].

Liegen für die zu vergleichenden Messmethoden jeweils zwei Messwiederholungen bei jeder Messeinheit vor, so lässt sich die Bland-Altman-Methode zunächst jeweils für die wiederholten Messungen der beiden Messmethoden anwenden. Mit Messwiederholungen sind hier Messungen an derselben Messeinheit unter

identischen Bedingungen gemeint, die unabhängig voneinander sind und sich nicht gegenseitig beeinflussen. In diesem Fall würde man eine durchschnittliche Differenz der Messwiederholungen von Null erwarten [5]. Zum Vergleich der beiden Messmethoden hinsichtlich der Wiederholbarkeit der Messungen kann die Standardabweichung s der Messwertdifferenzen herangezogen werden. Alternativ kann der so genannte ***Wiederholbarkeitskoeffizient*** verwendet werden. Dieser berechnet sich als WK = 2×s (oder genauer 1,96×s, wobei 1,96 das 97,5%-Quantil der Normalverteilung ist) und gibt die Differenz an, die von 95% der wiederholten Messungen an einer Messeinheit nicht überschritten wird [5]. Das heißt, Veränderungen zwischen den Messungen, die über diese Differenz hinausgehen, können dann (mit 5%-igem Irrtumsvorbehalt) als „echte" Veränderungen interpretiert werden, die nicht allein durch den Messfehler erklärbar sind. Die Intervalle von –WK bis +WK für die beiden Messmethoden und die Übereinstimmungsgrenzen des Vergleichs der Messmethoden können dann verwendet werden, um die Übereinstimmung der Messungen innerhalb der Messmethoden mit der Übereinstimmung der Messungen zwischen den Messmethoden zu vergleichen [5]. Bei mehr als zwei wiederholten Messungen können Verfahren der Varianzanalyse zur Berechnung der Streuung der Messungen herangezogen werden [2]. Für den Vergleich der Messmethoden kann der mittlere Wert der wiederholten Messungen pro Messmethode verwendet werden. Die Schätzung der Verzerrung zwischen den Messmethoden bleibt bei diesem Vorgehen erhalten, die zugehörige Standardabweichung wird jedoch unterschätzt. Bland & Altman [2] geben entsprechende Korrekturverfahren für diesen Fall an.

Weitere Anmerkungen

Die Übereinstimmung von Messmethoden ist mit einer einzelnen statistischen Maßzahl nicht umfassend zu beschreiben. Neben der Verzerrung, d. h. der durchschnittlichen Abweichung der Methoden, ist vor allen Dingen die Abweichung der individuellen Messungen und somit die Streuung der Abweichungen von entscheidender Bedeutung. Sind z. B. die Abweichungen der individuellen Messungen nach Methode 1 konsistent größer als nach Methode 2, so kann – bei geringer Streuung – trotz einer großen Verzerrung durch Subtraktion des Bias-Wertes d von den Messungen nach Methode 1 eine gute Übereinstimmung mit den Messungen nach Methode 2 erzielt werden.

Das Hauptinteresse bei Methodenvergleichen liegt in der Quantifizierung des Unterschiedes und nicht im Testen statistischer Hypothesen auf Gleichheit der Methoden. Die grafische Methode nach Bland-Altman liefert eine einfache Möglichkeit, die Übereinstimmung von Messmethoden anschaulich darzustellen und zu quantifizieren. Die Interpretation der ermittelten Übereinstimmungsgrenzen ist jedoch eine klinische und keine statistische Frage. Es bedarf einer sachwissenschaftlichen Bewertung, ob der Bereich zwischen den Übereinstimmungsgrenzen von einer klinisch bedeutsamen Größenordnung ist oder nicht.

Wenn kein „Goldstandard" (d. h. ein „wahrer" Messwert) vorhanden ist, so lassen sich nur Aussagen zur Vergleichbarkeit der Methoden machen, nicht aber darüber, welche der Methoden die bessere ist bzw. ob überhaupt eine der Messmethoden adäquate Werte liefert.

Das einfache und anschauliche Verfahren nach Bland und Altman hat sich als Verfahren der Wahl zur Auswertung von Methodenvergleichsdaten etabliert. Die Relevanz der Methode lässt sich auch daran ablesen, dass die Arbeit der Autoren aus dem Lancet [10] zu den zehn meist zitierten statistischen Arbeiten gehört [13].

Tab. 2 Übersetzung (deutsch – englisch)

Übereinstimmung	agreement
Kappa-Koeffizient	kappa coefficient
Wiederholbarkeit	repeatability
Zuverlässigkeit	reliability
Wiederholbarkeitskoeffizient	repeatability coefficient
Verzerrung	bias
Konkordanz-Korrelationskoeffizient	concordance correlation coefficient
Streudiagramm	scatter plot
Intraklassen-Korrelationskoeffizient	intra-class correlation coefficient
Gemischte Modelle	mixed models
Übereinstimmungsgrenzen	limits of agreement

kurzgefasst

Bei der Beurteilung der Übereinstimmung zweier auf einer kontinuierlichen Skala erfassten Messmethoden spielen zwei Aspekte eine Rolle: 1) die *durchschnittliche* Übereinstimmung der Verfahren (Bias) und 2) die *Streuung der individuellen* Messwertdifferenzen. Die Bland-Altman-Methode berücksichtigt beide Aspekte im Rahmen eines einfachen grafischen Verfahrens, das sich als Verfahren der Wahl zum Vergleich von Messmethoden etabliert hat.

Literatur

1 *Grouven U, Bender R, Ziegler A, Lange S.* Der Kappa-Koeffizient. Dtsch Med Wochenschr 2007; 132: e65–e68
2 *Bland JM, Altman DG.* Measuring agreement in method comparison studies. Stat Meth Med Res 1999; 8: 135–160
3 *Lin LI.* A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255–268
4 *Atkinson G, Nevill A.* Comment on the use of concordance correlation to assess the agreement between two variables. Biometrics 1997; 53: 775–777
5 *Bland JM, Altman DG.* Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22: 85–93
6 *Müller R, Büttner P.* A critical discussion of intraclass correlation coefficients. Stat Med 1994; 13: 2465–2476
7 *Bland JM, Altman DG.* A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990; 20: 337–340
8 *Bartko JJ.* Measures of agreement: a single procedure. Stat Med 1994; 13: 737–745
9 *Fleiss JL, Levin B, Paik MC.* Statistical Methods for Rates and Proportions. Third edition. Wiley, 2003
10 *Bland JM, Altman DG.* Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i: 307–310
11 *Bland JM, Altman DG.* Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995; 346: 1085–1087
12 *Bahmaie A, Hughes SW, Clark T et al.* Serial fetal lung volume measurement using three-dimensional ultrasound. Ultrasound Obstet Gynecol 2000; 16: 154–158
13 *Ryan TP, Woodall WH.* The most-cited statistical papers. J Appl Stat 2005; 32: 461–474