ÜBERSICHTSARBEIT

Konkordanzanalyse

Teil 16 der Serie zur Bewertung wissenschaftlicher Publikationen

Robert Kwiecien, Annette Kopp-Schneider, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Dieser Artikel beschreibt Methoden zum qualitativen und quantitativen Vergleich von Messverfahren oder Beurteilern. Ziel ist es beispielsweise, die Übereinstimmung von Mess- oder Beurteilungsverfahren zu zeigen, um so entsprechende Methoden zu etablieren.

Methode: Es wird eine Auswahl einfacher Methoden zum Vergleich von Mess- beziehungsweise Beurteilungsverfahren anhand eines Beispiels veranschaulicht und jeweils die dabei zugrundeliegende Idee anhand der Herleitung dieser Methoden erläutert. Basierend auf einer selektiven Literaturrecherche werden exemplarische Beispiele aus der medizinischen Forschung genannt.

Ergebnisse: Bei den Methoden zum Vergleich von Mess- beziehungsweise Beurteilungsverfahren unterscheidet man Techniken, deren Ausprägungen ein stetiges Skalenniveau haben von solchen mit einem nominalen Skalenniveau. Hierbei beschränkt sich der Artikel auf den Vergleich von je zwei Messverfahren beziehungsweise von je zwei Beurteilern. Es werden zudem übliche fehlerhafte Ansätze zur Beurteilung von Übereinstimmungen aufgezeigt.

Schlussfolgerung: Wenn beispielsweise im Bereich der Diagnostik ein neues Mess- oder Beurteilungsverfahren etabliert werden soll, oder wenn im Sinne der Qualitätssicherung die näherungsweise Übereinstimmung vieler Mess- beziehungsweise Beurteilungsverfahren dargelegt werden soll, sind Analysen zur Bewertung von Übereinstimmungen, sogenannte Konkordanzanalysen, notwendig. Fehlerhafte Ansätze können zur falschen Annahme in Bezug auf eine Übereinstimmung verschiedener Mess- oder Beurteilungsverfahren führen.

► **Zitierweise**
Kwiecien R, Kopp-Schneider A, Blettner M: Concordance analysis—part 16 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(30): 515–21. DOI: 10.3238/arztebl.2011.0515

Institut für Biometrie und Klinische Forschung (IBKF), Westfälische Wilhelms-Universität Münster: Dr. rer. nat. Kwiecien

Abteilung Biostatistik, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg: Prof. Dr. rer. nat. Kopp-Schneider

Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg Universität Mainz: Prof. Dr. rer. nat. Blettner

Bei vielen diagnostischen Methoden in der Medizin ist die Möglichkeit einer Fehldiagnose einzuräumen. Die Diagnose eines Arztes weist eine gewisse Fehlerrate auf, Diagnosen verschiedener Ärzte stimmen nicht immer überein, technische Messungen sind nicht beliebig genau. Sowohl die Diagnose des Arztes, als auch die technische Messung haben gemein, dass sie in der Regel fehlerbehaftet sind oder sein können. Diagnosestellende Personen sowie Diagnose- beziehungsweise Messmethoden sollen im Folgenden als Beurteiler, und Diagnosen beziehungsweise Messungen als Beurteilungen bezeichnet werden.

Falls eine Methode die interessierende Größe tatsächlich fehlerlos messen kann, wird diese Methode üblicherweise als Goldstandard bezeichnet. Wenn nun aber ein neues Verfahren, das beispielsweise das Tumorvolumen mit weniger Aufwand oder schonender für den Patienten misst, eingeführt werden soll, ist zu prüfen, ob die damit erzielten Ergebnisse mit denen der etablierten Methode übereinstimmen. Im Fall einer quantitativen Größe mit stetigem Skalenniveau (beispielsweise Tumorvolumen) ist es verbreitet, die Korrelation zwischen den beiden Messmethoden zu berechnen.

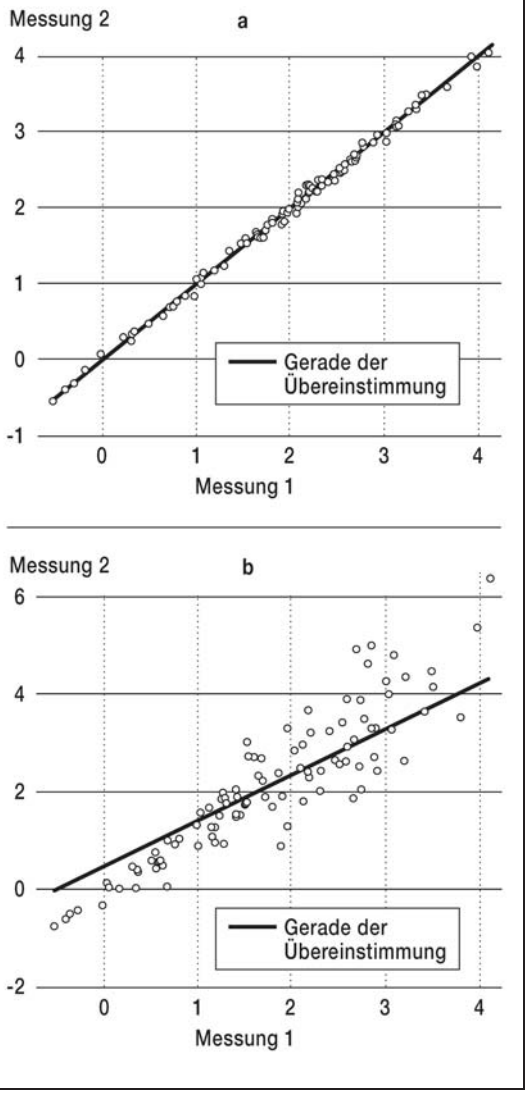
Dass dieses Vorgehen zur Prüfung der Übereinstimmung ungeeignet ist, wird hier dargelegt. Übereinstimmungen von quantitativen Messmethoden lassen sich visuell veranschaulichen und anhand der grafischen Darstellung kann die Güte der Übereinstimmung vom Arzt beurteilt werden.

Die Übereinstimmung einer neuen Methode mit dem Goldstandard beinhaltet die Bewertung des Messfehlers der neuen Methode. Vom Prinzip her gibt es hier keinen Unterschied zur Beurteilung der Übereinstimmung zweier fehlerbehafteter Messmethoden.

Die Übereinstimmung von Beurteilungen ist aber auch in Situationen von Interesse, in denen nominale Merkmale (zum Beispiel „Grippe“, „grippaler Infekt“, „Sonstiges“) oder nominal ordinale Merkmale (beispielsweise „gut“, „mittel“, „schlecht“) erhoben werden. Bei der Notenvergabe in Abiturklausuren könnte man etwa untersuchen, inwieweit zwei Korrektoren zum selben Ergebnis kommen, bei ärztlichen Diagnosen mit den möglichen Diagnosen „krank“ oder „gesund“, inwieweit die Diagnosen bei zwei Ärzten übereinstimmen.

In der vorliegenden Arbeit geht es somit nicht darum, ob Beurteiler richtig beurteilen, sondern inwie-

GRAFIK 1



Direkter Vergleich zweier Messmethoden mittels Punktwolke und Winkelhalbierender; Beispiel a) Messung 1 versus Messung 2; Beispiel b) Messung 1 versus Messung 2

weit zwei Beurteiler übereinstimmend beurteilen. Die Situation wird komplizierter, wenn es um die Übereinstimmung einer Vielzahl von Beurteilern geht. Die Autoren reduzieren in dieser Arbeit ihre Betrachtungen auf den Vergleich von zwei Beurteilern.

Dieser Artikel beschäftigt sich mit deskriptiven Methoden, um die Übereinstimmung von zwei Beurteilern visuell und quantitativ zu bewerten. Entsprechende Untersuchungen fallen unter den Oberbegriff Konkordanzanalyse. In dieser Arbeit werden dazu vorrangig Bland-Altman-Diagramme und Cohens Kappa behandelt. Dazu werden zwei Situationen unterschieden. In einer Situation sollen zwei Beurteiler eine Stichprobe vom Umfang n von zu beurteilenden Personen oder Objekten bezüglich einer nominalen Variablen mit Ausprägungen wie zum Beispiel „krank", „gesund" (dichotom) oder mit Ausprägungen wie zum Beispiel „grippaler Infekt", „Grippe", „Sonstiges" beurteilen. In der anderen Situation erfolgt von zwei verschiedenen Beurteilern für eine Stichprobe je eine Beurteilung über eine stetige Messgröße.

Beurteilungen mit stetigen Ausprägungen

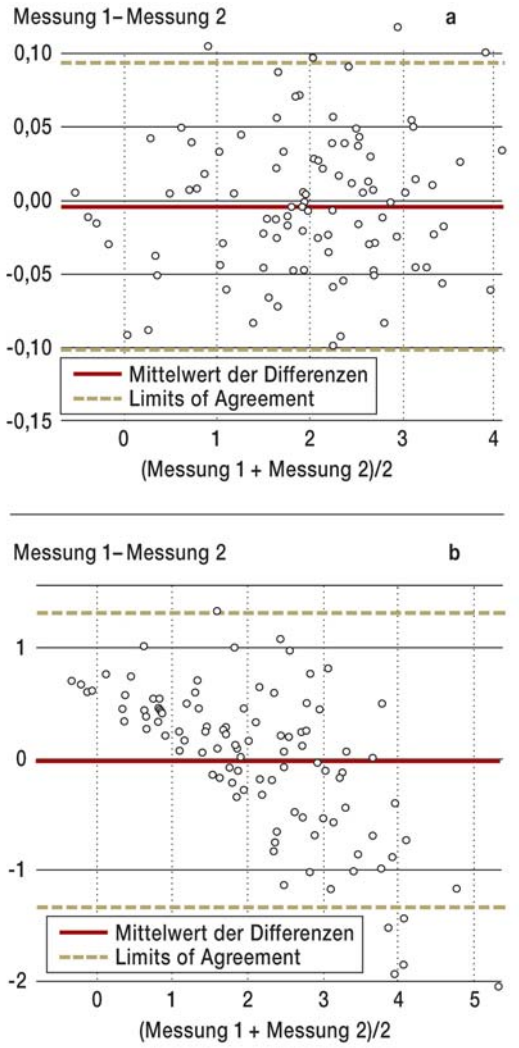
Eine Beurteilung mit stetiger Ausprägung tritt in der Regel bei physikalischen Messungen auf. Oftmals gibt es zu einzelnen Messvariablen verschiedene Methoden oder auch Geräte, um die Messungen durchzuführen, die dann einander gegenübergestellt werden sollen (1). Ist es erstrebenswert, bezüglich einer medizinischen Variablen eine neue Messmethode zu etablieren, so sollte die Güte der neuen Messmethode durch einen Vergleich mit einer etablierten Methode oder mit einem Goldstandard geprüft werden.

In diesem Abschnitt sollen anhand fiktiver Beispiele statistische Verfahren zum Vergleich zweier Messmethoden vorgestellt werden. Ausgegangen wird von einer gewissen Anzahl (zum Beispiel n = 100) verschiedener Personen oder Objekten, an denen pro Messmethode je einmal eine entsprechende Messung vorgenommen wird. Dies liefert n Paare von Messungen. Ein erster Schritt ist es, die Messungen der beiden Messmethoden in einem Diagramm gegeneinander abzutragen. Wenn die Messmethoden weitestgehend übereinstimmen, dann sollten sich die entsprechenden Punkte in der Nähe der Geraden, die die Übereinstimmung kennzeichnet, befinden. Diese Gerade wird auch häufig Winkelhalbierende genannt.

In *Grafik 1* stellen die *Beispiele a* und *b* recht eindeutige Situationen dar. Punkte, bei denen die Paare aus Messung 1 und Messung 2 absolut übereinstimmen, müssen auf der eingezeichneten Geraden liegen. Die *Grafik 1 a* (Beispiel a) spiegelt eine gute Übereinstimmung der beiden Messmethoden wider, die *Grafik 1 b* (Beispiel b) hingegen zeigt, dass die Streuung der Differenz zwischen Messmethode 1 und Messmethode 2 für größere Werte augenscheinlich zunimmt, und insgesamt größer ist, als im Beispiel a.

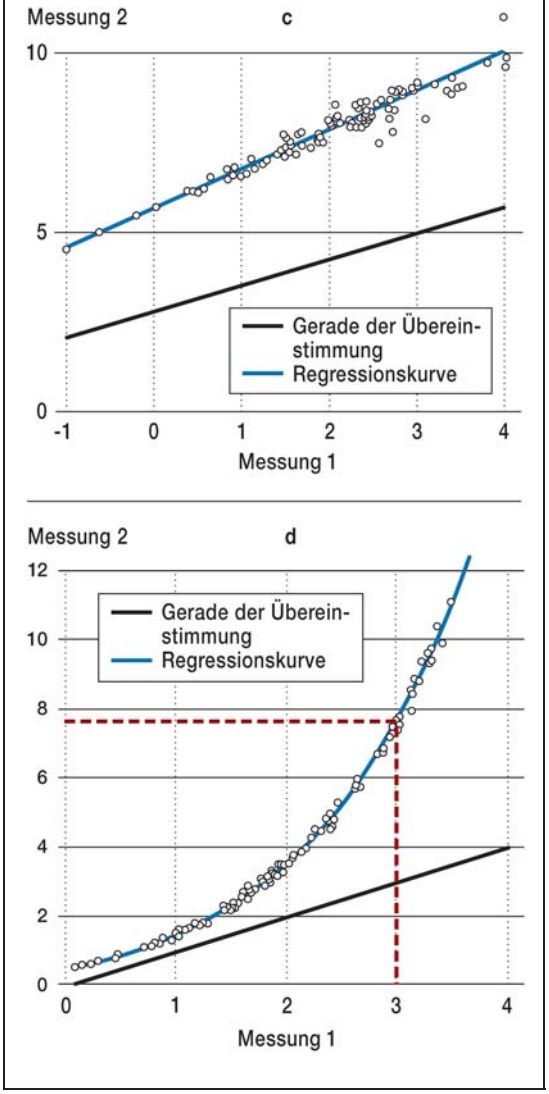
Um diese Zusammenhänge genauer zu beleuchten, wird für beide Fälle (Beispiel a und b) jeweils ein Bland-Altman-Diagramm erstellt *(Grafik 2)*. In einem Bland-Altman-Diagramm werden jeweils zu den Messpaaren die Mittelwerte der Messungen gegen die Differenzen der Messungen abgetragen. Zusätzlich wird der Mittelwert aller Differenzen als horizontale Linie sowie diese Mittelwertlinie ± 1,96 × Standardabweichung der Differenzen eingezeichnet (gestrichelte Linien). Der durch diese Linien eingegrenzte Bereich wird als Übereinstimmungsbereich („limits of agreement") bezeichnet. Die Mittelwertlinie beschreibt eine in der Regel systematische, korrigierbare Abweichung, der Übereinstimmungsbereich (beziehungsweise die „limits of agreement") eine in der Regel nicht korrigierbare Abweichung. Unter der Annahme einer Normalverteilung liegen schätzungsweise 5 % der Differenzen aus der Gesamtpopulation

GRAFIK 2



Vergleich zweier Messmethoden mittels Bland-Altman-Diagramm; Beispiel a) Bland-Altman-Plot; Beispiel b) Bland-Altman-Plot

GRAFIK 3



Punktwolke zum Vergleich zweier Messmethoden zwischen denen offenbar ein funktionaler Zusammenhang besteht; Beispiel c) Messung 1 versus Messung 2; Beispiel d) Messung 1 versus Messung 2

außerhalb des Übereinstimmungsbereichs, also außerhalb der 1,96 × Standardabweichungsschranken (2). Oftmals wird anstelle des Faktors 1,96 einfach nur mit 2 multipliziert. Der Wert 1,96 gilt als exakter, da 1,96 dem 97,5-%-Quantil der Standardnormalverteilung entspricht. Damit eignet sich ein Bland-Altman-Diagramm gut, um die Messdifferenzen visuell zu bewerten.

Das Bland-Altman-Diagramm in *Grafik 2* zum Beispiel a bestätigt eine gute Übereinstimmung. Die Mittelwertlinie der Differenzen liegt nahezu bei 0, es gibt also keinen Hinweis auf systematische Abweichungen zwischen den beiden Methoden. Die Standardabweichung der Differenzen liegt bei etwa 0,05. Bei einer angenommenen Normalverteilung ist davon auszugehen, dass die Differenzen in 95 % der Fälle betragsmäßig kleiner sind als 0,1 – im Verhältnis zu den Messwerten also klein sind. Der Abstand zwischen den „limits of agreement“ beziehungsweise die Größe des Übereinstimmungsbereichs beträgt hier 0,1. Bei der konkreten Anwendung von Bland-Altman-Diagrammen in der Praxis kann allerdings die Güte der Übereinstimmung nicht losgelöst von der fachlichen Fragestellung beurteilt werden. Der Anwender muss festlegen, wie groß die Übereinstimmung bezüglich der Größe des Übereinstimmungsbereichs beziehungsweise bezüglich der „limits of agreement“ zwischen den Messwerten im Hinblick auf die klinische Relevanz sein muss. Tetzlaff et al. (1) haben beispielsweise die zwei Messmethoden

KASTEN 1

Cohens Kappa – Rechenbeispiel

	Beurteiler 2		
Beurteiler 1	**gesund**	**krank**	**Randhäufigkeit**
gesund	50 (0,45)	10 (0,09)	60 (0,54)
krank	30 (0,27)	20 (0,18)	50 (0,45)
Randhäufigkeit	80 (0,73)	30 (0,27)	110

In 70 von 110 Fällen haben die beiden Ärzte (Beurteiler 1 und Beurteiler 2) übereinstimmend geurteilt. Obige Kontingenztafel enthält die entsprechenden Angaben sowohl zu den absoluten als auch zu den relativen Häufigkeiten.

Für die Beurteilung „gesund" ist die geschätzte Wahrscheinlichkeit für eine übereinstimmende Beurteilung etwa 45 %, für die Beurteilung „krank" etwa 18 %. Die geschätzte Wahrscheinlichkeit für eine übereinstimmende Beurteilung liegt also bei $p_0 = 70/110 = 45\ \% + 18\ \% = 63\ \%$. Bei eventuell völliger Willkür von einem oder beiden Beurteilern ist eine gewisse Zahl an Übereinstimmungen zu erwarten, die von den gegebenen Randhäufigkeiten abhängt. Dies entspricht im mathematischen Sinn der stochastischen Unabhängigkeit, womit durch die Beurteilung des einen Beurteilers keine zusätzliche Information zum Beurteilungsergebnis des anderen Beurteilers zu gewinnen wäre.

Damit die folgende Berechnung des unter Willkür beziehungsweise Unabhängigkeit zu erwartenden Anteils an Übereinstimmungen übersichtlicher wird, werden die relativen Anteile vorrübergehend in nicht prozentualer Form dargestellt, also schreibt man beispielsweise 0,54 anstelle von 54 %.

Bei der Willkür eines Beurteilers ist mit einer Wahrscheinlichkeit von etwa $0{,}54 \times 0{,}73 = 0{,}39$ eine übereinstimmende Beurteilung mit Urteil „gesund" zu erwarten, für das Urteil „krank" liegt diese Wahrscheinlichkeit bei etwa $0{,}45 \times 0{,}27 = 0{,}12$. Man rechnet bei Willkür also mit einer Wahrscheinlichkeit von etwa $p_e = 0{,}54 \times 0{,}73 + 0{,}45 \times 0{,}27 = 0{,}52$ (57,2 von 110 Fällen) an Urteilsübereinstimmung, das heißt 52 % gegenüber der beobachteten prozentualen Übereinstimmung von 63 %. Damit ist die beobachtete Wahrscheinlichkeit einer Urteilsübereinstimmung nur um $11\ \% = p_0 - p_e = 63\ \% - 52\ \%$ gegenüber der Urteilsübereinstimmung bei völliger Willkür bei (mindestens) einem Beurteiler erhöht.

Der Anteil dieser erhöhten prozentualen Übereinstimmung wird zur Normierung noch in Relation zum höchsten, theoretisch denkbaren Wert gesetzt, durch den noch dividiert wird. Dieser liegt hier bei $100\ \% - 52\ \%$ (beziehungsweise bei $1 - p_e$), also bei 100 % beobachteter Übereinstimmung, was natürlich nicht mehr steigerbar ist. Die normierte Größe $K_2 = (p_0 - p_e)/(1 - p_e)$ wird als Cohens Kappa bezeichnet. Im obigen Beispiel nimmt Cohens Kappa den Wert $11\ \% / (100\ \% - 52\ \%) = 0{,}23$ an.

Für Cohens Kappa wird genau dann ein Wert von 1 erreicht, wenn die Urteile der zwei Beurteiler vollständig übereinstimmen, ein Wert von 0 bedeutet, dass die Übereinstimmungen sich nicht von der zu erwartenden Urteilsübereinstimmung bei willkürlicher Beurteilung von (mindestens) einem Beurteiler abhebt. Negative Werte bedeuten, dass die Urteilsübereinstimmung gar noch geringer ist, als bei einer willkürlichen Beurteilung. Ein Wert von –1 kann im Allgemeinen nicht erreicht werden.

Magnetresonanztomographie (MRT) und Spirometrie, unter anderem Bland-Altman-Diagramme, verglichen und den Übereinstimmungsbereich als zufriedenstellend bewertet.

Das Bland-Altman-Diagramm zum Beispiel b *(Grafik 2)* zeigt gleich mehrere Mängel bei der Übereinstimmung auf. Im Mittel weichen die beiden Messmethoden zwar kaum voneinander ab, aber der Übereinstimmungsbereich ist durch das Intervall [-1,4; 1,4] gegeben, also werden etwa 95 % der zukünftig zu messenden Differenzen im Intervall [-1,4; 1,4] erwartet. Es ist vom Mediziner zu entscheiden, ob diese Abweichung akzeptabel ist. Die ungleichmäßige Verteilung der Punkte in diesem Bild weist auf eine systematische Verzerrung hin.

Man muss allerdings beachten, dass eine neue Messmethode nicht vorschnell zu verwerfen ist, wenn mit der Punktwolke und dem Bland-Altman-Diagramm eine schlechte Übereinstimmung belegt ist. In *Grafik 3* werden zwei weitere Fälle (Beispiel c und d) dargelegt, bei denen zwar offensichtlich eine schlechte Übereinstimmung vorliegt (die Punkte liegen weit ab von der Geraden der Übereinstimmung), aber bei denen zwischen den beiden Messmethoden ein funktionaler Zusammenhang besteht, der jeweils durch eine Regressionskurve veranschaulicht wird. Im Beispiel c *(Grafik 3 c)* liegt ein linearer, im Beispiel d *(Grafik 3 d)* ein nichtlinearer Zusammenhang vor.

Oftmals kann eine Messung über einen deutlichen funktionalen Zusammenhang aus einer Messung mit einer anderen Messmethode gut geschätzt werden, auch wenn die Messmethoden zunächst sehr unterschiedliche Messwerte ergeben. Für den Messwert von 3,0 in *Grafik 3 d* würde beispielsweise für die Messung unter Messmethode 2 der Wert 7,65 geschätzt werden. Die Unstimmigkeit zwischen beiden Messmethoden scheint also größtenteils korrigierbar zu sein. Um dann die mittels des geschätzten funktionalen Zusammenhangs „korrigierte" Messmethode 2 mit der Messmethode 1 zu vergleichen, kann wiederum zu den genannten Methoden wie zum Beispiel dem Bland-Altman-Diagramm gegriffen werden. Die Eichung von Messgeräten entspricht im Grundprinzip dieser Vorgehensweise. Für die Schätzung des in *Grafik 3* eingezeichneten funktionalen Zusammenhangs (Erstellung einer Regressionskurve) gibt es vielfältige statistische Methoden wie etwa die lineare oder nichtlineare Regression, die hier nicht näher besprochen werden sollen.

Oftmals wird der Pearson-Korrelationskoeffizient (2) zwischen den beiden Messungen betrachtet, um einen linearen (also speziellen funktionalen) Zusammenhang zwischen beiden Methoden zu belegen. Ein betragsmäßig großer Korrelationskoeffizient (nahe bei 1 oder –1) ist ein guter Hinweis auf einen linearen Zusammenhang. Ein häufiger Irrtum besteht dabei in der Fehlinterpretation von Signifikanztests in Bezug auf Korrelationskoeffizienten. Der Befund, dass die Korrelation von zwei Messmethoden signifikant von 0 verschieden ist, reicht nicht aus, die Übereinstim-

mung der Methoden zu belegen. Ein signifikantes Resultat wird bereits bei irrelevanten Zusammenhängen erreicht. Dies beinhaltet aber noch keinerlei Aussage darüber, wie groß die Abweichungen zwischen den beiden Methoden sind (3, 4).

Beurteilungen mit nominalem Skalenniveau – Cohens Kappa

In diesem Abschnitt werden Beurteilungen mit nominalen Ausprägungen besprochen. In der medizinischen Forschung wird zur Bewertung von Urteilsübereinstimmungen oftmals ein Maß herangezogen, das als Cohens Kappa bezeichnet wird. Song et al. (5) verglichen beispielsweise zwei Methoden zur Identifizierung von Knochenmetastasen miteinander, die eine gute Übereinstimmung erreichten (Kappa = 0,732). Cohens Kappa misst anschaulich gesprochen die normierte Differenz zwischen dem Anteil an beobachteten Urteilsübereinstimmungen und dem Anteil an Urteilsübereinstimmung, der durch reinen Zufall zu erwarten wäre.

Was bedeutet dies konkret? Es soll zunächst aus Gründen der Übersichtlichkeit der Fall dichotomer Beurteilungen anhand eines fiktiven Beispiels behandelt werden. Die Herleitung von Cohens Kappa ist in *Kasten 1* anhand dieses Beispiels detaillierter wiedergegeben. Dazu sollen zwei Ärzte bei n = 110 Patienten (n steht in diesem Artikel für Stichprobenumfänge) bezüglich eines Krankheitsbildes beurteilen, ob diese Patienten krank oder gesund sind. Im Zentrum der Betrachtung steht Urteilsübereinstimmung beziehungsweise Urteilskonkordanz der beiden beurteilenden Ärzte. Die je 110 Beurteilungen der beiden Ärzte sind in der *Tabelle in Kasten 1* dargestellt.

In 70 von 110 Fällen haben die beiden Ärzte das Krankheitsbild übereinstimmend beurteilt. Allerdings wird diese Zahl allein einen wenig brauchbaren Blick auf die Urteilskonkordanz der beiden Beurteiler liefern, da auch bei eventuell völliger Willkür bei einem der Beurteiler (oder gar beider Beurteiler) eine gewisse Zahl an Übereinstimmungen zu erwarten ist. Im Durchschnitt erwartet man bereits etwa 57 Übereinstimmungen durch reinen Zufall *(Kasten 1)*. Cohens Kappa bewertet den Unterschied zwischen der Zahl von 57 zufällig zu erwartenden Übereinstimmungen und der erreichten Zahl von 70 Übereinstimmungen bezogen auf die Gesamtzahl der Fälle (= 110). In diesem Beispiel nimmt Cohens Kappa den Wert 0,23 an. Bei völliger Übereinstimmung nimmt Cohens Kappa den Wert 1 an. Ein Wert von 0 bedeutet, dass die Übereinstimmungen der Zahl der zu erwartenden zufälligen Urteilsübereinstimmungen entsprechen und ist somit ein miserabler Wert. Negative Werte bedeuten, dass die Urteilsübereinstimmung noch geringer ist als durch reinen Zufall zu erwarten wäre, dass also die Beurteiler gegenläufig urteilen. Ein Wert von –1 kann im Allgemeinen nicht erreicht werden.

Die Interpretation einer Kenngröße wie Cohens Kappa ist letztendlich willkürlich. In der Arbeit von Altman (2) wird die Bewertung aus *Tabelle 1* vorgeschlagen. Im obigen Beispiel wäre ein Wert von $K_2 = 0{,}23$ als „hinreichend“ einzustufen.

In *Kasten 2* wird weiterführend die allgemeinere Situation mit 2 oder mehr als 2 ($k \geq 2$) möglichen Urteilsausprägungen behandelt.

Cohens Kappa ist ein Mittel, um das Ausmaß an Übereinstimmung zweier Beurteiler quantitativ zu bewerten, aber über die Zuverlässigkeit dieser Bewertung sagt diese Größe allein noch nichts aus. Bei einer kleinen Anzahl von Patienten ist Cohens Kappa wenig aussagekräftig, daher sollte – wie in vielen anderen Fällen auch – ein Konfidenzintervall *(Kasten 3)* berechnet werden (6).

Oftmals wird in der Praxis mittels Cohens Kappa einseitig getestet, ob die beobachtete Urteilsüberein-

TABELLE 1

Kategorisierung von Cohens-Kappa-Werten (2)

Wert von K_k	Ausmaß der Übereinstimmung
< 0,20	nicht ausreichend (poor)
0,21–0,40	hinreichend (fair)
0,41–0,60	moderat (moderat)
0,61–0,80	gut (good)
0,81–1,0	sehr gut (very good)

KASTEN 2

Cohens Kappa für k Urteilsausprägungen

In der unteren Kontingenztafel wird die allgemeinere Situation mit $k \geq 2$ möglichen Urteilsausprägungen in abstrakter Form dargestellt. Hier werden die Randhäufigkeiten mit $a_{i.} := a_{i1} + \ldots + a_{ik}$ und $a_{\cdot i} := a_{1i} + \ldots + a_{ki}$ notiert. Der Anteil übereinstimmender Beurteilungen ist $p_0 := 1/n \times (a_{11} + a_{22} + \ldots + a_{kk})$, was der Summe der Diagonaleinträge der allgemeinen Kontingenztafel dividiert durch die Anzahl n der Beurteilungsobjekte entspricht. Die zu erwartende Wahrscheinlichkeit einer Urteilsübereinstimmung bei gegebenen Randhäufigkeiten bei stochastisch unabhängigen Beurteilern ist durch $p_e := a_{1.}/n \times a_{.1}/n + a_{2.}/n \times a_{.2}/n + \ldots + a_{k.}/n \times a_{.k}/n$ gegeben, also der Summe der multiplizierten relativen Randhäufigkeiten bezüglich der Diagonaleinträge der unteren allgemeinen Kontingenztafel. Cohens Kappa wird wie schon im Fall von nur zwei möglichen Urteilsausprägungen definiert durch $K_k := (p_0 - p_e)/(1 - p_e)$.

	1	2	. . .	k	
1	a_{11}	a_{12}	. . .	a_{1k}	$a_{1.}$
2	a_{21}	a_{22}	. . .	a_{2k}	$a_{2.}$
. . .	. . .	. . .	. . .	. . .	. . .
k	a_{k1}	a_{k2}	. . .	a_{kk}	$a_{k.}$
	$a_{.1}$	$a_{.2}$	. . .	$a_{.k}$	n

KASTEN 3

Cohens Kappa – Konfidenzintervall

Die bloße Angabe einer deskriptiven Kennzahl (wie zum Beispiel dem Mittelwert) enthält meist keine Information darüber, in welchem Ausmaß diese auf die Gesamtpopulation übertragbar ist. Hierzu werden deskriptive Kennzahlen üblicherweise durch Konfidenzintervalle ergänzt. Ein approximatives (1-α)-Konfidenzintervall für Cohens Kappa der Gesamtpopulation ist durch folgende Formel gegeben:

$$KI := \kappa_k \pm z_{1-\alpha/2} \times \sqrt{\frac{p_0 \times (1-p_0)}{n \times (1-p_e)^2}}$$

Hier bezeichnet $z_{1-\alpha/2}$ das (1-α/2)-Quantil der Standardnormalverteilung, dessen Werte aus statistischen Tabellen entnommen werden können (9). Im vorliegenden Fall wird eine Approximation an eine Normalverteilung verwendet. Als Faustregel wird oftmals herangezogen: Die Approximation ist hinreichend gut, wenn $n \times p_0 \geq 5$ und $n \times (1-p_0) \geq 5$.

Zur Veranschaulichung wird das Zahlenbeispiel aus dem Abschnitt „Beurteilungen mit nominalen Skalenniveau – Cohens Kappa" verwendet. Der Stichprobenumfang ist n = 110, zudem wurde $p_e = 0{,}52$ und $p_0 = 0{,}63$ und $K_2 = 0{,}23$ berechnet. Es wird α = 5 % festgelegt, und aus einer statistischen Tabelle kann $z_{0{,}975} = 1{,}96$ abgelesen werden. Obere Formel liefert

$$KI = 0{,}23 \pm 1{,}96 \times 0{,}0959$$

beziehungsweise ein 95-%-Konfidenzintervall [0,042; 0,418] für Cohens Kappa der Gesamtpopulation.

stimmung stark genug ist, um auszuschließen, dass (mindestens) einer der Beurteiler willkürlich beurteilt. Irrtümlicherweise wird ein signifikantes Testresultat als objektives Zeichen für eine Urteilsübereinstimmung interpretiert. Für die Bewertung der Urteilsübereinstimmung ist ein solches Testresultat aber kaum aussagekräftig, da bei großem Stichprobenumfang auch ein sehr kleiner positiver Wert für Cohens Kappa zu einem signifikanten Resultat führen kann. Ein Signifikanztest ist hierbei unangebracht. Für Cohens Kappa sind noch Verfeinerungen und Verallgemeinerungen möglich. Bei ordinalen Ausprägungen kann es von Bedeutung sein, die Unterschiede zwischen zwei Ausprägungen unterschiedlich zu bewerten. Dafür kann das gewichtete Kappa verwendet werden. Die Bewertung von Übereinstimmungen von Beurteilungen kann auch in noch weiteren Situationen zum Beispiel für mehr als 2 Beurteiler erfolgen (7).

Zum Vergleich eines Beurteilers mit einem Goldstandard bei dichotomen Ausprägungen der Beurteilungen werden oftmals Sensitivität und Spezifität betrachtet (8), die das Maß der Übereinstimmung in beiden durch den Goldstandard definierten Teilpopulationen separat angeben. Cohens Kappa hingegen ermöglicht eine zusammenfassende Bewertung der Übereinstimmung zwischen Beurteiler und Goldstandard.

Diskussion

Statistische Methoden zur Bewertung von Übereinstimmungen von Beurteilungen zweier Beurteiler beziehungsweise von Messungen zweier Messmethoden unterscheiden zwischen zwei Situationen:

- Beurteilungen mit stetigen Ausprägungen
- Beurteilungen mit kategorialen Ausprägungen.

Für den ersten Fall ist anzuraten, deskriptive und grafische Methoden einzusetzen wie die Darstellung der Punktwolke, zusammen mit der Geraden der Übereinstimmung, und das Bland-Altman-Diagramm. Die Punktwolke ist die intuitivere und anschaulichere Methode, das Bland-Altman-Diagramm erlaubt aber eine differenziertere Analyse, um die Unterschiede teils auch quantitativ zu bewerten. Die „limits of agreement" beim Übereinstimmungsbereich in den Bland-Altman-Diagrammen können ungeeignet sein, um Abweichungen zweier Messmethoden zu beurteilen, wenn die Verteilung der Differenzen von einer Normalverteilung abweicht. Empirische Quantile können hierbei aber eine Alternative liefern.

Um die Verteilung der Differenzen zwischen zwei Messmethoden genauer zu untersuchen, können diese auch in einem Histogramm dargestellt werden (3). In vielen Fällen kann bei einem guten linearen, oder allgemeiner einem guten funktionalen, Zusammenhang das Messergebnis einer Methode umgerechnet werden, um das Messergebnis mit der anderen Methode vorherzusagen, auch wenn die zwei Messmethoden zunächst deutlich verschiedene Resultate liefern. Als weiteres deskriptives Verfahren dient der Pearson-Korrelationskoeffizient, der Hinweise auf einen linearen Zusammenhang gibt. Ein signifikant von 0 verschiedener Korrelationskoeffizient hingegen kann nicht im Sinne einer Übereinstimmung von Messmethoden bewertet werden, da die Abweichungen der Methoden voneinander trotzdem erheblich sein können.

Für die quantitative Bewertung von Übereinstimmungen bei Beurteilungen mit kategorialen Ausprägungen eignet sich die Berechnung von Cohens Kappa, für das ein Konfidenzintervall angegeben werden kann.

KERNAUSSAGEN

- Ein Beleg dafür, dass ein Korrelationskoeffizient signifikant verschieden von 0 ist (wie oftmals üblich), ist für Konkordanzanalysen in der Regel völlig ungeeignet.
- Die Methode der Konkordanzanalyse hängt vom Skalenniveau der jeweiligen zu untersuchenden Mess- beziehungsweise Beurteilungsverfahren ab.
- Punktwolke, Bland-Altman-Diagramm und Cohens Kappa bilden geeignete Methoden bei Konkordanzanalysen.
- Konkordanzanalysen bilden nicht die Richtigkeit von Mess- beziehungsweise Beurteilungsverfahren ab, sondern legen Übereinstimmungen verschiedener Mess- beziehungsweise Beurteilungsverfahren dar.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt besteht.

Manuskriptdaten
eingereicht: 22. 11. 2010, revidierte Fassung angenommen: 11. 5. 2011

LITERATUR

1. Tetzlaff R, Schwarz T, Kauczor HU, Meinzer HP, Puderbach M, Eichinger M: Lung function measurement of single lungs by lung area segmentation on 2D dynamic MRI. Acad Radiol. 2010; 17: 496–503.
2. Altman DG: Practical statistics for medical research. 1st edition. Oxford: Chapman and Hall 1991; 1–611.
3. Altman DG, Bland JM: Measurement in medicine: the analysis of method comparison studies. The Statistician 1983; 32: 307–17.
4. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–10.
5. Song JW, Oh YM, Shim TS, Kim WS, Ryu JS, Choi CM: Efficacy comparison between (18)F-FDG PET/CT and bone scintigraphy in detecting bony metastases of non-small-cell lung cancer. Lung Cancer 2009; 65: 333–8.
6. du Prel JB, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(19): 335–9.
7. Bortz J, Lienert G A, Boehnke K: Verteilungsfreie Methoden in der Biostatistik. 3rd Edition. Heidelberg: Springer 2008; 1–929.
8. Hilgers R D, Bauer P, Scheiber V: Einführung in die Medizinische Statistik. 2nd edition. Heidelberg: Springer 2007.
9. Altman DG, Machin D, Bryant TN, Gardner MJ: Statistics with confidence. 2nd edition. London: BMJ Books 2000.

Anschrift für die Verfasser
Dr. rer. nat. Robert Kwiecien
Institut für Biometrie und Klinische Forschung (IBKF)
Westfälische Wilhelms-Universität Münster
Albert-Schweitzer-Campus 1 – Gebäude A11, 48149 Münster
robert.kwiecien@ukmuenster.de

SUMMARY

Concordance Analysis—Part 16 of a Series on Evaluation of Scientific Publications

Background: In this article, we describe qualitative and quantitative methods for assessing the degree of agreement (concordance) between two measuring or rating techniques. An assessment of concordance is particularly important when a new measuring technique is introduced.

Methods: We give an example to illustrate a number of simple methods of comparing different measuring or rating techniques, and we explain the underlying principle of each method. We also give further illustrative examples from medical research papers that were retrieved by a literature search.

Results: Methods of comparing different measuring or rating techniques are of two kinds: those with a nominal rating scale and those with a continuous rating scale. We only discuss methods for comparing one measuring or rating technique with another one. Moreover, we point out some common erroneous approaches to concordance analysis.

Conclusion: Concordance analysis is needed to establish the validity of a new diagnostic measuring or rating technique or to demonstrate the near-equivalence of multiple measuring or rating techniques. Erroneous approaches to concordance analysis can lead to false conclusions.

Zitierweise
Kwiecien R, Kopp-Schneider A, Blettner M: Concordance analysis—part 16 of a series on evaluation of scientific publications.
Dtsch Arztebl Int 2011; 108(30): 515–21. DOI: 10.3238/arztebl.2011.0515

The English version of this article is available online:
www.aerzteblatt-international.de