tastdiagnostik

Inter-Rater Reliabilität der Tastdiagnostik

Eine kontrollierte und verblindete crossover-Studie zur Objektivität der Tastdiagnostik

Strebel W, Hebeisen J, Schraner T, Sidler S, Seichert N

Zusammenfassung (Abstract)

Einleitung: Von klinischen Testverfahren wird gefordert, dass sie bei wiederholter Anwendung vergleichbare Ergebnisse liefern (Reliabilität). Die Inter-Rater Reliabilität beschreibt die Unabhängigkeit des Befundergebnisses vom Untersucher (Rater). Bisherige Studienergebnisse verschiedener Autoren deuten auf eine schlechte Reliabilität klinischer, insbesondere manueller Testverfahren.

Ziel der Studie: Bestimmung der Inter-Rater Reliabilität des Befundes „Asymmetrie des Gewebespannungszustandes" gemäss der Tastdiagnostik nach W. Strebel.

Methodik: Randomisierter, verblindeter Crossover-Vergleich mit drei Ratern und 51 gesunden Probanden. Standardisierte Tastung der Gewebespannung an zwei Lokalisationen der Rückenhaut („cranial" und „caudal"). Bestimmung der überzufälligen Übereinstimmung zwischen allen Kombinationen von zwei Ratern mittels des Kappa-Wertes nach Cohen.

Ergebnisse: Die Kappa-Werte für die Inter-Rater Reliabilität liegen zwischen 0.03 und 0.22; als Mittelwert ergibt sich Kappa = 0.10. Die Ergebnisse sind für beide Lokalisationen und alle drei Rater vergleichbar.

Diskussion: Die gefundenen Kappa-Werte liegen alle deutlich unter der Schwelle von Kappa > 0.40, die als Minimum für eine klinisch brauchbare Reliabilität gefordert wird. Somit kann der Befund „Asymmetrie des Gewebespannungszustandes" an Gesunden nicht objektiv erhoben werden. In Studien anderer Autoren zur Reliabilität manueller Testverfahren resultierten ähnliche Ergebnisse.

Im Artikel werden mögliche Ursachen für die unbefriedigende Reliabilität diskutiert.

Einleitung und Fragestellung

Diagnostische Instrumente müssen bestimmte Qualitätseigenschaften aufweisen, damit sie diskriminativ (Unterscheidung von Kranken und Gesunden) und/oder evaluativ (Erfassung einer Veränderung im Zeitverlauf) eingesetzt werden können. Die zwei wichtigsten Qualitätseigenschaften sind:

- Der Test muss bei wiederholter Anwendung – entweder verschiedene Tester oder verschiedene Zeitpunkte – vergleichbare Ergebnisse liefern (Reliabilität).
- Der Test muss das messen, was er zu messen vorgibt (Validität).

Die vorliegende Studie untersucht die Reliabilität einer speziellen Technik der Tastdiagnostik nach W. Strebel. Wie bei anderen klinischen Testverfahren können zwei Arten von Reliabilität definiert werden:

a) Die „Intra-Rater Reliabilität" bewertet die Übereinstimmung des Tests bei Wiederholung durch den selben Beobachter („Rater"). Sie ist ein Mass für die Reproduzierbarkeit des Testergebnisses.
b) Die „Inter-Rater Reliabilität" bewertet die Übereinstimmung des Tests bei Wiederholung durch einen anderen Beobachter. Sie ist ein Mass für die Objektivität des Testverfahrens (Unabhängigkeit vom Untersucher) bzw. für die Übertragbarkeit des Ergebnisses zwischen verschiedenen Untersuchern oder Institutionen.

Für die Anwendung in der Praxis ist die Inter-Rater Reliabilität wichtiger. Ein Testverfahren mit einer guten Inter-Rater Reliabilität hat im Allgemeinen auch eine gute Intra-Rater Reliabilität; umgekehrt ist es aber möglich, dass zwar die Intra-Rater Reliabilität gut, aber die Inter-Rater Reliabilität schlecht ist.

In den letzten Jahren haben mehrere Studien gezeigt, dass manuelle Testverfahren selten eine gute Reliabilität besitzen, insbesondere gilt das für die Inter-Rater Reliabilität. Dies bedeutet, dass viele manuelle Testverfahren nicht objektiv sind, also das Ergebnis von der individuellen Ausführung durch den Rater abhängt.

Diese Studie soll die folgende **Frage** beantworten:

Wie gut ist die Inter-Rater Reliabilität der Tastdiagnostik am Beispiel der Asymmetrie des paravertebralen Gewebespannungszustandes bei gesunden Probanden?

Klinische Ausgangshypothesen:

- Auch sich gesund fühlende Personen weisen tastbare Asymmetrien des Gewebespannungszustandes auf, diese sind jedoch klinisch stumm.
- Eine beim Gesunden bestehende Asymmetrie des Gewebespannungszustandes bleibt lange genug bestehen, um eine Reliabilitätsprüfung mit drei Ratern durch zu führen.
- Bei Patienten ist die Asymmetrie des Gewebespannungszustandes manifester und zeitlich stabiler. Trotz dieser dritten Ausgangshypothese wurde die Studie mit Gesunden durchgeführt, da der Aufwand mit Patienten zu gross erschien.

Studiendesign / Methodik

Studiendesign: Randomisierter, verblindeter Crossover-Vergleich mit drei Ratern und 51 Probanden.

Crossover: Jeder Proband wird von allen drei Ratern untersucht.

Randomisiert: Die Reihenfolge der Untersuchung durch die drei Rater A, B und C wurde für jeden Probanden zufällig festgelegt.

Verblindet: Die Rater hatten keine Informationen über den Zustand der Probanden und über das Ergebnis der anderen Rater. Die Probanden wurden nicht über Zweck und Resultat der Befundung informiert.

Intervention und Rater: Alle 51 Probanden wurden im Abstand von 5 Minuten in unterschiedlicher Reihenfolge von den drei Ratern untersucht. Die Rater beurteilten an zwei vorher definierten Lokalisationen den paravertebralen Spannungszustand des subcutanen Gewebes mit der Hautfaltentechnik. Die zwei Lokalisationen waren „cranial“ (Bereich obere Brustwirbelsäule) und „caudal“ (Bereich Lendenwirbelsäule). Der Rater hatte jeweils zu entscheiden, ob eine Asymmetrie der Gewebespannung vorlag. Im bejahenden Fall musste entschieden werden, auf welcher Seite die Gewebespannung höher war.

Die drei Rater hatten jahrelange Erfahrung mit der Tastdiagnostik; der Ablauf der Untersuchung war vor der Studie standardisiert worden.

Die Studie fand an einem einzigen Tag statt, damit die Rater die Probanden in möglichst unverändertem Zustand untersuchen konnten. Die Probanden wurden in 3er-Gruppen einbestellt, durch den verblindeten Studienleiter informiert, an den beiden Lokalisationen „cranial“ und „caudal“ markiert und randomisiert dem ersten Rater zugewiesen. Danach erfolgte eine 5minütige Erholung und die randomisierte Zuweisung zum nächsten Rater. Die Rater befanden sich in drei unabhängigen Untersuchungsräumen; sie konzentrierten sich allein auf die Untersuchung. Die Instruktion der Probanden und die standardisierte Protokollierung der Ergebnisse wurden von drei unparteiischen „Controllern“ übernommen.

Probanden: 51 Probanden (Alter, Geschlecht) nahmen freiwillig an der Studie teil. Sie waren alle „gesund“ in dem Sinne, dass sie nicht akut in ärztlicher Behandlung waren. Es war zu erwarten, dass der Tastbefund mit etwa gleicher Häufigkeit als „asymmetrisch“ bzw. als „symmetrisch“ beurteilt werden würde.

Medizinische Ethik: Alle Probanden hatten eine schriftliche Einwilligung zur Teilnahme unterschrieben, nachdem sie über Studienablauf und –ziel informiert worden waren („written informed consent“). Darin wurden sie aufgeklärt, dass sie an einer Reliabilitätsstudie teilnehmen würden und hatten ihr Einverständnis gegeben, dass drei Rater nacheinander eine manuelle Befundung am Rücken vornehmen würden. Sämtliche Daten wurden sofort nach Erfassung anonymisiert. Das Studiendesign wurde der zuständigen kantonalen Ethikkommission (Aargau) vorgelegt und ohne weitere Prüfung gut geheissen.

Messgrössen und Statistik

An jeder der beiden Lokalisationen waren drei Befunde möglich: „symmetrisch“, „links mehr als rechts“, oder „rechts mehr als links“. Die Übereinstimmung der drei Rater wurde paarweise überprüft: „A mit B“; „A mit C“ und „B mit C“.

Diese Befunde sind nominale Messwerte, da sie weder ordinal noch rational skalierbar sind. Es sind daher statistische Testverfahren anzuwenden, die für nominale Messgrössen geeignet sind.

Statistische Auswertung: Die Reliabilität wurde mittels dem Kappa-Wert nach Cohen berechnet. Der Kappa-Wert bewertet nur die Anzahl an Übereinstimmungen zwischen zwei Ratern, welche die Anzahl der „zufällig“ erwarteten Übereinstimmungen übersteigt. „Zufällige“ Übereinstimmungen entstehen durch das nicht kausale zusammen Treffen gleicher Beurteilungen bei demselben Probanden. Die Zahl der „zufälligen“ Übereinstimmungen ist nur durch den Anteil an unterschiedlichen Befunden der Rater bestimmt. Das ist ganz analog zum Werfen von Münzen, wobei auch häufig „zufällige“ Übereinstimmungen entstehen.

$$\text{Definition: Kappa} = (\text{Üb} - \text{ÜbZ}) / (\text{Max} - \text{ÜbZ})$$

Üb = Anzahl der beobachteten Übereinstimmungen
ÜbZ = Anzahl der zufälligen Übereinstimmungen
Max = Maximal mögliche Übereinstimmungen

Zahlenbeispiel: Zwei Rater haben bei 43 von 51 Probanden denselben Befund. Anhand der von den beiden Ratern vergebenen Befunde erwartet man 28 „zufällige“ Übereinstimmungen. Dann gilt: $\text{Kappa} = (43 - 28) / (51 - 28) = 15/23 = 0.65$.

Interpretation von Kappa: Der Betrag von Kappa kann zwischen 0.0 (keinerlei Übereinstimmung) und 1.0 (vollständige Übereinstimmung) liegen. In Anlehnung an die international gültigen Empfehlungen gilt für die meisten klinischen Testverfahren „Kappa < 0.4“ als untaugliche, „0.4 < Kappa < 0.6“ als mässige, „0.6 < Kappa < 0.8“ als brauchbare und „Kappa > 0.8“ als gute Reliabilität.

Ergebnisse

Die folgenden Tabellen zeigen die Häufigkeit der Befunde bei den drei Ratern, die Zahl der gezählten („Treffer“) und der zufälligen („Zufall“) Übereinstimmungen sowie der Kappa-Werte für alle möglichen Kombinationen der Rater A, B und C.

Reliabilität der Tastdiagnostik: Cranialer Punkt (n=51)							
	Rater A	**Rater B**	**Rater A**	**Rater C**	**Rater B**	**Rater C**	**Präval**
Links	40	26	40	39	26	39	68.6%
Rechts	3	2	3	0	2	0	3.3%
Null	8	23	8	12	23	12	28.1%
Treffer	**25** (48%)		**34** (67%)		**30** (59%)		
Zufall	24.1		32.5		25.3		
Kappa	**0.03**		**0.08**		**0.18**		

Reliabilität der Tastdiagnostik: Caudaler Punkt (n=51)							
	Rater A	**Rater B**	**Rater A**	**Rater C**	**Rater B**	**Rater C**	**Präval**
Links	17	18	17	22	18	22	37.3%
Rechts	8	4	8	10	4	10	14.4%
Null	26	29	26	19	29	19	48.4%
Treffer	**28** (55%)		**20** (39%)		**21** (41%)		
Zufall	21.4		18.6		19.4		
Kappa	**0.22**		**0.04**		**0.05**		

Kappa cranial = 0.10	**Rater A: Kappa = 0.10**
Kappa caudal = 0.10	**Rater B: Kappa = 0.12**
Kappa gesamt = 0.10	**Rater C: Kappa = 0.09**

Diskussion, Schlussfolgerung

Das grösste überhaupt gemessene Kappa beträgt 0.22, im Durchschnitt über die drei Rater und zwei Lokalisationen ergibt sich Kappa = 0.10.
Die Qualität der drei Rater liegt zwischen 0.09 und 0.12, ist also vergleichbar.

Kappa = 0.10 bedeutet, dass die Übereinstimmung zwischen zwei Ratern bei dieser Untersuchung an Gesunden nur 10% über der zufälligen Übereinstimmung liegt.

Mit einem derart niedrigen Kappa-Wert verbietet sich die diagnostische Interpretation des Tastbefundes, der Befund „Asymmetrie des Gewebespannungszustandes" hat sich bei Gesunden als nicht objektiv erwiesen.

Damit ist die Inter-Rater Reliabilität des Befundes „Asymmetrie des Gewebespannungszustandes" bei Gesunden klinisch ungenügend. Ein solcher Befund kann nicht objektiv erhoben werden, d.h. er ist nicht von einem Rater auf den anderen übertragbar. Verschiedene neuere Studien zur Reliabilität manueller diagnostischer Tests erzielten ähnlich schlechte Ergebnisse (Literatur bei den Verfassern). Offensichtlich sind viele manuelle Befunde schlecht übertragbar, d.h. nicht objektiv. Sie hängen wahrscheinlich in unkontrollierter Art von den unterschiedlichen Situationsumständen ab.

Hypothesen für mögliche Ursachen der unbefriedigenden Reliabilität:

1. *Der Tastbefund ist bei Gesunden zeitlich nicht stabil und damit nicht reliabel*
   Möglicherweise kann sich eine tastbare Asymmetrie der Gewebespannung beim Gesunden innerhalb von Minuten ändern, so dass eine Reliabilitätsprüfung am Gesunden schlechte Resultate liefert.
2. *Die Kategorien „symmetrisch" versus „asymmetrisch" sind ungeeignet für die Beantwortung der Fragestellungt*
3. Eventuell sind andere tastdiagnostische Kriterien wie z.B. die „Bestimmung des therapeutischen Zugangsortes" reliabler.
4. *Nicht einheitliche Durchführung des Tests durch die drei Rater*
   Als mögliche Ursachen für schlechte Reliabilität kommen in Frage:
   - die ergonomische Ausgangsstellung der Probanden war zu wenig standardisiert
   - die Rater hatten nicht dieselbe Schichttiefe der Gewebe getroffen
   - Unterschiede in der Griffdistanz und im erfassten Gewebevolumen
   - unterschiedliche Griffstärke bei der Hautfaltentechnik
5. *Die Befunde folgten zu rasch aufeinander (Einfluss der vorangegangenen Befundung)*
   Möglicherweise wurde der Einfluss der Tastdiagnostik auf den Ausgangszustand unterschätzt. Es ist nicht auszuschliessen, dass der Gewebespannungszustand des Probanden beim 2. und 3. Rater sich wegen der vorgängigen Tastbefunde verändert hatte.
6. *Einfluss des Studiendesigns auf Vegetativum und Psyche der Probanden*
   Es könnte sein, dass einige Probanden von der Tatsache, durch mehrere Rater untersucht zu werden, so beeindruckt waren, dass sie mit Veränderungen des Gewebespannungszustandes reagierten (vergleiche dass bekannte Phänomen der „Praxishypertonie" beim Messen des Blutdrucks).
7. *Ermüdung bzw. Konzentrationsschwäche der Rater*
   Möglicherweise ist ein Rater überfordert, wenn er 50 Tastbefunde innerhalb weniger Stunden erheben muss.

Verbesserungsvorschläge für zukünftige Studien zur Reliabilität der Tastdiagnostik:

1. Das prinzipielle Design „randomisierter crossover-Test“ scheint gut geeignet. Auch die Zahl der Probanden und Rater hat sich bewährt.
2. Studie mit Patienten statt mit Gesunden.
   Ein derartiges Design drängt sich auf, auch wenn der Aufwand sehr gross ist.
   Möglicherweise sind die Tastbefunde nur bei tatsächlichen Patienten zeitlich stabil.
3. Strenge Standardisierung der Tasttechnik und der Ausgangsstellung des Probanden. Ein mögliches Design wäre, dass der Proband/Patient im selben Raum bleibt, und statt dessen die Rater den Raum wechseln.
4. Durchführung der Studie über einen längeren Zeitraum (nur wenige Probanden am selben Tag)
   Dadurch wäre eine eventuelle „Ermüdung“ der Rater vermeidbar.

Dr. rer. nat. Niko Seichert
Wissenschaftlicher Mitarbeiter
in der
Arbeitsgemeinschaft Tastdiagnostik
www.tastdiagnostik.ch **Postkonto 20 - 407719 - 3**
Präsident: Werner Strebel Haltenstr. 1 5444 Künten - Sulz Tel 056 / 496 15 72
Sekretariat: Linda Hämmerle Rooswiesenstr. 40 8155 Niederhasli Tel./Fax 01 / 850 05 34 info@tastdiagnostik.ch
Kurswesen: Jürg Hebeisen Vreniken 4 5454 Bellikon