Forschende Komplementärmedizin
Wissenschaft • Praxis • Perspektiven

Originalarbeit · Original Article

Forsch Komplementmed 2009;16:000–000
DOI: 10.1159/000209858

Published online: April 3, 2009

Zur Reliabilität der Beschreibung morphologischer Merkmale in der traditionellen chinesischen Zungendiagnostik

Yanqing Li[a] Klaus Linde[a] Jingzhang Dai[b] Jisheng Zhang[b] Stefan Hager[b] Reinhard Saller[c] Dieter Melchart[a,c]

[a] Zentrum für naturheilkundliche Forschung, II. Medizinische Klinik und Poliklinik, Klinikum rechts der Isar, Technische Universität, München, Deutschland
[b] Klinik für Traditionelle Chinesische Medizin, Kötzting, Deutschland
[c] Institut für Naturheilkunde, Department für Innere Medizin, Universitätsspital Zürich, Schweiz

Schlüsselwörter
Traditionelle chinesische Medizin · Zungendiagnostik · Reliabilität

Key Words
Traditional Chinese medicine · Tongue diagnostics · Reliability

Zusammenfassung

Hintergrund und Fragestellung: Die Zungenbeurteilung ist ein zentrales diagnostisches Verfahren der traditionellen chinesischen Medizin. In der vorliegenden Studienserie wurde untersucht, inwieweit unterschiedliche Beurteiler bei der morphologischen Beschreibung von Zungenmerkmalen übereinstimmen. **Methoden:** In zwei Pilotstudien (jeweils n = 15 Patienten) und einer größeren Untersuchung (Hauptstudie; n = 101) wurden Digitalfotografien der Zungenbefunde von Klinikpatienten durch zwei bzw. drei erfahrene Beurteiler mithilfe von Formularen verblindet bewertet. Das wichtigste Zielkriterium war die Übereinstimmung über die erwartete zufällige Übereinstimmung hinaus (Cohens Kappa). **Ergebnisse:** In der ersten Pilotstudie wurden in hohem Maße variable Kappa-Werte beobachtet (–0,15 bis 0,76). Bei zahlreichen Merkmalen zeigten sich jedoch nur geringe bis mäßige Übereinstimmungen. In der zweiten, mit einer verbesserten Methode durchgeführten Pilotstudie zeigten sich zwar wiederum in hohem Maße variable Kappa-Werte (–0,10 bis 1,00), bei 7 der 18 erhobenen Merkmale war die überzufällige Übereinstimmung jedoch sehr gut (Kappa ≥ 0,75). In der Hauptstudie lagen die Kappa-Werte zwischen 0,15 und 0,83. **Diskussion:** In den vorliegenden Studien wurden erste Schritte hin zu einer systematischen Untersuchung der Reliabilität der Zungendiagnose im Rahmen der traditionellen chinesischen Medizin unternommen. Die Ergebnisse deuten darauf hin, dass die Beurteilung morphologischer Merkmale eine befriedigende Reliabilität haben dürfte.

Summary

On the Reliability of the Description of Morphological Characteristics in Traditional Chinese Tongue Diagnostics

Background and Objective: The assessment of the tongue is a crucial diagnostic tool of traditional Chinese medicine. In a series of studies we aimed to investigate to what extent independent raters agree in the description of morphological tongue characteristics. **Methods:** In two pilot studies (n = 15 each) and one larger study (n = 101) two to three physicians experienced in traditional Chinese medicine assessed morphological characteristics in digital photos of tongues by use of a rating form and under blind conditions. The primary outcome measure was agreement beyond chance (Cohen's kappa). **Results:** Kappa values varied strongly in the first pilot study (–0.15 to 0.76) and for many items, agreement was weak or moderate. In the second pilot study which used improved methods kappa values still varied considerably (–0.10 to 1.00), but for 7 out of the 18 items assessed there was an excellent agreement (kappa ≥ 0.75). In the confirmatory study, kappa values ranged between 0.15 and 0.83. **Discussion:** The performed studies have to be seen as a first attempt to develop adequate methods in order to systematically investigate the reliability of traditional Chinese tongue diagnostics. The study findings suggest that the description of morphological characteristics within traditional Chinese tongue diagnostics has acceptable reliability.

KARGER

Fax +49 761 4 52 07 14
Information@Karger.de
www.karger.com

Accessible online at:
www.karger.com/fok

PD Dr. med. Klaus Linde
Zentrum für naturheilkundliche Forschung – Forschungsstelle
II. Medizinische Klinik und Poliklinik, Klinikum rechts der Isar
Technische Universität, Wolfgangstr. 8, 81667 München, Deutschland
Klaus.Linde@lrz.tum.de

Einleitung

Die Beurteilung der Zunge spielt in der traditionellen chinesischen Medizin (TCM) im Rahmen der Syndromdiagnostik eine wichtige Rolle. Eine kritische empirische Evaluation der Zungendiagnostik erfolgte bisher jedoch kaum. Eine Evaluation, die sowohl westlichen wissenschaftlichen Ansprüchen als auch der Theorie und Praxis der TCM gerecht wird, ist nicht einfach zu bewerkstelligen. Aus wissenschaftlicher Sicht muss ein diagnostisches Verfahren reliabel und valide sein. Als reliabel wäre die Zungendiagnose zu bewerten, wenn mehrere (angemessen ausgebildete und erfahrene) Ärzte bei der Beurteilung zur gleichen Bewertung kämen. Ein diagnostischer Test wird als valide bezeichnet, wenn er das Vorhandensein bzw. das Nicht-Vorhandensein einer Krankheit weitgehend sicher erkennt [1].

Insbesondere die Prüfung der Validität ist im Falle der Zungendiagnostik (und vieler anderer komplementärer Diagnoseverfahren) problematisch. In der westlichen Medizin erfolgt die Prüfung der Validität durch einen Vergleich des jeweiligen Tests mit dem sogenannten Goldstandard, also der besten und sichersten Möglichkeit, eine Diagnose zu stellen [1]. So würde man z.B. ein einfaches Screeninginstrument für Depressionen mit einer ausführlichen psychiatrischen Exploration vergleichen oder ein Belastungs-EKG zur Untersuchung einer koronaren Herzerkrankung mit einer Koronarangiographie. Im Fall der Zungendiagnostik würde man untersuchen, inwieweit eine alleinige Beurteilung der Zunge bereits die vollständige Syndromdiagnose vorhersagt bzw. inwieweit die Bewertung der Zunge zur Erstellung einer sicheren Syndromdiagnose beiträgt (im Vergleich zu einem diagnostischen Vorgehen ohne Zungendiagnose). Die chinesische Syndromdiagnose, die eher eine funktionale Zustandsbeschreibung des Organismus unter Berücksichtigung konstitutioneller Aspekte ist als eine «feste» Diagnose im westlichen Sinne, eignet sich für eine solche Vorgehensweise nur eingeschränkt. Dementsprechend fehlt eine eindeutige diagnostische Klassifikation, die die terminologischen Voraussetzungen für eine angemessene Validitätsprüfung schaffen würde.

Es ist daher forschungstechnisch einfacher, sich zunächst auf die Prüfung der Reliabilität zu konzentrieren. Dabei sind zwei Ebenen denkbar. Zum einen kann man überprüfen, ob verschiedene Beurteiler die einzelnen morphologischen Merkmale von Zungen übereinstimmend *beschreiben* (z.B. ob mehrere Beurteiler die Zungenfarbe als auffällig rot bewerten). Eine solche Prüfung scheint unproblematisch, ist jedoch nur von eingeschränkter klinischer Relevanz. Zum anderen kann man auf einer komplexeren Ebene untersuchen, ob die *Interpretation* der einzelnen Merkmale oder auch des Gesamtbildes der Zunge übereinstimmt. Diese Ebene hat eine deutlich größere klinische Relevanz. Für die notwendige terminologische Standardisierung sind jedoch umfangreiche Vorarbeiten und Vorprüfungen erforderlich.

In China scheinen sich die Forschungsbemühungen im Zusammenhang mit der Zungendiagnostik auf die Digitalisierung und Automatisierung der Beurteilung von Zungenbildern zu konzentrieren (z.B. [2–4]). Recherchen in PubMed fanden keinerlei Untersuchungen zur Reliabilität oder Validität. Den Autoren sind zwei lediglich als Abstracts verfügbare westliche Untersuchungen zu Aspekten der Reliabilität bekannt [5, 6]. Es wurde daher beschlossen, im Rahmen der Forschungsbegleitung der TCM-Klinik Kötzting – eines seit 1991 bestehenden Krankenhauses mit Akutversorgungsvertrag, in dem chinesische und deutsche Ärzte vorwiegend chronisch kranke Patienten gemeinsam betreuen – zu versuchen, Methoden für eine angemessene wissenschaftliche Überprüfung der Zungendiagnostik zu entwickeln. Das vorliegende Manuskript beschreibt die ersten im Rahmen einer Dissertation erarbeiteten Studien. Ziel der vorliegenden Studienreihe war eine explorative Überprüfung der Übereinstimmung mehrerer Beurteiler (Inter-Rater-Reliabilität) bezüglich der Beschreibung morphologischer Merkmale im Rahmen der Zungendiagnostik.

Methodik

Rahmenbedingungen. Es wurden drei Studien durchgeführt: zwei Pilotserien mit jeweils 15 Patienten bzw. Zungenfotografien und eine Hauptserie mit 101 Zungenfotografien. Die Durchführung der Studien erfolgte in der TCM-Klinik Kötzting. In dieser Klinik mit 75 Betten werden pro Jahr rund 1000 Patienten mit meist chronischen Erkrankungen stationär (mittlere Liegedauer 26 Tage) von 10 chinesischen und 6 deutschen Ärzten primär mit den Methoden der TCM (Akupunktur, chinesische Arzneimitteltherapie, Tuina und Qi Gong) behandelt. Zwischen 2003 und 2005 wurde routinemäßig von allen stationären Patienten der Klinik am Morgen nach der Aufnahme vor dem Frühstück im gleichen Raum unter gleichen Lichtbedingungen von einer deutschen Laborkraft mit einer Digitalkamera (Fuji Film Smart Media, Digital Camera, Fine Pix 2600 Zoom 3V) eine Fotografie der Zunge erstellt. Die Patienten saßen dabei immer auf dem gleichen Stuhl, um den Winkel, in dem fotografiert wurde, möglichst konstant zu halten. Die Fotografien wurden digital und kodiert elektronisch abgelegt. Für die Untersuchung wurden von einer an der Studie unbeteiligten Verwaltungskraft aus dem Pool von Fotodateien zufällig Fotografien gezogen.

Beurteiler. Die Zungenfotos wurden von zwei bis drei Beurteilern streng unabhängig voneinander (Blindbedingungen) mithilfe von Formularen bewertet. Die Bewertung erfolgte ausschließlich auf Basis der Zungenfotos ohne weitere Informationen. Insgesamt waren vier Beurteiler beteiligt. Beurteiler 1 (JD) und Beurteiler 2 (JZ) wirkten an allen drei Teilstudien mit, Beurteiler 3 (YL) ausschließlich an der ersten Pilotstudie und Beurteiler 4 (SH) ausschließlich bei der Hauptstudie. Die Beurteiler 1, 2 und 3 sind chinesische Ärzte mit abgeschlossenem TCM-Studium und langjähriger praktischer Erfahrung. Beurteiler 4 ist der deutsche Chefarzt der Klinik, der über 10 Jahre TCM-Erfahrung verfügt.

Beurteilung der Zungenfotografien. Auf der Basis einer Probeauswertung der Routinedokumentation in der elektronischen Patientenakte der Klinik und einer Konsensusdiskussion im Studienteam wurde ein erstes Formular erstellt, auf dem folgende Merkmale als vorhanden (ja) oder nicht vorhanden (nein) bewertet und dokumentiert werden sollten: Zunge rot, Zunge blass, Zunge bläulich, Zunge rötlich-purpur; Vorhandensein von Flecken, Zahneindrücken, Rissen; dicker Belag, klebriger Belag, sich auflösender Belag, dünner Belag. Diese Aufteilung erfolgte, um in allen Fällen eine klare Ja-Nein-Bewertung zu ermöglichen (z.B. Nichtvorhandensein eines dicken Belags kann sowohl kein Belag als auch dünner

Die einfachste Maßzahl der Übereinstimmung der Bewertung eines diagnostischen Merkmals von zwei Beurteilern als vorhanden oder nicht vorhanden ist die prozentuale Übereinstimmung. Diese berücksichtigt jedoch nicht, dass ein Teil dieser Übereinstimmung zufällig ist bzw. sein kann. Werden z.B. 2 Münzen geworfen, erwartet man in 50% der Würfe, dass beide übereinstimmend Kopf oder Zahl zeigen. Ein Maß zur Quantifizierung der Übereinstimmung jenseits des Zufalls ist der Kappa-Index nach Cohen [7]. Das Maß berücksichtigt außerdem auch, ob ein Merkmal (z.B. eine auffällig rote Zunge) sehr häufig oder sehr selten auftritt. Die Bestimmung erfolgt in 3 Schritten. Im 1. Schritt wird die tatsächlich beobachtete Übereinstimmung ermittelt. Im folgenden Beispiel beträgt diese 0,9 oder 90% (in 5 von 100 Fällen sagen beide Beurteiler, die Zunge sei rot, in weiteren 85 sagen beide, sie sei nicht rot).

	Beurteiler 2		Summe
Beurteiler 1	Rot	Nicht rot	
Rot	5	5	10
Nicht rot	5	85	90
Summe	10	90	100

Im 2. Schritt werden für die vier Kombinationsmöglichkeiten die erwarteten Häufigkeiten berechnet, indem die jeweilige Zeilensumme mit der Spaltensumme multipliziert und dann durch die Gesamtzahl beurteilter Fälle geteilt wird. Für die Beurteilung rot durch beide Beurteiler ist die Zeilensumme 10, die Spaltensumme ebenso 10; die bei nur zufälliger Übereinstimmung erwartete Häufigkeit der Bewertung rot durch beide Beurteiler wäre demnach 1 (10 × 10 / 100). Entsprechend ergibt sich für die beiden Nichtübereinstimmungsmöglichkeiten jeweils die erwartete Häufigkeit 9 (10 × 90 / 100) und für das Urteil nicht rot durch beide Beurteiler die erwartete Häufigkeit 81 (90 × 90 / 100). Die erwartete Gesamtübereinstimmung beträgt somit 0,82 bzw. 82%.

Beispiel mit den erwarteten Häufigkeiten in Klammern

	Beurteiler 2		Summe
Beurteiler 1	Rot	Nicht rot	
Rot	5 (1)	5 (9)	10
Nicht rot	5 (9)	85 (81)	90
Summe	10	90	100

Im 3. Schritt wird der eigentliche Kappa-Index berechnet, indem die erwartete von der beobachteten Übereinstimmung abgezogen (0,9 – 0,82 = 0,08) und dann durch 1 – erwartete Beobachtung (1 – 0,82 = 0,18) geteilt wird. Das errechnete Kappa für das Beispiel ist somit 0,08 / 0,18 = 0,44. Kappa-Werte < 0,40 werden als geringe überzufällige Übereinstimmung klassifiziert, Werte zwischen 0,40 und 0,59 als moderate Übereinstimmung, Werte zwischen 0,60 und 0,74 als gute und Werte ≥0,75 als sehr gute Übereinstimmung [7].

Abb. 1. Beispiel zur Bewertung von Übereinstimmungen mit Cohens Kappa.

Belag bedeuten). Die Beurteilung der Zungenfotos (n = 15) erfolgte ohne weitere Absprache (Rater-Training) zwischen den Beurteilern am gleichen Computerbildschirm zu unterschiedlichen Zeitpunkten.

Aufgrund der Ergebnisse der ersten Pilotstudie wurde das Vorgehen bei der Beurteilung der Zungenmerkmale für die zweite Pilotstudie modifiziert. Um die starke Beeinflussung des Farbeindrucks auf dem Computerbildschirm durch den Blickwinkel des Beurteilers zu vermindern (die Farbe erscheint auf dem Bildschirm umso dunkler, je höher die Augen des Beobachters liegen, d.h. einem größeren oder höher sitzenden Betrachter erscheint die Zunge dunkler als einem kleineren bzw. tiefer sitzenden), wurde ein Gestell gebaut, auf das alle Beurteiler bei der Bewertung der Bilder ihr Kinn legten. Damit wurde erreicht, dass der Blickwinkel für alle Beurteiler weitgehend identisch war. Als weiteres Problem bei der Bewertung der Farbe erwies sich die Tatsache, dass diese in unterschiedlichen Regionen der Zunge variieren kann. Für die Beurteilung aus Sicht der TCM ist es von Bedeutung, ob ein auffälliger Farbbefund z.B. die Zungenspitze oder deren Ränder betrifft. Das Formular wurde daher entsprechend überarbeitet. Ein drittes Problem der ersten Pilotstudie war, dass unklare Befunde unterschiedlich beurteilt wurden. Während ein Beurteiler z.B. nur pathologisch relevante Zahneindrücke als «Zahneindrücke vorhanden» bewertete, zählte ein anderer Beurteiler auch weniger deutliche, aber sichtbare Zahneindrücke. Vor der Bewertung der 15 Fotografien der zweiten Pilotstudie erfolgte daher eine Abstimmung zwischen den Beurteilern auf der Basis einer gemeinsamen Durchsicht der 15 Bilder der ersten Pilotstudie.

Die Beurteilung der Fotografien in der Hauptstudie erfolgte wie in der zweiten Pilotstudie. Es erfolgte lediglich eine Vereinfachung des Formulars beim Item «Zungenfarbe», da einzelne in der Pilotstudie abgefragte Merkmale extrem selten auftreten und daher von geringer praktischer Relevanz schienen.

Statistik. Die erhobenen Daten wurden in der begleitenden Forschungseinrichtung in eine SPSS-Datei eingegeben (Einfacheingabe mit Plausibilitäts- und Stichprobenchecks). Für jedes Beurteilerpaar wurde die prozentuale Übereinstimmung (Zahl der gleichen Bewertungen / Gesamtzahl der Bewertungen × 100) berechnet. Als Maß für die überzufällige Übereinstimmung zwischen einzelnen Paaren von Beurteilern wurde der Kappa-Index nach Cohen berechnet (Abb. 1). Für die Hauptstudie wurden mit StatsDirect statistical software Version 2.5.7 (07.08.2006) Kappa-Indizes mit dem zugehörigen 95%-Konfidenzintervall (95% CI) bestimmt. Zusätzlich wurde mit der Fleiss-Cusick-Erweiterung [7] ein Kappa (im Folgenden als Fleiss' Kappa bezeichnet) mit 95%-CI für die Übereinstimmung zwischen allen drei Beurteilern in der Hauptstudie berechnet.

Ergebnisse

Pilotstudie 1. Die Übereinstimmung variierte zwischen den einzelnen Bewertungskriterien und zum Teil zwischen den einzelnen Beurteilern für ein einzelnes Kriterium deutlich (Tab. 1).

Tab. 1. Ergebnisse der ersten Pilotstudie (n = 15), Übereinstimmungen sind fettgedruckt

Item		Beurteiler 1 vs. 2				Beurteiler 1 vs. 3				Beurteiler 2 vs. 3			
		nein	ja	%	Kappa	nein	ja	%	Kappa	nein	ja	%	Kappa
Zunge rot	nein	**1**	1	47	–0,02	**2**	0	40	0,11	**7**	1	67	0,31
	ja	7	**6**			9	**4**			4	**3**		
Zunge blass	nein	**11**	2	73	–0,15	**7**	6	60	0,24	**7**	6	60	0,24
	ja	2	**0**			0	**2**			0	**2**		
Zunge bläulich	nein	**3**	5	53	0,08	**7**	1	47	–0,13	**5**	0	40	0,07
	ja	2	**5**			7	**0**			9	**1**		
Zunge rötlich-purpur	nein	**14**	1	93	n.b.	**15**	0	100	n.b.	**14**	0	93	n.b.
	ja	0	**0**			0	**0**			1	**0**		
Flecken	nein	**4**	11	27	n.b.	**14**	1	93	n.b.	**4**	0	33	0,05
	ja	0	**0**			0	**0**			10	**1**		
Gedunsen	nein	**7**	0	73	0,48	**5**	2	87	0,73	**5**	6	60	0,31
	ja	4	**4**			0	**8**			0	**4**		
Zahneindrücke	nein	**2**	0	40	0,11	**2**	0	67	0,30	**7**	4	73	0,48
	ja	9	**4**			5	**8**			0	**4**		
Risse	nein	**9**	0	67	0,19	**8**	1	73	0,41	**11**	3	80	0,33
	ja	5	**1**			3	**3**			0	**1**		
Dicker Belag	nein	**11**	0	73	n.b.	**9**	2	87	0,71	**9**	6	60	n.b.
	ja	4	**0**			0	**4**			0	**0**		
Klebriger Belag	nein	**5**	1	73	0,47	**5**	1	73	0,47	***6***	0	93	0,86
	ja	3	**6**			3	**6**			1	**8**		
Auflösender Belag	nein	**12**	0	93	0,76	**12**	0	87	0,44	**13**	0	93	0,63
	ja	1	**2**			2	**1**			1	**1**		
Dünner Belag	nein	**0**	4	73	n.b.	**0**	0	87	n.b.	**1**	3	73	0,19
	ja	0	**11**			2	**13**			1	**10**		

n.b. = Nicht berechenbar (bei ≥2 leeren Feldern in der 4-Felder-Tafel).

So war z.B. die Übereinstimmung bezüglich der Zungenfarbe (mit Ausnahme der Bewertung rötlich-purpur: ja/nein) zwischen den drei Gutachtern gering. Das Vorhandensein von Flecken wurde von den Beurteilern 1 und 3 sehr ähnlich bewertet, Beurteiler 2 kam zu deutlich abweichenden Ergebnissen. Bei den Merkmalen «Zahneindrücke» und «Risse» war die Übereinstimmung gering bis mäßig. Bezüglich der Beurteilung der Beläge stimmten die Bewertungen zum Teil besser überein.

Pilotstudie 2. In der zweiten Pilotstudie zeigten sich zum Teil sehr gute Übereinstimmungen (Tab. 2). Bei der Farbbeurteilung kam es jedoch immer noch zu deutlichen Diskrepanzen, insbesondere bei der Frage, ob Zungenränder und Zungenkörper blass seien. Da einzelne Merkmale wie eine blasse Zungenspitze oder Risse sehr selten vorkamen, sind die entsprechenden Bewertungen kaum zu interpretieren.

Hauptstudie. Die Patienten, deren Zungenfotos zur Auswertung herangezogen wurden, waren zu 63% weiblich, der Altersmedian lag bei 52 Jahren (Minimum 20, Maximum 87 Jahre). 52% der Patienten litten an chronischen Schmerzsyndromen, 12% an Erkrankungen des Nervensystems und 9% an Erkrankungen des Magen-Darm-Systems (27% sonstige Erkrankungen). Die Ergebnisse der Zungenbewertung sind in Tabelle 3 dargestellt. Bei der Beurteilung der Farbmerkmale war die überzufällige Übereinstimmung zwischen den Beurteilern gering bis moderat (Kappa-Werte 0,19–0,56), ebenso bei den Merkmalen «Flecken» (0,15–0,30), «dicker Belag» (0,26–0,40), «klebriger Belag» (0,17–0,49) und «dünner Belag» (0,34–0,51). Gute bis sehr gute Übereinstimmungen ergaben sich insbesondere bei Zahneindrücken (0,65–0,83) und bei der Bewertung der Belagfarbe (0,64–0,83). Die Übereinstimmung bei den Merkmalen «gedunsene Zunge» (0,56–0,65), «Risse» (0,43–0,60), und «auflösender Belag» (0,55–0,78) schwankte zwischen moderat und sehr gut. Wurden die Angaben aller drei Beurteiler mit Fleiss' Kappa bewertet, ergab sich für fünf Merkmale eine geringe, für sechs eine moderate und für drei Merkmale eine gute Übereinstimmung.

Diskussion

Zusammenfassung der Hauptergebnisse. Die vorliegende Untersuchungsreihe zeigt, dass eine Untersuchung der Reliabilität der morphologischen Beschreibung von Zungenmerkmalen möglich, aber mit erheblichen methodischen Schwierigkeiten verbunden ist. In den Voruntersuchungen wurde deutlich, dass ohne eine angemessene Standardisierung der Beobachtungsbedingungen und ohne Vorabsprachen zwischen den Beurteilern über die Bewertung unklarer Befunde keine sinnvolle Studiendurchführung möglich ist. Die Ergebnisse der Hauptstudie zeigen erwartungsgemäß mit wenigen Ausnahmen für alle Items statistisch signifikante Übereinstimmungen über den Zufall hinaus. Allerdings ist das Aus-

Tab. 2. Ergebnisse der zweiten Pilotstudie, Übereinstimmungen sind fettgedruckt

Item		Beurteiler 1 vs. 2			
		nein	ja	%	Kappa
Zungenspitze rot	nein	**1**	1	93	0,63
	ja	0	**13**		
Zungenspitze blass	nein	**15**	0	100	n.b.
	ja	0	**0**		
Zungenspitze bläulich	nein	**15**	0	100	n.b.
	ja	0	**0**		
Zungenränder rot	nein	**12**	2	87	0,44
	ja	0	**1**		
Zungenränder blass	nein	**5**	8	47	0,14
	ja	0	**2**		
Zungenränder bläulich	nein	**14**	1	93	n.b.
	ja	0	**0**		
Zungenkörper rot	nein	**14**	0	100	1,00
	ja	0	**1**		
Zungenkörper blass	nein	**6**	6	60	0,29
	ja	0	**3**		
Zungenkörper bläulich	nein	**6**	0	67	0,39
	ja	5	**4**		
Flecken	nein	**12**	0	87	0,44
	ja	2	**1**		
Zunge gedunsen	nein	**9**	0	100	1,00
	ja	0	**6**		
Zahneindrücke	nein	**9**	0	100	1,00
	ja	0	**6**		
Risse	nein	**14**	0	100	1,00
	ja	0	**1**		
Dicker Belag	nein	**12**	2	80	–0,10
	ja	1	**0**		
Klebriger Belag	nein	**8**	0	100	1,00
	ja	0	**7**		
Auflösender Belag	nein	**14**	0	100	1,00
	ja	0	**1**		
Dünner Belag	nein	**5**	0	80	0,61
	ja	3	**7**		
		gelb	weiß		
Belagfarbe	gelb	**2**	0	100	1,00
	weiß	0	**13**		

n.b. = Nicht berechenbar (bei ≥2 leeren Feldern in der 4-Felder-Tafel).

maß dieser Übereinstimmung (quantifiziert durch den Kappa-Index) bei der verwendeten Untersuchungsmethode für die einzelnen Items sehr variabel und in vielen Fällen nur moderat. Bei Merkmalen wie z.B. «Zahneindrücken» ergaben sich hohe Übereinstimmungen, während sich z.B. die Farbbeurteilung der Zungen nicht zuletzt aus technischen Gründen bei der Bewertung am Computerbildschirm als schwierig erwies.

Stärken und Schwächen der verwendeten Methode. Das methodische Vorgehen in den Pilotstudien und der Hauptstudie wurde in enger Absprache mit Ärzten entwickelt, die in der TCM sehr erfahren sind. Sowohl die beiden Pilotstudien wie auch die Hauptstudie wurden unter strikten Blindbedingungen durchgeführt, d.h. die Beurteiler bewerteten die Zungenfotografien streng unabhängig voneinander und hatten keine weiteren Informationen zu den Patienten. Da die Bilder aus einem Pool von mehreren Tausend Fotografien ausgewählt wurden, ist es – vielleicht mit Ausnahme weniger sehr auffälliger Zungen – sehr unwahrscheinlich, dass sich die Beurteiler an die Patienten erinnerten. Dateneingabe und -auswertung wurden in einem unabhängigen Universitätsinstitut durchgeführt. Während in den Pilotstudien aufgrund der begrenzten Ressourcen nur 15 Fotografien verwendet wurden, ist die Hauptstudie mit 101 Fotografien angemessen umfangreich und erlaubt somit eine zuverlässige Aussage hinsichtlich der Übereinstimmung zwischen den teilnehmenden Beurteilern unter den gegebenen Studienbedingungen.

Trotz des systematischen Ansatzes trat bei der Untersuchungsreihe eine Vielzahl von Problemen auf, die zum Teil auch in zukünftigen Studien schwer zu lösen sein dürften. In den vorliegenden Studien erfolgte die Beurteilung aus Gründen der Durchführbarkeit und Standardisierbarkeit auf der Basis von digitalen Zungenfotografien und nicht direkt am Patienten. Die Verwendung von Fotografien und deren Beurteilung am Bildschirm erwies sich jedoch als problematisch. Obwohl die Zungenfotografien unter weitgehend standardisierten Bedingungen entstanden (gleiche Tageszeit, gleicher Raum, gleiche Sitzposition), war ihre Qualität nicht einheitlich. So führen z.B. bereits geringe Lageveränderungen oder die Art und das Ausmaß, wie weit der Patient die Zunge herausstreckt, zu deutlichen Unterschieden in der Ausleuchtung. Ein grundsätzliches Problem bei der Verwendung einzelner Zungenfotografien ist außerdem, dass im Gegensatz zur Beurteilung am realen Patienten, nur ein Eindruck aus einer einzigen Perspektive möglich ist. Die Entscheidung für eine Beurteilung am Bildschirm war getroffen worden, um eine mögliche weitere Verfälschung durch den Ausdruck der Fotografien zu vermeiden. Die Beurteilung am Bildschirm erwies sich jedoch durch den enormen Einfluss der Bildschirmqualität und des Blickwinkels als unbefriedigend. Um die Fehlerquellen zu minimieren, mussten daher alle Beurteilungen am gleichen Bildschirm und in Pilotstudie 2 und der Hauptstudie in einer extra angefertigten Vorrichtung zur weitgehenden Standardisierung des Blickwinkels erfolgen.

In der Praxis der Zungendiagnose spielt der Gesamteindruck der Zunge eine große Rolle. Eine detaillierte morphologische Beschreibung der Zunge erfolgt in der Routinedokumentation selten. Vielmehr erfolgt eine Interpretation der Zungenzeichen unter Einbezug der übrigen Informationen, Eindrücke und Befunde am Patienten. In der vorliegenden Studienserie wurde der Großteil dieses Vorgehens ausgeblendet. Um eine einfache Auswertung der Daten zu gewährleisten, waren auf dem Beurteilungsformular einzelne Merkmale als vorhanden oder nicht vorhanden zu beurteilen. Da die Auswahl der Zungenfotografien zufällig erfolgte, ist davon auszugehen, dass eine erhebliche Zahl uneindeutiger

Tab. 3. Ergebnisse Hauptstudie, Übereinstimmungen sind fettgedruckt

Item		Beurteiler 1 vs. 2				Beurteiler 1 vs. 4				Beurteiler 2 vs. 4				3 Beurteiler
		nein	ja	%	Kappa (95%-CI)	nein	ja	%	Kappa (95%-CI)	nein	ja	%	Kappa (95%-CI)	Fleiss' Kappa (95%-CI)
Zungenspitze rot	nein	**7**	8	85	0,39	**12**	3	80	0,43	**12**	2	80	0,44	0,41
	ja	7	**78**		(0,15 bis 0,64)	17	**68**		(0,24 bis 0,63)	18	**69**		(0,25 bis 0,63)	(0,30 bis 0,52)
Zungenränder rot	nein	**63**	13	84	0,56	**72**	7	84	0,50	**64**	2	83	0,56	0,51
	ja	2	**19**		(0,39 bis 0,73)	9	**12**		(0,29 bis 0,71)	15	**17**		(0,38 bis 0,74)	(0,40 bis 0,62)
Zungenkörper rot	nein	**82**	14	85	0,24	**90**	5	93	0,33	**81**	1	87	0,42	0,32
	ja	1	**3**		(–0,01 bis 0,48)	2	**2**		(–0,04 bis 0,7)	12	**6**		(0,18 bis 0,67)	(0,20 bis 0,43)
Zungenkörper blass	nein	**50**	16	75	0,45	**60**	6	82	0,56	**53**	5	74	0,44	0,46
	ja	8	**23**		(0,27 bis 0,63)	12	**20**		(0,39 bis 0,74)	20	**20**		(0,27 bis 0,62)	(0,34 bis 0,57)
Zungenkörper bläulich	nein	**36**	18	76	0,52	**16**	36	57	0,19	**17**	24	70	0,36	0,30
	ja	6	**39**		(0,36 bis 0,68)	6	**39**		(0,03 bis 0,34)	6	**52**		(0,18 bis 0,53)	(0,18 bis 0,41)
Flecken	nein	**55**	2	70	0,3	**38**	16	62	0,15	**51**	25	66	0,18	0,14
	ja	27	**12**		(0,13 bis 0,46)	18	**17**		(0,04 bis 0,35)	6	**8**		(–0,01 bis 0,36)	(0,03 bis 0,25)
Zunge gedunsen	nein	**30**	2	83	0,65	**24**	8	81	0,56	**30**	14	81	0,61	0,60
	ja	15	**54**		(0,50 bis 0,80)	11	**56**		(0,39 bis 0,73)	5	**50**		(0,46 bis 0,77)	(0,49 bis 0,72)
Zahneindrücke	nein	**59**	4	92	0,83	**49**	12	83	0,65	**50**	11	85	0,69	0,70
	ja	4	**34**		(0,72 bis 0,94)	5	**33**		(0,51 bis 0,80)	4	**34**		(0,55 bis 0,84)	(0,59 bis 0,81)
Risse	nein	**79**	4	91	0,60	**68**	15	82	0,48	**68**	17	80	0,43	0,50
	ja	5	**11**		(0,39 bis 0,81)	3	**14**		(0,29 bis 0,68)	3	**12**		(0,24 bis 0,63)	(0,39 bis 0,61)
Dicker Belag	nein	**79**	17	83	0,32	**90**	6	91	0,26	**78**	1	84	0,40	0,32
	ja	0	**5**		(0,1 bis 0,53)	3	**2**		(–0,07 bis 0,6)	15	**7**		(0,17 bis 0,62)	(0,21 bis 0,43)
Klebriger Belag	nein	**47**	12	73	0,42	**45**	14	62	0,17	**53**	8	77	0,49	0,36
	ja	15	**26**		(0,24 bis 0,60)	24	**17**		(–0,02 bis 0,36)	15	**23**		(0,32 bis 0,67)	(0,25 bis 0,48)
Auflösender Belag	nein	**88**	1	97	0,78	**84**	4	93	0,55	**89**	4	94	0,59	0,50
	ja	2	**6**		(0,55 bis 1,0)	3	**5**		(0,25 bis 0,85)	2	**5**		(0,30 bis 0,89)	(0,39 bis 0,62)
Dünner Belag	nein	**26**	8	78	0,53	**20**	14	70	0,34	**24**	16	72	0,41	0,43
	ja	14	**53**		(0,36 bis 0,70)	16	**51**		(0,15 bis 0,54)	12	**49**		(0,23 bis 0,60)	(0,32 bis 0,54)
		gelb	weiß			gelb	weiß			gelb	weiß			
Belagfarbe	gelb	**7**	1	95	0,71	**6**	2	94	0,64	**8**	1	97	0,83	0,71
	weiß	2	**88**		(0,48 bis 0,95)	2	**88**		(0,37 bis 0,90)	0	**90**		(0,64 bis 1,00)	(0,61 bis 0,82)

n.b. = Nicht berechenbar (bei ≥2 leeren Feldern in der 4-Felder-Tafel).

Befunde selektiert wurde. In der ersten Pilotstudie erfolgten keinerlei Absprachen, wie «Grenzwertbefunde» zu bewerten seien. In der zweiten Pilotstudie und in eingeschränktem Ausmaß nochmals vor der Hauptstudie erfolgten mündliche Absprachen auf der Basis von Bildern aus den jeweiligen Vorstudien. Derartige «begrenzte» Absprachen scheinen für eine sinnvolle Studiendurchführung unumgänglich. Allerdings erfolgten die Festlegungen in den vorliegenden Studien nicht schriftlich.

Andere Untersuchungen zum Thema. Wie bereits angemerkt, scheint der Untersuchung der Reliabilität der Zungendiagnose bisher kaum Aufmerksamkeit geschenkt worden zu sein. Entsprechende Arbeiten aus China konnten auch mithilfe chinesischer Fachleute nicht identifiziert werden. Die beiden nur als Abstracts publizierten westlichen Untersuchungen [5, 6] können nur als Pilotstudien interpretiert werden. Biltstein und Bonzak [5] ließen drei erfahrene Akupunkteure 31 Merkmale bei 20 Zungenbildern beurteilen. Die ermittelten Kappa-Werte lagen zwischen 0,21 und 0,6. Unbefriedigende Kappa-Werte betrafen besonders die Beurteilung der Zungenspitze. Gareus et al. [6] versuchten in einem ersten Schritt, möglichst optimale und standardisierte Bedingungen für die Zungenfotografien zu erreichen, z.B. um immer einen möglichst identischen Lichteinfall zu gewährleisten. Zwei Beurteiler bewerteten dann den Zungenkörper und den Belag direkt am Patienten sowie am Computerbild und nach einer längeren Zeitspanne nochmals das Computerbild. Die Kappa-Indizes beim Vergleich von Computerbild und Direktbeurteilung am Patienten waren 0,35 in Bezug auf den Zungenkörper und 0,34 in Bezug auf den Zungenbelag beim ersten und 0,65 bzw. 0,62 beim zweiten Beurteiler. Die Werte für die Test-Retest-Reliabilität betrugen 0,53 bzw. 0,48 für den ersten und 0,65 bzw. 0,62 für den zweiten Beurteiler. Über das Internet konnte ein Artikel gefunden werden, der eine Übersicht über den Stand der Zungendiagnoseforschung in China gibt [8]. Demnach untergliedert sich die Forschung in diesem Bereich

dort in folgende Bereiche: «Die Systematisierung der schriftlichen Informationen zur Zungendiagnose, die Untersuchung von gesunden Probanden zur Aufstellung eines Standards, die experimentelle Erforschung der Zungendiagnose, die Methodik der Diagnose des Zungengrundes, die Erforschung des Zungenfarbgrades, die Untersuchung der Zungenwärme mit Infrarotmessungen, die Entwicklung von Diagnosegeräten zur Zungendiagnose und schließlich die experimentelle und klinische Untersuchung der Zungenveränderung bei Krankheiten.» Untersuchungen zur Reliabilität werden in der Übersicht nicht genannt.

Schlussfolgerung und Empfehlungen für zukünftige Forschungsvorhaben. Die vorliegenden Ergebnisse deuten darauf hin, dass die Beurteilung morphologischer Merkmale im Rahmen der Zungendiagnose eine befriedigende Reliabilität haben dürfte. Eine weitergehende Schlussfolgerung ist angesichts der geringen Zahl vorliegender Untersuchungen, aufgrund der methodischen Schwierigkeiten und aufgrund der unklaren Übertragbarkeit der Ergebnisse auf andere Personen, die TCM praktizieren, nicht vertretbar. Angesichts der Verbreitung der TCM und der Relevanz der Zungendiagnostik im Rahmen dieses Therapiesystems ist eine weitere Erforschung wünschenswert. Studien zur Reliabilität der morphologischen Beurteilung sind sinnvoll, wenn auch nur als erster Schritt hin zu Untersuchungen, die die Reliabilität der Interpretation der Zungenzeichen überprüfen, sowie Studien, die sich mit der Validität der Zungediagnose im Rahmen der TCM auseinandersetzen. Angesichts der offensichtlich umfangreichen Bemühungen zur Digitalisierung und Objektivierung der Dokumentation in China [1–3, 8] muss versucht werden, eine bessere Übersicht über den Stand der Forschung zu erlangen und Erkenntnisse für weitere Untersuchungen auch im Westen zu nutzen.

Dank

Die Autoren danken Prof. Dr. S. Wagenpfeil (Institut für medizinische Statistik und Epidemiologie der Technischen Universität München) für die Berechnung der Kappa-Indizes für die Hauptstudie.

Conflict of Interests

Keine.

Literatur

1 Fletcher RW, Fletcher SW: Clinical Epidemiology, ed 4. Philadelphia, Lippincott, Williams and Wilkins, 2005.
2 Chiu CC: A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue. Comput Methods Prog Biomed 2000;61:77–89.
3 Wei BC, Shen LS, Wang YQ, Wang YG, Wang AM, Zhao ZX: A digital tongue image analysis instrument for traditional Chinese medicine [in Chinese]. Zhongguo Yi Liao Qi Xie Za Zhi 2002; 26:164–9.
4 Zhou Y, Yang J, Shen L: Methodological study on digitalization of tongue image in traditional Chinese medical diagnosis [in Chinese]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi 2004;21:917–20.
5 Blitstein R, Bonzak S: Interobserver reliability in traditional Chinese medicine tongue observation. J Altern Complement Med 2006;12:206.
6 Gareus IK, Tan L, Lüdtke R, Niggemeier C, Li FM, Bäcker M, Lehmann C, Klose P, Spahn G, Dobos GJ: Introducing a computer-assisted, digital tongue-imaging device to standardise and evaluate traditional Chinese tongue diagnosis. Focus Alter Complement Ther 2005;10(suppl 1):20–21.
7 Fleiss JL, Levin B, Paik MC: Statistical methods for rates and proportions, ed 3. New York, Wiley, 2003.
8 Zhang BL: Erforschung der Systematisierung und Objektivierung der TCM-Zungendiagnose [ohne Jahreszahl]. *www.tcminter.net/Artikel/Zungendiagnose.htm.*