7 Diskussion

7.1 Faltenmessung und Bedeutung der Reproduzierbarkeit

Falten und ihre Ausprägung zu messen ist schwierig. Zwar gibt es verschiedene Ansätze für ein objektives Messverfahren, entsprechende Methoden sind aber entweder sehr aufwendig und/ oder teuer und daher im medizinischen Alltag kaum einsetzbar. Hierzu zählt z.B. die Methode von Hatzis (Hatzis 2004), bei der man einen Silikonabdruck der relevanten Faltenregion macht, diesen mit einem andersfarbigen Silikon ausgießt, das ganze in fünf Millimeter dicke Scheiben schneidet, anschließend unter einem Mikroskop fotografiert oder scannt, exakt vermisst und zuletzt das Faltenvolumen berechnet. Ebenso aufwendig ist die Messung der Muskelaktivität (Lowe et al. 1996) oder des Augenbrauenabstandes bei der Behandlung der Glabella mit BoNT (Heckmann et al. 2001).

In den meisten klinischen Studien zur Therapie von mimischen Falten wird daher ein subjektives Verfahren, eine klinische Skala, verwendet. Klinische Skalen sind in der Medizin weit verbreitet. Sie sind meist einfach anzuwenden und nicht kostenintensiv. Das Problem bei solchen Verfahren ist, dass jeder Beobachter die Wirklichkeit (hier: die Falten) mehr oder weniger unterschiedlich wahrnimmt und Bewertungen folglich mehr oder weniger voneinander abweichen. Selbst bei ein und demselben Beobachter kann es in Abhängigkeit von inneren und äußeren Einflüssen zu erheblichen Schwankungen kommen. Das macht einen Vergleich der getroffenen Bewertungen schwierig und kann in der klinischen Anwendung weit reichende Folgen haben. Es ist also nötig, ein solches subjektives Verfahren vor seinem Gebrauch auf seine Leistungsfähigkeit, d.h. auf seine Reproduzierbarkeit hin zu untersuchen. Nur wenn ein gewisses (Mindest-) Maß an Konstanz in der Anwendung bei den Beobachtern vorhanden ist, werden die getroffenen Bewertungen ausreichend objektivierbar und damit auswert- und vergleichbar. Nur ein Verfahren, das ausreichend reproduzierbare Ergebnisse sowohl für den einzelnen Beobachter als auch für mehrere unterschiedliche Beobachter liefert, kann zu qualitativ hochwertigen und aussagekräftigen Ergebnissen führen (Bland u. Altman 2002).

7.2 Instrumente zur Messung der Reproduzierbarkeit

Als einfachstes Instrument zur Überprüfung der Reproduzierbarkeit kann man die prozentuale Übereinstimmung der getroffenen Bewertungen unter mehreren Beobachtern zu demselben Zeit-

punkt (interindividuelle Reproduzierbarkeit) oder bei einem Beobachter zu zwei unterschiedlichen Zeitpunkten (intraindividuelle Reproduzierbarkeit) benutzen. Dabei werden jedoch die zufällig übereinstimmenden Bewertungen nicht erfasst. Darin liegt ein erheblicher methodischer Nachteil. So kann unter Umständen eine hohe Übereinstimmung gemessen werden, obwohl diese in Wahrheit tatsächlich wesentlich geringer ist. Um diese Ungenauigkeit auszuschließen, wurde in der vorliegenden Arbeit die wesentlich genauere Kappa-Statistik verwendet. Diese berücksichtigt im Gegensatz zu der prozentualen Übereinstimmung die zu erwartenden zufällig übereinstimmenden Bewertungen und vermittelt so ein deutlich genaueres Bild von der tatsächlichen Reproduzierbarkeit (siehe unter Material und Methoden 5.2). Auf die zusätzliche Berechnung der qualitativ schlechteren und letztlich nicht aussagekräftigen prozentualen Übereinstimmung wurde verzichtet.

7.3 Bedeutung der Wichtung

Bei polytomen Daten offenbart sich eine Schwäche des Kappa-Koeffizienten: er misst nur exakt konkordante Urteile ungeachtet des Grades der ungefähren Übereinstimmung.

Kappa behandelt alle Diskordanzen als identisch. Diskordante Bewertungen können aber unterschiedlich stark voneinander abweichen.

Um diese Schwäche auszugleichen wurde das gewichtete Kappa entwickelt. Das Prinzip des gewichteten Kappa ist es, dass, wenn wie in unserem Falle Skalen-Werte zwischen 0 und 3 zur Verfügung stehen, eine Diskordanz zwischen den Werten 0 und 1 als weniger gravierend zu bewerten ist als eine Diskordanz zwischen 0 und 2 oder gar 0 und 3. Entsprechend weist man geringen Diskordanzen kleine Gewichte und starken Diskordanzen große Gewichte zu. Dabei sollte man willkürliche Gewichtungen unbedingt vermeiden und stattdessen auf Standardgewichtungen zurückgreifen. Eine dieser Standardgewichtungen wurde von Cicchetti und Allison vorgeschlagen (Cicchetti u. Allison 1971). Diese wurde hier verwendet.

Die Wichtung wirkt sich wie folgt auf den Kappa-Wert aus: wenn die Abweichungen hauptsächlich aus geringer gewichteten Werten bestehen, ergibt sich ein höherer Kappa-Wert. Bestehen sie hauptsächlich aus stärker gewichteten Werten, resultiert ein niedriger Kappa-Wert.

Ein Vergleich von gewichtetem und ungewichtetem Kappa zeigt demnach, in welchem Bereich sich die vorhandenen Abweichungen konzentrieren.

7.4 Reproduzierbarkeit der klinischen Skalen unter den Prüfärzten

Die Ergebnisse unter den Prüfärzten zeigten insgesamt eine mäßige interindividuelle und eine gute intraindividuelle Reproduzierbarkeit für den Gebrauch der Vier-Punkte-Skalen zur klinischen Beurteilung periokulärer Falten, sowohl in entspanntem Zustand als auch in kontrahiertem Zustand.

Bei der interindividuellen Reproduzierbarkeit zeigten sich kaum Unterschiede in den Ergebnissen für den entspannten und den kontrahierten Zustand. Die größte Reproduzierbarkeit zeigte sich in beiden Fällen für den niedrigsten Skalen-Wert 0 (keine Faltenbildung), die niedrigste Reproduzierbarkeit jeweils für den Skalen-Wert 2 (mäßige Faltenbildung), sie war im entspannten und im kontrahierten Zustand nur mäßig. In entspanntem Zustand war die Reproduzierbarkeit für den Skalen-Wert 3 (starke Faltenbildung) größer als für den Skalen-Wert 1 (leichte Faltenbildung). In kontrahiertem Zustand verhielt es sich genau anders herum. Das lässt vermuten, dass es nach der Schulung am schwersten war, zwischen leichter und mäßiger Faltenbildung in entspanntem und zwischen mäßiger und starker Faltenbildung in angespanntem Zustand zu unterscheiden.

Die Übereinstimmung in den Beurteilungen war bei den weiblichen Teilnehmern größer als bei ihren männlichen Kollegen.

Die intraindividuelle Reproduzierbarkeit war bei allen Teilnehmern gut bis sehr gut, unabhängig davon, ob die Faltenbildung in entspanntem oder in angespanntem Zustand begutachtet wurde. Einzige Ausnahme war das ungewichtete Kappa für den entspannten Zustand bei einem der männlichen Teilnehmer, das mit einem Wert von 0,47 nur mäßig war. Die gewichteten Kappa-Werte waren allesamt höher als die ungewichteten. Das lässt darauf schließen, dass die Abweichungen meist nur gering waren.

Ebenso wie bei der interindividuellen Reproduzierbarkeit war auch bei der intraindividuellen Variabilität die Konstanz bei den weiblichen Teilnehmern tendenziell höher als bei den männli-

chen Teilnehmern. So erreichten die Frauen im Gegensatz zu den Männern ausschließlich gute bis sehr gute Reproduzierbarkeitswerte.

7.5 Reproduzierbarkeit der klinischen Skalen unter den Experten

Unter den Experten zeigte sich insgesamt eine mäßige bis gute Reproduzierbarkeit für den Gebrauch der Vier-Punkte-Skalen zur klinischen Beurteilung periokulärer Falten. Die Ergebnisse für den entspannten Zustand waren besser als die für den kontrahierten Zustand.

Die interindividuelle Reproduzierbarkeit war für den entspannten Zustand bei dem Skalen-Wert 0 (keine Faltenbildung) am höchsten, bei 1 (leichte Faltenbildung) am niedrigsten. Für den kontrahierten Zustand wies der Skalen-Wert 1 (leichte Faltenbildung) die höchste Reproduzierbarkeit auf, der Skalen-Wert 2 (mäßige Faltenbildung) dagegen die niedrigste. Für die Experten war es offenbar am schwierigsten, zwischen leichter und mäßiger Faltenbildung in entspanntem und zwischen mäßiger und starker Faltenbildung in angespanntem Zustand zu unterscheiden. Dies bestätigte die Ergebnisse aus der vorangegangenen Reproduzierbarkeits-Studie mit den Prüfärzten (siehe oben 7.3).

Mit Ausnahme einer Expertin, die mit einem ungewichteten Kappa-Wert von 0,35 für den angespannten Zustand eine nach Altman nur schlechte Reproduzierbarkeit erreichte, lagen die übrigen intraindividuellen Kappa-Werte im Bereich mäßiger bis guter Reproduzierbarkeit.

Dabei waren die gewichteten Kappa-Werte ausnahmslos größer als die ungewichteten. Hieraus lässt sich folgern, dass die Abweichungen in den Beurteilungen auch hier meist klein und von geringem Gewicht waren.

Die Ergebnisse der Untersuchung bei den Prüfärzten wurden somit durch die Ergebnisse unter den Experten im Wesentlichen bestätigt.

7.6 Vergleich der Ergebnisse von Prüfärzten und Experten

Insgesamt betrachtet sind die Ergebnisse unter den Prüfärzten tendenziell etwas besser als die Ergebnisse unter den Experten. Es fällt u.a. auf, dass die niedrigste intraindividuelle Reproduzierbarkeit in den beiden Studien überhaupt, die zugleich auch die einzige schlechte war, von einer der Expertinnen erzielt wurde. Dies ist insofern unerwartet, weil man vermuten könnte,

dass die Experten aufgrund ihrer umfangreichen Vorkenntnisse in der Beurteilung von Gesichtsfalten gegenüber den erstmalig geschulten Prüfärzten eine höhere Reproduzierbarkeit ihrer Bewertungen erreichen. Offenbar haben die Experten jedoch aus ihren Vorkenntnissen bei der Faltenbewertung keinen Vorteil in Bezug auf die Reproduzierbarkeit ziehen können. Eine mögliche Erklärung für das etwas schlechtere Abschneiden der Experten könnte es sein, dass sie wegen ihrer Vorkenntnisse eine individuell bereits gefestigte Vorstellung von der Klassifizierung in "keine", "leichte", "mäßige" und "starke Faltenbildung" besaßen, die nicht exakt den hier verwendeten Ausprägungsgraden entsprachen, und infolgedessen die Falten trotz Schulung mehr aufgrund ihrer eigenen Vorstellungen und weniger anhand des Atlas bewerteten.

Bei dem Vergleich der Ergebnisse von den Experten einerseits und den Prüfärzten andererseits muss allerdings auch bedacht werden, dass die Anzahl der Experten aus organisatorischen Gründen mit nur vier Teilnehmern relativ klein war und deshalb eine vergleichende Darstellung nur von begrenztem Aussagewert sein kann.

Die tendenziell besseren Ergebnisse der Prüfärzte könnten so gedeutet werden, dass gerade auch Nicht-Experten (oder "Laien") in der Beuteilung von periokulären Falten bei der klinischen Anwendung der beiden hier untersuchten Skalen imstande sind, Bewertungen von guter Reproduzierbarkeit abzugeben.

7.7 Vergleich mit anderen Studien

Die Ergebnisse sind vergleichbar mit denen anderer Studien (vgl. Tabelle 5), wenngleich diese sich mit anderen Bereichen des Gesichts befassten und teilweise anstatt zweier getrennter Skalen nur eine gemeinsame Skala für den entspannten und den angespannten Zustand verwendeten (Lemperle et al. 2001, Honeck et al. 2003, Day et al. 2004, Kim et al. 2004).

Da die Anwendung zweier getrennter Skalen gegenüber der Verwendung einer einzelnen gemeinsamen Skala die Reproduzierbarkeit nicht verschlechtert, sollten bei zukünftigen Studien getrennte Skalen für den entspannten und den angespannten Zustand verwendet werden. Diese getrennten Skalen ermöglichen eine genauere Differenzierung des Schweregrades von periokulären Falten und damit der Abschätzung der Wirksamkeit des zu untersuchenden Behandlungsverfahrens.

Tabelle 37: Reproduzierbarkeit klinischer Skalen zur Beurteilung mimischer Falten

Author/Jahr	Skala	Lokalisation der beurteilten Falten und Zustand der Muskulatur	Anzahl der Beurteiler	Anzahl der gezeigten Fotografien	Interindividuelle Variabilität	Intraindividuelle Variabilität	Anmerkung
Honeck et al. 2003	0 - 3	Glabella in entspanntem und in maximal kontrahiertem Zustand	28 (17 weibliche und 11 männliche)	50	Ungewichtetes Kappa: 0.62 (0.63 bei den weiblichen und 0.59 bei den männlichen Teilnehmern)	Ungewichtetes Kappa: 0.57 - 0.91; Gewichtetes Kappa: 0.68 - 0.94	Die in entspanntem und in kontrahiertem Zustand aufgenommenen Fotografien wurden mit einer Skala beurteilt
Day et al. 2004	1 - 5	Nasolabialfalten in entspanntem Zustand	5	30	Prozentuale Übereinstimmung: 0.68 (linke Gesichtshälfte) und 0.72 (rechte Gesichtshälfte); Gewichtetes Kappa: 0.65 - 0.85 (linke Gesichtshälfte) und 0.63 - 0.89 (rechte Gesichtshälfte)	Prozentuale Übereinstimmung: 0.69 (linke Gesichtshälfte) und 0.73 (rechte Gesichtshälfte); Gewichtetes Kappa: 0.77(linke Gesichtshälfte) und 0.81 (rechte Gesichtshälfte)	Prozentuale Übereinstimmung und gewichtetes Kappa wurden berechnet; Nasolabialfalten der linken und rechten Gesichtshälfte wurden getrennt beurteilt
Kim et al. 2004	0 - 4	Hyperkinetische Falten im Bereich des M. frontalis, des M. corrugator supercilii und des M. orbicularis oculi in entspanntem und in kontrahiertem Zustand	11	120	Ungewichtetes Kappa: 0.64 (M. frontalis), 0.52 (M. corrugator supercilii) und 0.43 (M. orbicularis oculi)	-	Die jeweils in entspanntem und in kontrahiertem Zustand aufgenommenen Fotografien wurden nebeneinander beurteilt
Lemperle et al. 2001	0 - 5	Jeweils rechts- und linksseitige Glabellafalten, Nasolabialfalten, Oberlippenfältchen und Marionettenfalten; Zustand ist nicht angegeben	17 (9 zur direkten Beurteilung der Patienten und 8 zur Beurteilung der Fotografien)	76 mimische Falten in den Gesichtern von 32 Personen (direkte Beurteilung) und 130 Gesichtsfalten auf 80 Fotografien	Prozentuale Übereinstimmung: 0.93 (direkte Beurteilung) und 0.89 (Beurteilung der Fotografien)	-	Es wurde lediglich die prozentuale Übereinstimmung berechnet, Kappastatistik wurde nicht verwendet

7.8 Möglichkeiten zur Steigerung der Reproduzierbarkeit

Ungeachtet der festgestellten ausreichend guten Reproduzierbarkeit stellt sich die Frage nach Möglichkeiten, diese für zukünftige Studien weiter zu verbessern.

Die Teilnehmer der beiden Reproduzierbarkeitsstudien erhielten aus organisatorischen Gründen nur eine relativ kurze Schulung zur Anwendung der Skalen anhand des Atlas. Da eine ausführlichere Schulung der Beobachter über einen längeren Zeitraum auch mit einer höheren Reproduzierbarkeit einhergehen könnte, sollte überlegt werden, im Vorfeld künftiger Studien eine längere Schulungsphase einzuplanen.

In beiden Beurteilergruppen zeigte sich, dass es für die Beobachter am schwierigsten war, zwischen leichter und mäßiger Faltenbildung in entspanntem und zwischen mäßiger und starker Faltenbildung in angespanntem Zustand zu unterscheiden. Hier wäre zu überlegen, ob dem Schulungsatlas für die betreffenden Skalen-Werte weitere Beispielfotos hinzugefügt werden sollten, um eine noch höhere Reproduzierbarkeit zu erreichen.

Im (theoretischen) Idealfall könnte man ferner eine möglichst große Anzahl an Beobachtern schulen, sie auf die Reproduzierbarkeit ihrer Beobachtungen hin untersuchen und anschließend diejenigen mit den besten Reproduzierbarkeitswerten selektieren. Dies dürfte in der Praxis aufgrund des dazu nötigen zusätzlichen personellen und zeitlichen Aufwandes allerdings zumeist schwer durchführbar sein.

Die Ergebnisse der beiden Studien suggerieren, dass weibliche Beobachter tendenziell höhere Reproduzierbarkeitswerte erreichen als ihre männlichen Kollegen. Aufgrund der geringen Anzahl der untersuchten Beobachter lässt sich dies jedoch nicht abschließend beurteilen. Weitere Untersuchungen diesbezüglich wären nötig. Sollte sich diese Vermutung bestätigen, wäre eine Auswahl der Beobachter nach ihrem Geschlecht durchaus denkbar und könnte zu einer Steigerung der Reproduzierbarkeit beitragen. Allerdings dürfte die weibliche Geschlechtszugehörigkeit alleine kein Garant für eine hohe Reproduzierbarkeit sein.

7.9 Zusammenfassung der Diskussion

Die beiden hier untersuchten klinischen Faltenskalen mit den Ausprägungen von 0 bis 3 für den entspannten und den angespannten Zustand zeigen eine mäßige bis gute Reproduzierbarkeit.

Eine Steigerung der Reproduzierbarkeit durch verstärkte Schulung in der Anwendung und durch Selektion der Beurteiler erscheint möglich und sollte bei einer kleinen Gruppengröße von Beurteilern in Erwägung gezogen werden.