1. Zusammenfassung

Tibiakopffrakturen stellen im klinischen Alltag häufige und ernst zu nehmende Frakturen dar. Das primäre radiologische Erscheinungsbild ist nicht unbedingt aussagekräftig bezüglich des tatsächlichen Frakturausmaßes. Dies lässt sich oft erst intraoperativ feststellen. Um prä-operativ eine Entscheidung über den weiteren therapeutischen Weg treffen zu können, wurden im Laufe der Jahre multiple Klassifikationssysteme entwickelt, deren Unterscheidungsmerkmal die jeweils zur Verfügung stehenden diagnostischen Methoden darstellen.

Die Klassifikation der Tibiakopffrakturen nach radiologisch-morphologischen Kriterien (AO) und funktionellen Gesichtspunkten (Moore) ist heute Grundlage jeder Therapie.

In dieser retrospektiven Studie wurde die AO-Klassifikation mit der Moore-Klassifikation verglichen. Ziel der Studie war es, die Interobserver Zuverlässigkeit und die Intraobserver Reproduzierbarkeit im direkten Vergleich beider Klassifikationssysteme zu untersuchen.

Es wurden für diese Studie 25 Patienten-Bildersätze (nativ radiologisch und CT) mit Tibiakopffrakturen unterschiedlichen Schweregrades herausgesucht. Die Bildersätze wurden von 21 Auswertern mit unterschiedlichem Ausbildungsstand klassifiziert. Diese wurden in 4 Gruppen je nach Ausbildungsstand und Zugehörigkeit eingeteilt. Es wurde zum Zeitpunkt 0 und nach 3 Monaten klassifiziert. Die Auswertung der Daten erfolgte mittels des Kappa Koeffizienten nach der von FLEISS et al 1979 beschriebenen Methode. Zur Interpretation der gewonnen Kappa Werte wurden die von LANDIS und KOCH 1977 vorgeschlagenen Richtlinien herangezogen.

Nach Berechnung der Kappa Werte ergaben sich für die Interobserver Zuverlässigkeit Werte zwischen 0,20 und 0,41. Nach der Interpretationstabelle für Kappa Werte nach Landis und Koch 1977 zeigt dies eine ordentliche Übereinstimmung. Für die Intraobserver Zuverlässigkeit wurden Werte zwischen 0,17 und 0,46. Dies entspricht einer ordentlichen bis mäßigen Übereinstimmung.

Weiterhin konnten Intraobserver Kappa Werte mit Standardfehlern berechnet werden. Die Werte liegen zwischen 0,13 und 0,46. Nach Landis und Koch 1977 stellt dies eine schlechte bis mäßige Übereinstimmung dar.
Die Ergebnisse zeigten, dass sowohl die Interobserver Zuverlässigkeit als auch die Intraobserver Reproduzierbarkeit bei der AO-Klassifikation im Vergleich zur Moore-Klassifikation keine eindeutigen Unterschiede zeigen. Der Ausbildungsstand ist für die Auswertung bei keinem der beiden Klassifikationssysteme relevant. Selbst eine durchgeführte Fortbildungsmaßnahme in einer der 4 Gruppen konnte keinen Unterschied in der Auswertung zeigen.

Somit stellte sich die Frage nach der Validität und der Zuverlässigkeit der hier verglichenen Klassifikationssysteme. Beide wurden entwickelt um Diagnostik, Therapie und Therapieergebnisse standardisiert beurteilen zu können. Beide Klassifikationssysteme werden nebeneinander benutzt und sind international anerkannt.

Fazit:
Bei den Tibiakopffrakturen sind sowohl die AO- als auch die Moore-Klassifikation bezüglich Inter- und Intraobserver Reliability schlecht,
Tibiakopffrakturen sind schwierig zu klassifizierende Frakturen,
weitere neue Merkmale zum sicheren Klassifizieren müssen erarbeitet werden.

2. Summary

Interobserver Reliability and Intraobserver Reproducibility in two compared classification systems of fractures of the tibial plateau A retrospective study

Fractures of the tibia plateau are frequently seen in clinical everyday work. True fracture dimension only intra-operative look shows. During the years many classification system have being developed to be able to a pre-operative therapeutical decision. Their distinctive marks are built by the diagnostic possibilities used.

Classification of fractures of the tibial plateau is based on radiological-morphological criteria according to AO/ASIF or on functional criteria according to Moore.

These retrospective study compared the AO-classification with the MOORE classification. Aim of these study was the direct in-between comparison of interobserver reliability and intraobserver reproducibility.

25 cases of tibial plateau fractures imaged with plain films and CT scans of different fracture morphology were choosen. Fracture images were presented independently to 21 investigators. They were divides into 4 groups depending to their level of experience. and clinical membership. Classification was repeated as a second reading after 3 month to obtain intraobserver data. Classification categories were treated as nominal data. The interobserver and intraobserver reliabilities were evaluated for any possible pair of observers with the use of the weighted kappa coefficients as described by FLEISS et al 1979. Interpretation of kappa coefficient based on guidelines provided by LANDIS and KOCH 1977.

Kappa coefficients evaluated for Interobserver reliability ranged between 0,20 and 0,41. Based on the interpretation guidelines by Landis and Koch 1977 this shows a fair agreement. Intraobserver reproducibility ranged between 0,17 and 0,46 showing a slight to moderate agreement.

Intraobserver reproducibility including standard error of the mean shows a kappa coefficient by 0,13 to 0,46 with a slight to moderate agreement.
Interobserver reliability and intraobserver reproducibility between first and second readings showed no considerable differences between the compared classification system. Experience did not change or improve agreement at all. Even by special training of one group before classification could not improve interobserver agreement.

Validity and reliability of compared classification systems is to be questioned. Both were developed to enable a standardized diagnostic and therapy. AO and Moore classification are used simultaneously and are international accepted.

To sum it up:
For the tibia plateau fractures are both AO- and Moore-classification according to their intra- and interobserver reliability bad,
tibia plateau fractures are difficult to classify,
new and further characteristics for the classification have to be worked out.

3. Einleitung

3.1 Einteilung in verschiedene Klassifikationssysteme

„Eine Klassifikation ist nur nützlich, wenn sie sich auf den Schweregrad der Fraktur bezieht und als Grundlage sowohl für die Behandlung als auch für die Beurteilung der damit erreichten Ergebnisse dient“

Maurice E. Müller

Im Laufe der Jahre gab es verschiedene Klassifikationssysteme die sich je nach zur Verfügung stehenden diagnostischen Methoden unterscheiden lassen. Am Anfang standen die rein auf die klinisch-anatomische Beurteilung beruhenden Methoden, die ihrerseits entsprechende für die jeweilige Zeit durchaus moderne Einteilungssysteme entwickelten. Mit der Erfindung und zunehmenden Verbreitung der Röntgentechnik hatte man eine Methode, die dem Operateur ein verbessertes Diagnostik-Fenster sowie sinnvolle Klassifikationsmöglichkeiten in die Hand gab.
Die diagnostischen und therapeutischen Möglichkeiten der jeweiligen Zeit sind der begrenzende Faktor jeder Klassifikation. Mit der Zunahme der therapeutischen und diagnostischen Verfahren wachsen auch die Ansprüche an Klassifikationssysteme. Klassifikation ist immer Kodierung von Daten, und somit auch Informationsverlust. Je komplexer eine Klassifikation wird, desto mehr Informationen werden übermittelt. Von einer Klassifikation wird erwartet, dass sie als ein Hilfsmittel für die Entscheidungsfindung und Therapieplanung dient [ACKERMANN et al 1986]. Idealerweise sollte sie ein Kommunikationsmittel zwischen den Operateuren darstellen, Richtlinien für die Behandlung sowie eine eingeschränkte Prognosemöglichkeit geben und schließlich eine Methode darstellen zur Berichterstattung sowie Ergebnis- und Behandlungsvergleiche in der Literatur ermöglichen [BURSTEIN 1993, FRADSEN et al 1988, LINDSJO 1985, MARTIN und MARSH 1997, THOMSEN et al 1991, WALTON et al 2003]. All diese Anforderungen sind schwierig zu erreichen [WALTON et al 2003]. In der klinischen Praxis zeigt es sich, dass mittels einer Beschreibung der Seitenangabe, Muster, Dislokation und Weichteilbeteiligung ausreichende Angaben zur Verständigung gemacht werden

[WALTON et al 2003]. Bei genauerer Betrachtung aber ist Ungenauigkeit vorprogrammiert und die Reproduzierbarkeit im Rahmen von Forschungs- oder Publikationsbestrebungen nicht durchführbar [WALTON et al 2003].

Zu den vielen Klassifikationsmöglichkeiten, die im Laufe der Jahre entwickelt wurden, zählen auch die AO-Klassifikation [MÜLLER et al 1977] sowie im Bereich der proximalen Tibia die Klassifikation nach TM Moore [MOORE 1980].

Beide wurden entwickelt in dem Bestreben die Diagnostik, Behandlung sowie die Ergebnisse standardisiert beurteilen zu können. Beide Klassifikationssysteme werden nebeneinander benutzt und sind aus dem klinisch-chirurgischen Alltag nicht mehr wegzudenken.

Da aber für klinische Untersuchungen, gerade bei geringen Fallzahlen und im multizentrischen Bereich der korrekten Einteilung der Studienarme eine essentielle Bedeutung zukommt, wird in zunehmenden Maße die „Inter-„ und „Intraobserver Reliability“ derartiger Systeme hinterfragt. K.A. Siebenrock et al 1992 und K.A. Siebenrock et al 1993 zeigen in Ihren Arbeiten über Klassifikationssysteme des proximalen Humerus, dass Klassifikationssysteme lediglich als Hilfsmittel zum Entscheid über Prognose und therapeutisches Vorgehen angesehen werden dürfen. Die geringe und schlechte interobserver Zuverlässigkeit die Sie in der 1993 publizierten Studie vorstellten, stellt Ihrer Meinung nach ein wichtiges Argument bezüglich der kontroversen Diskussion bezüglich einer Behandlungsempfehlung.

Da Fragestellungen in der „Versorgungsforschung“ nicht zuletzt aus ökonomischen Gründen eine zunehmende Rolle spielen, muß auch dieser bisher weitgehend als „schicksalhaft“ hingenommene Nachteil der Klassifikation hinterfragt werden.

3.2 AO-Klassifikation

Professor Maurice E. Müller aus Bern beschäftigte sich seit Jahren mit Klassifikationssystemen. Er wollte ein einheitliches Einteilungssystem für Frakturen aufstellen. Ein „Esperanto“ für Frakturen [Colton 1991]. Er beriet sich mit seinen Kollegen in und außerhalb der AO-Gruppe und fing an, ein System zur Codierung von Frakturen zu entwickeln, welches eine Vielzahl von Kriterien erfüllen sollte:

es sollte logisch und konsequent

es sollte das Ausmaß der Fraktur wiederspiegeln

es sollte relativ einfach zu merken und nachvollziehbar sein

und es sollte international verständlich sein unabhängig von der Muttersprache des Anwenders. Nicht zu vergessen sollte es auch noch kompatibel sein mit den zur Verfügung stehenden technischen Möglichkeiten [COLTON 1991].

Obwohl manche am Erfolg zweifelten, liegt die „Klassifikation der langen Röhrenknochen“ nun vollständig vor [MÜLLER, NAZARIAN, KOCH und SCHATZKER 1990: Seite 527]. Die Publikation entstand in Zusammenarbeit mit der AO-Stiftung und der Internationalen Gesellschaft für Orthopädische Chirurgie [SICOT].

Die AO [Arbeitsgemeinschaft für Osteosynthese] / OTA [Orthopaedic Trauma Association] stellte 1987 zum ersten Mal ein zusammenhängendes, nach Knochensegmenten unterteiltes System der Öffentlichkeit vor, welches eine möglichst genaue Unterteilung der Frakturen der langen Röhrenknochen erlaubt.

Daraus formulierte die „Swiss Association for the Study of the Problems of Internal Fixation“ [AO] ein praktikables und verständliches Klassifikationssystem für lange Röhrenknochen.

Das AO/OTA Klassifikationssystem nach MÜLLER, NAZARIAN, KOCH und SCHATZKER 1990 wurde von den Redakteuren des Journal of Bone and Joint Surgery befürwortet und wird sowohl bei den Treffen der Orthopaedic Trauma Association als Standard benutzt, als auch vom Journal of Orthopaedic Trauma im Rahmen ihrer Veröffentlichungen[WALTON et al 2003] verwendet.

Die Vorzüge dieses standardisierten und verständlichen Klassifikationssystems im Vergleich mit früheren Klassifikationsmethoden wurde bereits herausgearbeitet [COLTON 1991, JOHNSTONE 1993]. Sie wird von der AO als einheitliches, logisches, einfach reproduzierbares und in jeder Lokalisation anwendbares Klassifikationssystem nach vorwiegend morphologischen und lokalisatorischen Gesichtspunkten weltweit propagiert und findet entsprechend Verbreitung [COLTON 1991, LICHTENSTEIN et al 1991, LINDSJÖ 1985, SMITH 2000].

Das grundlegende Prinzip dieser Klassifikation ist die Unterteilung aller Frakturen eines Knochensegmentes in drei Typen, die ihrerseits in drei Gruppen und dann in je drei Untergruppen unterteilt sind. Weiterhin sind die Frakturen nach zunehmendem Schweregrad entsprechend ihrer morphologischen Komplexität, dem Schwierigkeitsgrad ihrer Behandlung und ihrer Prognose geordnet. Somit ergeben sich insgesamt 27 Untergruppen für jedes Knochensegment.

Den langen Knochen wurde eine numerische Ziffer zugeordnet:

Humerus 1
Radius/Ulna 2
Femur 3
Tibia/Fibula 4

Die nachfolgende Unterteilung bezieht sich auf die 3 Unterteilungsregionen der Knochen:

Proximale Metaphyse/Gelenk 1
Diaphyse 2
Distale Metaphyse/Gelenk 3

Somit stellt z.B. 4.1 eine Fraktur der Tibia/Fibula im proximalen Bereich dar.

Weiterhin sind Frakturtypen definiert:

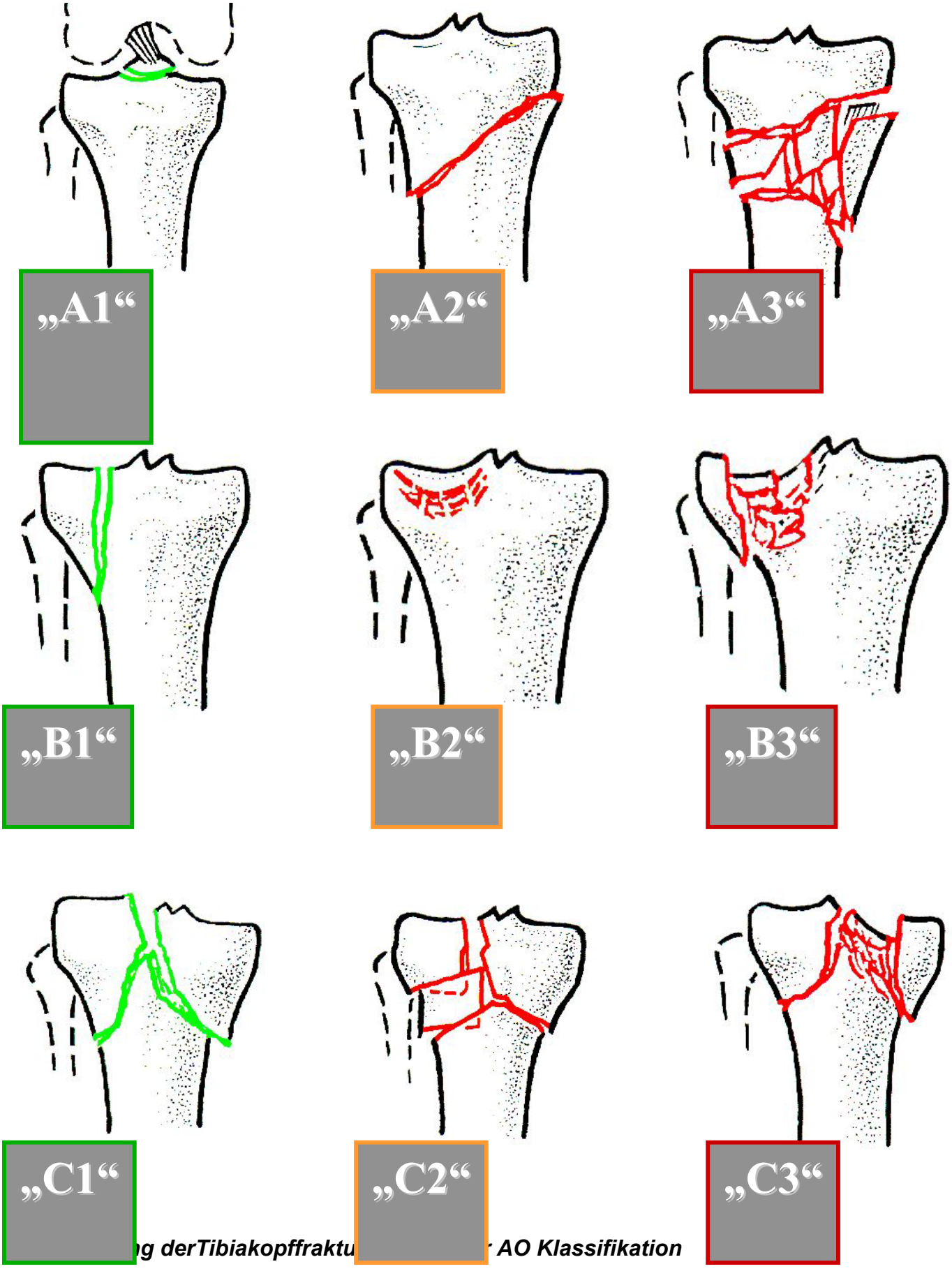
Typ A einfach oder oder bei Gelenkfrakturen extraartikulär
Typ B keilförmig, bei Gelenkfrakturen partiell intraartikulär
Typ C komplex, bei Gelenkfrakturen intraartikulär

Ausnahmen mit eigenen Kodierungen stellen der proximale Humerus, der proximale Femur und die Malleolen dar.

Die Klassifikation einer Fraktur legt deren Schweregrad fest und gibt Hinweise für deren beste Behandlungsmöglichkeiten. Das System will, wie von ihrem Begründer klar ausgedrückt, Schweregrad, Therapierichtlinien und Prognose beinhalten. Seit

über 30 Jahren wird das System weiterentwickelt, die AO-Dokumentationszentrale in Bern überblickt riesige Serien einheitlich klassifizierter Frakturen, die wissenschaftlich verwertet werden (LICHTENSTEIN et al 1991).

Abbildung Nr.1



ng derTibiakopffraktu r AO Klassifikation

3.3 Klassifikation nach Moore

Als Tillmann M. Moore 1980 sein System der Einteilung der Luxations-Frakturen der proximalen Tibia veröffentlichte, stellte er damit eine für den Alltag praktikable, auf klinischen und radiologischen Gesichtspunkten basierende Klassifikation vor.
Er unterschied generell zwischen Luxations- und Plateaufrakturen je nach Unfallmechanismus. Die Luxationsfrakturen sind immer instabil, verbunden mit femorotibialer Luxation oder Subluxation. Somit ist die Fraktur immer mit einer Bandverletzung verbunden [HERTEL 1997]. Hiermit sollte eine bessere präoperative Einschätzung der erwarteten ligamentären-, neurovaskulären- und Meniskus-assoziierten Schäden am Kniegelenk gelingen. Er teilte die Luxations-Frakturen des Kniegelenkes in 5 Typen ein, wobei Typ 2 und 4 nochmals in eine mediale und eine laterale Form unterschieden werden.
Es ist jedoch nicht gesagt, dass die Häufigkeit des Auftretens von ligamentären-, neurovasculären und Meniskus-assoziierten Schäden mit der aufsteigenden Reihenfolge der Typeneinteilung positiv korrelieren.
Die Klassifikation nach Tillmann M. Moore stellt eine einfache, und für den klinischen Alltag sehr praktikable Möglichkeit der Fraktureinteilung dar. Sie wird benutzt, um sich einen schnellen Überblick über die Verletzungsart und das weitere Procedere im Rahmen der Frakturversorgung zu machen. Die nachfolgende Abbildung Nr. 2 zeigt eine Darstellung der Klassifikation nach Tillmann M. Moore 1980.

Abbildung Nr. 2

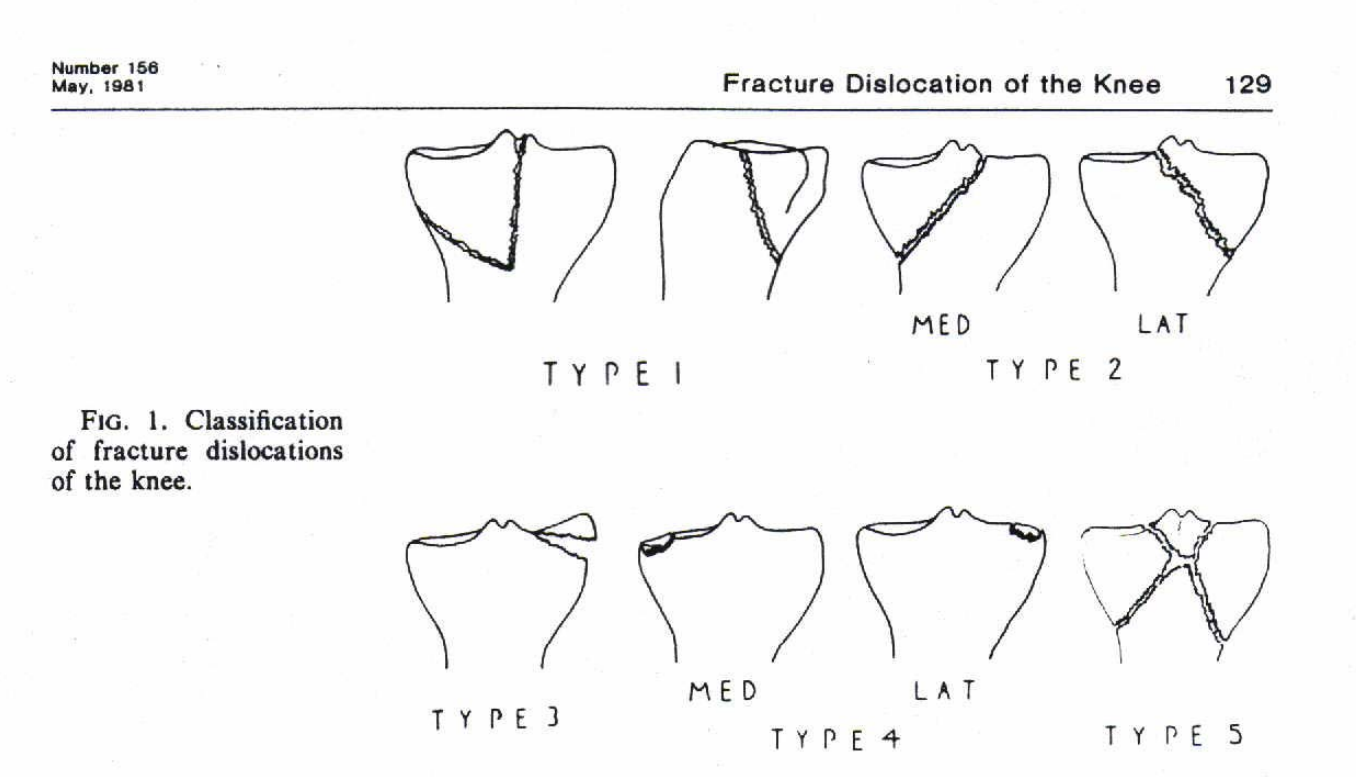


FIG. 1. Classification of fracture dislocations of the knee.

- Typ I: dorsaler Kondylenspaltbruch durch luxierenden Femurkondylus
- Typ II: kompletter Kondylenbruch, erreicht das kontralaterale Tibiaplateau
- Typ III: Kapselbandabriß (Typ Segond), meist verbunden mit vorderem Kreuzbandriß

  Typ IV: Kantenimpression, verbunden mit schwerer Bandverletzung der Gegenseite (Aufklappmechanismus)

  Typ V: Trümmerbruch mit Eminentiaausriß

3.4 Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit

Eine Frakturklassifikation sollte zuverlässig und valide (genau) sein [AUDIGÈ et al 2001, BURSTEIN et al 1993, FLEISS et al 1979, MARTIN et al 1997]. Die Zuverlässigkeit kann auf verschiedene Art und Weise beurteilt werden. Eine Möglichkeit stellt ein Vergleich zwischen verschiedenen Auswertern sowie der Vergleich zwischen ein und demselben Auswerter nach einem definierten Zeitabstand dar. Als Zuverlässig ist eine Klassifikationsmethode zu bezeichnen, wenn unterschiedliche Anwender zum gleichen Ergebnis kommen, oder der gleiche Anwender die genau gleiche Einteilung nach einem definierten Zeitabstand macht [BURSTEIN 1999]. Eine sehr gute Zusammenfassung für die Begriffe der Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit geben SVANHOLM et al 1989 in Ihrer Arbeit über die „Reproduzierbarkeit histomorphologischer Diagnosen mit speziellen Verweis auf die Kappa Statistik". Hierin wird die „Intraobserver Einigkeit" als der Bereich gekennzeichnet wo der gleiche Beobachter/Auswerter eine Probe an zwei oder mehr verschiedenen Zeitpunkten beurteilt. Dieser Vergleich ermöglicht eine Aussage zwischen den Klassifikationen eines Auswerters zu unterschiedlichen Zeitpunkten. Die „Interobserver Einigkeit" wird als derjenige Bereich definiert, wo zwei oder mehr Beobachter/Auswerter die gleiche Probe beurteilen [AUDIGÈ et al 2004]. Hiermit wird ein Vergleich zwischen den Auswertern möglich. Es kann fast als Qualitäts-Management bezeichnet werden, denn es zeigt inwiefern ein Klassifikationssystem reproduzierbar ist. Wenn alle Auswerter der gleichen Meinung sind, ist ein Klassifikationssystem zu 100% reproduzierbar, die Einteilung somit nicht vom Betrachter abhängig. Eine Klassifikationsmethode ist umso valider (genauer), je näher sie der tatsächlichen Wahrheit kommt. Ein „Gold-Standard" wurde in unserer Studie nicht definiert, somit werden die Ergebnisse der Auswerter/Anwender nicht damit verglichen. Als „Gold-Standard" bezeichnet man eine intra-operative Fraktureinteilung aufgrund des tatsächlichen Frakturausmaßes und Weichteilschadens [GOLDBERG et al 1988]. Der „Gold-Standard" wird willkürlich festgelegt und stellt damit eine subjektive Sichtweise dar. Da unsere Arbeit eine retrospektive Studie darstellt, ist die Festlegung eines „Gold-Standards" post-

operativ nicht mehr möglich. Die Definition eines Auswerters als „Gold-Standard“ beinhaltet die Behauptung, dass alle anderen Auswerter per definitionem nicht besser sein können.

Somit existiert auch keine absolute Frakturklassifikation, da die Klassifikation jeglicher Frakturen eine jeweils subjektive Interpretation darstellt [SIDOR et al 1993]. Wir wollen lediglich eine Aussage über Zuverlässigkeit zwischen den einzelnen Anwendern, Interobserver Reliability, sowie zwischen den Anwendern, Intraobserver Reliability, nach einem definierten Zeitraum treffen. Die „commonly“/üblicherweise verwendeten Fraktur-Klassifikationssysteme wurden nur ungenügend auf Inter- und Intraobserverzuverlässigkeit hin überprüft [DIRSCHL et al 1997, FRADSEN et al 1988, NIELSEN et al 1990, SIEBENROCK 1993, SIDOR et al 1993, THOMSEN et al 1991], was manche Autoren zu der Behauptung verleitete, die Anwendung eines Frakturklassifikationssystem vor Prüfung auf Zuverlässigkeit sei unangebracht [BURSTEIN et al 1981, DIRSCHL et al 1997, SIDOR et al 1993, THOMSEN et al 1991]. Im vorliegenden Literaturmaterial wurden unterschiedliche Ergebnisse bezüglich der Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit dargestellt. MARTIN et al 1997 zeigten in Ihrer Studie an Frakturen der distalen Tibia, dass die Übereinstimmung zwischen den erfahrenen Fachärzten bezüglich der Interobserver Zuverlässigkeit größer war als zwischen den wenig erfahrenen Assistenten. Dies bezog sich auf alle in der Studie gemachten Aussagen, es konnte jedoch keine statistische Signifikanz erreicht werden. Im Gegensatz zu der von WEN-JIE JIN et al 2005 in Ihrer Studie getroffenen Aussage, konnten MARTIN et al 1997 zeigen, dass die Erfahrung der Auswerter keinen Einfluss auf die Intraobserver Reproduzierbarkeit hat. Somit kann aus der Arbeit von MARTIN et al 1997 der Schluss gezogen werden, dass die Erfahrung der Auswerter dazu tendiert die Interobserver Zuverlässigkeit zu erhöhen, auf die Intraobserver Reproduzierbarkeit keinerlei Auswirkung hat. Ganz im Gegenteil scheint es interessanterweise so zu sein, dass die Reproduzierbarkeit der Ergebnisse der 2. Auswertungsreihen bei den erfahrenen Auswertern niedriger war als bei den Assistenten.

SIDOR et al 1993, THOMSEN et al 1991 und WALTON et al 2003, konnten in Ihren Studien zeigen, dass die Erfahrung der Auswerter keine nennenswerte Auswirkung auf die Interobserver-Zuverlässigkeit der Frakturklassifikationen hatte.

Jedoch stellten SIDOR et al 1993 fest, dass die Intraobserver-Zuverlässigkeit mit dem Grad der Erfahrung der Auswerter zunahm. DIRSCHL et al 1997 zeigten in Ihrer Studie dass die Interobserver-Zuverlässigkeit bezogen auf Frakturen des distalen Tibiaplateaus gering bis mittelmäßig ist und die Ergebnisse sich somit mit denen vorangegangener Studien decken [BRIEN et al 1995, FRADSEN et al 1988, KRISTIANSEN et al 1988, NIELSEN et al 1990, SIEBENROCK et al 1993, SIDOR et al 1993, THOMSEN et al 1991].

3.5 Fragestellung

Wir untersuchten in unserer retrospektiv angelegten Studie die Übereinstimmung in der Klassifizierung von Tibiakopffrakturen durch verschiedene Untersucher (Zuverlässigkeit) sowie die Übereinstimmung durch denselben Untersucher (Reproduzierbarkeit) bei Auswertung von Bilderreihen. Nach Durchsicht der bis zum Beginn der Studie bereits veröffentlichten Arbeiten, stellten wir fest, dass ein Vergleich der AO/OTA Klassifikation der Tibiakopffrakturen mit der MOORE–Klassifikation bis dato noch nicht verfolgt wurde, obwohl in mehreren Studien bereits unterschiedliche Klassifikationssysteme miteinander verglichen wurden [DIRSCHL et al 1997, MARTIN et al 1997, SIEBENROCK et al 1993, SWIONTKOWSKI et al 1999, THOMSEN et al 1991, WALTON et al 2003, WEN-JIE JIN et al 2005].

Das Ziel unserer retrospektiven Studie war es die Interobserver Zuverlässigkeit sowie die Intraobserver Reproduzierbarkeit für die zwei gegenüber gestellten Frakturklassifikationen am Tibiakopf darzustellen.
Hierfür wurden in der AO/OTA – Klassifikation die Einteilung nach Typ und Gruppe ausgewertet. Wir sahen von einer weiteren Beurteilung der Subgruppen ab, da bereits die Auswertung von Typ und Gruppe in anderen Regionen schlechte Ergebnisse erzielt. Weiterhin verfügten wir nicht über eine ausreichende Anzahl an Frakturen hatten, um eine aussagekräftige Schlussfolgerung bezüglich von Typ, Gruppe und Subgruppe ziehen zu können [MARTIN et al 1997].
Bei der Klassifikation nach Moore wurden alle Stufen inklusive die Einteilung lateral und medial in der Gruppe II und IV beibehalten. Es konnte bereits in vorangegangen Studien dargelegt werden, dass die Zuverlässigkeit abnimmt, wenn im Rahmen der Auswertung die Unterteilung in Subgruppen im Rahmen der AO/OTA – Klassifikation einbezogen wird [MARTIN et al 1997, PERVEZ et al 2002, SCHIPPER et al 2001, WEN-JIE JIN et al 2005]. Soweit uns bekannt ist, ist dies die erste Studie, die einen Vergleich zwischen der AO/OTA-Klassifikation und der Moore-Klassifikation darstellt.
Obwohl einfach in der Handhabung, hat sich die Klassifikation nach Moore in der Fachliteratur nicht durchgesetzt. Somit ist die Fragestellung des Vergleiches bezüglich der Inter- und Intraobserver-Zuverlässikeit berechtigt.

Nach Durchsicht der vorhandenen Literatur zum Thema Inter- und Intraobserverzuverlässigkeit in verschiedenen Klassifikationssystemen scheinen die Autoren zum Teil getrennter Auffassung zu sein was die Reliability (Zuverlässigkeit) angeht, und ob diese von Aus- und Weiterbildungsstand des jeweiligen Untersuchers abhängt [BORRELLI et al 2002, KRISTIANSEN et al 1988, SCHIPPER et al 2001, SIDOR et al, MARTIN et al 1997, WEN-JIE JIN et al 2005].
Da die Zuverlässigkeit somit in den meisten Studien als unzuverlässig dargestellt wird [AUDIGÉ et al 2001], sollte laut AUDIGÉ et al 2001 ein besonderer Augenmerk auf die Studienmethoden gelegt werden.

Im Rahmen dieser Dissertationsarbeit sollten folgende Fragen beantwortet werden:

1. In wie weit beeinflusst die Zugehörigkeit zu einem Krankenhaus der Maximalversorgung (Universitätsklinik) im Vergleich zu einem peripheren Krankenhaus die Reliability/Zuverlässigkeit?
2. Beurteilen Chefärzte von Krankenhäusern der peripheren Versorgung die Frakturen anders als Oberärzte an einer Universitätsklinik als Traumazentrum?
3. In wie weit unterscheiden sich die Auswertungen zum Zeitpunkt 0 von den erneuten Auswertungen 3 Monate später (Inter- und Intraobserverzuverlässigkeit)?
4. Wie beurteilen Assistenten im 1. Jahr die Frakturen?
5. Verbessert gezieltes Training die Intraobserver-Reliability bei der Gruppe der Assistenten im 1.Jahr signifikant?

4. Material und Methode

4.1 Erstellen der Bildersätze

Zu Beginn unserer Studie wurden Eckpunkte festgelegt und ein Arbeitsplan aufgestellt. Seitens der Abteilung für Unfall-, Hand- und Wiederherstellungschirurgie der Universitätsklinik Homburg wurden die OP-Aufzeichnungen für die Jahre 2000 bis 2003 zur Verfügung gestellt. Daraus wurde alle in diesem Zeitraum operativ versorgten Tibiakopffrakturen herausgesucht und im Röntgenarchiv der Chirurgischen Abteilung die dazugehörigen Röntgenbilder gesichtet. Im Jahr 2000 wurden in der Abteilung für Unfall-, Hand- und Wiederherstellungschirurgie 25 Tibiakopffrakturen operiert, 2001 21 Tibiakopffrakturen, 2002 26 Tibiakopffrakturen und 2003 44 Tibiakopffrakturen. Es wurden nur diejenigen Patienten einbezogen, bei denen die a.p. und die seitlichen Aufnahmen gemeinsam vorhanden waren [MARTIN et al 1997, WALTON et al 2003]. Anschließend wurden im CT-Archiv der Radiologischen Klinik die Datensätze verglichen, um festzustellen, ob zu den ausgesuchten Röntgen-Bildern auch die passenden prä-operativen Computertomographien der Frakturzone vorhanden waren. Anschließend wurden alle so erhaltenen Bilderreihen nochmals gesichtet und die Bildqualität überprüft. Nur qualitativ einwandfreies Bildmaterial wurde verwendet. Somit blieb ein Datensatz bestehend aus 25 Fällen übrig. Diese Vorgehensweise wird von Audigé et al 2001kritisch bewertet. Audigé et al 2003 empfehlen in ihrer Studie alle vorhandenen Bilder, ungeachtet der Vollständigkeit in die Studie einzubeziehen.

Aus der Fülle des CT-Dokumentation wurden nur die unmittelbar die Frakturzone betreffenden Bilderreihen selektiert. Auf der sagittalen Aufnahme wurde das Bild markiert, welches das Mittelbild darstellt, und davon ausgehend die Schichten 8 Millimeter nach ventral und 8 Millimeter nach dorsal gesichtet. Die Schichtdicke der CT-Bilder betrug jeweils 2 Millimeter. Ziel war es die Frakturzone von ventral nach dorsal darzustellen. Es wurde versucht, 2 axiale, 2 frontal-coronare und 2 sagitale Darstellungen zu erhalten. Somit lag bei jedem der 25 Patienten folgender Datensatz vor: Nativ-Röntgenaufnahmen in ap- und axialer Form, CT-Aufnahmen in axialer, coronarer und sigitaler Form. Dieses zeitaufwendige und sorgfältige Vorgehen war notwendig, da die so gewonnen Daten nun im Foto-Labor der Universitätsklinik

eingescannt, digital überarbeitet und auf CD gebrannt wurden. Alle personenbezogenen Daten wurde vorher ausgeblendet.
Somit wurde sichergestellt, dass die Frakturbilder nicht mit Patientennamen in Verbindung gebracht werden könnten. Aus Praktikabilitätsgründen entschlossen wir uns nur die oben genannten vollständigen Bildersätze in unserer Studie einzubeziehen.
Nach sorgfältiger Überarbeitung wurden die auf CD gebrannten Datensätze gemeinsam mit den ausgearbeiteten Auswertungsbögen den jeweiligen Kollegen zur Auswertung zugesandt.
Beispiel eines Bildersatzes siehe Anhang 1
Auswertungsbogen siehe Anhang 2

4.2 Auswahl der Studienteilnehmer

Es wurden im Rahmen des ersten Auswertungsdurchlaufes 16 ärztliche Kollegen unterschiedlichen Ausbildungsstandes angeschrieben und um die Auswertung des zugesandten Bildmateriales gebeten.
Weiterhin wurden 5 Assistenten im 1. Weiterbildungsjahr der Abteilung für Unfall-, Hand- und Wiederherstellungschirurgie der Universitätskliniken des Saarlandes um die Auswertung der gleichen Frakturen gebeten.
Durch die breite Fächerung unserer Zuschriften, gelang es uns sowohl Assistenzärzte als auch Fachärzte und Chefärzte auswärtiger Kliniken für unserer Studie zu gewinnen. Der unterschiedliche Ausbildungsstand der Assistenzärzte ermöglicht eine differenzierte Aussage über die Zuverlässigkeit der Studienergebnisse. Durch die Einbeziehung von Assistenten im 1. Weiterbildungsjahr ist weiterhin eine Aussage bei „Klassifikationsanfängern“ möglich.
Weiterhin können somit der Vergleich zwischen den unterschiedlichen Ausbildungslevels, den verschiedenen zugehörigen Ausbildungsstätten sowie zwischen Chefärzten von Krankenhäusern der peripheren Versorgung und Fachärzten/Oberärzten eines Traumazentrums bezüglich Ihrer Auswertung getroffen werden.
Unter den 21 auswertenden ärztlichen Kollegen befanden sich 6 Chefärzte aus Krankenhäusern der peripheren Versorgung, 5 Oberärzte eines Traumazentrums, 5 Weiterbildungsassistenten im fortgeschrittenen Stadium der Weiterbildung sowie 5 Assistenten im 1. Weiterbildungsjahr.

Daraus ergibt sich folgende Einteilung:

6 Chefärzte – alle aus Abteilungen für Unfallchirurgie und seit mindestens 12 Jahren Fachärzte für Unfallchirurgie
5 Oberärzte – alle aus dem Traumazentrum Homburg/Saar und seit mindestens 2 Jahren Fachärzte für Unfallchirurgie
5 Weiterbildungs-Assistenzärzte – unterschiedlichen Ausbildungsstandes
5 Assistenten im 1. Weiterbildungsjahr

Die Auswerter erhielten die Unterlagen zum Zeitpunkt 0 und 3 Monate nach Eingang der 1. Auswertung. Um einem Wiedererkennungsfaktor zu vermeiden, wurden die Bildersätze bei der 2. Auswertung in einer geänderten Reihenfolge dargeboten.
Die Bildersätze wurden jedes Mal von einer detaillierten Anleitung zur Auswertung begleitet (siehe Anhang 1), sowie von einer Kopie der Klassifikationen nach MÜLLER 1987 (AO-Klassifikation) und der Klassifikation nach MOORE 1980. Die Auswertung erfolgte auf den beigefügten Auswertungsbogen. Abbildung Nr.3 zeigt einen entsprechenden Auswertungsbogen.
Um einen zusätzlichen Aussagegewinn zu erzielen erhielten die 5 teilnehmenden Assistenten im 1. Weiterbildungsjahr nach der ersten Auswertung eine gezielte Fortbildung zum Thema Klassifikationen nach dem AO-System und nach Moore 1. Die anschließende 2. Auswertung erfolgte direkt im Anschluss an die Fortbildung um die vermittelten Erkenntnisse direkt anwenden zu können.
Nach Erhalt aller Unterlagen wurden die Datensätze als Excel-Datei angelegt und für die weitere rechnergestützte Auswertung bearbeitet.

4.3 Dateneingabe

Nach Erhalt der Auswertungen erfolgte die Eingabe der Klassifikationsergebnisse in eine Excel-Tabelle (Excel 4.0) und wurde anschließend nach den Richtlinien der Abteilung für Medizinische Biometrie und Statistik der Universität des Saarlandes bearbeitet. Die Bearbeitung erfolgte mit dem Ziel, die erhaltenen Daten für die rechnergestützte Auswertung mit SAS Prozedur Proc Freq vorzubereiten. Dies ist Teil des Programms 2004 SAS Online Doc © 9.1.3 der Firma SAS Institute Inc. Cary. North Carolina. Die Berechnung der Multirater Kappa mit Standardfehler erfolgte mit dem selben Programm.
Nach entsprechender Vorbereitung und Anlage einer speziellen Tabelle, konnten die Datensätze ausgewertet werden. Dies war notwendig, da das Programm nur entsprechend vorbereitete Datensätze bearbeiten kann.

4.4 Statistische Auswertung – Kappa Werte

Die Klassifikationskategorien wurden aufgrund des Fehlens eines absoluten „Gold-Standards“ als nominale Daten betrachtet. Für die Berechnung der Interobserver Zuverlässigkeit sowie der Intraobserver Reproduzierbarkeit wurde die Kappa Statistik verwendet. Die Kappa Statistik erlaubt eine zufallskorrigierte korrekte Berechnung der Übereinstimmung von Nominaldaten. Diese Methode wurde zuerst von Cohen beschrieben [COHEN 1960]. Es vergleicht ein beobachtetes Übereinstimmungsmaß P(A) mit dem rein zufälligen Maß der Übereinstimmung P(E). Die original Kappa Statistik war nur auf die Situation anwendbar, in welcher zwei Auswerter die gleichen Daten in einander ausschließenden Kategorien klassifizieren sollten. Diese heißen Cohens Kappa. Übereinstimmungen werden als 100% richtig und Nichtübereinstimmungen als falsch gewertet.

Cohens Kappa für den einfachen Kappa Koeffizienten

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e}$$

wobei

$$P_o = \sum_i p_{ii}$$

und

$$P_e = \sum_i p_{i.} p_{.i}$$

ist [Cohen 1960].

Po = relative Häufigkeit der Übereinstimmung (=Anzahl Übereinstimmungen/Anzahl Patienten)

Pe = erwarteter Anteil an Übereinstimmungen bei unabhängigen Befundern mit denselben Randverteilungen

Hierbei entspricht ein Kappa Koeffizient von +1 einer 100%en Übereinstimmung zwischen allen Auswertern. Wenn die beobachtete Übereinstimmung die rein zufällige überschreitet, somit Kappa positiv ist, reflektiert die Höhe des Wertes das Maß an Übereinstimmung. Obwohl in der Praxis unüblich, zeigt ein negativer Kappa Wert ein beobachtetes Maß an Übereinstimmung, welches geringer als das rein zufällige ist. Der minimale Kappa Wert bewegt sich zwischen –1 und 0, abhängig von den Randbedingungen.

Cohens Kappa für den gewichteten Kappa Koeffizienten

$$\hat{\kappa}_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}}$$

wobei

$$P_{o(w)} = \sum_i \sum_j w_{ij} p_{ij}$$

und

$$P_{e(w)} = \sum_i \sum_j w_{ij} p_{i\cdot} p_{\cdot j}$$

ist [Cohen 1960].

Gewichtete Kappa Koeffizienten werden für Ordinalskalen verwendet. Bei 5 geordneten Kategorien würde :

Übereinstimmung [i-j]=0 mit 1 gewichtet (also perfekt)

beinahe Übereinstimmung [i-j]=1 mit 0,75

[i-j]=2 mit 0,5

[i-j]=3 mit 0,25 und [i-j]=4 mit 0

Dieses nennt man auch lineare Gewichtung.

Die in dieser Dissertationsarbeit verwendeten Klassifikationssysteme entsprechen nominalen Skalen und somit wird für diese Berechnung der einfache Kappa Koeffizient verwendet.
Seitdem wurden mehrere Statistiken beschrieben die eine größere Flexibilität zuließen [EDWARDS et al 2002, FLEISS 1971, FLEISS, NEE und LANDIS 1979, FLEISS 1981, FLEISS und SHROUT 1994, KREDER et al 1996, O'CONNELL und DOBSON 1984, POSNER et al 1990, SHROUT und FLEISS 1979]. Eine Übersichtsarbeit zum Thema Kappa Koeffizienten in der medizinischen Forschung stellt die Studie von KRAEMER et al 2002 dar.
LANDIS und KOCH 1977 entwickelten eine willkürliche Einteilung der Übereinstimmungsgrade von Kappa Koeffizienten. Diese wird in Tabelle 1 dargestellt.
Nach KORAN 1975 [first and second of two parts] fallen die meisten klinischen Studienergebnisse in die Kategorie „mäßig" [KREDER et al 1996]. Die Interobserver Zuverlässigkeit erhielten wir durch Ergebnisvergleich zwischen den 21 Auswertern. Der Grad der Interobserver Zuverlässigkeit und der Intraobserver Reproduzierbarkeit wurde bestimmt durch den Vergleich der zugesandten Datensätze der einzelnen Auswerter in den zwei Auswertungsdurchgängen.

Die Auswertung erfolgte mittels des computergenerierten Kappa Koeffizienten mit der SAS Proc. Freq Prozedur. Hierbei werden wie nach FLEISS et al 1979 beschrieben Kappa Werte für jedes mögliche zu bildende Ergebnispaar berechnet. Die zwei Auswertungsdurchgänge wurden unabhängig voneinander bearbeitet. Die Kappa Werte wurden nach der von FLEISS et al 1979 beschriebenen Methode entsprechend der zu beantwortenden Fragen ausgerechnet. Der Kappa-Wert ist +1 wenn alle Auswerter in allen Fällen übereinstimmen. Ein Kappa-Wert von 0 zeigt, dass die Übereinstimmung der Auswerter zufällig ist, ein Wert zwischen 0 und -1 gibt eine Übereinstimmung an, die noch geringer ist als die zufällig erwartete. Mehrere Autoren haben bereits Richtlinien für die Interpretation des Kappa Koeffizienten publiziert [CHAN et al 1997, KRISTIANSEN et al 1988, LANDIS et at 1977, FLEISS et al 1979, RASMUSSEN et al. 1993, SIEBENROCK et al 1993]. Die Interpretation der Kappa Werte basiert auf die von LANDIS und KOCH 1977 vorgeschlagenen Richtlinien. Sie werden in Tabelle 1 dargestellt.

Tabelle 1

Interpretation der Kappa Werte basierend auf Landis und Koch 1977

Kappa Werte	Übereinstimmung
<0,0	schlecht
0,00 bis 0,20	gering
0,21 bis 0,40	ordentlich
0,41 bis 0,60	mäßig
0,61 bis 0,80	gut
0,81 bis 1,00	exzellent

Die Inter- und Intraobserver Kappa Koeffizienten wurden anschließend verglichen. Das Programm gibt sowohl das einfache als auch das gewichtete Kappa mit den entsprechenden 95% Konfidenzintervallen an. Da die Klassifikationskategorien als nominale Daten betrachtet werden, ist das einfache Kappa zu verwenden. Das 95% Konfidenzintervall darf nur dann verwendet werden, wenn ein Auswerter mit einem anderen Auswerter oder ein Auswerter mit sich selbst verglichen wird. In unserer Arbeit werden jedoch gemittelte Werte über jeweils mehrere Auswerter verwendet. Somit sind die 95% Konfidenzintervalle nicht zu verwenden.

Die erhaltenen Auswertungsbögen waren zum Teil unvollständig. Somit mussten die erhaltenen Tabellen entsprechend bearbeitet werden, um eine statistische Auswertung möglich zu machen. Die unvollständigen Datensätze mussten für die Berechnung der Multirater Kappa mit Standardfehler ausgeklammert werden. Somit erhielten wir für die Intraobserver Reproduzierbarkeit Kappa Werte die sich geringgradig unterscheiden, auf der einen Seite Kappa Werte unter Berücksichtigung aller Auswerter ohne Standardfehler, und auf der anderen Seite Kappa Werte nur aus den vollständigen Datensätzen mit Berechnung der Kappa Werte. Diese werden als Kappa Wert plus minus 2 Standardfehler angegeben.

Die Berechnung einer statistischen Signifikanz konnte beim Vergleich der AO-Klassifikation mit der Moore-Klassifikation nicht erfolgten. Die AO-Klassifikation hat 27 Möglichkeiten der Einteilung einer Fraktur. Wir beschränkten uns wie bereits aufgeführt auf 9 Möglichkeiten A1 bis C3 (Typ und Gruppe). Bei der Moore-Klassifikation sind es 6 Möglichkeiten. Somit kann keine statistische Signifikanz berechnet werden, da in einer 9-stufigen Klassifikation aufgrund der größeren Auswahlmöglichkeiten die Einteilung einfacher erfolgen kann als in einer 6-stufigen.

5. Ergebnisse

5.1 Interobserver Zuverlässigkeit aller Auswerter

Eine Einteilung der Interobserver-Zuverlässigkeit nach Auswertern, den zwei unterschiedlichen Zeiträumen (Zeitpunkt 0 und 3 Monate), und im direkten Vergleich AO-Klassifikation gegen Moore-Klassifikation wird in Tabelle 2 und 3 dargestellt. Zum Zeitpunkt 0 Monate ließ sich ein Kappa-Wert von 0,29 berechnen .Zum Zeitpunkt 3 Monate betrug der Kappa-Wert 0,34. Dies stellt nach den Interpretationsrichtlinien von Landis und Koch 1977 eine geringe Übereinstimmung dar.

Tabelle 2 einfaches Kappa

AO Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
alle	0,29	0,34

Tabelle 3 einfaches Kappa

Moore Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
alle	0,23	0,31

Es zeigen sich vergleichbare Werte bei der Auswertung nach der AO-Klassifikation und bei der Klassifikation nach Moore. Die Tatsache, dass die Auswerter im Rahmen des zweiten Durchganges nach 3 Monaten, sowohl was die AO-Auswertung als auch die Moore-Auswertung betrifft, etwas besser abschneiden, bleibt ohne statistische Signifikanz. Dies ist dennoch ein Trend der in der gesamten Auswertung zu beobachten ist.

5.2 Intraobserver Zuverlässigkeit alle Auswerter

Tabelle 4 zeigt die Kappa-Werte nach der Auswertung des gesamten Datenpools bezogen auf die AO-Klassifikation. Die Darstellung der Ergebnisse der Moore-Klassifikation zeigt Tabelle 5. Hierbei werden die Zeiträume 0 und 3 Monate (jeweils gesamte Datenmenge AO und gesamte Datenmenge Moore) miteinander verglichen.

Tabelle 4 einfaches Kappa

AO Klassifikation

Auswerter	Zeitpunkt 0 > 3 Monaten
alle	0,40
Chefärzte	0,38
Oberärzte	0,37
Assistenzärzte	0,46
Assist. 1.WJ	0,20

Somit kann auch hier gezeigt werden, dass die berechneten Werte kleiner als 0,5 sind und folglich eine schlechte Übereinstimmung aufweisen. Die in Tabelle 5 dargestellten Werte für die Moore-Klassifikation liegen im gleichen Bereich. Lediglich der Kappa Wert der Chefärzte der auswertigen Krankenhäuser ist geringfügig niedriger. Eine statistische Signifikanz liegt nicht vor, da die Konfidenzintervalle aller 4 Gruppen einen ähnlichen Bereich überspannen.

Tabelle 5 einfaches Kappa

Moore Klassifikation

Auswerter	Zeitpunkt 0 > 3 Monaten
alle	0,42
Chefärzte	0,33
Oberärzte	0,46
Assistenzärzte	0,48
Assist. 1. WJ	0,29

5.3 Interobserver Zuverlässigkeit im Vergleich

Tabelle 6 und Tabelle 7 zeigen die Kappa-Werte im Vergleich der 4 Gruppen jeweils zu den Zeitpunkten 0 und 3 Monate. Die Kappa Werte sind zum Zeitpunkt 3 Monate etwas besser als zum Zeitpunkt 0, liegen jedoch unter 0,5 sind und zeugen somit von einer schlechten Übereinstimmung. Auch hier kann keine statistische Signifikanz nachgewiesen werden.

Tabelle 6

Interobserver Kappa

AO-Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
Oberärzte > Chefärzte	0,29	0,33
Oberärzte > Assistenten	0,26	0,30
Chefärzte > Assistenten	0,33	0,41

Tabelle 7

Interobserver Kappa

Moore-Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
Oberärzte > Chefärzte	0,27	0,33
Oberärzte > Assistenten	0,25	0,31
Chefärzte > Assistenten	0,20	0,31

5.4 Intraobserver Reproduzierbarkeit im Vergleich

Tabelle 8 und 9 zeigen die einfachen Kappa Werte im Vergleich. Hierbei wird zum Teil sichtbar, dass das Kappa bei der Auswertung zum Zeitpunkt 3 Monate bei allen Auswertern besser war als zum Zeitpunkt 0. Eine statistische Signifikanz liegt nicht vor.

Tabelle 8

Intraobserver im Vergleich

AO-Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
Oberärzte > Oberärzte	0,23	0,19
Chefärzte > Chefärzte	0,31	0,40
Assistenten > Assistenten	0,30	0,36

Tabelle 9

Intraobserver im Vergleich

Moore-Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
Oberärzte > Oberärzte	0,27	0,26
Chefärzte > Chefärzte	0,17	0,35
Assistenten > Assistenten	0,19	0,25

5.5 Intraobserver Kappa mit Standardfehler

Wie bereits im Abschnitt 4.4 Statistische Auswertung – Kappa Werte beschrieben, werden im Folgenden die Multirater Kappa (=Intraobserver Kappa) mit Angabe der Standardfehler dargestellt. Diese Kappa Werte konnten nur aus vollständigen Datensätzen gewonnen werden. Somit musste die Auswertung Nr. 12 hiervon ausgenommen werden. Aus diesem Grund wird in der Tabelle jeweils angegeben welcher Auswerter hierfür nicht bewertet werden konnte. Es wird jeweils der Wert plus minus 2 Standardfehlern angegeben. Tabelle 10 zeigt die Werte für die Auswertung der AO-Klassifikation und Tabelle 11 die Werte für die Auswertung nach der Moore-Klassifikation.

Tabelle 10

Intraobserver Kappa mit Standardfehler

AO-Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
alle ohne # 12	0,25 +/- 0,01	0,32 +/- 0,01
OÄ ohne # 12	0,23 +/- 0,07	0,19 +/- 0,07
CÄ ohne # 12	0,31 +/- 0,07	0,46 +/- 0,07
ASS ohne # 12	0,30 +/- 0,07	0,36 +/- 0,07
Ass 1.WJ ohne # 12	0,23 +/- 0,06	0,21 +/- 0,06

Tabelle 11

Intraobserver Kappa mit Standardfehler

Moore-Klassifikation

Auswerter	Zeitpunkt 0	Zeitpunkt 3 Monate
alle ohne # 2,12,16	0,24 +/- 0,02	0,31 +/- 0,02
OÄ ohne # 2,12,16	0,26 +/- 0,09	0,25 +/- 0,08
CÄ ohne # 2,12,16	0,16 +/- 0,07	0,32 +/- 0,08
ASS ohne # 2,12,16	0,13 +/- 0,09	0,26 +/- 0,09
Ass 1.WJ ohne # 2,12,16	0,31 +/- 0,06	0,26 +/- 0,06

zum Zeitpunkt 3 Monate ohne Auswerter Nr. 2,12,16

6. Diskussion

Wie bereits in der Einleitung beschrieben, erhielten die Studienteilnehmer die Unterlagen persönlich zugesandt und es wurde kein Zeitlimit für die Auswertung gegeben. Da wir für die weitere Verwendung die Datensätze einscannten und auf CD brannten, wäre bei Verwendung von Bilder unterschiedlicher Qualität ein nur eingeschränkt auswertbarer Datensatz entstanden. Wir sahen hiervon für die Verwendung im Rahmen dieser Dissertationsarbeit ab. Sicherlich ist dies im Rahmen einer großangelegten Multi-Center-Studie sinnvoll.

Die Auswertung der AO-Klassifikation erfolgte nach Typ, Gruppe und Untergruppe. In Rahmen der Statistischen Auswertung zeigte sich, dass der Erhalt von 27 Einteilungsmöglichkeiten die Berechnung des Kappa Wertes nach FLEISS 1979 fast unmöglich machte. Da das Weglassen der Untergruppen die getroffene Aussage nicht veränderte und im klinischen Alltag keine Rolle spielt, wurden die entsprechenden Kappa Werte somit durch Typ und Gruppe gebildet.

ANDERSEN et al 1995 zeigten in Ihrer Arbeit, dass die Reduzierung der AO-Klassifikation auf 9 Gruppen ein nur geringfügiges Ansteigen des Kappa Wertes zur Folge hat. Nach weiterer Vereinfachung auf die Typen (A,B,C) konnten Sie gute bis fast exzellente Werte erreichen. Ähnliche Ergebnisse erreichten auch BEAULÉ et al 2003, DAI et al 2005, MARTIN et al 1997, SCHIPPER et al 2001, SIEBENROCK et al 1993, WALTON et al 2003, WEN-JIE JIN et al 2005 und WOOD et al 2005. Diese und ähnliche Ergebnisse werden in Tabelle 12 dargestellt.

Tabelle Interobserver und Intraobserver Unterschiede in der Literatur

Publikation	Klassifikationssystem	Kappa Statistik	
Siebenrock und Gerber	Neer: Proximaler Humerus	0,53	
Siebenrock und Gerber	AO: Proximaler Humerus subgruppen	0,42	
Siebenrock et al 1993	Neer: Schulter	0,30	
Kristiansen et al 1988	Neer: Schulter	0,30	
Sidor et al	Neer: Proximaler Humerus	0,52	
Brumback and Jones	Gustillo: Offene Frakturen, Weichteile	0,60	
Kreder et al	AO: Distaler Radius (Typ)	0,48	
Thomsen et al 1991	Lauge-Hansen (OSG)	0,55	
	Weber (OSG)	0,57	
Thomsen et al	Garden: Femurhals	0,39	
Smith et al	Ficat: Osteonekrosen	0,46	
Horn and Rettig 1993	Gustillo-Andersen: Offene Frakturen, Weichteile	0,53	
Dirschl und Adams 1997	Rüdi-Algower (OSG, distale Tibia)	0,48	
Martin et al 1997	Rüdi-Algower (OSG, distale Tibia)	0,46	
Martin et al 1997	distale Tibia: AO/ASIF segment 43	0,60	Typ
		0,38	Gruppe
Craig et al 1998	Sprunggelenk: AO/ASIF segment 44	0,77	Typ
		0,61	Gruppe
Dai and Jin 2005	Thoraco-lumbale WK-Frakturen	0,16 - 0,50	interobserver
		0,33 - 0,76	intraobserver
Humphrey et al 2005	Sanders: Calcaneusfrakturen	0,41	
Beaulé et al 2003	Letournel: Acetabulumfrakturen	0,51 - 0,74	interobserver
		0,64 - 0,83	intraobserver

Tabelle 12

Typ: bei der Auswertung der Ergebnisse wurden für die Kappa Statistik nur die Daten bezüglich Typ verwendet. Gruppe sowie Untergruppe wurden nicht berücksichtigt.

Gruppe: bei der Auswertung der Ergebnisse wurden für die Berechnung die Daten bezüglich Typ und Gruppe verwendet. Die Untergruppe wurde nicht berücksichtigt.

Somit ließ sich erkennen, dass unsere Werte mit denen der über die letzten Jahrzehnte hinweg publizierten Studien vergleichbar sind. Und wie schon von KORAN 1975 [first and second of two parts] beschrieben, bewegen sich die meisten klinischen Studienergebnisse im mäßigen Bereich.

Die Liste der Autoren, die über die Jahre Arbeiten über die Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit unterschiedlicher Klassifikationssysteme veröffentlicht haben ist lang. Die verschiedensten Klassifikationssysteme wurden im Rahmen von Studien über Frakturen des proximalen Humerus [KRISTIANSEN et al 1988, SIDOR et al 1993, SIEBENROCK et al 1993], des Schenkelhalses [FRADSEN et al 1988], der intertrochantären Schenkelhalsregion [ANDERSEN et al 1990] und des Sprunggelenkes [RASMUSSEN et al 1993, THOMSEN et al 1991] gegeneinander sowie untereinander bewertet

Die Studien im Bereich des proximalen Humerus zeigten moderate Ergebnisse. Indem er das gewichtete Kappa zur Auswertung verwendete, kam KRISTIANSEN et al 1988 auf nicht akzeptable Kappa-Werte (0,07-0,48) bezüglich der Interobserver-Zuverlässigkeit bei der Klassifikation der distalen Humerusfrakturen nach Neer. Daraus schloss er, dass die Beurteilung schwieriger Fälle erfahrenen Chirurgen beziehungsweise Radiologen überlassen werden sollte.

In der Arbeit von SIDOR et al 1993, welche ebenfalls die Einteilung von proximalen Humerusfrakturen in der Klassifikation nach Neer zum Thema hatte, bewegten sich die Interobserver Kappa Werte zwischen 0,37 bis 0,62, während die Intraobserver Kappa Werte den Bereich 0,50 bis 0,83 einnahmen.

SIEBENROCK et al 1993 untersuchten die Interobserver Zuverlässigkeit und die Intraobserver Reproduzierbarkeit bezogen auf proximale Humerusfrakturen als Vergleich zwischen der Neer-Klassifikation und der AO-Klassifikation. Ihre Schlussfolgerung besagt, dass keines der beiden Systeme ausreichend reproduzierbar ist um einen aussagekräftigen Vergleich von ähnlich klassifizierten Frakturen in unterschiedlichen Studien zu erlauben.

WAINWRIGHT et al 2000 zeigten in Ihrer Studie über die Interobserver und Intraobserver Variationen in Klassifikationssystemen für distale Humerusfrakturen, dass die angewendeten Systeme eine nur geringe bis mäßige Übereinstimmung zeigen. Somit stellten Sie in Frage, dass die benutzten Klassifikationssysteme in

Bezug auf Entscheidungsfindung und Ergebnisvergleich zu Rate gezogen werden. Die Ergebnisse waren nicht so gut, als dass die Klassifikationssysteme für beide genannte wichtige klinische Kriterien, Entscheidungsfindung und Ergebnisvergleich, eingesetzt werden sollten.

EDWARDS et al untersuchten 2002 die Interobserver und Intraobserver Reliability bei der Messung der Innenrotation der Schulter bezogen auf die Wirbelsäulenebene. Sie fanden eine geringe Interobserver Zuverlässigkeit und mäßige Intraobserver Reproduzierbarkeit. Obgleich die Messung der Schulterinnenrotation bezogen auf die Wirbelsäulenebene das Standardverfahren darstellt, ist die Reproduzierbarkeit zwischen den 13 Auswertern der Studie nicht gegeben.

ANDERSEN et al 1991 untersuchten die Interobserver Zuverlässigkeit und die Intraobserver Reproduzierbarkeit der Klassifikation nach Older bei der Einteilung von 185 distalen Radiusfrakturen. Die 4 Auswerter zeigten in beiden Untersuchungskriterien gute Übereinstimmung (Kappa 0,69-0,79). Es wird explizit darauf verwiesen, dass es keinen Unterschied zwischen den erfahrenen und unerfahrenen Auswertern gab. Die Auswerter erhielten eine Fortbildung auf Wunsch.

ANDERSEN et al 1995 verglichen in Ihrer Studie über distale Radiusfrakturen gleich vier Klassifikationssysteme miteinander: die Frykman, Melone, Mayo und die AO-Klassifikation. Es zeigten sich weder signifikante Unterschiede zwischen den erreichten Kappa Werten in den zwei dargestellten Durchgängen, noch in der Interobserver Zuverlässigkeit oder der Intraobserver Reproduzierbarkeit zwischen den Handchirurgen und den Radiologen als Auswerter der Studie. Daraus schlussfolgerten ANDERSEN et al 1995, dass aufgrund der geringen Interobserver Reproduzierbarkeit und Intraobserver Zuverlässigkeit die Rolle der in der Arbeit verglichenen Frakturklassifikationssysteme als alleinig entscheidend für den weiteren Therapieweg und den direkten Vergleich zwischen unterschiedlichen Studien nicht haltbar ist.

GEHRCHEN et al 1993 untersuchten die Interobserver Zuverlässigkeit und die Intraobserver Reproduzierbarkeit anhand von Trochanterfrakturen mittels der Evans-Klassifikation. Sie konnten eine gute Interobserver Zuverlässigkeit (Kappa 0,69-0,81) und eine ebenfalls gute Intraobserver Reproduzierbarkeit (Kappa 0,41-0,77) nachweisen. Es wurden die Interpretationsrichtlinien nach LANDIS und KOCH 1977 verwendet. Es sollte bei der Interpretation dieser Studie bedacht werden, dass die 52

ausgewerteten Röntgenbilder von nur 4 Auswertern gesichtet wurden. Es wird nicht erwähnt, ob die 4 Auswerter die Klassifikation unabhängig voneinander oder gemeinsam durchführten und welchen Ausbildungsstand sie angehörten. Somit ist die Aussagekraft sicherlich eingeschränkt und mit unserer nicht unbedingt zu vergleichen.

In Ihrer 1997 veröffentlichten Studie, wurden von GEHRCHEN et al 1997 all diese Studienbedingungen veröffentlicht, die eine Vergleichbarkeit ermöglichen. Sie untersuchten die Interobserver Zuverlässigkeit und die Intraobserver Reproduzierbarkeit von subtrochantären Frakturen anhand der Klassifikation nach Seinsheimer. Die Interobserver Zuverlässigkeit bewegte sich zwischen einem Kappa Wert von 0,20 und 0,57, was nach LANDIS und KOCH 1977 einer mäßigen Übereinstimmung entspricht. Die Intraobserver Reproduzierbarkeit schwankte zwischen Kappa Werten von 0,37 bis 0,72. Dies entspricht einer Übereinstimmung von ordentlich bis gut. Als Schlussfolgerung wird empfohlen die Klassifikation nach Seinsheimer für die Klassifizierung von subtrochantären Frakturen nicht zu verwenden, da diese ungenaue Ergebnisse liefert. Es wird vorgeschlagen die Ergebnisse mittels Fortbildungsmaßnahmen zu verbessern, wie dies RASMUSSEN et al 1993 zeigten. Diese konnten in Ihrer Studie belegen, dass eine gezielte Fortbildung mit dem Thema des in der Studie untersuchten Klassifikationssystems keinerlei Verbesserung des Ergebnisses brachte. Der Kappa Wert der Auswerter die eine Fortbildung erhielten lag bei 0,44, jener der Auswerter ohne Fortbildung bei 0,52. Somit lässt sich darlegen, dass der von GEHRCHEN et al 1997 vorgeschlagene Versuch der Verbesserung von Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit mittels Fortbildung nicht zum gewünschten Erfolg führt.

WEN-JIE JIN et al 2005 konnten in Ihrer Studie an intertrochantären Frakturen des proximalen Femurs eine sehr hohe Intra- und Interobserver Zuverlässigkeit nachweisen. Bei genauerer Betrachtung des Studienaufbaus zeigt sich, dass die Auswertung durch lediglich 5 hoch qualifizierte Fachärzte vorgenommen wurde. Somit stellt sich die Frage, ob bei Auswertung durch Ärzte unterschiedlichen Weiterbildungsstandes dasselbe Ergebnis erzielt worden wäre. Die Antwort geben WEN-JIE JIN et al 2005 indem sie darauf hinweisen, dass, unabhängig vom

Klassifikationssystem, die Erfahrung der Auswerter als einer der wichtigsten Faktoren die Zuverlässigkeit beeinflussen. In ihren Arbeiten kamen ACKERMANN et al 1986 und KRISTIANSEN et al 1988 die beide Frakturen des proximalen Humerus untersuchten, zum gleichen Ergebnis.

DIRSCHL und ADAMS 1997 untersuchten in Ihrer Studie die Interobserver Zuverlässigkeit an Tibiaplateaufrakturen mittels der Rüedi-Allgöwer und einem binären Klassifikationssystem. Sie mussten feststellen, dass es keinen Unterschied in der Interobserver Zuverlässigkeit zwischen beiden Systemen gab. Sie schlussfolgerten, das dies mitunter an der Komplexität der Tibiakopffrakturen liegt, und empfahlen weitere Forschungsarbeit bevor ein binäres Klassifikationssystem im klinischen Alltag Einzug erhält. Die beschriebenen Probleme der Übereinstimmung zwischen den Auswertern sind nicht auf Frakturklassifikationssysteme beschränkt. Viele im medizinischen Alltag verwendete und klinisch etablierte Tests haben ähnliche Probleme. VERESS et al 1993 untersuchten in Ihrer Studie die Interobserver-Zuverlässigkeit bezüglich der Ermittlung der Todesursache in Autopsiebefunden. Die Kappa-Werte zeigten nur moderate bis gute Ergebnisse (0,43-0,75). SMITH et al 1995 fanden in Ihrer Arbeit über die Interobserver Zuverlässigkeit bezüglich rektal-digitaler Untersuchungen im Rahmen der Prostatakrebsvorsorge eine schlechte Übereinstimmung mit einem Kappa Wert von 0,22. Fast gleiche Ergebnisse erhielten SNOEY et al 1994 (Kappa 0,32) beim Vergleich von Kardiologen und Notfallambulanzärzten in der Beurteilung von Elektrokardiogrammen.

Keine der zur Erstellung dieser Dissertationsarbeit durchgearbeiteten Publikationen oder die darin enthaltenen Querverweise konnten die Überlegenheit eines Klassifikationssystems den anderen Systemen gegenüber bezogen auf die Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit nachweisen.

Es konnte nur mehrfach gezeigt werden, dass diejenigen Systeme, welche in der klinischen Alltagspraxis verwendet werden, und mit denen der Anwender durch den täglichen Gebrauch Übung im Umgang entwickelt, einen Vorteil gegenüber rein akademisch gebrauchter Klassifikationssysteme haben.

Dies konnte in dieser Arbeit jedoch nicht belegt werden. Die hier miteinander verglichen Klassifikationssysteme – AO-Klassifikation und Moore-Klassifikation –

haben ähnlich schlechte Kappa Werte gezeigt. Dies gilt sowohl für die Interobserver Zuverlässigkeit, als auch für die Intraobserver Reproduzierbarkeit.
Aber nicht nur die Klassifikationssysteme werden kontrovers diskutiert. Auch die Kappa Statistik wird in der Literatur aus mehreren Gesichtswinkeln betrachtet. THOMSEN et al zeigten 2002 in Ihrer Studie, dass die Kappa Statistik schwierig zu interpretieren ist, abhängig von der Anzahl der Beobachtungen (Bilderdatensätze) und der Anzahl der Auswerter, und somit nicht alleinig als einfache Beurteilungsmethode der unterschiedlichen Auswertungsergebnisse dienen kann. Sie empfahlen die Anwendbarkeit der Kappa Statistik einer erneuten Evaluierung zu unterziehen.
Um die Aussagekraft unserer Ergebnisse zu steigern, erhielten die Assistenten im 1. Weiterbildungsjahr vor Beginn der 2. Auswertung eine ausführliche Fortbildung zu beiden Klassifikationssystemen. Wie unsere Ergebnisse zeigten, konnte keine Steigerung bezüglich der Interobserver Zuverlässigkeit oder der Intraobserver Reproduzierbarkeit erreicht werden. Somit kann der Ausbildungsstand als entscheidender Faktor zum Auswertungsergebnis im Rahmen dieser Studie ausgeschlossen werden.
Im dieser Dissertationsarbeit lässt sich somit festhalten, dass der Ausbildungsstand der Auswerter, und somit die Übung im Umgang mit den Frakturklassifikationssystemen im Bereich der proximalen Tibia, einen vernachlässigbaren Faktor darstellt. Weiterhin lässt sich auch der positive Einfluss einer Fortbildungsmaßnahme ausschließen Dies stimmt mit den Ergebnissen anderer Studien überein [ANDERSEN et al 1991, ANDERSEN et al 1995, JOHNSTONE et al 1993, MARTIN et al 1997, RASMUSSEN et al 1993].
In unserer Studien konnten wir unter Verwendung der gewichteten Kappa Statistik nach FLEISS et al 1979 zeigen, dass die Interobserver Zuverlässigkeit und die Intraobserver Reproduzierbarkeit bei der AO-Klassifikation im Vergleich mit der Moore Klassifikation keine eindeutigen Unterschiede aufweisen. Dies konnte sowohl für den direkten Vergleich aller Auswerter untereinander als auch beim auf die einzelnen Gruppen bezogenen Vergleich gezeigt werden. Der Ausbildungsstand der Auswerter ist bei keiner der beiden Klassifikationssysteme von Relevanz.

Selbst nach Durchführung einer Fortbildungsveranstaltung und direkt anschließender Auswertung konnte keine relevante Verbesserung der Inter- oder Intraobserver Werte in der Gruppe der Assistenten im 1. Weiterbildungsjahr erzielt werden.

Eine statistische Signifikanz zwischen beiden untersuchten Klassifikationssystemen kann nicht angegeben werden, da sich statistisch nur Systeme mit gleich vielen Einteilungsmöglichkeiten vergleichen lassen. Die AO-Klassifikation hat 9 Einteilungsmöglichkeiten (von uns werden nur Typ und Gruppe ausgewertet) und die Moore-Klassifikation nur 7 Einteilungsmöglichkeiten. Somit verbietet sich ein statistischer Vergleich zwischen beiden.

Nachdem viele Autoren Richtlinien für die Interpretation von Kappa-Werten veröffentlicht haben [CHAN et al 1997, KRISTIANSEN et al 1988, MARTIN et al 1997, RASMUSSEN et al 1993, SIEBENROCK et al 1993,], sollten wir uns für die in dieser Dissertationsarbeit vorgestellten Werte auf die Interpretation fokussieren, welche im Durchschnitt als Standard benutzt wird. Tabelle 1 zeigt eine Übersicht verschiedener Interpretationsrichtlinien für Kappa-Werte.

Beim direkten Vergleich der einzelnen Gruppen gegeneinander (Interobserver Zuverlässigkeit) bezogen auf die Einteilung nach der AO-Klassifikation bewegte sich der Kappa Wert zwischen 0,26 und 0,41. Dies stellt eine geringe Übereinstimmung dar. Wenn die Kappa-Werte untereinander verglichen werden (Intraobserver Reproduzierbarkeit), also der Kappa-Wert zum Zeitpunkt 0 mit dem zum Zeitpunkt 3 Monate, lassen sich die Werte bestätigen. Kappa 0,20 bis 0,46. Dies stellt eine Übereinstimmung von gering bis ordentlich dar.

Die Ergebnisse der Moore-Klassifikation zeigen Kappa Werte die zwischen 0,20 und 0,33 schwanken. Dies ist ebenfalls eine geringe Übereinstimmung dar. Dies gilt für die Auswertung bezogen auf die einzelnen Gruppen gegeneinander (Interobserver Zuverlässigkeit).

Beim Vergleich aller Auswerter untereinander (Intraobserver Reproduzierbarkeit) zum Zeitpunkt 0 und 3 Monate liegt der Kappa-Wert zwischen 0,26 und 0,48. Die Übereinstimmung bewegt sich zwischen gering und ordentlich.

7. Schlussfolgerung

ANDERSEN et al 1995, JOHNSTONE et al 1993 und MARTIN et al 1997 zeigten in ihren Arbeiten, dass der Ausbildungsstand und somit die damit erhaltene Übung im Umgang mit den Klassifikationssystemen keinen Einfluss auf die Auswertungsergebnisse hat. Geringe Unterschiede zwischen den einzelnen auswertenden Gruppen kommen vor, diese erreichen jedoch keine statistisch signifikante Werte. RASMUSSEN et al 1993 zeigten in Ihrer Studie, dass eine Fortbildungsmaßnahme keine positive Wirkung auf die Ergebnisse hatte. Somit muss auch in dieser Arbeit festgestellt werden, dass trotz unterschiedlichern Ausbildungsstands, unterschiedlicher klinischer Zugehörigkeit und allen Versuchen, durch vorangehender Fortbildung keine signifikante Verbesserung der Klassifikationseinteilung erreicht werden konnte. Ein Unterschied in der Einteilung nach der AO- oder Moore- Klassifikation konnten zu keinem Zeitpunkt festgestellt werden. Obwohl die AO-Klassifikation als Standard in Deutschland eingesetzt wird, ist eine gute (0,61-0,80) bis exzellente (0,81-1,0) Klassifikation bei keiner der 4 auswertenden Gruppen zu erzielen. Diese Ergebnisse werden durch alle in dieser Dissertationsarbeit vorgestellten Publikationen bestätigt und führen nochmals den akademischen Charakter beider Klassifikationssysteme vor Augen. Obwohl für die klinische Alltagspraxis entwickelt, zeigt sich, dass die Reproduzierbarkeit nicht ausreichend gegeben ist. Diese Aussage wurde auch von ANDERSEN et al 1995, SIEBENROCK und GERBER 1992 und SIEBENROCK et al 1993 bestätigt, die Rolle der Frakturklassifikationssysteme als alleinig entscheidend für den weiteren Therapieweg ablehnten.

Wir konnten in dieser Dissertationsarbeit keine Überlegenheit eines Klassifikationssystems gegenüber dem anderen feststellen. Es bleibt zu sagen, dass die verglichenen Klassifikationssysteme nützliche Gegenstände sind, aber wie auch von anderen Autoren beschrieben, nicht alleinig entscheidend für den weiteren therapeutischen Weg sein dürfen.

Aus den herausgearbeiteten Studienergebnissen lassen sich mehrere Schlussfolgerungen ableiten:

1. Die Ergebnisse für die Einteilung der Tibiakopffrakturen nach der AO- oder Moore-Klassifikation zeigten gleich schlechte Werte bezüglich Interobserver Zuverlässigkeit und Intraobserver Reproduzierbarkeit.
2. Bei Betrachtung der Intraobserver Reproduzierbarkeit mit Standardfehler fanden sich bei der Gruppe der auswärtigen Chefärzte und in der Gruppe der Assistenzärzte im Verlauf (Vergleich 3 Monate versus 0 Monate) bessere Werte. Eine statistische Signifikanz konnte jedoch nicht erreicht werden.
3. Die Oberärzte hingegen klassifizierten zu beiden Zeitpunkten gleich.
4. Der Ausbildungsstand und somit die Jahre der Erfahrung haben im klinischen Alltag keine statistisch signifikant nachweisbare Bedeutung. Dies gilt für beide Klassifikationssysteme.
5. Die in der Gruppe der Assistenten im 1. Weiterbildungsjahr durchgeführte Fortbildungsveranstaltung konnte die Ergebnisse der Intraobserver Reproduzierbarkeit nicht steigern.
6. Weitere Kriterien sind zu erarbeiten für ein suffizientes, therapiebezogenes Klassifikationssystem im Bereich der proximalen Tibia.

8. Anhang

Abbildung Nr. 3

Patient 21

AO-Klassifikation

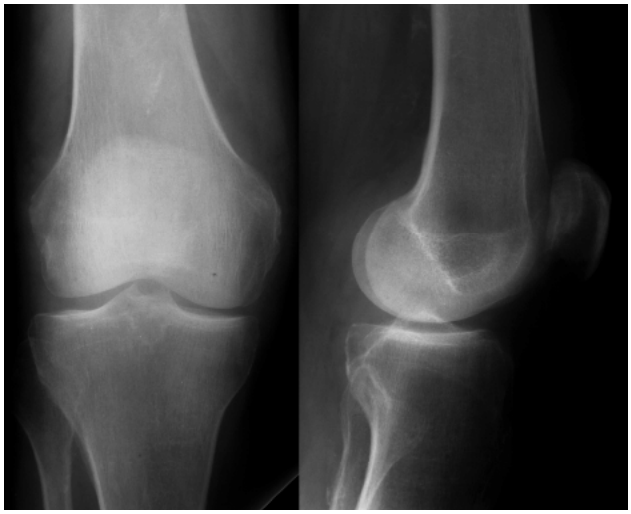
A1	A1.1	☐
	A1.2	☐
	A1.3	☐
A2	A2.1	☐
	A2.2	☐
	A2.3	☐
A3	A3.1	☐
	A3.2	☐
	A3.3	☐
B1	B1.1	☐
	B1.2	☐
	B1.3	☐
B2	B2.1	☐
	B2.2	☐
	B2.3	☐
B3	B3.1	☐
	B3.2	☐
	B3.3	☐
C1	C1.1	☐
	C1.2	☐
	C1.3	☐
C2	C2.1	☐
	C2.2	☐
	C2.3	☐
C3	C3.1	☐
	C3.2	☐
	C3.3	☐

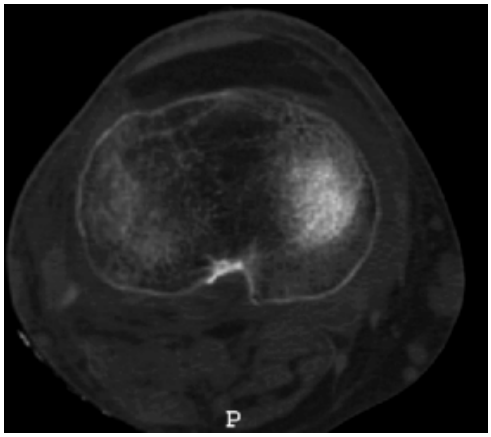
Klassifikation nach Tile and Moore

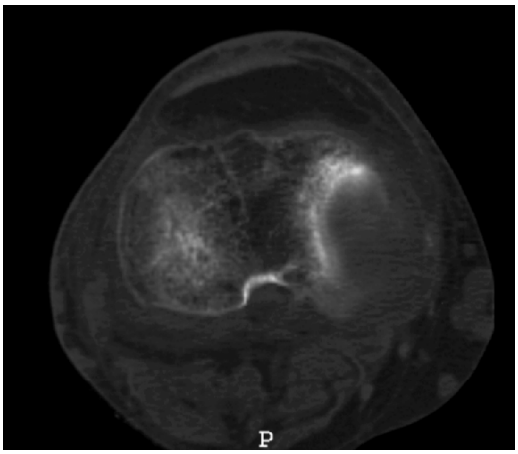
Type I		☐
Type II	medial	☐
	lateral	☐
Type III		☐
Type IV	medial	☐
	lateral	☐
Type V		☐

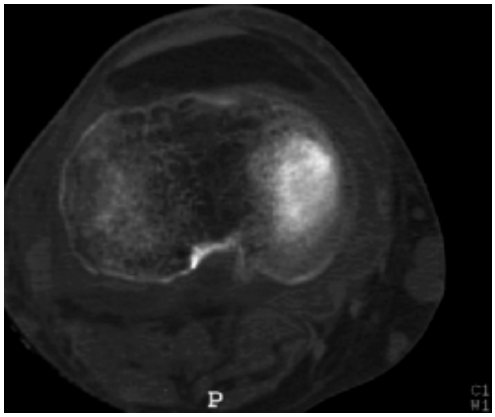
Auswertungsbogen

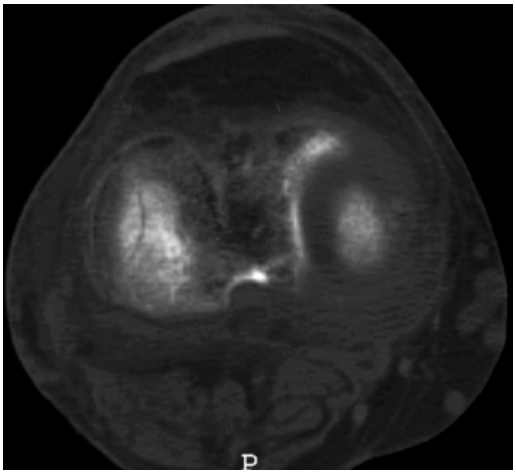
Beispiel Patientenbildersatz

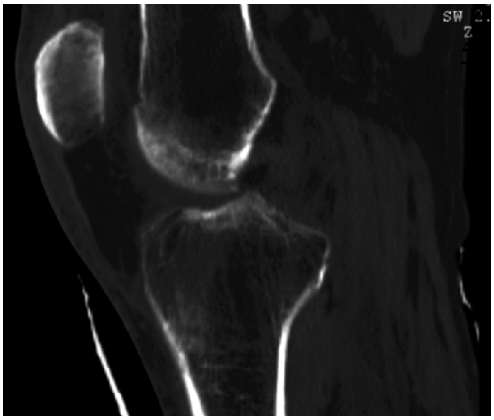


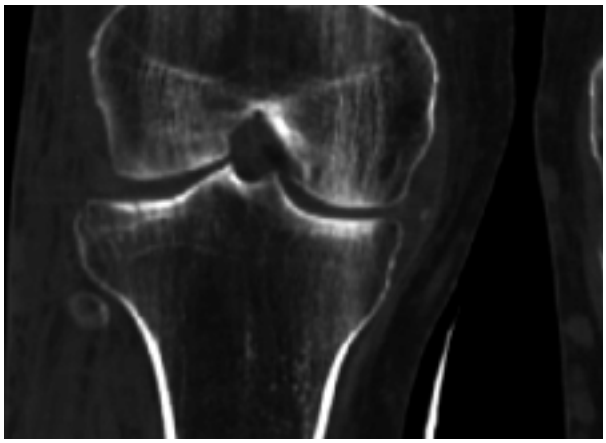


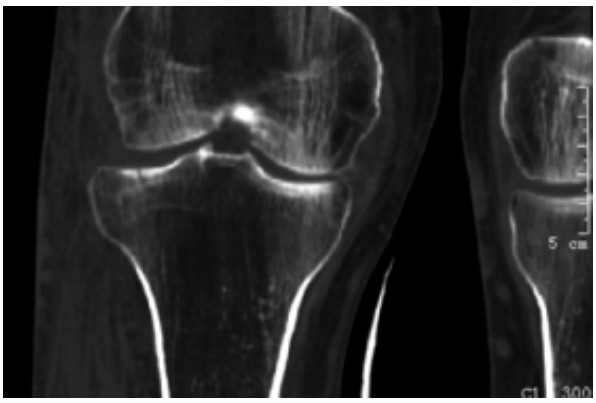












9. Literaturliste

1 Ackermann C, Lam Q, Linder P, Kull C, Regazzoni P (1986) Zur Problematik der Frakturklassifikation am proximalen Humerus. Zeitschrift für Unfallchirurgie 79:209-215

2 Andersen DJ, Blair WF, Steyers CM, Adams BD, El-Khouri GY, Brandser EA (1995) Classification of Distal Radius Fractures: An Analysis of Interobserver Reliability and Intraobserver Reproducibility. The Journal of Hand Surgery 21A:574-582

3 Andersen E, Jorgensen LG, Hededam LT (1990) Evans' Classification of trochanteric fractures: an assessment of the interobserver and intraobserver reliability. Injury 21:377-378

4 Andersen GR, Rasmussen J-B, Dahl B, Solgaard S (1991) Older's Classification of Colle's fractures. Acta Orthopaedica Scandinavia 62:463-464

5 Audigé L, Bhandari M, Kellam J (2001) How reliable are reliability studies of fracture classifications?. Acta Orthopedica Scandinavia 75:184-194

6 Beaulé PE, Dorey JD, Matta JM (2003) Letournel classification for acetabular fractures: assessment of interobserver and intraobserver reliability. Journal of Bone and Joint Surgery Am. 85: 1704-1709

7 Borelli JB Jr., Goldfarb C, Catalano L, Evanoff BA (2002) Assesment of Articular Fragment Displacement in Acetabular Fractures: A Comparison of Computerized Tomography and Plain Radiographs. Journal of Orthopaedic Trauma 16:449-456

8 Brien H, Noftall F, MacMaster S, Cummings T, Landells C, Rockwood P (1995) Neer's classification System: a critical Appraisal. The Journal of Trauma 38:257-260

9 Burstein AH Ph. D. (1993) Editorial Fracture Classification Systems: Do they work and are they useful? The Journal of Bone and Joint Surgery 75:1743-1744

9 Chan PSH, Luchetti JJ, Wayne T, Esterhai JL, Kneeland JB, Murray K, Heppenstall BR (1997) Impact of CT Scan on Treatment Plan and Fracture Classification of Tibial Plateau Fractures. Journal of Orthopaedic Trauma 11:484-489

10 Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37-46

11 Colton CL (1991) Editorial Telling the bones. Journal of bone and Joint Surgery 73:362-364

12 Craig WL III, Dirschl DR (1998) Effects of Binary Decision Making on the Classification of Fractures of the Ankle. Journal of Orthopaedic Trauma 12:280-283

13 Dai L-Y, Jin W-J (2005) Interobserver and Intraobserver reliability in the Load Sharing Classification of the assessment of thoracolumbal burst fractures. Spine 30 (3): 354-358

14 Dirschl DR, Adams GL (1997) A Critical Assessment of Factors Influencing Reliability in the Classification of Fractures, Using Fractures of the Tibial Plafond as a Model. Journal of Orthopaedic Trauma 11:471-476

15 Edwards TB, Bostick RD, Green CC, Baratta RV, Drez D (2002) Interobserver and intraobserver reliability of the measurement of shoulder rotation by vertebral level. Journal of Shoulder and Elbow Surgery 11:40-42

16 Fleiss JL (1971) Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin 76:378-382

17 Fleiss JL, Nee JCM, Landis JR (1979) Large Sample Variance of Kappa in the Case of Different Sets of Raters. Psychological Bulletin 86:974-977

18 Fleiss JL, Shrout PE (1994) Intraclass Correlations: Uses in assessing rater Reliability. Psychological Bulletin 86:124-132

19 Fleiss JL (1981) Statistical Methods for Rates and Proportions, Second Edition. New York

20 Fradsen PA, Andersen E, Madsen F, Skjodt T (1988) Gardens Classification of femoral neck fractures. Journal of Bone and Joint Surgery 70:588-590

21 Gehrchen PM, Nielsen JO, Olesen B (1993) Poor reproducibility of Evan's classification of the trochanteric fracture. Acta Orthopaedica Scandinavica 64:71-72

22 Gehrchen PM, Nielsen JO, Olesen B, Andresen BK (1997) Seinsheimer's classification of subtrochanteric fractures. Acta Orthopaedica Scandinavica 68:524-526

23 Hertel P (1997) Tibiakopffrakturen. Der Unfallchirurg 100:508-523

24 Horn BD, Rettig ME (1993) Interobserver Reliability in the Gustilo and Anderson Classification of Open Fractures. Journal of Orthopaedic Trauma 7:357-360

25 Humphrey CA, Dirsch DR, Ellis TJ (2005) Interobserver Reliability of a CT-based fracture classification system. Journal of Orthopaedic Trauma 19 (9): 616-622

26 Hüfner T, Pohlemann T, Gänsslen A, Assassi P, Prokop M, Tscherne H (1999) Die Wertigkeit der CT zur Klassifikation und Entscheidungsfindung nach Acetabulumfrakturen. 102:124-131

27 Johnstone DJ, Radford WJP, Parnell EJ (1993) Interobserver variation using the AO/ASIF classification of long bone fractures. Injury 24:163-165

28 Koran L. M.(1975) The Reliability of clinical Methods, Data and Judgments (first of two parts). The new England Journal of Medicine 25: 642-646

29 Koran L. M.(1975) The Reliability of clinical Methods, Data and Judgments (second of two parts). The New England Journal of Medicine 25: 695-701

30 Kraemer HC, Periyakoil VS, Noda A (2002) Tutorial in Biostatistics: Kappa coefficients in medical research. Statistics in medicine 21:2109-2129

31 Kreder HJ, Hanel DP, McKee M, Jupiter J, McGillivary G, Swiontkowski MF (1996) Consistency of AO fracture classification for the distal radius. Journal of Bone and Joint Surgery 78:726-731

32 Kristiansen B, Christensen SW (1987) Proximal Humeral Fractures. Acta Orthopaedica Scandinavica 58:124-127

33 Kristinasen B, Andersen ULS, Olsen CA, Varmarken J-E (1988) The Neer Classification of fractures of the proximal humerus. Skeletal Radiology 17:420-422

34 Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159-174

35 Lichtenhahn P, Fernandez DL, Schatzker J (1991) Analyse zur „Anwenderfreundlichkeit“ der AO-Klassifikation für Frakturen. Helvetica Chirurgica Acta 58:919-924

36 Lindsjö U (1985) Classification of Ankle Fractures: The Lauge-Hansen or AO System? Clinical Orthopaedics and Related Research 199:12-16

37 Martin JS, Marsh JL, DeCoster T, Bonar S, Found E, Brandser E (1996) Assessment of the AO/ASIF Fracture Classification for the tibial plafond. Orthopaedic Transplantation 20:98

38 Martin JS, Marsh JL, Susan K, DeCoster TA, Found EM, Brandser EA (1997) Assessment of the AO/ASIF Fracture Classification for the Distal Tibia. Journal of Orthopaedic Trauma 11:477-483

39 Martin JS, Marsh JL (1997) Current Classification of Fractures. Radiologic Clinics of North America 35:491-506

40 Moore TM (1981) Fracture Dislocation of the Knee. Clinical Orthopaedics and Related Research 150:128-140

41 Müller M.E., Nazarin S., Koch P., Schatzker J (eds) (1990) The Comprehensive Classification of Fractures of the Long Bones. Springer, Berlin Heidelberg New York

42 Nelitz M, Guenther K-P, Gunkel S, Puhl W (1999) Reliability of radiological measurements in the assessment of hip dysplasia in adults. The British Journal of Radiology 72:331-334

43 Niels OB, Overgaard S, Olsen LH, Hansen H, Nielsen ST (1991) Observer Variation in the Radiographic Classification of Ankle Fractures. Journal of Bone and Joint Surgery 73:676-678

44 Nielsen JO, Dons-Jensen H, Sorensen HAT (1990) Lauge-Hansen classification of malleolar fractures. Acta Orthopaedica Scandinavica 61:385-387

45 Petrisor BA, Bhandari M, Orr RD, Mandel S, Kwok DC, Schemitsch EH (2003) Improving reliability in the classification of fractures of acetabulum. Archives of Orthopaedic and Trauma Surgery 2003

46 Rasmussen S, Madsen PV, Bennicke K (1993) Observer Variation in the Lauge-Hansen classification of ankle fractures. Acta Orthopaedica Scandinavica 64:693-694

47 Schatzker J, F.R.C.S., McBroom R, Bruce D (1979) The Tibial Plateau Fracture: The Toronto Experiment. Clinical Orthopaedics and Related Research 138:94-104

48 Schipper IB, Steyerberg EW, Castelein RM, Vugt van AB (2001) Reliability of the AO/ASIF classification for pertrochanteric femoral fractures. Acta Orthopaedica Scandinavica 72:36-41

49 Shrout PE, Fleiss JL (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 86:420-428

50 Sidor ML , Zuckerman JD, Lyon T, Kenneth K, Coumo F, Schoenberg N (1993) The Neer Classification System for Proximal Humerus Fractures. Journal of Bone and Joint Surgery 75:1745-1750

51 Siebenrock KA, Gerber C (1992) Frakturklassifikation und Problematik bei proximalen Humerusfrakturen. Der Orthopäde 21:98-105

52 Siebenrock KA, Gerber C, Schweiz B (1993) The Reproducibility of Classification on Fractures of the Proximal End of the Humerus. Journal of Bone and Joint Surgery 75:1751-1755

53 Smith RM (2000) Editorial: The Classification of Fractures. Journal of Bone and Joint Surgery 82:625-626

54 Svanholm H, Starklint H, Gundersen HJG, Fabricus J, Barlebo H, Olsen S (1989) Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS 97:689-698

55 Swiontkowski MF, Sands AK, Agel J, Diab M, Schwappach JR, Kreder H.J. (1997) Interobserver Variation in the AO/OTA Fracture Classification System for Pilon Fractures: Is there a Problem. Journal of Orthopaedic Trauma 11:467-470

53 Thomsen NOB, Jensen CM, Skovgaard N, Pedersen MS, Pallesen P, Soe-Nielsen NH, Rosenklint A (1996) Observer Variation in the radiographic classification of fractures of the neck of the femur using Garden's system. International Orthopaedics 20:326-329

54 Thomsen NOB, Olsen LH, Nielsen ST (2002) Kappa statistics in the assessment of observer variation: the significance of multiple observers classifying ankle fractures. Journal of Orthopaedic Science 7:163-166

55 Tscherne H, Lobenhoffer P (1993) Tibial Plateau Fractures. Clinical Orthopaedic and Related Research 292:87-100

56 Wainwright AM, Williams JR, Carr AJ (2000) Interobserver and intraobserver variation in classification systems for fractures of the distal humerus. Journal of Bone and Joint Surgery 82:636-642

57 Walton NP, Harish S, Roberts C, Blundell C (2003) AO or Schatzker? How reliable is classification of tibial plateau fractures. Archive of Orthopaedic Trauma Surgery 123:396-398

58 Wen-Jie J, Li-Yang D, Yi-Min C, Qing Z, Lei-Sheng J, Hua L (2005) Reliability of classification systems for intertrochanteric fractures of the proximal femur in experienced orthopaedic surgeons. Injury 36:858-861

59 Wood KB, Khanna G, Vaccaro AR, Arnold PM, Harris MB, Mehbod AA (2005) Assessment of two thoracolumbar fracture classification systems as used by multiple surgeons. Journal of Bone an Joint Surgery Am. 87: 1423-1429

10. Danksagung

Mein respektvoller Dank gilt meinem Betreuer Hr. Dr. Reiner Wirbel, der immer Zeit und ein offenes Ohr für meine Probleme hatte. Weiterhin danke ich Ihm für die Mühe und Zeit die er bei der Korrekturarbeit aufgewendet hat.

Die Statistische Auswertung entstand am Institut für Medizinische Biometrie und Statistik der Universitätskliniken des Saarlandes in Homburg/Saar unter besonderer Hilfestellung durch Hr. Dr. sc. hum König. Er half mir stets bei allen Fragen die Statistik betreffend und sorgte dafür, dass ich das schwierige Thema in verständlichen Worten ausdrücken konnte. Hierfür möchte ich mich bei Ihm ganz herzlich bedanken.

Ich möchte mich außerdem noch bei meinem Ehemann Rüdiger Vrabac für seinen Beistand während der mühevollen Zeit der Entstehung dieser Dissertationsarbeit.

Die Doktorarbeit hat viel Zeit und Mühe gekostet, trotzdem empfinde ich als persönliche Bereicherung.

Vielen Dank

11. Tabellarischer Lebenslauf

Persönliche Daten

Name, Vorname:	Vrabac, Cristina Elena
Anschrift:	An der Ziegelhütte 23
	66484 Schmitshausen
Geburtsdatum:	18.03.1973
Geburtsort:	Bukarest/Rumänien
Familienstand:	verheiratet

Schulausbildung

Sept. 1979- April 1984	Grundschule Bukarest
April 1984- Juni 1985	Hauptschule Lebach
Aug. 1985- Juli 1987	Marienschule Saarbrücken
Sept. 1987- Juni 1989	Abschluß Realschule Saarbrücken
Juni1992	Abitur am Wirtschaftswissenschaftlichen Gymnasium Saarbrücken

Hochschulausbildung

Okt. 1992- Feb. 1993	Studium der Chemie
Okt. 1993- 2001	Studium der Medizin an der Universität des Saarlandes in Homburg/Saar
April 2000	Abschluß des Medizinischen Staatsexamens

Facharztanerkennung

21 Juni 2006	Fachärztin für Arbeits- und Betriebsmedizin

Arbeitsverhältnisse

Mai 2000- Okt. 2001	Ärztin im Praktikum in der Abteilung für Allgemeine, Viszeral- und Gefäßchirurgie der Chirurgischen Universitätskliniken des Saarlandes in Homburg
April 2002- März 2003	Assistenzärztin in der Inneren Abteilung des Kreiskrankenhauses Blaubeuren
Seit April 2003	Weiterbildungsassistentin im Fachbereich Arbeits- und Betriebsmedizin bei der B.A.D GmbH in Kaiserslautern
Seit 21 Juni 2006	Fachärztin für Arbeits- und Betriebsmedizin