Leistungserfassungsinstrumente der Pflege

Literaturarbeit

E. Näf
ID Nr. 983764
Fakultät der Gesundheitswissenschaften
Universität Maastricht NL / Aarau CH
Master of Nursing Science

Mai 2003

Begleitung: W. Fischer, lic. oec. HSG, Zentrum für Informatik und wirtschaftliche Medizin, Wolfertswil

Ernst Näf, Eptingerstr. 26, 4052 Basel
E-Mail G: ernst.naef@bethesda.ch / E-Mail P: ernst@email.ch
Tel. G: +41 (61) 315 22 35 / Tel. P: +41 (61) 311 36 57

Inhaltsverzeichnis

1 Einleitung

In den letzten Jahren ist die Forderung, Pflege zu beschreiben immer häufiger und vernehmbarer geäussert worden. Die Ansätze, welche dafür gewählt werden, sind durchaus unterschiedlich. Die beschreibende Vorgehensweise und damit das qualitative Paradigma hat eine lange Tradition und wurde gerade auch in den deutschsprachigen Ländern bevorzugt. In einer Zeit, in der aber zunehmend Zahlen bestimmend sind, ist der Ruf nach quantifizierten und damit so genannt „objektiven“ Grundlagen für die Beschreibung der Pflege immer lauter geworden. Die von Norma Lang (2003) geprägte Aussage „If we cannot name it (nursing), we cannot control it, practice it, research it, teach it, finance it, or put it into public policy“ dient(e) vielen Pflegewissenschaftlerinnen und Pflegewissenschaftern als Ansporn für die Entwicklung von Methoden, mit welchen Pflege strukturiert benannt, gemessen und dargestellt werden kann.
Seit den Fünfzigerjahren hat der „Pflegeprozess“ in den USA eine wichtige Struktur für das Denken in der Pflege vorgegeben (Pesut & Herman, 1999). Die Schweizerinnen Fiechter und Meier machten in Europa 1981 den Pflegeprozess zum Thema (Linck, 1995). Der Pflegeprozess hat in seiner Entwicklung verschiedene Formen angenommen. Isfort (2001) beschreibt drei verschiedene Typen; das 4-, das 5- und das 6-Schritt-Modell. Die folgenden Prozessschritte sind dabei aber allen Modellen gemeinsam (auch wenn die Benennung nicht einheitlich ist):

1. eine Form der Informationssammlung
2. Planung von Pflegehandlungen
3. Durchführung der Pflegehandlungen
4. Evaluation der Wirkung

Methoden und Instrumente, welche der Aufforderung nach strukturierter Benennung und Messung von Pflege nachkommen, lassen sich denn auch alle mindestens einem der obigen Prozessschritte zuordnen.

Pflege-Kosten machen die grösste Portion des Personalbudgets eines Spitals aus (Ebener, 1985, S.324; JPPC, S.4). Gemessen an den Gesamtkosten eines Spitals werden, je nach Quelle, zwanzig bis sechzig Prozent vom Pflegebereich verbraucht (Fischer, 2001; Phillips, Castorr, Prescott, & Soeken, 1992, S.46; Sherman, 1990, S.12) (die Berechnungsweise ist dabei offensichtlich unterschiedlich). Angesichts dieser Zahlen und im Wissen um die Notwendigkeit einer ständigen Steuerung der Ressourcen ist klar, dass es sich lohnt, genau zu verfolgen, an welchen Orten und unter welchen Bedingungen diese Kosten entstehen.
Anfang 2000 hat der Arbeitgeber des Autors dieser Arbeit ein Projekt zur Einführung einer Pflegeleistungserfassung mit der Methode LEP (Leistungserfassung in der Pflege) gestartet. Die Einführung der Pflegeleistungserfassung erfolgte neben den oben erwähnten Gründen auch im Bewusstsein, dass gemäss Artikel 76 der Verordnung zum Schweizerischen Krankenversicherungsgesetz die Versicherer vom Leistungserbringer Angaben über Art und Umfang der erbrachten Leistungen einfordern können. Mit LEP wird der dritte Prozessschritt dokumentiert; es geht also um die Pflegeleistungen.
Das Projekt der LEP-Einführung ist zur Zeit in der abschliessenden Phase. Der Autor besetzt die Funktion der Projektleitung. Im Rahmen dieser Projektleitung hat sich die Frage gestellt, wie weit die Methode LEP reliabel ist.

2 Zweck

Der Zweck dieser Literaturarbeit ist die Besprechung verschiedener Formen von Pflegeleistungserfassungsinstrumenten[1], sowie der Möglichkeiten, bei solchen die Validität und insbesondere die Reliabilität zu bestimmen. Dem Instrument LEP wird dabei besonderes Augenmerk geschenkt, da die Literaturarbeit als Hintergrund für eine Reliabilitätsstudie dieses Instrumentes dienen soll.

3 Methode der Literatursuche und -auswahl

In einem ersten Schritt wurden die Datenbanken Medline und CINAHL (Cumulative Index to Nursing and Allied Health Literature) nach Publikationen mit Erscheinungsjahr 1990 oder jünger durchsucht. Folgende MeSH (Medical Subject Headings) und Subheadingkombinationen wurden dabei - teilweise einzeln, teilweise kombiniert - verwendet:

Medline
- Inpatients/classification
- Nursing Acuity
- Nursing Administration Research
- Nursing Administration Research/methods
- Nursing Assessment
- Nursing Intensity
- Nursing Staff, Hospital/psychology/supply & distribution
- Nursing Staff/supply & distribution
- Personnel Staffing and Scheduling/organization & administration
- Personnel Staffing and Scheduling/standards
- Reproducibility of Results
- Time and Motion Studies
- Workload
- Workload/standards

CINAHL
- Instrument-Validation
- Interrater-Reliability
- Intrarater-Reliability
- Patient-Classification
- Personnel-Staffing-and-Scheduling
- Reliability-and-Validity
- Workload-Measurement

Gesucht wurden damit speziell Artikel zu folgenden Themenbereichen:
- Reliabilität / Validität von Pflegeaufwandmessinstrumenten allgemein
- Pflegeaufwandmessinstrumente im stationären Akutbereich
- Pflegeaufwandmessinstrumente ausserhalb des stationären Akutbereiches, sofern ein Hauptteil des Artikels der Bestimmung von Reliabilität / Validität gewidmet ist

[1] Hauptgewichtig interessierten Instrumente in der Akutpflege. Die vielen Instrumente der Geriatrie- und Langzeitpflege wurden nicht berücksichtigt.

Folgende Artikel wurden vorerst ausgeschlossen:

- Sprache nicht Englisch oder nicht Deutsch
- kein Abstract
- besonders aufwändiger / teurer Beschaffungsweg

Das Auffinden relevanter Artikel erwies sich als schwierig, da zum gewünschten Thema in den Siebzigerjahren einerseits viel publiziert wurde, die Verschlagwortung in Medline sich aber als sehr inkonsistent und unzuverlässig zeigte und zudem Abstracts in den früheren Jahren in Pflegezeitschriften selten waren. CINAHL erwies sich in der Verschlagwortung als wesentlich besser, beginnt aber einerseits erst mit dem Jahr 1982 und enthält andererseits nicht mehr Abstracts als Medline. Somit konnten einige relevante Artikel erst mittels Schneeballsuche gefunden werden. Der Einsatz von „Related Articles“ in PubMed (Internetbasierte Form von Medline) brachte ebenfalls einige Publikationen zu Tage. Einige Arbeiten konnten über eine Suche in der Bibliothek am WE’G und eine Internetsuche mittels der Suchmaschine „Google“ gefunden werden.
Publikationen über Instrumente, welche nur den autonomen Bereich der Pflege erfassen (und damit weisungsgebundene Tätigkeiten ausschliessen) wurden nachträglich ausgeschlossen.
Es zeigte sich, dass wenig ganz aktuelle Literatur vorhanden ist; das Thema scheint im Moment im angelsächsischen Raum nicht brennend zu sein. Neuere Publikationen wie etwa Hernandez (1996a) fassen vieles zusammen, was in früheren Jahren schon publiziert wurde, ohne wesentlich Neues dazuzufügen.

4 Geschichtliche Entwicklung und Verbreitung

Der Versuch, pflegerische Leistungen mit Aufwandmesssystemen zu erfassen, begann schon 1947 (Abdellah & Levine, 1979, S.477). Im Bereich der Pädiatrischen Pflege wurde ein erstes Instrument entwickelt, das die Menge an benötigter Pflegezeit für die einzelnen mit dem Instrument fest zu legenden Kategorien aber erst sehr grob zuwies, und so noch ein zu wenig sensitives Instrument darstellte. Trotz dieser ersten Ansätze wurde bis in die Sechzigerjahre praktisch überall nur mit durchschnittlichen Stunden pro Pflegetag gerechnet (O'Brien-Pallas, Giovannetti, Peereboom, & Marton, 1995, S.9). Die Ergebnisse der Dissertation von Robert Connor Anfangs der Sechzigerjahre zeigten (die heute als höchst selbstverständliche betrachtete Tatsache), dass der Pflegeaufwand nicht primär von der Anzahl belegter Betten, sondern sehr stark von weiteren Faktoren abhängt (McHugh & Dwyer, 1992, S.21). Diese für die damalige Zeit (vermutlich nur für das Management, nicht aber für die praktisch tätigen Pflegenden) neue Erkenntnis führte vor allem in den USA zur Entwicklung einer Vielzahl von Aufwandmessinstrumenten (Edwardson & Giovannetti, 1994, S.100, S.104).
Die Instrumente scheinen im US-amerikanischen und kanadischen Raum häufig eingesetzt zu werden. Malloch und Conovaloff (1999, S.49) berichten von einer weiten Verbreitung schon in den Sechzigerjahren. Gemäss Edwardson und Giovannetti (1994) wird generell angenommen, dass die meisten Spitäler der USA mit mehr als fünfzig Betten ein Pflegeaufwandmesssystem einsetzen. Eine Erhebung in Kanada Ende der Achtzigerjahre zeigte, dass 67% der Spitäler mit mehr als fünfzig Betten und über achtzig Prozent der Spitäler mit mehr als dreihundert Betten ein entsprechendes Instrument einsetzen. 1996 setzten 89% aller Spitäler in Ontario ein Pflegeaufwandmessinstrument ein, wobei der Grossteil auf einige we-

nige Systeme fällt (53% GRASP, 19% NISS, 12% Medicus, 5% andere)[2] (JPPC, S.3). Aufhorchen lässt die nicht speziell vertrauenerweckende Aussage, dass viele grössere Einrichtungen im Gesundheitswesen innerhalb weniger Jahre ihr System auf der Suche nach dem richtigen drei bis vier Mal ausgewechselt hätten (Edwardson & Giovannetti, 1994, S.117). Eine ähnliche Aussage macht Botter (2000, S.544), welche eine Studie von 1988 zitiert, wonach mehr als fünfzig Prozent der befragten Institutionen angaben, dass sie ihr Instrument noch kein Jahr hätten.
In den nordischen Ländern begann die Entwicklung von Pflegeaufwandmessinstrumenten gemäss Fagerström (1998) Ende der Sechzigerjahre. Fagerström vermutet, dass seit Anfang der Neunzigerjahre in fast allen Gemeinden Finnlands ein Instrument gebraucht wird.
In Deutschland wurde 1992 der Einsatz der Pflegepersonalregelung PPR bundesweit verordnet, um den Bedarf an Pflegepersonal auf einheitlicher Basis zu gewährleisten (Klee, 1993). Die PPR wurde 1996 zwar wieder zurück genommen (Höfert, 2003), wird aber teilweise weiter verwendet (Fischer, 2002, S.140).
In der Schweiz wurde in den Jahren 1965 – 1973 die später auch in Deutschland und Österreich verbreitete „Exchaquet-Methode“ entwickelt und breit angewandt (Exchaquet & Züblin, 1975; Güntert & Maeder, 1994). Bei der Methode nach Exchaquet und Züblin handelt es sich um ein einfach zu handhabendes Prototypenmodell (vergleiche Abschnitt 6.1) für den stationären Akutbereich mit ursprünglich drei Kategorien, welches aber bald von einzelnen Institutionen verändert wurde (Güntert & Maeder, 1994). Später wurden weitere Methoden im stationären Akutbereich der Schweiz eingesetzt: RME, Pflegeanalyse (Thurgau), VBK-Methode, Psych-PV, LEP, PRN [3] (Fischer, 1995), wobei LEP mit 106 Anwendungen das weitaus verbreitetste Instrument ist, gefolgt von PRN, das in zehn Institutionen, primär in der französischen Schweiz Anwendung findet. Das Instrument LEP scheint sich zu einer Art Quasi-Standard durchzusetzen; es ist auch in den Sprachen Französisch und Italienisch erhältlich.

5 Leistungserfassungsinstrumente der Pflege – verwendete Begriffe

In der deutschsprachigen Literatur sind verschiedenste Begriffe anzutreffen, wenn es darum geht, pflegerische Leistungen – für welchen Zweck auch immer - zu messen. Folgende Beispiele mögen dies verdeutlichen:
„Pflegeleistungserfassung“ (Brosziewski & Brügger, 2001), “Pflegeleistungsmessung“ (Fischer, 2002), “Messinstrument für die Pflegearbeit“ (Brosziewski & Maeder, 2001), “Personalbemessung“, “Messung vom Pflegezeitbedarf“ (Bartholomeyczik & Hunstein, 2000), “Patientenklassifikation“ (Maeder, Brügger, Longerich, & Güntert, 1992), “Pflegeaufwandmessung“ (Fischer, 1995; Willems, 1992).
Teilweise sind die Begriffe Ausdruck dafür, dass Erfassungsinstrumente für einen bestimmten oder mehrere Zwecke eingesetzt werden, teilweise sind sie Ausdruck für den Zeitpunkt im Pflegeprozess, zu welchem die Erfassung vorgenommen wird. Die genaue Bedeutung der Begriffe und der Zusammenhang mit ähnlichen Begriffen muss teilweise aus dem Kontext gelesen werden, in welchem sie verwendet werden, weil entsprechende Definitionen fehlen. Eine Ausnahme bildet Fischer (1995; 2002).

[2] GRASP: GRACE Reynolds Application of the Study PETO (ausführlich siehe Abkürzungsverzeichnis S. 37) / NISS: Nursing Information System Saskatchewan
[3] RME: Référentiels médico-économiques; PRN: Project de Recherche en Nursing; Pflegekategorisierung Verband Bernischer Krankenhäuser); Psych-PV: Psychiatrische Personalverordnung.

In der englischsprachigen Literatur wird ebenfalls eine unübersichtliche Menge verschiedener Begriffe verwendet:
„Patient Classification System“ respektive „Patient Classification Instrument“ (Botter, 2000; DeGroot, 1994a; Giovannetti & Johnson, 1990; Van Slyck, 1991), „Workload Measurement System“ respektive “Workload Measurement Instrument“ (Hernandez & O'Brien-Pallas, 1996a, 1996b; O'Brien-Pallas, 1988), “Instrument Workload Methodology“ (Anderson, 1997), “Nursing Intensity“ (Phillips et al., 1992), “Patient Acuity“ (McHugh & Dwyer, 1992), “Acuity Classfication System“ (Detwiler & Clark, 1995), “Nurse Demand Methods“ (Arthur & James, 1994).
Wie in der deutschsprachigen Literatur wird oft nicht auf die genaue Bedeutung des Begriffes und die Abgrenzung zu anderen im selben Zusammenhang verwendeten Termini eingegangen. Edwardson und Giovannetti (1994, S.98) sowie Hernandez und O’Brien-Pallas (1996a, S.34) kritisieren insbesondere die Verwendung des häufig eingesetzten Begriffs „Patient Classification System“ (PCS) wegen der Gefahr der Verwechslung mit anderen Patientenklassifikationssystemen, wie beispielsweise Diagnosis Related Groups (DRGs), welche aber nicht zur Steuerung der Pflegepersonaldotation eingesetzt werden. Sie propagieren die Verwendung des Begriffs „Nursing Workload Measurement Systems“, was in etwa mit Pflege-Arbeitsbelastungsmesssystem übersetzt werden könnte (Edwardson & Giovannetti, 1994; Hernandez & O'Brien-Pallas, 1996a). Auch der Begriff „Patient Acuity System“ birgt grössere Verwechslungsgefahr in sich, da er als Begriff auch gebraucht wird, um Systeme zur Bestimmung der rein medizinischen Akutheit darzustellen (McHugh & Dwyer, 1992, S.20).

Allen aufgezählten Begriffen in der deutsch- und englischsprachigen Literatur ist gemeinsam, dass sie unter anderem im Zusammenhang mit der Steuerung der Personaldotation verwendet werden.
In dieser Literaturarbeit wird der Begriff Pflegeleistungserfassungsinstrumente als Sammelbegriff für alle Formen der Erfassung pflegerischer Arbeit verwendet, unabhängig von der Nutzung der erhobenen Daten oder Einteilung in eine Typologie (siehe die folgenden Abschnitte).

6 Typen von Pflegeleistungserfassungsinstrumenten

Pflegeleistungserfassungsinstrumente lassen sich nach ganz unterschiedlichen Kriterien einteilen, wie sich in der Literatur zeigt (Arthur & James, 1994, S.560; Bigbee, Collins, & Deeds, 1992, S.32; Malloch & Conovaloff, 1999, S.51; O'Brien-Pallas, 1988, S.8ff). Eine Erläuterung dieser individuellen Typologien würde den Rahmen dieser Arbeit sprengen; im weiteren wird nur auf die beiden Gliederungen näher eingetreten, welche in der Literatur häufig verwendet werden.

6.1 Prototypenmodelle versus Faktorenmodelle

Die in der Literatur am häufigsten anzutreffende Einteilung von Pflegeleistungserfassungsinstrumenten in Prototypenmodelle und Faktorenmodelle stammt ursprünglich aus einem Text von Abdellah und Levine (1979, S.475). Bei den Prototypenmodellen werden die Patienten anhand von Kriterien *direkt* in bestimmte Kategorien eingeteilt, welche ihrerseits zeitgewichtet sind. Bei den Faktorenmodellen werden die Patienten mittels einer Liste von Faktoren (benötigte Hilfe in bestimmten pflegerischen Bereichen und / oder Patientenzustände) beurteilt, welche ihrerseits ebenfalls wieder zeitgewichtet sind. Bei den Faktorenmodellen bleibt die Information der zu einem bestimmten Patienten zugewiesenen Faktoren erhalten, wäh-

renddessen bei den Prototypenmodellen die Kriterien, welche zur Einteilung in die Kategorien führte, verloren gehen. Maeder et al. benennen die Prototypenmodelle als „Vorabkategorisierungs-Modelle“ (1992, S.70). Edwardson und Giovannetti (1994, S.101) nennen sie „Kritische-Indikatoren-Modelle“ und fügen einen weiteren Modelltyp hinzu: „Pflegetätigkeitenmodelle“ (Fischer (2002, S.140) übersetzt mit „Modelle mit Pflegetätigkeitslisten). Bei diesen werden ganz konkrete Pflegetätigkeiten benannt, mit Zeitwerten versehen und den Patienten zugewiesen. Hughes (1999, S.317 f) spricht von „abhängigkeitsbasierten Modellen“ und „aktivitätsbasierten Modellen“, bezieht sich dabei aber ebenfalls auf die Faktorenmodelle und Pflegetätigkeitenmodelle von Edwardson und Giovannetti. In dieser Arbeit wird im weiteren von Prototypenmodellen und Faktorenmodellen gesprochen.
Faktorenmodelle listen spezifische Elemente der Pflege oder Indikatoren auf, welche Cluster von Pflegeaktivitäten bilden. Aktivitäten resp. darauf hinweisende Indikatoren, welche nötige Unterstützung für Essenseinnahme, Körperpflege, Mobilisation anzeigen, kommen praktisch überall vor, da sie grosse Bedeutung haben für die Unterscheidung von „aufwändigen“ und „einfachen“ Patienten. Die Anzahl und Art der Indikatoren variiert stark und basiert nicht nur auf dem Beitrag für die statistische Validität und Reliabilität, sondern auch auf ihrem Zutun zur Akzeptanz und Augenscheinvalidität (siehe Abschnitt 10.2.1). Sie reflektieren zudem unterschiedliche klinische Spezialgebiete (Edwardson & Giovannetti, 1994, S.101).
Der Vorteil von Prototypenmodellen wird in ihrer Einfachheit der Entwicklung und Anwendung gesehen, was dazu führt, dass sie für die Einführung weniger Schulung brauchen und im alltäglichen Gebrauch weniger Zeit benötigen. Prototypenbasierte Systeme scheinen aber eine grössere Varianz bezüglich benötigter Zeit innerhalb einer Kategorie zu haben, was deren Aussagekraft im Gegensatz zu Faktorenmodellen schwächt (Algera-Osinga, Halfens, Hasman, & Wiersma, 1994, S.38). Zudem wird davon ausgegangen, dass Faktorenmodelle grössere Objektivität erreichen, da die Gefahr kleiner ist, durch einen Faktor überproportional beeinflusst zu werden, wenn mehrere Faktoren simultan eingeschätzt werden müssen[4] (Abdellah & Levine, 1979, S.483; Ebener, 1985, S.326).

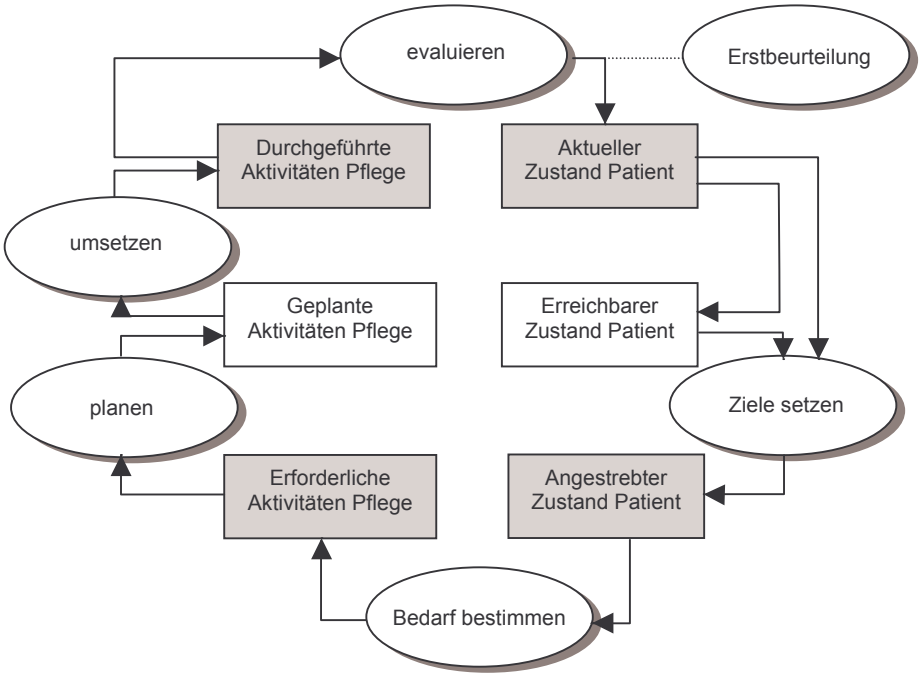
6.2 Prospektive versus retrospektive Modelle

Eine wichtige Einteilung erfolgt mit dem Kriterium der Zeitperspektive: Einschätzungen, welche vorausschauend auf das erfolgen, was geleistet werden sollte resp. geleistet werden kann, werden prospektiv genannt (gelegentlich taucht auch der Begriff prediktiv auf; bspw. bei Van Slyck (1991)). Messungen die zurückblickend erfassen, welche Leistungen erfolgten oder hätten erfolgen müssen, werden retrospektiv genannt. Die Modelle mit prospektivem Fokus sind in der grossen Überzahl (Edwardson & Giovannetti, 1994, S.98). Gewisse Publikationen negieren sogar das Vorhandensein retrospektiver Modelle, indem sie Pflegeleistungserfassungsinstrumenten allgemein Attribute zuweisen, welche sich ausschliesslich auf prospektive Instrumente beziehen; bspw. Haas (1988) oder Hernandez (1996a).
Zusätzlich erwähnt werden muss eine Art „Zwitteransatz“: das schwedische Zebra-System verfolgt gleichzeitig ein prospektives und retrospektives Vorgehen. Die Einschätzung wird um 12.00 Mittags vorgenommen und umfasst die Zeit von 07.00 des aktuellen bis 07.00 Uhr des darauf folgenden Tages .
Vor- und Nachteile der unterschiedlichen Zeitperspektive kommen primär im Zusammenhang mit der Nutzung der Instrumente zum Tragen und werden deshalb im Kapitel 7 diskutiert.

Aus Sicht des Autoren dieser Arbeit greift die Unterscheidung in prospektive versus retrospektive Modelle noch zu kurz, weil sie den Zeitpunkt, zu welchem die Einschätzung innerhalb des Pflegeprozesses statt findet, zu wenig genau einordnet. Es scheint dem Autor un-

[4] Dieses Phänomen ist unter dem Begriff „halo effect“ bekannt.

abdingbar zu sein, dass bei der Kommunikation über und der Anwendung von Pflegeleistungserfassungsinstrumenten allen jederzeit ganz klar ist, was genau erfasst wird resp. werden soll. Das von Fischer (2002, S.78) übernommene und leicht modifizierte Schema (Figur 1), welches den Pflegeprozess abbildet, könnte dazu als Basis dienen.



Figur 1: Schematische Darstellung des Pflegeprozesses

7 Nutzung von Pflegeleistungserfassungsinstrumenten

7.1 Steuerung der Personaldotation

Pflegeleistungserfassungsinstrumente wurden ursprünglich in der Regel primär für die Steuerung der Personaldotation entwickelt (Edwardson & Giovannetti, 1994; Hlusko & Nichols, 1996; Van Slyck, 1991). JCAHO (Joint Commission an Accreditation of Healthcare Organizations), die Organisation, welche seit 1951 Einrichtungen des Gesundheitswesens in den USA beurteilt und akkreditiert, empfahl 1983 die Personaldotation auf Grund von Pflegeleistungserfassungsinstrumenten zu planen (Lovett, Reardon, Gordon, & McMillan, 1994, 1709). Die meisten in Amerika und Kanada entwickelten Systeme zielen auf eine Einsatzsteuerung auf kurzfristiger Basis, also von Schicht zu Schicht (McHugh & Dwyer, 1992). Aber auch die längerfristige Planung wird als Grund für den Einsatz erwähnt (Güntert & Maeder, 1994; Needham, 1997). Alward (1983, S.15) berichtet, der Einsatz von Pflegeleistungserfassungsinstrumenten führe zu grösserer Bereitschaft Pflegender, auf anderen Stationen auszuhelfen, wenn Bedarf nachgewiesen sei.
Jene vielen Instrumente, welche primär für die kurzfristige Personalsteuerung von Schicht zu Schicht entwickelt wurden, verwenden normalerweise den prospektiven Ansatz.

7.2 Ermittlung von Pflegekosten

Durch den im Gesundheitswesen zunehmenden ökonomischen Druck wird die Pflegeleistungserfassung vermehrt auch genutzt, um die Pflegekosten zu ermitteln, da die teilweise auch heute noch übliche Verwendung von Pflegetagen für die Kostenkalkulation ebenso mit Mängeln behaftet ist wie bei der Anwendung für die Personalplanung (Cockerill, O'Brien-Pallas, Bolley, & Pink, 1993). Die Kostendaten ihrerseits werden vielfältig gebraucht, etwa zur Formulierung und Verteidigung von Ressourcen innerhalb des Spitals (O'Brien-Pallas, Irvine, Peereboom, & Murray, 1997, S.171), zur Budgeterstellung und Rechnungsstellung (Botter, 2000, S.544), zur Produktivitätsüberwachung oder für Vertragsverhandlungen mit HMOs[5] (Giovannetti & Johnson, 1990, S.33). In einer Klinik in Schweden bestand der Hauptgrund für die Einführung eines Instrumentes in der Datengenerierung für Kostenkalkulationen der Pflege, und erst in zweiter Linie für Personaldotationszwecke (Levenstam & Engberg, 1997, S.106).

Van Slyck (1991, S.25) bemängelt die häufige Verwendung von Daten für andere Zwecke als die Steuerung der Personaldotation, für welche die Instrumente ursprünglich meistens entwickelt worden seien. Trotz dieser Kritik und im vollen Bewusstsein, dass eine Art Zweckentfremdung statt findet, wird im „Ontario Guide to Case Costing“ fest gestellt, dass die existierenden Pflegeleistungserfassungsinstrumente innerhalb der momentan verfügbaren Möglichkeiten als die beste Basis für die Messung der Verteilung der Pflegekosten auf die Patienten akzeptiert sind (*Ontario Guide to Case Costing*, S.75 f). In ihrer Übersicht über unterschiedliche Ansätze in der Erhebung von Pflegekosten bestätigen sowohl Sovie (1988, S.135) als auch Sherman (1990, S.14), dass Pflegeleistungserfassungsinstrumente die häufigsten dafür verwendeten Methoden waren.

Gemäss Sherman (1990, S.15) verwendeten die meisten Forschungen, welche die Pflegekosten untersuchten, prospektive Pflegeleistungserfassungsinstrumente und massen damit eigentlich den Pflegebedarf und nicht die tatsächlich durchgeführte Pflege. Die allgemeine Sorge besteht darin, dass bspw. auf Grund zu knapper Ressourcen weniger an Pflege tatsächlich erbracht wird, als im Voraus als nötig eingeschätzt wird, was bei entsprechender Verwendung der Daten zu überhöhten Forderungen an die Finanzierer führen könnte. Entgegen dieser Befürchtung steht das Ergebnis einer Studie, welche bei der gleichzeitigen prospektiven und retrospektiven Anwendung desselben Instrumentes eine leicht höhere Einschätzung (vier bis fünf Prozent) aus retrospektiver Sicht kam (Hlusko & Nichols, 1996).

Es herrscht jedenfalls grosse Übereinstimmung, dass für Kostenkalkulationen retrospektiv ausgerichtete Instrumente zum Einsatz kommen sollten, für die Personalplanung hingegen prospektive (Hlusko & Nichols, 1996, S.44; JPPC, Appendix B: ii; O'Brien-Pallas et al., 1995, S.16; Sherman, 1990, S.15). Die Tatsache, dass neuere Instrumente, welche stärker auch mit dem Ziel der Kostenkalkulation entwickelt wurden, retrospektiv ausgerichtet sind (bspw. OPC (Fagerström, Rainio, Rauhala, & Nojonen, 2000, S.483) / ARIC (Giovannetti & Johnson, 1990, S.35) / PINAC und PINI (Prescott & Phillips, 1988; Prescott et al., 1991; Prescott & Soeken, 1996b))[6], zeigt, dass dieser Forderung Rechnung getragen wird.

7.3 Allgemeine Managementzwecke

Eine qualitative Fallstudie von Botter (2000, S.546), welche die Nutzung des in den USA über vierhundert Mal eingesetzten Instrumentes Medicus in einem einzelnen Spital untersuchte, zeigte, dass die Daten neben den oben bereits aufgeführten Zielen zu verschiedensten weiteren Managementzwecken eingesetzt werden. Botter führt dazu beispielsweise fol-

[5] HMOs: Health Maintenance Organizations; auch Gesundheitskassen genannt

[6] OPC: Oulu Patient Classification / ARIC: Allocation, Resource Identification and Costing / PINAC: Patient Intensity for Nursing: Ambulatory Care / PINI: Patient Intensity for Nursing Index

gende Begriffe auf: Leiten; Trends analysieren; Arbeitsbelastung überprüfen, Erklären; Rechtfertigen; Verteidigen; Vergleichen; Projektieren; Rapportieren; Schulen; Neugier befriedigen; Qualitätssicherung; Forschung. Gemäss Edwardson und Giovannetti (1994, S.115) werden die Daten auch für medizinisch-juristische Zwecke eingesetzt.
Überschaut man die in diesem Kapitel 7 beschriebenen Einsatzgebiete, wird klar, dass die direkte Nutzung der Daten fast ausschliesslich beim Management liegt, jedoch weniger bei den Pflegenden für ihre direkte Arbeit mit und am Patienten. Es sind jedoch genau diese Pflegenden, welche die nötigen Daten im Alltag erheben müssen. Es ist nachvollziehbar, dass die Motivation für eine zuverlässige tägliche Anwendung der Instrumente einerseits abhängt von einer gehörigen Portion Verständnis für die Datenbedürfnisse des Managements, andererseits von der Erfahrung, dass die Daten vom Management adäquat verwendet werden. Gleichzeitig ist die Einsicht der praktisch tätigen Pflegenden nötig, dass die Zuteilung nötiger Ressourcen für eine qualitativ hochstehende Pflege von der Argumentationskraft der von ihnen generierten Zahlen abhängen kann. Fehlt die Einsicht in diese Zusammenhänge, besteht die Gefahr einer schlechten Datenqualität mangels Reliabilität. Auch Hughes äussert den Verdacht, dass Pflegende die Anwendung der Systeme möglicherweise nicht als ihre Aufgabe erachten und die Datenerfassung dadurch evtl. leidet (1999, S.319).

Aus der Erkenntnis, dass die Anwendung der Instrumente im Alltag für die Pflegenden einen Zusatzaufwand bedeutet, gehen immer mehr Bestrebungen in die Richtung, die Daten direkt elektronisch aus der Pflegedokumentation zu generieren (Edwardson & Giovannetti, 1994, S.115; LEP-AG, 2002a; SAHO, 2003). Malloch und Conovaloff haben sogar schon Visionen von einem zukünftigen elektronischen Pflegeleistungserfassungssystem, welches eine automatische zeitnahe (real-time) Verfolgung der Patienten-Pflegenden Interaktionen vornimmt. Diese würden von einer Aufsichtsperson überwacht, welche eingreift, wenn die geplante Pflege nicht entsprechend erfolgt (Malloch & Conovaloff, 1999, S.53).

8 Zeitquantifizierung und Zuordnung zu den Kategoriensystemen

Damit die beiden Hauptnutzungen von Pflegeleistungserfassungsinstrumenten (Steuerung der Personaldotation / Kostenkalkulation) überhaupt möglich sind, werden den einzelnen abschliessenden Kategorien (Prototypenmodell) oder den Indikatoren (Faktorenmodelle), wiederum Zeitwerte in irgend einer Form zugeordnet. (Die Kostenkalkulation ist eng gekoppelt an die bezahlten Löhne, und diese hängen - mindestens in den meisten allgemein akzeptierten Entlöhnungssystemen - wiederum stark von der Menge der gearbeiteten Zeit ab.) Die Zeitzuordnung erfolgt teilweise direkt über Minutenangaben, häufig aber auch über indirekte Punktwerte, welche zu einem späteren Zeitpunkt in Zeit umgerechnet werden.

Unterschiedliche Instrumente erfassen auch unterschiedliche Anteile der Pflegearbeit. So wird häufig zwischen direkter und indirekter Pflege unterschieden, wobei die Definitionen leider teilweise voneinander abweichen (Fischer, 2001, S.112 / S.142; O'Brien-Pallas, Leatt, Deber, & Till, 1989, S.18). Mehrmals angetroffen wurde in der Literatur folgende Definition: Direkte Pflege sind alle pflegerischen Aktivitäten, welche im Beisein des Patienten oder von dessen Angehörigen erfolgen. Indirekte Pflege sind alle Aktivitäten, welche nicht *beim* Patienten, aber *für* den Patienten erfolgen. Zusammen genommen, also für die Zeit der direkten + indirekten Pflege, wird sodann der Begriff “dem Patienten zuweisbare (patient assignable) Zeit” verwendet. Die dem Patienten zuweisbare Zeit hebt sich ihrerseits ab von

der Zeit, welche für die Aufrechterhaltung des Betriebes sowie für Pausen, Toilettenbesuche und dergleichen (personal time) benötigt wird (Prescott et al., 1991, S.219; Sovie, 1988, S.141). Im Gegensatz zur obigen Definition, zählen gewisse Autoren sämtliche Zeit, welche nicht direkt beim Patienten verbracht wird, als indirekte Zeit (Willems, 1992, S.40; Williams, 1977, S.15). Gerade in diesem Punkt unterscheiden sich auch die beiden in der Schweiz hauptsächlich eingesetzten Instrumente LEP und PRN.
Ob die Menge der Zeit, die dem Patienten nicht zuweisbar ist, sowie die Menge der indirekten Pflege proportional zur direkten Pflege variiert, ist nicht klar (Williams, 1977, S.15). Beim Zebra-System (Schweden) werden diese Zeiten auf alle Patienten gleich verteilt, da man der Ansicht ist, die Pflegekategorie habe darauf keinen Einfluss (Levenstam & Engberg, 1997, S.112).

Es existieren mehrere Methoden, um die Standard-Zeitwerte zu bestimmen. Alle einigermassen präzisen Techniken um den Verbrauch an Pflegezeit in einer Einheit zu bestimmen, sind aufwändig bezüglich Finanzen und menschlicher Ressourcen (McDaniel, 1994, S.25; McHugh & Dwyer, 1992, S.24). Bartholomeyczik et al. zeigen, wie schwierig Zeiterhebungen in pflegerischem Umfeld sind, weil die Tätigkeiten stark zerstückelt erfolgen und Anfangs- und Endpunkte der Messung schwierig zu definieren sind (2001). Aus diesem Grund wird häufig die Methode der Expertenschätzung eingesetzt (Edwardson & Giovannetti, 1994, S.112). Beispiele dafür sind zu finden bei Sarnecki , Haas, Stevens und Willemsen (1998, S.38) oder Brosziewski und Brügger (2001, S.65). Studien, welche die Genauigkeit solcher Expertenschätzungen überprüfen, sind gemäss Edwardson und Giovannetti selten.
Es gibt vielfältige Methoden, um die Zeit nicht zu schätzen, sondern empirisch zu messen. Eine weitere Diskussion dieser Ansätze würde den Rahmen dieser Arbeit allerdings sprengen. Informationen darüber können nachgelesen werden unter Edwardson & Giovannetti (1994, S.112 f), McHugh und Dwyer (1992, S.23 ff), Willems (1992, S.14 ff).
Obwohl die exakteren Methoden der Zeitbestimmung so aufwändig sind, herrscht in der Literatur grosse Einigkeit darüber, dass die den Kategorien zugewiesenen Zeiten oder Quantifizierungskoeffizienten in jeder Organisation neu bestimmt werden müssen, da bspw. anderes Equipment und viele andere Faktoren auf die Zeitwerte einen Einfluss haben (Alward, 1983, S.16; Giovannetti, 1979, S.6; McHugh & Dwyer, 1992, S.24; Trivedi & Hancock, 1975, S.371). Carr Hill und Jenkins Clark gehen sogar noch weiter und schlagen vor, dass die Analysen auf jeder Station separat durchgeführt werden müssten (1995, S.222).
Da die die Zeitwerte beeinflussenden Faktoren nicht unbedingt stabil sind, sollten die Zeitwerte in gewissen Abständen zusätzlich überprüft und allenfalls neu bestimmt werden. Ein eindrückliches Beispiel für diese Notwendigkeit beschreiben Churness, Kleffel, und Onodera, welche zeigten, wie stark sich die prädiktive Validität in einer späteren Phase der Instrumententwicklung verschlechterte durch eine Veränderung im Lohnsystem einer Gemeindekrankenpflege: Es wurde von der Bezahlung der gearbeiteten Zeit gewechselt zur Bezahlung der Anzahl durchgeführter Besuche (Churness, Kleffel, & Onodera, 1991, S.20).

Eine grundsätzliche Kritik an der Zeitbestimmung mittels „Operational Research“, also Zeitbestimmungsstudien erfolgt von Procter (1991) auf Grund von Beobachtungen innerhalb einer Untersuchung. Durch Zeitbestimmungsstudien würde nur das vorherrschende Personaldotationsniveau der gemessenen Betriebe repliziert. Diese wiederum wären ihrerseits stark durch Budget und lokale Arbeitsmarktsituation beeinflusst, nicht aber primär durch einen theoretischen Rahmen der Pflege, welcher auch die Qualität der geleisteten Pflege gebührend berücksichtigt. Diese Einschätzung wird unterstützt durch eine Behauptung von Unger (1985, S.19), dass ein PCS primär kompatibel sein müsse mit etablierten Personalschlüsseln.

9 Die Theoretische Basis von Pflegeleistungserfassungsinstrumenten

Die Theoretische Basis der Pflegeleistungserfassungsinstrumente ist ein in der Literatur viel diskutiertes Thema. Es sind viele kritische Anmerkungen diesbezüglich zu finden. Hughes meint, Pflegephilosophie lege Wert darauf, den Patienten als Individuum zu betrachten, Messsysteme klassifizierten ihn aber in Kategorien (1999, S.319). Needham meint, es existiere ein grosser Graben zwischen Pflegeaufwandmesssystemen und dem theoretischen Verständnis der Pflege (1997, S.83); bevor darüber diskutiert werde, was Pflegeaufwandmessung sei, müsste Pflege exakt definiert werden (S. 86). Edwardson und Giovannetti stellen fest, dass die meisten Instrumente nicht auf einer bestimmten Pflegetheorie basierten; die Klassifikation der Variablen zeige bloss den Versuch, Muster zu erkennen (1994, S.111). Forchuk, (1996) kritisiert, in der Psychiatriepflege müsste ein valides System auf der Messung der Beziehung Klient-Pflegender basieren, um Pflege richtig abbilden zu können, nicht auf den in den Instrumenten sonst üblichen supportiven Aufgaben oder Aktivitäten. Needham meint, die Essenz und Komplexität der Pflege gingen verloren durch Aktivitätsanalysen (1997, S.85). Haas kritisiert, die verwendeten Indikatoren sollten auf einer Theorie basieren; die ohne Theorie „wahrgenommenen“ Indikatoren würden nicht den gesamten Pflegeprozess spiegeln (1988, S.57).

Diesen Vorwürfen gegenüber stehen einige verteidigende Statements. De Groot (1989a, S.31) erklärt, dass die Indikatoren nicht dazu da seien, eine vollständige Liste von möglichen Patientenbedürfnissen zu präsentieren, sondern das Ziel sei, die Patientenbedürfnisse mit möglichst wenigen Indikatoren vorauszusagen. Der Qualitätsaspekt der Sparsamkeit (parsimony) stehe hier dem Mythos vom „mehr“ (je mehr Indikatoren, desto besser) gegenüber. Auch Edwardson und Giovannetti (1994, S.115) meinen, wenn es darum gehe, Voraussagen zu machen bezüglich Personalbedarf für kommende Schichten, brauche es Variablen, die leicht verfügbar, objektiv und für Voraussagen statistisch mächtig seien; also Körperpflege, Ernährung, Bewegung, Behandlung, Anleitung. Weil kein Unterschied gemacht werde zwischen Variablen, welche mit der Arbeitsbelastung korrelieren, und solchen, welche die Pflege beschreiben und erklären, werde die „taskiness“ der Indikatoren bemängelt sowie das Fehlen von Tätigkeiten, welche kognitive Aspekte der Handlungsfreiheit aufzeigten. Neuere Instrumente, welche beinahe alle Aktivitäten enthielten, seien populär geworden, weil geglaubt werde, durch mehr Variablen werde die Messung genauer (S.111 f). Williams (1977, S.15) anerkennt von Anfang an, dass Pflegeleistungserfassungsinstrumente das Meiste an Assessment, Planung, Evaluation, Beurteilung und Caring nicht messen und betrachtet dies auch nicht als deren Aufgabe. Auch für die Methode LEP wird klar auf Einschränkungen hingewiesen: Es könnten eine ganze Reihe von pflegerisch relevanten Tätigkeiten, die für den Spitalbetrieb insgesamt von zentraler Bedeutung seien, nicht erfasst werden. Etwa Handlungen, die auf die Sicherheit, das Wohlbefinden oder die Gefühlslage von Patienten und Patientinnen abzielten. Diese Leistungen seien kaum quantifizierbar und liessen sich mit einem Instrument wie LEP nur schwer erfassen (Brosziewski & Brügger, 2001, S.64; LEP-AG, 2002b, Abschnitt 2.6.7).

Es zeigt sich, dass das Problem der theoretischen Basis klar mit dem Einsatzbereich der Instrumente zusammen hängt. „Die mangelnde theoretische Basis ist akzeptabel, wenn es nur um die Pflegebedarfsbestimmung geht. Wenn die Instrumente aber die gesamte Pflege beschreiben und dokumentieren sollen, könnte es angebracht sein, einen theoretischeren Ansatz zu verwenden“ (Edwardson & Giovannetti, 1994, S.112). Die weiter oben schon

einmal aufgeführte Kritik von Van Slyck (1991, S.25), welcher die häufige Verwendung von Daten für andere Zwecke als die Steuerung der Personaldotation bemängelt - für welche die Instrumente ursprünglich meistens entwickelt worden seien - kommt auch hier wieder zum Tragen. Wenn beispielsweise als einer der Gründe für den Einsatz von LEP angegeben wird, mit dem Instrument könnten die Pflegenden ihre Arbeit für sich, aber auch für Dritte (Verwaltung, Ärzte, Kostenträger) innerhalb und ausserhalb des Spitals transparent machen (Brügger, Bamert, & Maeder, 2001, S.5), so ist verständlich, dass die Pflegenden möglichst alle Aspekte ihrer Arbeit mit dem Instrument dokumentieren wollen.
Haas (1988) warnt davor, dass wenn gegen aussen nicht ganz klar gemacht werde, dass mit Pflegeleistungserfassungsinstrumenten nicht die gesamte Pflege abgebildet wird, dies zu einer von aussen durch Zeitbeschneidung auferlegten Beschränkung der Pflege führen könnte, nämlich auf die Interventionen, welche im Instrument vorkommen, währenddem die anderen Aspekte des Pflegeprozesses ausserhalb der Intervention verloren gehen könnten.

Bezüglich der Diskussion um die Suche nach einer definierten theoretischen Basis von Pflegeleistungserfassungsinstrumenten muss erwähnt werden, dass einzelne Instrumente sich bewusst um die theoretische Basis bemühten: Das Instrument von Reitz (1985a, S. 28+30) versuchte den Pflegeprozess und die NANDA-Diagnosen als Entwicklungsbasis für das Instrument zu nehmen. OPC (Oulu Patient Classification) wurde auf Roper, Tierny und Logans Pflegemodell basierend entwickelt (Fagerström & Engberg, 1998). Gemäss O'Brien-Pallas (1988, S.11) wurde auch ein Instrument auf der Basis von Johnsons Verhaltenssystemmodell entwickelt. Da kein gemeinsam anerkanntes übergreifendes Pflegemodell für die Pflege vorhanden ist (und evtl. auch nicht wünschbar ist), werden gemäss O'Brien-Pallas weitere Schritte in diese Richtung aber eingeschränkt sein.

10 Reliabilität und Validität von Pflegeleistungserfassungsinstrumenten

Die Diskussionen der vorherigen Kapitel „Nutzung von Pflegeleistungserfassungsinstrumenten“ / „Zeitquantifizierung und Zuordnung zu den Kategoriensystemen“ / „Die Theoretische Basis von Pflegeleistungserfassungsinstrumenten“ haben einen direkten Zusammenhang mit Aspekten der Validität und Reliabilität. Deren Inhalt muss deshalb auch in diesem Kapitel ständig im Hintergrund mit berücksichtigt werden.

Die Bestimmung der Reliabilität und der Validität von Pflegeleistungserfassungsinstrumenten wird als wichtig erachtet, obwohl diese nicht primär für Forschungszwecke, sondern für den täglichen Gebrauch eingesetzt werden (Edwardson & Giovannetti, 1994). Allgemein wird kritisiert, dass Reliabilität und Validität von Pflegeleistungserfassungsinstrumenten viel zu selten oder mit mangelnder wissenschaftlicher Stringenz nachgewiesen wurden (Edwardson & Giovannetti, 1994, S.96; Hernandez & O'Brien-Pallas, 1996a, S.34; Hughes, 1999).
Unger (1985) kann aus Sicht des Autors dieser Arbeit als Vertreterin eines der vielen handgestrickten Instrumente gelten, welche gemäss Edwardson (1994, S.99) existieren, über welche bloss anekdotenhaft berichtet wird und für deren behauptete Reliabilität und Validität keine echten Nachweise geliefert werden. So zeigt Unger bspw. mit der Ansicht, dass nicht jeder das Rad neu erfinden müsse, kein Verständnis für die Forderung, dass jedes Spital eigene Arbeitszeitstudien machen müsste (vgl. Kapitel 8). Das Hauptziel und auch der einzige Beweis für so genannte Reliabilität des von ihr entwickelten Instrumentes besteht für die Autorin darin, dass es dieselbe Anzahl Pflegender prognostiziert, welche auch durch die Sta-

tionsleitungen aus fachlichem Ermessen nötig sind.
Als Kuriosum erscheint dem Autor dieser Arbeit auch das Vorgehen von Malloch et al. (1999, S.38), welche so genannte „qualitative Reliabilitäts- und Validitätskonzepte“, die aus der qualitativen Forschung bekannt sind, für ein quantitatives Instrument eingesetzt haben (3PCS; Third-Generation Patient Classification System).

O’Brien-Pallas forderte schon 1988 (1988, S.11), dass Reliabilität und Validität in Zukunft auf einem viel anspruchsvolleren Niveau angegangen werden müssten als dies damals der Fall war. Akzeptanz der Validität auf der Stufe der Augenschein- oder Inhaltsvalidität könnten nicht mehr toleriert werden. 1994 stellen Edwardson und Giovannetti (1994, S.96) weiterhin fest, dass in der publizierten Literatur Berichte über die Entwicklung und Testung einiger der am häufigsten eingesetzten urheberrechtlich geschützten Instrumente wie Medicus und GRASP vermisst wurden. Hughes (1999, S.319) kommt 1999 mit ihren Aussagen, Pflegeaufwandmessmethoden lieferten keine reliablen Informationen; Reliabilitätswerte seien tief; entweder seien die Methoden nicht valide, oder die Methodenanwender müssten angezweifelt werden, noch immer zu keinem besseren Schluss.

10.1 Reliabilität

Die Reliabilität von Pflegeleistungserfassungsinstrumenten wird als sehr wichtig erachtet. Weil diese dafür geschaffen sind, um von vielen Pflegenden verwendet zu werden, ist die konsistente Anwendung entscheidend, damit die Messung genau sein kann (Giovannetti, 1979, S.6; McKenzie, 1991, S.522).
Es werden drei Hauptformen der Reliabilität unterschieden: die Stabilität, die Homogenität, und die Äquivalenz.

10.1.1 Stabilität

Die Stabilität misst den Grad, mit dem die selben Ergebnisse bei wiederholter Anwendung desselben Instrumentes erzielt werden (Polit & Hungler, 1999, S.412). Gewöhnlich wird die Test-Retest-Methode angewendet. Prinzipiell kommen dieselben statistischen Tests in Frage, wie für die weiter unten besprochene Interraterreliabilität; auf diese wird dort weiter eingegangen.
Ein Problem bei der Test-Retest-Methode besteht darin, dass das zu messende Merkmal zwischen den Messpunkten stabil bleiben muss, was in der Situation von Pflegeleistungserfassungsinstrumenten meistens nicht der Fall ist. Verkürzt man die Zeit zwischen den Messpunkten, ist aber mit einem Erinnerungseffekt zu rechnen. Somit muss auf „konservierte“ resp. standardisierte Situationen ausgewichen werden, etwa durch schriftlich dokumentierte, evtl. konstruierte Fälle. Aus diesen Gründen wird in der Literatur eher selten über die Test-Retest-Methode berichtet. Das Instrument von Reitz (1985b) ist grundsätzlich für eine Klassifikation anhand der Dokumentation vorgesehen. Insofern wäre dieses Instrument geradezu prädestiniert für die Anwendung der Test-Retest-Methode. Leider scheint diese nicht angewendet worden zu sein.

- Für ein niederländisches Prototypensystem für die Gemeindekrankenpflege wurde ein Kappa-Wert berechnet (Algera-Osinga et al., 1994, S.36).
- Santamaria et al. (2001, S.12). errechneten eine Pearson Korrelation nach zweimaliger Anwendung des Instrumentes mit einem Monat Abstand. Die Kritik der Verwendung der Korrelation erfolgt weiter unten.
- Turner Stokes et al. (1998, S.311) wendeten ihr Instrument in einer Praxissituation (Rehabilitation) innerhalb von zwei Tagen zwei Mal an. Die Autoren glauben, keinen Erinnerungseffekt zu haben, da die Pflegenden im Arbeitsalltag ständig alle Patienten einschät-

zen mussten und nach zwei Tagen nicht mehr gewusst hätten, was sie bei der ersten Einschätzung für Werte zugewiesen hätten. Neben prozentualer Übereinstimmung der ordinalen Kategorien wurden auch Differenzen der Erst- und Zweitbestimmung errechnet und statistisch geprüft. Damit wurde implizit von äquidistanten ordinalen Daten ausgegangen, was etwas fragwürdig ist.

10.1.2 Interne Konsistenz oder Homogenität

„Ein Instrument kann in dem Grad als intern konsistent oder homogen bezeichnet werden, wie alle seine Unterteile dieselbe Charakteristik messen“ (Polit & Hungler, 1999, S.414). Eine globale Messung der Homogenität ist nach Ansicht des Autors dieser Arbeit nicht angezeigt, da Pflegeleistungserfassungsinstrumente *kein* einheitliches globales Charakteristikum messen. So kann ein Patient bspw. bezüglich den Aktivitäten des täglichen Lebens völlig selbständig sein, aber viel Pflegezeit für Instruktion (bspw. Diabetes) benötigen. Die Messung der Homogenität (von Subdimensionen) macht also nur dann Sinn, wenn die Subdimensionen des Instrumentes ihrerseits wieder mit mehreren Items erfasst werden. Pflegeleistungserfassungsinstrumente sind aber auf den täglichen Gebrauch hin konstruiert, so dass sie möglichst wenig Zeit beanspruchen. Deshalb werden die Subdimensionen häufig nur mit einem Item gemessen. Der Autor nimmt an, dass dies der Grund für die seltene Bestimmung der Homogenität ist.
Für die Instrumente PINI und PINAC wurden in Reliabilitätsstudien ein Kronbachs Alpha ermittelt (Prescott et al., 1991; Prescott & Soeken, 1996b); sie erreichten relativ ansehnliche Resultate von r=0.85 und r=0.75.

10.1.3 Äquivalenz

Das Ziel der Äquivalenz besteht darin, die Konsistenz oder Äquivalenz zu bestimmen, mit welcher das Instrument dasselbe Merkmal in denselben Situationen misst (Polit & Hungler, 1999, S.416). Die Parallelform-Methode ist für Pflegeleistungserfassungsinstrumente nicht praktikabel (Ebener, 1985, S. 326). Die Interraterreliabilität[7] hingegen als zweite Methode der Äquivalenzbestimmung wird als wichtigstes Reliabilitätsmass erachtet (Alward, 1983, S.16) und scheint auch am häufigsten ermittelt worden zu sein.

10.1.4 Interraterreliabilität

10.1.4.1 Praktisches Vorgehen (Design)

Bei der Interraterreliabilität wird ein Instrument in exakt derselben Situation von mehreren Personen unabhängig voneinander angewendet, worauf überprüft wird, wie gut die Ergebnisse übereinstimmen. Voraussetzung ist, dass die Rater (also jene Personen, welche das Instrument zu dessen Testung anwenden) denselben Informationsstand über die Situation resp. den Patienten haben. Dies ist in realen Praxissituationen oft schwierig zu erreichen. Pflegende können im Alltag für das Erfassen eines Pflegeleistungserfassungsinstrumentes auf eine Fülle von Informationen zurückgreifen, welche sie bei der Arbeit am und mit dem Patienten erhalten. Um echten Informationsgleichstand in Praxissituationen zu erreichen, müssten gleichzeitig mehrere Pflegende miteinander arbeiten. Dies ist im Normalfall schwierig zu organisieren und beschränkt sich wenn schon auf höchstens zwei Personen. Zudem entspricht dies dann doch wieder keiner echten Praxissituation.
In der Literatur sind zwei Ansatzpunkte erkennbar, mit denen versucht wurde, das Problem zu umgehen. Einerseits wurden Situationen resp. verfügbare Informationen darüber in irgend

[7] Wird gelegentlich auch als Beobachterübereinstimmung oder Interobserverreliabilität bezeichnet.

einer Form konserviert und mehreren Ratern immer dieselbe Standardsituation zur Einschätzung vorgelegt, oder die Einschätzung erfolgte durch praktisch tätige Pflegende hintereinander in der Hoffnung und Annahme, dass die Situation resp. der Patient für eine Interraterreliabilitätsbestimmung genügend stabil bleibt.
Die Methode „Standardsituation“ hat den Nachteil, dass es schwierig ist, die nötige Informationsmenge und Art zu bestimmen, welche die „Konserve“ enthalten muss. Zudem unterscheidet sich die Ratingsituation von der Praxissituation, indem sie in einem künstlichen Umfeld erfolgt. Der Vorteil der Methode „Standardsituation“ besteht darin, dass sie mehreren Ratern parallel vorgelegt und gleichzeitig auch für eine Homogenitätsüberprüfung verwendet werden kann (siehe Abschnitt 10.1.2). Zudem sind damit Interraterreliabilitätsüberprüfungen auch über mehrere Organisationseinheiten möglich, was für die Vergleichbarkeit bspw. von Stationen innerhalb eines Spitals wesentlich sein kann.
De Groot (1989b, S.34) bezeichnet den Ansatz, bei dem schriftliche hypothetische Situationen für das Rating konstruiert werden als „patient constant approach“. Castorr et al. (1990, S.314) empfehlen, dass hypothetische Situationen, welche für das Training und eine Interraterreliabilitätsbestimmung gleichzeitig genutzt werden könnten, vorgetestet und klarer gemacht werden sollten. Hier muss aus der Sicht des Autoren dieser Arbeit allerdings eingeschränkt werden, dass sich dieses „klarer machen“ darauf beschränken muss, dass die Standardsituation im Vergleich zu Praxissituation möglichst genügend und geeignete Information bietet. Dieses „klarer machen“ darf, mindestens für eine Interraterreliabilitätsstudie, hingegen nicht dazu führen, dass die Standardsituationen so weit von der Realität abweichend zurecht gebogen werden, bis sie schön in die Kategorien des Instrumentes passen. Dies würde zu zu hohen Interraterreliabilitätswerten im Vergleich zur Anwendung in der Praxis führen.
Die Verwendung von Videos für die Interraterreliabilitätsbestimmung von Pflegeleistungserfassungsinstrumenten wird zwar in der Literatur vorgeschlagen (Castorr et al., 1990, S.314; Haas, 1988, S.57), der Autor dieser Arbeit konnte aber keine Anwendung in Reliabilitätsstudien finden.

Berichte über die Bestimmung der Interraterreliabilität in der Literatur beziehen sich meistens auf den Zeitpunkt der Instrumentenentwicklung. Teilweise wird nicht genau berichtet, wie bei der Interraterreliabilitätsbestimmung vorgegangen wurde, sondern es werden nur Resultate erwähnt, bspw. bei Sarnecki et al. (1998).
Für die alltägliche Anwendung ist normalerweise die Interraterreliabilität über die abschliessende Klassifikation ausreichend. Vor allem während der Entwicklung des Systems macht es jedoch Sinn, die Übereinstimmung bei den Items zu berechnen um die Fehlerquellen zu eruieren (Edwardson & Giovannetti, 1994, S.114).
Bei der Bestimmung der Interraterreliabilität muss genau überlegt werden, ob Experten mit in den Ratingprozess einbezogen werden sollen und können oder nicht. Hohe Reliabilitätswerte welche erreicht werden, ohne dass Experten (quasi als Goldstandard) einbezogen sind, geben noch keine Garantie dafür, dass die Einschätzung korrekt erfolgte. Es ist möglich, dass alle Rater einem gemeinsamen, also systematischen Fehler unterliegen (Castorr et al., 1990, S.313; McHugh & Dwyer, 1992, S.28). Castorr et al. (1990) sprechen beim Einbezug von Experten von „criterion-related agreement“. Es liesse sich diskutieren, ob es hier wirklich noch um Reliabilität geht, oder eher um Validität; in der Literatur wird dieser Ansatz aber überall unter Reliabilität aufgeführt (Edwardson & Giovannetti, 1994, S.113; Giovannetti & Johnson, 1990, S.35; Giovannetti & Mayer, 1984, S.33; Noyes, 1994, S.7).

Beispiele aus der Literatur

- Edwardson und Giovannetti (1994, S.110) berichten von einem Instrument, welches für ein Ambulatorium entwickelt worden war, das anhand von Tonbandaufnahmen von Rapporten über die Patienten getestet wurde.
- Bei einem Instrument für die Gemeindepflege wurden die Ratings der Pflegenden, welche den Besuch machten, verglichen mit Ratings, welche andere Pflegende auf Grund der Dokumentation dieses Besuches erstellten (Churness et al., 1991, S.19). Es fragt sich, ob das Resultat nicht durch den unterschiedlichen Informationsstand beeinträchtigt wurde. Die Verwendung von Pflegedokumentationen scheint generell fragwürdig, da die Dokumentation nicht unbedingt das spiegelt, was tatsächlich an Pflege erfolgte (Björvell, Thorell-Ekstrand, & Wredling, 2000, S.8).
- Bei einem weiteren Instrument der Gemeindepflege wurden jeweils zwei Ratings von Pflegenden untersucht, welche hintereinander Patienten besuchten, die mindestens täglich oder mehrmals täglich Besuche brauchten (Anderson & Rokosky, 2001). Hier kann kritisiert werden, dass sich das Resultat nur auf akute Situationen bezieht, da Patienten, die weniger Betreuung brauchten, auf Grund des Untersuchungsdesigns nicht einbezogen werden konnten.
- Ein Instrument, das nach der Entlassung anhand der Dokumentation angewendet wird (hier muss die Validität entsprechend der vorgängig erwähnten Kritik auch in der täglichen Anwendung bezweifelt werden), wurde in einem Pilottest von einer Primary Nurse (also mit mehr Information) und einer Pflegenden, die die Patientin nicht gepflegt hat, beurteilt. In der Hauptstudie kommt nicht klar zum Ausdruck, welche Ratings von wem genau mit welchen Ratings von Anderen verglichen werden (Reitz, 1985b, S.32).
- Beim PINI-Instrument wurden die Ratings der Tagschicht-Pflegenden, vorgenommen gegen Ende der Schicht, verglichen mit den Ratings der Nachtschicht-Pflegenden, vorgenommen im ersten Teil der Schicht. So lagen etwa vier Stunden zwischen den Ratings (Prescott, Soeken, & Ryan, 1989, S.258).

Es zeigt sich, dass es beinahe unmöglich ist, ein praktisch durchführbares aber theoretisch trotzdem optimales Design zu erstellen.

10.1.4.2 Statistische Parameter für Interraterreliabilität

In der Literatur wird eine ganze Reihe von möglichen statistischen Parametern für die Ermittlung der Interraterreliabilität erwähnt. Da diese Parameter unter anderem vom Datentyp abhängen, welche durch die Anwendung des Instrumentes entstehen, schränken sich die Möglichkeiten für deren Einsatz aber jeweils ein.

Folgende Parameter werden in der Literatur erwähnt:

Prozentuale Übereinstimmung, Cohens Kappa, Korrelation, Phi (Topf, 1986), ANOVA (Bigbee et al., 1992; Soeken & Prescott, 1986), D-L-Test (Bigbee et al., 1992), Ar-Statistic (Reitz, 1985b). Die ersten drei Parameter sind mit keiner Literaturangabe versehen worden, weil sie sehr häufig erwähnt werden. Häufig existieren mehrere Unterformen der einzelnen Parameter. Die Parameter sind teilweise miteinander verwandt. In der weiteren Diskussion wird nur auf die ersten drei Parameter eingegangen, weil die restlichen sehr selten angewendet werden und weil dem Autor dieser Arbeit die nötigen statischen Kenntnisse fehlen.

Prozentuale Übereinstimmung

Die prozentuale Übereinstimmung ist der am häufigsten ermittelte Wert. Obwohl es verschiedene Unterformen gibt, ist in den Publikationen nicht ersichtlich, welche Form gewählt wurde.

Die einfache Situation um eine Interraterreliabilität zu errechnen kann entsprechend Figur 2

als Vierfeldertafel (teilweise auch als Kontingenztafel benannt) aufgezeichnet werden. N entspricht der Anzahl der gerateten Situationen. Die Buchstaben a bis d entsprechen den möglichen Übereinstimmungen resp. Nicht-Übereinstimmungen, wenn zwei Rater bewerten, ob ein Ereignis/Zustand vorhanden ist oder nicht.

		Rater 2 (oder Experte)		
		+	-	
Rater 1	+	a	b	g1 (=a+b)
	-	c	d	g2 (=c+d)
		f1 (=a+c)	f2 (=b+d)	N (=a+b+c+d)

Figur 2: Vierfeldertafel für prozentuale Übereinstimmung, Kappa, Sensitivität und Spezifität

Zelle a zeigt die Häufigkeit, bei der die beiden Rater übereinstimmen, dass ein Zustand vorhanden ist. Zelle d zeigt die Häufigkeit, bei der die beiden Rater übereinstimmen, dass ein Zustand nicht vorhanden ist. Zelle b zeigt die Häufigkeit, bei der Rater 1 den Zustand als vorhanden einschätzt, während ihn Rater 2 als nicht vorhanden einschätzt. Zelle c zeigt die Häufigkeit, bei der Rater 2 den Zustand als vorhanden einschätzt, während ihn Rater 1 als nicht vorhanden einschätzt.
Die Zellen a und d zeigen also Übereinstimmung an, die Zellen c und b Nicht-Übereinstimmung.
Die prozentuale totale Übereinstimmung errechnet sich folgendermassen: $(a+d) / N * 100$.
Gelegentlich wird die totale Übereinstimmung auch als p_o bezeichnet (das tief gesetzte o steht für observed agreement) und erscheint dann nicht als Prozentsatz. p_o errechnet sich durch $p_o=(a+d) / N$.

Die Hauptkritik am Parameter der prozentualen Übereinstimmung ist, dass eine allfällige zufällige Übereinstimmung nicht berücksichtigt ist und zu falsch hohen Ergebnissen führen könnte (Polit & Hungler, 1999, S.416). Uebersax (2002b) empfiehlt deshalb, nicht die totale Übereinstimmung zu dokumentieren, sondern die Werte der positiven Übereinstimmung (p_{pos}) und den der negativen Übereinstimung (p_{neg}) ebenfalls.
P_{pos} errechnet sich folgendermassen: $2a / (f1 + g1)$, p_{neg} folgendermassen: $2d / (N – (a – d))$.
Diese Werte können als geschätzte bedingte Wahrscheinlichkeiten interpretiert werden: p_{pos} zum Beispiel schätzt die bedingte Wahrscheinlichkeit, dass einer der beiden Rater, per Zufall ausgewählt, ein positives Rating setzt, welches der andere Rater ebenfalls tun wird (Uebersax, 2002b).

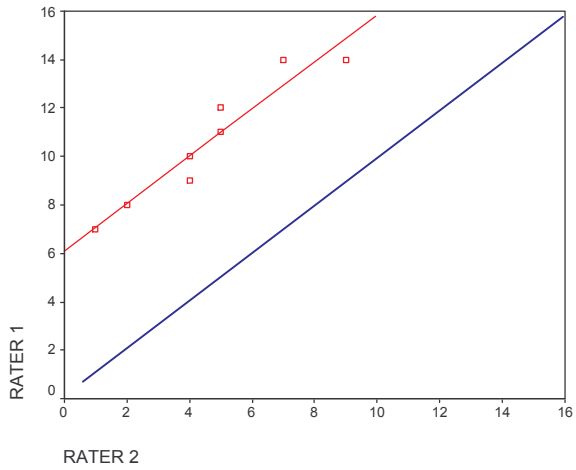
Cohens Kappa κ
Cohens Kappa korrigiert den Interraterreliabilitätswert und vergleicht die Übereinstimmung oberhalb Zufall mit dem, was potenziell möglich ist. Deshalb ist er einer der am meisten eingesetzten Werte für Interraterreliabilität bei nominalen oder Kategoriellen Daten (Soeken & Prescott, 1986, S.734).
Kappa errechnet sich unter Anwendung von Figur 2 folgendermassen: $\kappa = (p_o – p_c) / (1 - p_c)$, wobei p_o dieselbe Bedeutung hat wie weiter oben erklärt ($p_o=(a+d) / N$) und p_c für die zufällige Übereinstimmung steht (chance agreement). p_c errechnet sich folgendermassen: $(f1 * g1 + f2 * g2) / N^2$.

Die Anwendung von Kappa bringt allerdings ihrerseits wieder mehrere Probleme mit sich. Kappa setzt voraus, dass jeder Rater eine relativ fixierte vorausgehende Wahrscheinlichkeit hat, dass er positive oder negative Ratings macht. Diese Annahme dürfte für die meisten Situationen, in den Kappa errechnet wird nicht unbedingt zutreffen (Feinstein & Cicchetti, 1990, S.548). Im weiteren reagiert κ auf die Prävalenz der beobachteten Einheit mit Veränderungen, auch wenn die Treffsicherheit der beiden Rater (und damit deren Übereinstimmung) konstant bleiben. Dies kann zum Paradoxon führen, dass eigentlich eine hohe Übereinstimung vorhanden wäre, sich aber ein tiefer κ ergibt. Dies könnte nur ausgeschlossen werden, wenn die Prävalenz vor der Studie bestimmt und allenfalls korrigiert würde (Feinstein & Cicchetti, 1990, S.548), was im Normalfall nicht praktikabel ist. Cicchetti und Feinstein (1990, S.557) empfehlen, dass neben κ immer auch der Wert der positiven Übereinstimmung (p_{pos}) und der negativen Übereinstimung (p_{neg}) mit errechnet und angegeben werden, damit das Paradoxon bei κ entdeckt werden kann. Uebersax (2002b) ist der Meinung, dass die Angabe von p_{pos} und p_{neg} den Kappa-Wert unnötig macht.
Beim Vergleich von Ratings von Experten mit demjenigen von „normalen“ Anwendern macht es zudem keinen Sinn, eine Übereinstimmung durch Zufall von der gesamten Übereinstimmung zu berücksichtigen, da ein Vergleich mit einem Goldstandard möglich ist. In einer solchen Situation wäre die Berechnung der Spezifität und der Sensitivität (wie bei einem diagnostischen Test) angebracht (Feinstein & Cicchetti, 1990, S.544).

Sensitivität und Spezifität
Sensitivität und Spezifität (gelegentlich auch Spezifizität genannt) stammen ursprünglich aus der Epidemiologie und sind Gütekriterien für einen diagnostischen Test. Damit die Parameter berechnet werden können, muss eindeutig bekannt sein (also durch eine Art Goldstandard gesichert sein), ob eine Krankheit vorhanden ist oder nicht.
Sensitivität (Empfindlichkeit) in epidemiologischem Sinne meint den Anteil kranker Personen in der untersuchten Population, die durch den Test als krank erkannt wurde (Beaglehole, Bonita, & Kjellström, 1997, S.136). Umgemünzt auf eine Reliabilitätssituation könnte dann gesagt werden, dass Sensitivität der Anteil der Ereignisse/Zustände ist, welche eine Raterin (im Vergleich zum Experten) korrekt identifiziert.
Spezifität heisst in epidemiologischem Sinne: Anteil der gesunden Personen, die der untersuchte Test identifiziert. Umgemünzt auf eine Reliabilitätssituation könnte dann gesagt werden, dass Spezifität der Anteil der Situationen ist, bei welchen eine Raterin (im Vergleich zum Experten) ein Ereignis oder einen Zustand korrekt als nicht vorhanden identifiziert.
Unter Anwendung der Figur 2 und mit der Annahme, dass es sich beim Rater 2 um den Experten handelt, berechnen sich die beiden Parameter folgendermassen (Beaglehole et al., 1997, S.137):

- Sensitivität: a / f1
- Spezifität: d / f2.

Korrelation
Die Korrelation ist ein Mass für die Stärke eines linearen Zusammenhangs (Moore, 1995, S.114). Obwohl beispielsweise Polit und Hungler (1999, S.416) oder Haas (1988, S.57) die Korrelation als möglichen Parameter für Interraterreliabilität angeben, muss fest gestellt werden, dass dieser sich nicht eignet, da die Interraterreliabilität die *Übereinstimmung* messen sollte, und nicht den linearen Zusammenhang (Müller, 2002).



Figur 3: Korrelation versus Übereinstimmung

In Figur 3 beträgt die Korrelation der roten Ratingpunkte r = 0,96 und erreicht damit einen sehr hohen Wert. Gleichzeitig ist das Rating von Rater 1 aber konstant wesentlich höher als das von Rater 2. Es existiert somit eine schlechte Übereinstimmung, also eine schlechte Interraterreliabilität, obwohl sich ein hoher Korrelationswert ergibt.
Bei einer guten Interraterreliabilität müssten sich die Ratings entlang der blau eingezeichneten Winkelhalbierenden gruppieren.

10.1.4.3 In der Literatur verwendete Parameter für Interraterreliabilität

Folgende Tabelle zeigt, welche statistischen Parameter in der vom Autor bearbeiteten Literatur verwendet wurden. Bezüglich prozentualer Übereinstimmung bleibt anzumerken, dass diese in keiner einzigen Publikation näher definiert ist. Vermutlich handelt es sich meistens um die totale prozentuale Übereinstimmung.

Tabelle 1: verwendete Parameter für Interraterreliabilität in der recherchierten Literatur

	%-Übereinstimmung	Kappa	Korrelation	Anderes
Algera-Osinga et al., 1994		x		
Churness et al., 1991	x			
Cockerill et al., 1993	x			
Detwiler & Clark, 1995	x		x[8]	
Fagerström et al., 2000	x			
Levenstam & Engberg, 1993	x			
Lovett et al., 1994			x	
O'Brien-Pallas et al., 1989	x			
O'Brien-Pallas et al., 1997	x			
Phillips et al., 1992		x		
Prescott & Phillips, 1988; Prescott et al., 1991; Prescott & Soeken, 1996a, 1996b[9]	x	x		
Reitz, 1985b	x	x		Ar-Statistic
Santamaria et al., 2001			x	
Sarnecki et al., 1998			x	
Turner Stokes et al., 1998	x		x	

[8] Es ist nicht ganz klar, um welchen Parameter es sich handelt; es wird lediglich von „inter-rater-reliability coefficient of 0.946 gesprochen. Vermutlich handelt es sich um einen Korrelatinskoeffizienten.
[9] In allen Publikationen wurden beide Parameter angegeben.

10.1.4.4 Angestrebte Werte für Interraterreliabilitätsparameter

Es existiert eine gewisse Bandbreite bezüglich der Empfehlungen in der Literatur, bezogen auf die Werte, welche die Instrumente erreichen sollten.

prozentualen Übereinstimmung

Die meisten Empfehlungen werden bezüglich der prozentualen Übereinstimmung gegeben. Auch bezüglich empfohlener Werte wird keine Unterscheidung gemacht zwischen den unterschiedlichen Möglichkeiten der Prozentübereinstimmung. Der Bereich der Empfehlungen liegt zwischen 80% und 95%, wobei einige Autoren unterschiedliche Situationen unterscheiden.

Eher niedrigere tolerierte Werte werden angegeben für:

- Faktorenmodelle (im Gegensatz zu Prototypenmodellen) (*Ontario Guide to Case Costing*, S.77)
- Übereinstimmung auf einzelnen Items resp. Faktoren (im Gegensatz zur Übereinstimmung bezüglich einer abschliessenden Kategorie) (Noyes, 1994, S.8)
- erste Überprüfung nach der Instrumentenentwicklung (Anderson & Rokosky, 2001, S.61)
- Personaldotation (im Gegensatz zu Fallkostenberechnungen) (Hernandez & O'Brien-Pallas, 1996a, S.44)

Korrelation

Bezüglich Korrelation macht nur Haas (1988, S.58) eine Aussage; sie gibt einen Zielwert von r=0.9 an.

Kappa

Prescott et al. (1989) sowie Phillips et al. (1992) geben für Kappa die folgende Bewertungsskala:

- 0-0.20 armselig
- 0.21-0.40 mässig
- 0.41-0.60 mittel
- 0.61-0.80 substanziell
- 0.81 – 1.00 sehr gross.

Uebersax (2002a) betrachtet solche Bewertungsskalen jedoch als unangebracht und empfiehlt, sie nicht zu brauchen.

10.1.5 Häufigkeit der Bestimmung der Reliabilität der Instrumente im Alltagsgebrauch

Da einerseits Reliabilität und Validität auch nach erfolgtem Nachweis in einem bestimmten Setting immer nur bis zu einem gewissen Grad als gegeben angesehen werden kann und bei jeder neuen Verwendung von Instrumenten neu bestimmt werden sollte (De Groot, 1989b, S.24; Waltz, Strickland, & Lenz, 1991, S.23), und andererseits der schnelle Wandel auch in der pflegerischen Arbeitswelt eine einmal ermittelte Reliabilität resp. Validität gefährdet (Edwardson & Giovannetti, 1994), werden regelmässig wiederkehrende Reliabilitäts- und Validitätsbestimmungen gefordert. Die Empfehlungen sind recht unterschiedlich, der geforderte Standard erscheint aber insgesamt als verblüffend hoch:

- wöchentlicher oder mindestens monatlicher mit randomisierter Patientenauswahl (Alward, 1983, S.16)
- monatlich mit mindestens 10% der Patientenentlassungen oder mindestens 8 Patienten (*Ontario Guide to Case Costing*, S.77)
- vierteljährlich; häufiger wenn schlechte Resultate. Randomisiert 15-20 % der Patienten,

aber mindestens 5 Patienten (Giovannetti & Mayer, 1984, S.33)
- vierteljährlich; bei Verwendung der Daten nur für Personaldotation: 10% der Patienten. Bei Verwendung der Daten für Fallkostenberechnungen: 15-20% der Patienten, aber mindestens 5 (Hernandez & O'Brien-Pallas, 1996a, S.44)
- vierteljährlich jede Station 15-20% ihrer Patienten, aber nicht weniger als 5 Patienten. (Noyes, 1994, S.8)
- halbjährlich bis vierteljährlich (Haas, 1988, S.58)
- jährlich; halbjährlich wenn sich schlechte Resultate zeigen. Randomisiert 10% aller Patienten während einer 2-Wochen-Periode (Ebener, 1985, S.326)
- ein oder zweimal jährlich (Levenstam & Engberg, 1993, S.233)
- jährlich (McHugh & Dwyer, 1992, S.28)

10.1.6 Ursachen für schlechte Reliabilität

Bei mangelnder Reliabilität muss den Gründen nachgegangen werden (Giovannetti & Mayer, 1984, S.33). Die Literatur enthält verschiedenste Hinweise darauf, wo die Ursache für eine schlechte Reliabilität liegen könnte; die Gründe können beim Instrument selber liegen oder bei den Instrumentenanwendern.
Gemäss Walker und Avant (1998, S.89) muss eine operationale Definition so präzise sein, dass sie von verschiedenen Anwendern wiederholt verwendet werden kann und doch immer wieder zu den gleichen objektiven Ergebnissen führt. Nach McHugh und Dwyer (1992, S.29) sind mangelhaft definierte Begriffe ein wichtiger Grund für schlechte Reliabilität. Es darf keine Unklarheiten oder Überlappungen zwischen Kategorien geben, damit eine hohe Reliabilität erreicht werden kann (Alward, 1983, S.15). Die Klassen in einem Klassifikationssystem sollten unter anderem erschöpfend und sich gegenseitig ausschliessend sein (Aydelotte & Hope Peterson, 1987). Castorr et al. (1990, S.313) relativieren diese Ansprüche allerdings mit der Aussage, dass es unwahrscheinlich sei, dass ein Instrument so gut definierte und sich gegenseitig ausschliessende Kategorien haben könne, so dass jedes Ereignis im Feld perfekt in eine Kategorie passe.
Ein Instrument muss des weiteren praktikabel sein, also bspw. darf keine zu lange Liste von Indikatoren vorhanden sein und der Aufwand für die Anwendung des Instrumentes darf nicht zu hoch sein (Hernandez & O'Brien-Pallas, 1996a, S.44 f). Die Validität ihrerseits hat ebenfalls einen Einfluss auf die Reliabilität. Wenn das Instrument aus der Sicht der Anwender die Pflege nicht spiegelt (Face Validity) resp. den Arbeitsaufwand nicht korrekt aufzeigt (Predictive Validity), so werden es diese nicht mit genügend Ernsthaftigkeit anwenden.
Auf Seiten der Pflegenden muss, wie in Abschnitt 7.3 schon erwähnt, Einsicht in den Nutzen des Instrumentes vorhanden sein. Im weiteren ist wesentlich, dass die Pflegenden gründlich geschult und in der Anwendung zu Beginn begleitet werden (Castorr et al., 1990; Polit & Hungler, 1999, S.415). Um eine akzeptable Reliabilität zu erreichen braucht es gemäss Giovannetti (1984, S.32) mehrere Monate nach der eigentlichen Einführung. Es wird eine ausführliche theoretische Einführung mit Fallbeispielen empfohlen, danach erst die begleitete Anwendung in der Praxis. Poulson (1987) sowie Hernandez und O'Brien-Pallas (1996b) dokumentieren eingehend je den Prozess der Einführung des Instrumentes Medicus mit gleichzeitiger fortlaufender Überwachung der Interraterreliabilität in ein Spital.

10.2 Validität

Validität ist ein Mass dafür, wie gut ein Instrument wirklich das Konzept misst, welches es zu messen vorgibt (Polit & Hungler, 1999, S.418). Es werden folgende drei Hauptformen der Validität unterschieden: Inhaltsvalidität (Content Validity), Kriteriumsvalidität (Criterion-Related Validity) und Konstruktvalidität (Construct Validity) (Polit & Hungler, 1999, S.418 ff), welche weiter unten näher diskutiert werden.
Edwardson und Giovannetti kritisieren allgemein den Mangel an Validität der Instrumente, räumen aber ein, dass die Tatsache, dass Anbieter-entwickelte und implementierte Systeme wie GRASP, Medicus und PRN breit gestreut in Nordamerika angewendet werden, auf eine gewisse Akzeptanz unter den Anwendern hinweise (1994, S.100).
Validität wird üblicherweise bei der Implementierung in ein bestimmtes Setting bestimmt. Die schnellen Veränderungen in Technologie, Pflegepraxis und Pflegeorganisation stellen ein grösseres Problem für die Validität dar (Edwardson & Giovannetti, 1994, S.115). Aus diesem Grund wird empfohlen, wie bei der Reliabilität auch, in gewissen Abständen die Validität formell zu überprüfen (Fagerström et al., 2000, S.484). Dies betrifft nicht nur die im Kapitel 8 schon erwähnte Forderung, dass die den Kategorien zugewiesenen Zeiten oder Quantifizierungskoeffizienten in jeder Organisation neu bestimmt werden müssen. McHugh und Dwyer treten für eine Überprüfung mindestens alle drei Jahre ein. Das Ontario Joint Policy & Planning Committee empfiehlt die Überprüfung alle ein bis zwei Jahre, oder zusätzlich, wenn in der Praxis Situationen eintreten, welche darauf schliessen lassen, dass eine Anpassung nötig ist (JPPC, S.9).

10.2.1 Inhaltsvalidität

Inhaltsvalidität geht der Frage nach, ob ein Instrument die richtigen und die relevanten Aspekte abdeckt, welche zum zu messenden Konzept gehören. Hier besteht also eine wichtige Verbindung zur Frage des theoretischen Hintergrundes der Instrumente, welche im Kapitel 9 schon diskutiert wurde.
Mit der Augenscheinvalidität (Face Validity) als schwächster Form der Inhaltsvalidität wird eine Aussage darüber gemacht, ob ein Instrument auf den ersten Blick ungefähr etwa das Richtige misst, ob es plausibel ist. Augenscheinvalidität ist für die allererste Akzeptanz bei den Anwendern des Instruments wichtig, kann aber nicht quantifiziert werden. Damit Inhaltsvalidität auf einer höheren Ebene als blosser Augenscheinvalidität erreicht werden kann, identifizieren oder verifizieren klinische Experten während der Entwicklung des Instrumentes deren einzelne Anteile bspw. mit folgenden Techniken: Brainstorming, Nominal Group Process, Delphi-Technik, Q-sort (Edwardson & Giovannetti, 1994, S.115). Dabei können (bspw. bei der Delphi-Methode) auch statistische Methoden zum Einsatz kommen. Nach erfolgter Entwicklung kann die Inhaltsvalidität durch einen CVI (Content Validity Index) ermittelt werden, wo wiederum Experten die einzelnen Instrumententeile auf einer Vier-Punkte-Skala bezüglich Inhalt bewerten (Polit & Hungler, 1999, S.419).
In der vom Autor gelesenen Literatur zu einzelnen Instrumenten wird nur vereinzelt detailliert auf die Erreichung der Inhaltsvalidität eingegangen. Es wird nirgends erwähnt, dass ein CVI ermittelt wurde. Für das Instrument GRASP wurde ein modifizierter Delphiprozess angewendet ("WORKLOAD MEASUREMENT - THE GRASP® SYSTEM,", S.4). Lovett et al. (1994, S.1711) berichten von der Anwendung der Delphitechnik, aber nur für die Beurteilung, ob die Indikatoren in die vorgeschlagene Kategorie gehören, jedoch nicht für die Überprüfung der Vollständigkeit der Indikatoren.
Levenstam und Engberg (1993, S.233) geben an, dass sie Gruppendiskussionen mit Stationsleiterinnen für die Entwicklung vom Zebra-System einsetzten. Bei Reitz (1985b, S.34)

überprüfte ein Expertenpanel die Inhaltsvalidität in einem Pilottest; es scheint aber keine quantitative Auswertung statt gefunden zu haben. Santamaria et al. (2001, S.11) berichten, dass die Entwicklung ihres Instrumentes mit Experten erfolgte; bezüglich einer Überprüfung der Inhaltsvalidität wird aber nichts ausgesagt.

Inhaltsvalidität ist wichtig, damit Pflegende der Praxis genügend Vertrauen in das System haben und es damit richtig anwenden. Die Akzeptanz der Instrumente durch die Pflegenden wird gar als entscheidend für deren Erfolg erachtet (Schnetzler, 2002, S.17).
Die Erfahrung zeigt, dass gewisse Aktivitäten wie bspw. Krisenintervention den Pflegenden wichtig erscheinen, weil sie viel Zeit in Anspruch nehmen. Obwohl im Einzelfall viel Zeit verbrauchend, treten sie im Akutspital eher selten auf und sind damit nicht mächtig, die Zeit für das Gros der Patienten zu schätzen. Aus diesem Grund gilt, dass Inhaltsvalidität nötig, aber nicht genügend ist; die Einschätzung von Experten sollte durch andere Methoden wie der prognostischen Validität (siehe folgender Abschnitt) ebenfalls verifiziert werden (Edwardson & Giovannetti, 1994, S. 115 f).

10.2.2 Kriteriumsvalidität und Konstruktvalidität

Kriteriumsvalidität misst die Korrelation zwischen einer Messung des untersuchten Instrumentes und einem „Aussen-Indikator“ (Frank-Stromborg, 1988, S.38) und ist eher „praktisch orientiert“ (Courtens, 1999). Konstruktvalidität untersucht die Fähigkeit des untersuchten Instrumentes, Propositionen resp. Hypothesen zu messen, welche aus einer „Mini-Theorie“ im Zusammenhang mit dem gemessenen Phänomen abgeleitet werden (Courtens, 1999; Fagerström et al., 2000, S.485). Kriteriumsvalidität und Konstruktvalidität lassen sich zwar theoretisch unterscheiden, liegen aber ganz nahe beieinander (Courtens, 1999). Wenn die Kriteriumsvalidität mit einem „Aussen-Indikator“ gemessen wird, so muss dieser „Aussen-Indikator“ seinerseits ja eine theoretische Verbindung zum gemessenen Konstrukt aufweisen. Wie weit diese theoretische Verbindung eher „praktisch orientiert“ ist und ab wann man von der Prüfung eigentlicher Propositionen resp. Hypothesen sprechen kann, ist Ansichtssache.
Kriteriumsvalidität wird üblicherweise weiter unterteilt in prognostische Validität (Predictive Validity) und Konkurrenzvalidität (Concurrent Validity). Die Konstruktvalidität wird unterteilt in Kontrastgruppenansatz (Known-Groups Approach), Konvergenzvalidität (Convergence-Validity), Diskriminanzvalidität (Discriminant-Validity) und Faktorenanalyse aufgeteilt. Konstruktvalidität gilt allgemein als die am schwierigsten zu erreichende Validität; so wurde diesbezüglich hinsichtlich Pflegeleistungserfassungsinstrumenten entsprechend am wenigsten geforscht (Alward, 1983, S.327; Hernandez & O'Brien-Pallas, 1996a, S.41). Da die „richtige“ Zuteilung in Kriteriumsvalidität und Konstruktvalidität resp. deren Unterformen teilweise Ansichtssache ist, werden diese beiden Validitätsformen in diesem Kapitel gemeinsam diskutiert.

Prognostische Validität
Prognostische Validität ist die wichtigste Validitätsform, wenn das Instrument primär als Personaldotationssteuerungsinstrument eingesetzt wird (Edwardson & Giovannetti, 1994, S.116). Bezogen auf Pflegeleistungserfassungsinstrumente wird bei der prognostischen Validität der Korrelationskoeffizient zwischen vorausgesagtem und aktuell benötigtem Wert, oder der Prozentwert korrekter Voraussagen bestimmt (Ebener, 1985, S.327). Gemessen an der oben erwähnten Wichtigkeit dieser Validitätsform, scheint diese gemäss der vom Autor bearbeiteten Literatur relativ selten erhoben worden zu sein:

- Churness et al. (1991, S.20) ermittelten die Korrelation zwischen ihrem Instrument der Gemeindepflege und der Dauer der Einsätze; es konnten nur 46-64% der Variabilität (r^2) der Zeit durch das Instrument erklärt werden.
- Bei einer Untersuchung von Prescott und Soeken (1996b, S.91) konnte das Instrument PINAC nur 35% der Zeit erklären, welche per Selbstbeobachtung durch die Pflegenden erhoben wurde.

Konkurrenzvalidität
Konkurrenzvalidität vergleicht das zu untersuchende Instrument mit einem anderen Instrument, welches für denselben Zweck entwickelt wurde oder wie weiter oben aufgeführt, mit einem „Aussen-Indikator". Beim Vergleich mit einem anderen Instrument, welches für denselben Zweck entwickelt wurde, sollte idealerweise der Vergleich mit einem als Gold-Standard akzeptierten Instrument erfolgen, was aber mangels eines breit akzeptierten und validierten Pflegeleistungserfassungsinstruments schwierig ist (Alward, 1983, S.16). Zudem muss vor einer Messung die Forderung aus Kapitel 8 bedacht werden, dass die den Kategorien zugewiesenen Zeiten oder Quantifizierungskoeffizienten in jeder Organisation neu bestimmt werden müssen (Giovannetti, 1979, S.7). Zudem ist aus Sicht des Autoren dieser Arbeit sehr genau zu berücksichtigen, an welcher Stelle des Pflegeprozesses das zu überprüfende Instrument ansetzt, und an welcher Stelle jenes, welches zum Vergleich verwendet wird (vgl. Figur 1 S.9). So ist bspw. zu berücksichtigen, dass die geplanten Aktivitäten für den Patienten evtl. nicht ausschliesslich auf dem angestrebten Zustand basieren, sondern dass diese in bestimmten Fällen bspw. durch die Qualifikation des vorhandenen Personals beeinflusst werden könnte. Wenn solche Faktoren nicht berücksichtigt werden, sind ungerechtfertigt niedrige Resultate in Validitätsstudien zu erwarten.

Weil die benötigte Zahl Pflegender nicht nur eine Funktion des direkten Pflegebedarfs ist, und weil die Ergebnisse zweier kleinerer Instrumentenvergleichsstudien eine gute Vergleichbarkeit der Instrumente anzeigte, ging man bis 1991 davon aus, die verschiedenen Systeme würden zu vergleichbaren Ergebnissen in der Anwendung führen (Edwardson & Giovannetti, 1994, S.103). Ab der breit angelegten Studie von O'Brien-Pallas änderte sich dies aber.

In ihrer Dissertation berichtete O'Brien-Pallas et al. (1989) von signifikanten Unterschieden der Schätzung von benötigtem Pflegepersonal, als die Instrumente PRN 76, GRASP und Medicus auf verschiedenen Stationen gleichzeitig eingesetzt wurden. Es gab auf den meisten Stationen klinisch relevante, signifikante Unterschiede unter allen Instrumenten von bis zu fünfzig Prozent; auf einer Station ermittelte GRASP bspw. durchschnittlich 4.42 Stunden, PRN 76 hingegen 6.66 Stunden. Die Instrumente korrelierten aber hoch miteinander. In einer Replikationsstudie, bei der fünf Instrumente getestet wurden (PRN 76, PRN 80, Medicus, GRASP, NISS), zeigten sich sehr ähnliche Ergebnisse. Hier erreichte der grösste Unterschied auf einer Station sogar einhundert Prozent; bei NISS 3.69, bei PRN 80 hingegen 7.38 Stunden. Die Instrumente korrelierten erneut deutlich (O'Brien-Pallas, Cockerill, & Leatt, 1992). Die Resultate sind besonders bemerkenswert, da sie auf einem sehr fundierten Forschungsdesign basieren; die einzelnen Instrumente wurden mit aller Sorgfalt implementiert und die Reliabilität bestimmt, bevor die eigentliche Datenerhebung statt fand. Angesichts der Tatsache, dass die Instrumente für die Personaldotation und Kostenkalkulation eingesetzt werden, sind die Resultate Besorgnis erregend. Da die Instrumente miteinander korrelieren, lassen sich aber wenigstens Patienten mit grösserem von solchen mit kleinerem Aufwand unterscheiden, was eine nützliche Information sowohl für die Verteilung von Personal als auch von Kosten in einem Betrieb ist. Von der Verwendung der absoluten Zahlen müsste

aber angesichts der Resultate wohl abgeraten werden.

Cockerill et al. (1993) (O'Brien-Pallas war hier auch mitbeteiligt) verglichen in einer weiteren Studie die Kosten bei Case Mix Groups (CMGs)[10], wenn diese mit Hilfe der Instrumente PRN 76, PRN 80, Medicus, GRASP, NISS ermittelt wurden. Bei NISS fielen die vorausgesagten Kosten in einer bestimmten CMG um 75% höher aus als bei GRASP. Die Autoren schliessen aus den Resultaten, dass es zu grösseren Problemen kommen könnte, wenn Fallkosten in einem Pilottest mit einem bestimmten Instrument errechnet werden, und dieselben Preise nachher in einem Setting angewendet werden, welche mit einem anderen Pflegeleistungserfassungsinstrument gemessen werden. Bei der Studie kann allerdings kritisiert werden, dass die Daten anhand der Patientendokumentation erhoben wurden und nicht anhand der „realen" Patienten (siehe Kritik auf S.18). Es wurden zudem jeweils alle Instrumente von denselben Datensammlerinnen angewendet, wodurch eine gegenseitige Beeinflussung nicht auszuschliessen ist.

Phillips et al. (1992) verglichen das PINI-Instrument mit Medicus und GRASP, wobei sich Korrelationen von r=0.66 resp. r=0.69 (S.49) ergaben. Die Autoren erklären sich das Resultat damit, dass GRASP und Medicus die Zeit gemäss Time and Motion-Studien bestimmen würden, bei PINI die Zeit jedoch direkt im Instrument als ein Faktor von den Pflegenden eingeschätzt wird (S.51). Allerdings wird in einer späteren Version von PINI das Item Zeit entfernt (Prescott & Soeken, 1996a), vermutlich weil es mit den anderen Items zu stark korrelierte. Auf Grund der Resultate einer multiplen linearen Regression, mit der versucht wird, die GRASP und die Medicus-Scores durch die Items im PINI-Instrument zu erklären, kommen die Autoren zum Schluss, dass die Systeme GRASP und Medicus unterschiedliches messen, weil die Verteilung von R auf die PINI-Items für die beiden Vergleichsinstrumente unterschiedlich ist. Das kumulative R bei Einbezug aller PINI-Item in das Modell in beiden Vergleichsinstrumenten ist allerdings sehr ähnlich (r=0.72 resp. r=0.777). Interessant ist, dass O'Brien-Pallas et al. (1992, S.20) beim Vergleich von GRASP und Medicus eine Korrelation von r=0.92 nachwies, während dem die Korrelation zwischen PINI und GRASP und PINI und Medicus nur r=0.66 resp. r=0.69 erreichte. Dies lässt aus Sicht des Autors diese Arbeit eher darauf schliessen, dass PINI nicht dasselbe Konzept misst wie GRASP und Medicus. Die durch PINI mittels der multiplen linearen Regression erklärte Varianz (R^2) für GRASP und Medicus erreicht denn auch nur 51% resp. 60% (Phillips et al., 1992, S.51)[11].

Carr-Hill und Jenkins-Clarke (1995) haben vier wenig bekannte und wohl nur in England eingesetzte Instrumente miteinander verglichen. Auch hier korrelierten die Instrumente (zwischen r=0.83 und r=0.91, S.223), es waren aber wiederum sehr grosse Unterschiede ersichtlich (in einem Fall um einhundert Prozent, S.223). Die Autoren ziehen folgenden Schluss: „there is no evidence, that the nursing workload measurement systems deployed in the UK are anything more than an expensive numbers game" (S.221). Leider ist das Studiendesign sehr schlecht beschrieben; es werden bspw. keinerlei Angaben darüber gemacht, wer die Instrumente angewendet hat, wie diese geschult wurden und ob Reliabilitätswerte erhoben

[10] Kanadische Form der Diagnoses Related Groups (DRGs) (O'Brien-Pallas et al., 1995, S.9)

[11] Ein Vergleich von Korrelationen ist allerding nicht unproblematisch, da Werte in einem grösseren Skalenbereich zu höheren Korrelationen führen als bei einem schmaleren Bereich (Polit & Hungler, 1999, S.414); wenn also im einen Fall bspw. gleichzeitig Werte verglichen werden, die sowohl auf einer tagesklinischen Station als auch auf einer Palliativstation erhoben wurden, wo die Pflegezeiten pro Patient also weit auseinander liegen, so wird dies zu einem höheren Korrelationswert führen, als in einer Studie, wo bspw. nur Werte aus der Palliativstation verglichen werden.

wurden[12].

Das NPDS-Instrument (Northwick Park Dependency Score), welches in Neurorehabilitationsstationen eingesetzt werden soll, wurde mit dem Barthel-Index, einem Instrument zur Erfassung grundlegender Alltagsfunktionen resp. Behinderungen von Patienten mit neuromuskulären oder muskulo-skelettalen Erkrankungen verglichen und erzielte eine sehr hohe Korrelation von r=–0.91 (Turner Stokes et al., 1998).

Bei einem Vergleich eines Prototypen mit einem faktorenbasierten Instrument zeigte sich, dass die Anwesenheitszeit von Spitex-Pflegenden durch das Faktoreninstrument, welches die durchgeführten Tätigkeiten erhob, zu einem wesentlich grösseren Grad erklären konnte (39%) als das Prototypeninstrument (13%). Das Faktoreninstrument hatte einen retrospektiven Ansatz, das Prototypeninstrument einen prospektiven (Tiesinga, Halfens, Algera-Osinga, & Hasman, 1994).

Fagerström et al. verglichen das OPC-Instrument mit PAONCIL (Professional Assessment of Optimal Nursing Care Intensity Level), einer subjektiven globalen Einschätzung der Pflegeintensität von Pflegenden auf einer siebenstufigen Skala. Bei der linearen Regression ergab sich ein : R^2 von 0.366 bei einem r=0.605 (Fagerström et al., 2000, S.488). Für das Instrument PAONCIL bestehen ebenfalls keine Messungen zu Reliabilität und Validität, wie sie normalerweise für einen Goldstandard nötig wären. Allerdings lässt sich diskutieren, ob eine subjektive Einschätzung wie ein herkömmliches Messinstrument mit Reliabilitäts- und Validitätswerten bewertet werden kann oder darf.

Anderson und Rokosky (2001) verglichen ein Instrument der Gemeindepflege mit der Dauer der Pflege und der Anzahl der Besuche sowie mit ausgewählten Items eines Medicare Outcome- und Assessmentinstrumentes, welche primär Patientenzustände und Pflegebedarf darstellen. Von 27 zu testenden Hypothesen erwiesen sich jedoch nur fünf als statistisch signifikant. Es kann allerdings gefragt werden, ob dies an falschen Hypothesen, der mangelnden Power durch ein zu kleines Sample oder wirklich am zu testenden Instrument lag.

Das Instrument von Reitz (1985b) wurde in einer Studie mit „Severity of Illness“, der Aufenthaltsdauer (Length Of Stay, LOS) und verschiedenen Kostenstellen (ohne Pflege) korreliert. Gemäss der Autorin bestand das Hauptziel der Studie darin, ein Instrument zu entwickeln und zu testen, das homogene Patientengruppen bezüglich Pflegebedarf identifizieren kann. Es wird dem Autor dieser Arbeit nicht klar, weshalb obige Parameter zur Korrelation mit dem Instrument gewählt wurden, und welche Aussage die errechneten Korrelationswerte bezüglich der Instrumentenqualität machen sollen.

Kontrastgruppenansatz
Beim Kontrastgruppenansatz wird das Instrument bei zwei Gruppen eingesetzt, bei denen man davon ausgeht, dass sie sich bezüglich dem zu messenden Konzept unterscheiden; es wird untersucht, ob das Instrument in der Lage ist, den Unterschied zu erfassen (Polit & Hungler, 1999, S.421). Haas (1988, S.58) schlägt vor, Patienten, bei denen man erwartet, dass ihr Aufwandlevel wechselt, mit dem Instrument zu messen. Haas ordnet dies der Vorhersagevalidität zu; der Autor dieser Arbeit würde dieses Vorgehen eher als Kontrastgruppenansatz sehen.

[12] Dies ist umso erstaunlicher, als die Publikation im renommierten Journal of Advanced Nursing erschienen ist.

In einer Studie von Santamaria et al. (2001) wendeten zwanzig Pflegende das Instrument an sechs konstruierten Beispielen an, wobei diese so konstruiert waren, dass sie für einen bestimmten Komplexitätslevel stehen. Es wurde die Differenz zwischen beabsichtigtem und in der Studie erfasstem Komplexitätsgrad errechnet.

Konstruktvalidiät allgemein
Die Ergebnisse der Scores des PINI-Instrumentes wurden mit folgenden Variablen verglichen:

- Korrelation mit der Anzahl medizinischer Zweitdiagnosen: r=0.33 (es war ein relativ tiefer Wert erwartet worden, da PINI weitere Faktoren berücksichtigt)
- Korrelation mit der Anzahl Konsultationen bei medizinischem Spezialisten: r=0.17 (der Wert war tiefer als erwartet)
- Korrelation mit „Severity of Illness“ (Schwere der medizinischen Diagnosen; gemessen mit einem Instrument, das von Ärzten standardmässig angewendet wird): r=0.44,
- Korrelation mit der Aufenthaltsdauer (LOS): r=0.31
- Ort der Entlassung (in andere Institution oder nach Hause); Patienten die nach Hause entlassen wurden hatten einen tieferen PINI-Wert als jene, welche in andere Institution entlassen wurden
- Korrelation mit Medicus r=0.7, mit GRASP: r=0.54, und mit San Joaquin-Pflegeleistungserfassungsinstrument: r=0.55
- Prozentuale Übereinstimmung der Messung der tatsächlich beobachteten geleisteten Pflege (Prescott et al., 1991).

Die Autoren ordnen dies alles unter Konstruktvalidität ein; es scheint fraglich, ob dies nicht eher zur Kriteriumsvalidität gehören würde.

11 Grenzen von Pflegeleistungserfassungsinstrumenten

In den bisherigen Kaptiteln sind schon an mehreren Stellen Grenzen und Einschränkungen von Pflegeleistungserfassungsinstrumenten ersichtlich geworden. An dieser Stelle soll auf einschränkende Aspekte resp. Kritikpunkte eingegangen werden, welche in der Literatur zusätzlich genannt werden, und welche ebenfalls einen Einfluss auf die Reliabilität und Validität der Instrumente haben können.

11.1 Gleichzeitige Erbringung mehrerer Leistungen: Multitasking

Edwardson und Giovannetti (1994, S.112) warnen, dass die gleichzeitige Erbringung zweier oder mehrerer Aktivitäten zu zu hohen Schätzungen von benötigter Zeit führen, wenn die Zeit für die individuelle Pflege einfach summiert werde.
Hughes (1999, S.319) kritisiert an den Pflegeleistungserfassungsinstrumenten, sie würden die Fähigkeit von Pflegenden ignorieren, Pflegeleistungen simultan zu erbringen. So könne bspw. jemandem auf die Toilette geholfen und gleichzeitig deren Verhältnisse zu Hause im Zusammenhang mit der Entlassung besprochen werden. Diese Fähigkeit der simultanen Pflege sei stark situationsgebunden und damit schlecht voraussagbar.
Malloch und Conovaloff (1999, S.50) geben ein weiteres Beispiel: Beim Wechsel eines Venenkatheters kann die Pflegende gleichzeitig den physischen Zustand des Patienten beobachten, die Schmerzen einschätzen und Informationen bezüglich Medikation geben.
In der Studie zur Bestimmung von Zeitrichtlinien für die Pflege von Patienten im Laienbereich wurde ebenfalls wiederholt bestätigt, dass in Pflegesituationen verschiedene Handlungen

parallel laufen, z.B. Aushandeln der Essenspläne während der Ganzkörperwäsche (Bartholomeyczik & Hunstein, 2001, S.261). Auch in der Studie zur Dauer beruflicher ambulanter Pflege des Fraunhofer Instituts für Materialwirtschaft und Logistik (1991) wird offenbar die Überlagerung von Arbeitsabläufen als eines der herausragenden Probleme bei der Zeiterfassung von Pflege benannt (Bartholomeyczik et al., 2001, S.51).

11.2 Individuelle Leistungsfähigkeit / Anpassungsfähigkeit der Pflegenden

In der Pflege tragen bestimmte Tätigkeiten ein grösseres Risiko in sich, sind komplexer und brauchen deshalb höhere Kompetenz als andere, obwohl sie gleich viel Zeit verbrauchen wie einfachere Tätigkeiten. So braucht es bspw. weniger Kompetenzen, einem an beiden Händen operierten Patienten Essen einzugeben, als für ein Beratungsgespräch bezüglich einer Ernährungsumstellung (Van Slyck, 1991, S.23). Gemäss Phillips et al. (1992, S.47) sollte die Messung von Pflegeressourcenverbrauch die Menge der Leistungen *und* die nötige Qualifikation beinhalten; die meisten Systeme würden aber nur ersteres messen. Auch O'Brien-Pallas et al. (1995, S.17) fordern, dass zukünftige Instrumente den Mix der Pflegenden berücksichtigen müssten. Für Kostenberechnungen sollte die Qualifikation der beteiligten Personen laut Sherman (1990, S.15) bekannt sein. Dies gilt aber sicher auch für die Planung der Personaldotation.
Hughes (1999, S.320) sowie Prescott und Phillips (1988, S.17) kritisieren, dass keines der Messsysteme den Faktor mit einzubeziehen scheint, dass Pflegende bei erhöhtem Arbeitsanfall härter und schneller arbeiten können.

11.3 Qualität der Leistungen

Isfort (2001, S.49) bemerkt, dass handlungsbezogene Messverfahren keine Aussage über die Qualität der geleisteten Arbeit machen. Die Art der Intervention bleibt also nur summarisch gezählt, nicht inhaltlich begründet oder geklärt. Auch Needham (1997, S.84) stellt fest, dass Pflegeaufwandmessungen eher auf Quantität als auf Qualitätsaspekte des Gemessenen setzen. Es gibt keine Beweise dafür, dass genügend Personal durch Pflegeleistungserfassungsinstrumente auch die Pflegequalität garantiert. (Giovannetti, 1979, S.8; Haas, 1988, S.59). Sovie (1988, S.146) kritisiert in ihrer Übersicht, der Mangel an Daten über die Qualität der Pflege sei eine der ernsthaften Schwächen der Studien, welche die Kostenverteilung in der Pflege untersuchten.

11.4 Unkorrektes Handling / Missbrauch von Pflegeleistungserfassungsinstrumenten

Pflegeaufwandmesssysteme basieren auf professioneller Einschätzung; die dafür verwendeten Informationen müssen objektiv, relevant, konsistent und applikabel sein, sonst kann Misstrauen entstehen, sowohl beim Management als auch bei den Pflegenden selber (Needham, 1997, S.87). Fischer meint, in der Administration könnten Befürchtungen aufkommen, dass Instrumente ausgenutzt werden (1995, S.28).
Gemäss Krempels (2002, S.36) werden die meisten Versicherer parallel zu den Erhebungen der Praktiker eigene Kontrollmessungen durchführen – nicht weil sie die Wissenschaftlichkeit der Instrumente anzweifeln würden; vielmehr bezweifelten sie, dass die Instrumente in der Praxis auch wirklich korrekt gehandhabt werden. Diese Vorsicht scheint nicht übertrieben zu sein, denn Malloch und Conovaloff (1999, S.49) erwähnen, dass Missbrauch und Manipulation der Systeme nicht selten seien. Hierfür wurde der englische Begriff „Acuity creep“ geprägt, was die häufigste Form der „Nichtreliabiliät“ von Pflegeleistungserfassungsinstrumenten sei und darin bestehe, dass Patienten nicht korrekt eingeschätzt würden; beinahe immer in der Richtung eines höheren Pflegebedarfs (DeGroot, 1994a, S.48; Van Slyck, 1991, S.24). Giovannetti (1984, S.31) gibt einen möglichen Grund für Mogelei an: wenn ein Instrument von den Pflegenden nicht als zuverlässig angeschaut werde.

12 Die Methode LEP® Nursing 2

Ende der Achtziger- und Anfang der Neunzigerjahre wurden am Kantonsspital St. Gallen und am Universitätsspital Zürich, zuerst unabhängig voneinander, später aber teilweise gemeinsam und in Zusammenarbeit mit dem Interdisziplinären Forschungszentrum für Gesundheit St. Gallen und Wissenschaftlern des Soziologischen Seminars der Universität St. Gallen die Methoden PAMS (Patienten Aufwand Mess-System) und SEP-USZ (System zur Erfassung des Pflegeaufwandes am Universitätsspital Zürich) entwickelt (LEP-AG, 2002b; Maeder et al., 1992). 1996 wurden die beiden Methoden endgültig vereinheitlicht und erhielten die Bezeichnung LEP (Leistungserfassung in der Pflege). Die beiden Entwicklergruppen schlossen sich zusammen (Fischer, 2002).
Ab 1995 wurde die Methode auch ausserhalb der beiden Ursprungsspitäler eingesetzt; 1996 setzten schon 21 Institutionen LEP ein. Damit die Weiterentwicklung unabhängig von den beiden Ursprungsspitälern möglich war, erfolgte 2000 die Firmengründung der LEP-AG. Im Moment wird die Methode LEP in 106 Institutionen der deutsch- und französischsprachigen Schweiz, sowie in neun Betrieben in Deutschland (alle in Projektphasen) eingesetzt (Wirthner M., persönliches E-Mail, 30.4.2003). Da seit einiger Zeit auch ein Erfassungsmodul für physiotherapeutische Arbeit besteht, wurde der Name auf LEP® Nursing erweitert. Im Folgenden wird der Zusatz „Nursing“ nicht mehr gebraucht, ist aber bei jeder Nennung von LEP mit gemeint.
Es existieren zwei LEP-Versionsfamilien. Im Folgenden wird nur auf die aktuell propagierte Familie 2 eingegangen (obwohl neunzehn Betriebe teilweise aus EDV-technischen Gründen noch Versionen aus der alten Familie betreiben). Eine detaillierte Beschreibung von LEP Version 2 würde den Rahmen dieser Arbeit sprengen. Im Folgenden werden nur die wichtigsten Eigenschaften beschrieben; wo nichts anderes vermerkt ist, stammen die Informationen aus folgenden Quellen: (Brügger et al., 2001; Fischer, 2002; Güntert & Maeder, 1994; LEP-AG, 2002b).
Bei LEP handelt es sich um ein Pflegeleistungserfassungsinstrument mit retrospektivem Fokus für den stationären Akutbereich. Es wird erfasst, was tatsächlich an Leistungen erbracht wurde; gemäss Figur 1 S.9 also die durchgeführten Aktivitäten der Pflege. Eine prospektive Anwendung wäre theoretisch möglich, erfolgt aber gemäss LEP-AG in der Praxis nirgends (Wirthner M., persönliches E-Mail, 9.1.2003). LEP unterscheidet zwischen Tätigkeiten, welche einzelnen Patienten zuordenbar sind (im Folgenden als direkte Pflegetätigkeiten benannt), und allen andern Verrichtungen (bspw. Stationsunterhalts- oder Managementarbeiten; im Folgenden als indirekte Pflegetätigkeiten benannt). Die direkten Pflegetätigkeiten werden anhand einer Liste von 56 Verrichtungen mit jeweils einer bis vier Ausprägungen (auch als Aufwandstufen bezeichnet) von den Pflegenden möglichst fortlaufend elektronisch erfasst. Im LEP-Sprachgebrauch wird bei jeder Ausprägung von einer Variablen gesprochen (wenn im weiteren Verlauf des Textes der Begriff „LEP-Variable“ verwendet wird, ist damit die Verrichtung und deren Aufwandstufe gemeint). Pro LEP-Variable existiert im Normalfall eine Variablenbezeichnung, eine Beschreibung, Beispiele, eventuelle Bemerkungen und Anleitungen sowie ein Zeitwert (siehe Figur 4). Der Zeitwert wurde mit Experten der Pflege auf Grund von Schätzungen so fest gelegt, dass sie jene Zeit abbildet, welche ausgebildete und erfahrene Personen im Durchschnitt für die Erledigung der Tätigkeit in angemessener Qualität benötigen. Dabei ist die Vorbereitung, Durchführung und Nachbearbeitung respektive Entsorgung des Materials, sowie eine allfällig nötige Kommunikation sowie Dokumentation bezüglich der Tätigkeit mit enthalten, sofern sie sich im üblichen Rahmen hält.

Mobilisation aufwändig	31.03

Beschreibung	Der Patient / die Patientin erhält aufwändige Unterstützung für die Mobilisation.
Beispiele	- Mobilisation mit zwei Personen - Aufstehen mit Periduralanalgesie während der Geburt - Aufwändige Mobilisation aufgrund von therapeutischen Massnahmen oder neurologisch / motorisch bedingten Störungen - Aufwändige Mobilisation mit Prothese inkl. Stumpfbandage anlegen
Bemerkungen	Die Variable umfasst die ganze Mobilisation inklusive der Verwendung von eventuellen Hilfsmitteln. Sie beinhaltet die zielgerichtete Beobachtung, Begleitung und Unterstützung des Patienten / der Patientin zur Förderung der Selbständigkeit / Gesundheit.
Anleitung	Abgrenzung zu folgenden Variablen beachten: Variablen 54.17/18/19/20 Hilfsmittel herstellen / anpassen
Zeitwert	30 Minuten

Figur 4: Beispiel einer LEP-Variablen

Neben dem oben beschriebenen Normalfall existieren zusätzlich Informationsvariablen, welche gewisse Patientenzustände dokumentieren; diesen ist jedoch kein Zeitwert hinterlegt – die Informationen sollen lediglich die Interpretation der Auswertungen erleichtern. Speziell behandelt werden auch jene Tätigkeiten, welche keine Ausprägungen ausweisen, sondern denen ein Basiszeitwert zugewiesen ist. Erfasst wird die tatsächlich verbrauchte Zeit in Einheiten des Basiswertes. Dieser Tätigkeitstyp wurde gewählt bei erfahrungsgemäss äusserst variablem Zeitaufwand oder wenn mehrere Patienten gleichzeitig Empfänger einer einzigen Pflegeleistung sind (bspw. bei Beschäftigung und Freizeitaktivität in einer Gruppe). Die Pflegevariablen sind in vierzehn Variablengruppen eingeordnet. Da gewisse Verrichtungen respektive Ausprägungen nur in bestimmten Fachbereichen vorkommen, stellt jede Station einen für sie gültigen Variablenkatalog aus der Gesamtliste zusammen.
Die indirekten Pflegetätigkeiten werden rein kalkulatorisch ermittelt aus der prozentualen Differenz der Personalzeiten von den direkten Pflegetätigkeiten.
LEP 2 kann (im Gegensatz zu LEP 1) entsprechend der Modelltypologie nach Edwardson und Giovannetti (1994) als Pflegetätigkeitenmodell (und damit gleichzeitig auch als Faktorenmodell nach Abdellah (1979)) bezeichnet werden.
Für LEP 2 werden folgende Nutzungsformen propagiert: Planung, Steuerung und Auswertung pflegerischer Arbeit für die Pflege selber; Schaffung von Transparenz gegenüber nicht pflegerischen Bereichen - innerhalb und ausserhalb von Betrieben; Berechnung von Stellenplänen; Schaffung von Kostentransparenz (beispielsweise zur Fallkostenberechnung oder als Rechnungsstellungsbasis); Datenbasis für Pflegeforschung. Damit geht der Anspruch weit über eine reine Steuerung der Personaldotation hinaus.

12.1 Validität und Reliabilität von LEP

Gemäss Isfort und Klug (2002, S.23) liegen keine differenzierten Daten bezüglich Validität und Reliabilität von LEP vor, was auch vom Geschäftsführer der LEP-AG bestätigt wird (Bamert U., persönliches Telefonat , Dezember 2002). Brosziewski und Brügger (2001, S.65) lehnen es ab, die auf Grund von Konventionen den Variablen zugewiesenen Zeitwerte unabhängig empirisch zu überprüfen, mit der Begründung, bei LEP handle es sich nicht um eine

Methode zur Überprüfung wissenschaftlicher Hypothesen, sondern um ein sozialwissenschaftliches Instrument zur Instruktion von Managemententscheidungen. Dem gegenüber steht die Tatsache, dass als Grund für den Einsatz von LEP auch der mögliche Rückgriff von Pflegeforschung auf eine grosse Datenbasis angepriesen wird (Brügger et al., 2001, S.5), und dass LEP-Daten schon in mindestens zwei Studien verwendet wurden (Baumberger, 2001; Fischer, 1999). In der Schweiz wurde der den Variablen hinterlegte Zeitwert mittels Piloterhebungen validiert (LEP-AG, 2002b, 2.1 B, S.7), welche aber nicht publiziert und damit nicht zugänglich sind. In Deutschland wurden die hinterlegten Zeiten gewisser Variablen sowohl einer Einschätzungsüberprüfung als auch einer Ist-Zeiten-Messung unterzogen, wobei sich zeigte, dass Anpassungen nötig sind (Isfort & Klug, 2002, S.57 ff) (was noch keine Rückschlüsse auf die Verhältnisse in der Schweiz zulässt).
In einer multizentrischen Pilotstudie wird momentan die Validität von LEP bezogen auf die Situation von Intensivpflegestationen in Deutschland untersucht. Erste Teilergebnisse zeigen, dass für den Einsatz in Deutschland kleine sprachliche Anpassungen vorgenommen werden müssen. Expertenschätzungen der den Variablen zugeordneten Zeiten, sowie negative C-Werte im Praxiseinsatz deuten teilweise auf Abweichungen im Vergleich zu den in der Schweiz fest gelegten Zeiten. Die Ursache dafür wird noch genauer abgeklärt (Behrens & Horbach, 2002).
Brosziewski und Brügger (2001, S.63) postulieren eine „professionelle Validität“, welche eine Plausibilisierung im täglichen Gebrauch darstelle und die damit „schärfer“ sei als die Augenscheinvalidität (Brügger in Fischer, 2002, S.150). Die weite Verbreitung der Methode LEP sowie die Bemühungen der Methodenentwickler, das Instrument den Bedürfnissen und Veränderungen in der Praxis laufend anzupassen (sich entwickelnde Instrumentenversionen; jährliche Anwenderkonferenz; Internetauftritt mit Forum etc.), können neben ihrer Funktion zur Verbreitung des Instrumentes auch als Indiz für Augenscheinvalidität gewertet werden. Ein wissenschaftlicher Nachweis der Inhaltsvalidität, etwa mittels einer Delphi-Untersuchung wurde allerdings bisher nicht erbracht.
Im Kantonsspital Schaffhausen wurde an einer Stichprobe von zwanzig Prozent aller Patienten an einem bestimmten Tag die Plausibilität der erfassten LEP-Variablen anhand eines Vergleichs mit der Patientendokumentation überprüft. Da bei dieser Überprüfung aber gleichzeitig auch die Dokumentationsqualität anhand der LEP-Variablen kontrolliert wurde, kann bei Abweichungen die Ursache sowohl bei der Dokumentation als auch bei der LEP-Erfassung liegen. Diese Überprüfung erhebt keinen wissenschaftlichen Anspruch und ist nicht publiziert (Holenstein M., persönliches Telefonat, 30.12.2002).
Offenbar hat vom 15.1.03 bis 28.02.03 eine Studie im Johanna Etienne Krankenhaus (in Neuss, Deutschland) stattgefunden, wo auf einer orthopädisch / neurologisch gemischten Station parallel LEP, DTA (Diagnosebezogene Tätigkeitsanalyse) und FIM (Functional Independence Measure) eingesetzt wurde (Schulz, 2003). Nähere Informationen zu dieser Studie, welche vermutlich in Richtung Konkurrenzvaliditätsstudie geht, waren noch nicht erhältlich.
Wissenschaftliche Untersuchungen zu allen Formen der Reliabilität (Stabilität, Äquivalenz und Homogenität) scheinen bisher zu fehlen.

13 Schlussfolgerung

Das Instrument LEP reiht sich in das Glied der vielen anderen Pflegeleistungserfassungsinstrumente, für die keine oder kaum nachgewiesene Validität und Reliabilität vorhanden sind. Insbesondere bei der Äquivalenz, welche für Pflegeleistungserfassungen besonders wichtig ist, ist dieser Mangel bedeutend. Angesichts der Forderung der Literatur, dass Zeitbestimmungen für Pflegeleistungserfassungsinstrumente für jeden Einsatzort neu bestimmt werden müssten und der Tatsache, dass die Validierung der durch Experten fest gelegten Zeiten nur durch nicht publizierte Pilotstudie(n) unbekannter Grösse erfolgte, eröffnet sich hier ein grosser Forschungsbedarf. Validierungsstudien, die die Inhalts- und Kriteriumsvalidität messen, wären ebenfalls nötig. Der wohl grösste Mangel aber besteht in den fehlenden Reliabilitätsstudien, welche nicht nur bei der Entwicklung und Implementierung eines Instrumentes erfolgen, sondern auch nach deren Einführung regelmässig wiederholt werden sollten. Hier besteht wohl der grösste Forschungsbedarf. Unter Berücksichtigung der weiten Verbreitung des Instrumentes in der Schweiz und dem Interesse im benachbarten Ausland zeigt sich der Mangel nur noch um so stärker.

Hauptliteraturübersicht

Tabelle 2: Übersicht über die recherchierte Hauptliteratur

	Reliabilitäts- resp. Validitätstestung Einzelinstrument	Instrumenten- beschrieb / -entwicklung / -anwendung	Übersichtsartikel Pflegeleistungser- fassungsinstrumente	Übersichtsartikel Methoden der Reli- abilitäts- resp. Validitätstestung	Übersichtsartikel Kostenallokation in der Pflege	Anderes
Ontario Guide to Case Costing,						x
"WORKLOAD MEASUREMENT - THE GRASP® SYSTEM,"		x				
Abdellah & Levine, 1979			x			
Algera-Osinga et al., 1994	x					
Alward, 1983			x			
Anderson, 1997		x				
Anderson & Rokosky, 2001	x	x				
Arthur & James, 1994			x			
Bartholomeyczik et al., 2001	x					
Behrens & Horbach, 2002		x				
Bigbee et al., 1992				x		
Botter, 2000		x				
Brosziewski & Brügger, 2001		x				
Brügger et al., 2001		x				
Carr Hill & Jenkins Clarke, 1995	x					
Churness et al., 1991	x	x				
Cicchetti & Feinstein, 1990				(x)		
Cockerill et al., 1993	x					
De Groot, 1989b		(x)	x			
De Groot, 1989a		(x)	x			
DeGroot, 1994a			x			
DeGroot, 1994b		x				
Detwiler & Clark, 1995	x	x				
Ebener, 1985				x		
Edwardson & Giovannetti, 1994			x			
Fagerström & Engberg, 1998			x			
Fagerström et al., 2000	x					
Feinstein & Cicchetti, 1990				(x)		
Forchuk, 1996			x			
Giovannetti, 1979			x			
Giovannetti & Mayer, 1984				x		
Giovannetti & Johnson, 1990		x				
Güntert & Maeder, 1994		x				
Haas, 1988			x	x		
Hernandez & O'Brien-Pallas, 1996a				x		
Hlusko & Nichols, 1996	x					
Hughes, 1999			x			
McHugh & Dwyer, 1992						x
Isfort, 2001			x			
Isfort & Klug, 2002	x					

	Reliabilitäts- resp. Validitätstestung Einzelinstrument	Instrumenten-beschrieb / -entwicklung / -anwendung	Übersichtsartikel Pflegeleistungser-fassungsinstrumente	Übersichtsartikel Methoden der Reli-abilitäts- resp. Validitätstestung	Übersichtsartikel Kostenallokation in der Pflege	Anderes
JPPC,						x
Klee, 1993		x				
Krempels, 2002						x
LEP-AG, 2002b		x				
Levenstam & Engberg, 1993	x	x				
Levenstam & Engberg, 1997		x				
Lovett et al., 1994	x	x				
Maeder et al., 1992		x				
Malloch et al., 1999	(x)	x				
Malloch & Conovaloff, 1999			x			
McDaniel, 1994	x			x		
McKenzie, 1991	x	x				
Needham, 1997						
Noyes, 1994			x	x		
O'Brien-Pallas, 1988		x				
O'Brien-Pallas et al., 1989	x					
O'Brien-Pallas et al., 1992	x					
O'Brien-Pallas et al., 1994		x				
O'Brien-Pallas et al., 1995			x		x	
O'Brien-Pallas et al., 1997		x				
Phillips et al., 1992	x					
Prescott & Phillips, 1988		x				
Prescott et al., 1989	x					
Prescott et al., 1991	x					
Prescott & Soeken, 1996b	x					
Prescott & Soeken, 1996a		x				
Procter & Hunt, 1994						x
Reitz, 1985a		x				
Reitz, 1985b	x					
Santamaria et al., 2001	x	x				
Sarnecki et al., 1998	x	x				
Sherman, 1990					x	
Sovie, 1988					x	
Tiesinga et al., 1994	x					
Trivedi & Hancock, 1975	x					
Turner Stokes et al., 1998	x	x				
Unger, 1985	x	x				
Van Slyck, 1991		x				

Abkürzungsverzeichnis

3PCS	Third-Generation Patient Classification System
ARIC	Allocation, Resource Identification and Costing
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CMGs	Case Mix Groups
DRGs	Diagnosis Relatet Groups
DTA	Diagnosebezogene Tätigkeitsanalyse
FIM	Functional Independence Measure
GRASP	GRACE Reynolds Application of the Study PETO[13]
HMO	Health Maintenance Organization (= Gesundheitskasse)
JCAHO	Joint Commission an Accreditation of Healthcare Organizations
LEP	Leistungserfassung in der Pflege
LOS	Length Of Stay
MeSH	Medical Subject Headings
NISS	Nursing Information System Saskatchewan
NPDS	Northwick Park Dependency Score
OPC	Oulu Patient Classification
PAMS	Patienten Aufwand Mess-System
PAONCIL	Professional Assessment of Optimal Nursing Care Intensity Level
PCS	Patient Classification System
PINAC	Patient Intensity for Nursing: Ambulatory Care
PINI	Patient Intensity for Nursing Index
PPR	Pflegepersonalregelung
PRN	Project de Recherche en Nursing
Psych-PV	Psychiatrische Personalverordnung
RME	Référentiels médico-économiques
SEP-USZ	System zur Erfassung des Pflegeaufwandes am Universitätsspital Zürich
VBK	Verband Bernischer Krankenhäuser

[13] GRACE ist das erste Spital, welches das Instrument anwendete; Reynolds steht für „Reynolds Aluminum Company“, welche die Entwicklung des Instrumentes mitfinanzierte; PETO steht für die ersten Buchstaben der Namen jener Personen, welche erste Vorarbeiten für das Instrument leisteten (Polland,English,Thornton,Owen).

Literaturverzeichnis

Abdellah, F. G., & Levine, E. (1979). Patient Classification Methods, *Better Patient Care Through Nursing Research* (Second ed.). New York: Macmillian.

Algera-Osinga, J. T., Halfens, R., Hasman, A., & Wiersma, D. (1994). A Dutch patient classification system for community care. *The Journal of nursing administration, 24*, 32-38.

Alward, R. R. (1983). Patient classification systems: the ideal vs. reality. *The Journal of Nursing Administration, 13*, 14-19.

Anderson, K. L., & Rokosky, J. S. (2001). Evaluation of a home health patient classification instrument. *Western Journal of Nursing Research, 23*, 56-71.

Anderson, L. (1997). The role and resources required for the introduction of generic ward assistants using GRASP systems workload methodology: a quantitative study. *Journal of Nursing Management, 5*, 11-17.

Arthur, T., & James, N. (1994). Determining nurse staffing levels: a critical review of the literature. *Journal of Advanced Nursing, 19*, 558-565.

Aydelotte, M. K., & Hope Peterson, K. (1987). Nursing taxonomies-state of the art, *Classification of nursing diagnoses proceedings of the seventh conference*. St. Louis: Mosby.

Bartholomeyczik, S., & Hunstein, D. (2000). *Pflegezeitbedarf, Personalbemessung und Fachkraftanteil in vollstationären Einrichtungen*.Unpublished manuscript.

Bartholomeyczik, S., & Hunstein, D. (2001). Die Messung von Pflegezeiten - methodische und inhaltliche Probleme. *Pflege, 14*, 259-266.

Bartholomeyczik, S., Hunstein, D., Koch, V., & Zegelin-Abt, A. (2001). *Zeitrichtlinien zur Begutachtung des Pflegebedarfs: Evaluation der Orientierungswerte für die Pflegezeitbemessung*. Frankfurt: Mabuse.

Baumberger, D. (2001). *Pflegediagnosen als Indikator der Streuung des Pflegeaufwandes in DRGs.* Unpublished Master Thesis: Master of Nursing Science, Universität Maastricht NL, Maastricht NL / Aarau CH.

Beaglehole, R., Bonita, R., & Kjellström, T. (1997). *Einführung in die Epidemiologie* (A. Pause, Trans.). Bern: Verlag Hans Huber.

Behrens, J., & Horbach, A. (2002). *Von "L"EP zu LEP und zu LEP-D*. Halle-Wittenberg: Institut für Gesundeheits- und Pflegewissenschaft der Martin-Luther-Universität.

Bigbee, J. L., Collins, J., & Deeds, K. (1992). Patient classification systems: a new approach to computing reliability. *Applied nursing research : ANR, 5*, 32-38.

Björvell, C., Thorell-Ekstrand, I., & Wredling, R. (2000). Development of an audit instrument for nursing care plans in the patient record. *Quality in health care, 9*, 6-13.

Botter, M. L. (2000). The use of information generated by a patient classification system. *The Journal of Nursing Administration, 30*, 544-551.

Brosziewski, A., & Brügger, U. (2001). Zur Wissenschaftlichkeit von Messinstrumenten im Gesundheitswesen: Am Beispiel der Methode LEP. *Pflege, 14*, 59-66.

Brosziewski, A., & Maeder, C. (2001). *Produkte der Ethnographie in der Produktion des Unternehmens* [Online in Internet]. Retrieved 20.04.2002, 2002, from the World Wide Web: http://www.arbeitskulturen.de/down/091brosz_mae.htm

Brügger, U., Bamert, U., & Maeder, C. (2001). *LEP Beschreibung der Methode LEP Nursing 2*. St. Gallen: LEP AG.

Carr Hill, R. A., & Jenkins Clarke, S. (1995). Measurement systems in principle and in practice: the example of nursing workload. *Journal of Advanced Nursing, 22*, 221-225.

Castorr, A. H., Thompson, K. O., Ryan, J. W., Phillips, C. Y., Prescott, P. A., & Soeken, K. L. (1990). The process of rater training for observational instruments: implications for interrater reliability. *Research in nursing & health, 13*, 311-318.

Churness, V. H., Kleffel, D., & Onodera, M. (1991). Home health patient classification system. *Home Healthcare Nurse, 9*, 14-22.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*, 551-558.

Cockerill, R., O'Brien-Pallas, L., Bolley, H., & Pink, G. (1993). Measuring nursing workload for case costing. *Nursing Economic$, 11*, 342-349.

Courtens, A. (1999). *Vorlesung Reliabilität und Validität; Notizen Ernst Näf*.Unpublished manuscript, Maastricht.

De Groot, H. A. (1989a). Patient classification system evaluation. Part 1: Essential system elements. *The Journal of Nursing Administration, 19*, 30-35.

De Groot, H. A. (1989b). Patient classification system evaluation: Part 2, System selection and implementation. *The Journal of Nursing Administration, 19*, 24-30.

DeGroot, H. A. (1994a). Patient classification systems and staffing. Part 1, Problems and promise. *The Journal of Nursing Administration, 24*, 43-51.

DeGroot, H. A. (1994b). Patient classification systems and staffing. Part 2, Practice and process. *The Journal of Nursing Administration, 24*, 17-23.

Detwiler, C., & Clark, M. J. (1995). Acuity classification in the urgent care setting. *The Journal of nursing administration, 25*, 53-61.

Ebener, M. K. (1985). Reliability and validity basics for evaluating classification systems. *Nursing Economic$, 3*, 324-327.

Edwardson, S. R., & Giovannetti, P. B. (1994). Nursing workload measurement systems. *Annual Review of Nursing Research, 12*, 95-123.

Exchaquet, N. F., & Züblin, L. (1975). *Wegleitung zur Berechnung des Pflegepersonalbedarfs für Krankenstationen in Allgemeinspitälern*. Bern: O.V.

Fagerström, L., & Engberg, I. B. (1998). Measuring the unmeasurable: a caring science perspective on patient classification. *Journal of Nursing Management, 6*, 165-172.

Fagerström, L., Rainio, A., Rauhala, A., & Nojonen, K. (2000). Validation of a new method for patient classification, the Oulu Patient Classification. *Journal of Advanced Nursing, 31*, 481-490.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Fischer, W. (1995). *Leistungserfassung und Patientenkategorisierung in der Pflege*. Wolfertswil: ZIM.

Fischer, W. (1999). *Die Bedeutung von Pflegediagnosen in Gesundheitsökonomie und Gesundheitsstatistik* [Online in Internet]. Retrieved 30.07.2002, 2002, from the World Wide Web: http://www.fischer-zim.ch/artikel/Pflege-Diagnosen-9901-WEG.htm

Fischer, W. (2001, 10.04.2002). *Möglichkeiten zur Abbildung der Pflege in DRG-Systemen* [Online in Internet]. Retrieved 30.7.2002, 2002, from the World Wide Web: http://www.fischer-zim.ch/auszuege-drg-pflege/Pflege-in-DRG-Systemen-0109.htm#Note_1

Fischer, W. (2002). *Diagnosis Related Groups (DRGs) und Pflege; Grundlagen, Codierungssysteme, Integrationsmöglichkeiten*. Bern: Huber.

Forchuk, C. (1996). Workload measurement and psychiatric mental health nursing: mathematical and philosophical difficulties. *Canadian journal of nursing administration, 9*, 67-81.

Frank-Stromborg, M. (1988). Evaluating Instruments for Use in Clinical Nursing Research, *Instruments for Clinical Nursing Research* (pp. 3 - 18). Norwalk: Appleton & Lange.

Giovannetti, P. (1979). Understanding patient classification systems. *The Journal of Nursing Administration, 9*, 4-9.

Giovannetti, P., & Johnson, J. M. (1990). A new generation patient classification system. *The Journal of Nursing Administration, 20*, 33-40.

Giovannetti, P., & Mayer, G. G. (1984). Building confidence in patient classification systems. *Nursing Management , 15*, 31-34.

Güntert, B. J., & Maeder, C. (1994). *Ein System zur Erfassung des Pflegeaufwandes; Darstellung der Methode SEP des Universitätsspitals in Zürich*. Muri: Schweizerische Gesellschaft für Gesundheitspolitik SGGP.

Haas, S. A. (1988). Patient classification systems: a self-fulfilling prophecy. *Nursing Management, 19*, 56-58, 60-52.

Hernandez, C. A., & O'Brien-Pallas, L. L. (1996a). Validity and reliability of nursing workload measurement systems: review of validity and reliability theory. *Canadian journal of nursing administration, 9*, 32-50.

Hernandez, C. A., & O'Brien-Pallas, L. L. (1996b). Validity and reliability of nursing workload measurement systems: strategies for nursing administrators. *Canadian journal of nursing administration, 9*, 33-52.

Hlusko, D. L., & Nichols, B. S. (1996). Can you depend on your patient classification system? *The Journal of Nursing Administration, 26*, 39-44.

Höfert, R. (2003). *Die Situation der Pflege in Deutschland* [Online in Internet]. Deutscher Pflegeverband. Retrieved 30.4.03, 2003, from the World Wide Web: http://www.dpv-online.de/Informationen/Pflege_KONKRET/2003/Pflege_KONKRET_03_01.pdf

Hughes, M. (1999). Nursing workload: an unquantifiable entity. *Journal of Nursing Management, 7*, 317-322.

Isfort, M. (2001). *Pflegequalität und Pflegeleistungen I. Zwischenbericht zur ersten Phase des Projektes „Entwicklung und Erprobung eines Modells zur Planung und Darstellung von Pflegequalität und Pflegeleistungen“* [Online in Internet]. Deutsches Institut für angewandte Pflegeforschung / Katholischer Krankenhausverband Deutschlands. Retrieved Frühling, 2002, from the World Wide Web: http://www.dip-home.de/downloads/downloads.htm

Isfort, M., & Klug, E. (2002). *Pflegequalität und Pflegeleistungen 2. Zweiter Zwischenbericht zur zweiten Phase des Projektes: "Entwicklung und Erprobung eines Modells zur Planung und Darstellung von Pflegequalität und Pflegeleistungen"* [Online in Internet]. Katholischer Krankenhausverband Deutschlands e.V. Retrieved 09.04.2002, 2002, from the World Wide Web: http://www.dip-home.de/downloads/downloads.htm

JPPC. *Nursing Resource Consumption* [Online in Internet]. JPPC, Ontario Joint Policy & Planning Committee; Nursing Professional Advisory Working Group. Retrieved 24.08.2002, 2002, from the World Wide Web: http://www.jppc.org/library/mis/nursing.pdf

Klee, D. (1993). Erfahrungen mit der Umsetzung der Pflegepersonalregelung; Chancen für die interne Organisation. *Die Schwester / Der Pfleger, 32*, 1009-10016.

Krempels, J. (2002). Der Theorie-Praxis-Transfer bei der Instrumentenentwicklung- Erfahrungen eines Sozialwissenschafters. *Managed Care*, 35-37.

Lang, N. (2003). *Penn Nursing* [Online in Internet]. Retrieved 26.03.2003, 2003, from the World Wide Web: http://www.nursing.upenn.edu/faculty/profile.asp?pid=45

LEP-AG. (2002a). LEP Informationen Nr. 15. St. Gallen: LEP AG.

LEP-AG. (2002b). *LEP-Methoden-Handbuch*. St. Gallen: LEP-AG.

Levenstam, A. K., & Engberg, I. B. (1993). The Zebra system--a new patient classification system. *Journal of Nursing Management, 1*, 229-237.

Levenstam, A. K., & Engberg, I. B. (1997). How to translate nursing care into costs and staffing requirements: part two in the Zebra system. *Journal of Nursing Management, 5*, 105-114.

Linck, W. (1995). *Pflege als Problemlösungs- und Beziehungsprozess.* Unpublished Seminararbeit der Pflege- und Gesundheitswissenschaften, Darmstadt.

Lovett, R. B., Reardon, M. B., Gordon, B. K., & McMillan, S. (1994). Validity and reliability of medical and surgical oncology patient acuity tools. *Oncology Nursing Forum, 21*, 1709-1717.

Maeder, C., Brügger, U., Longerich, H., & Güntert, B. (1992). Patientenklassifikation und Arbeitsbelastung in der Pflege: Das Modell SEP-USZ. *Pflege, 5*, 63-73.

Malloch, K., & Conovaloff, A. (1999). Patient classification systems, Part 1: The third generation. *The Journal of Nursing Administration, 29*, 49-56.

Malloch, K., Neeld, A. P., McMurry, C., Meeks, L., Wallach, M., Williams, S., & Conovaloff, A. (1999). Patient classification systems, Part 2: The third generation. *The Journal of Nursing Administration, 29*, 33-42.

McDaniel, A. M. (1994). Using generalizability theory for the estimation of reliability of a patient classification system. *Journal of Nursing Measurement, 2*, 49-62.

McHugh, M. L., & Dwyer, V. L. (1992). Measurement issues in patient acuity classification for prediction of hours in nursing care. *Nursing Administration Quarterly, 16*, 20-31.

McKenzie, D. A. (1991). Proposed prototype for identifying and correcting sources of measurement error in classification systems. *Medical Care, 29*, 521-530.

Moore, D. S. (1995). *Basic Practice of Statistics*. New York: W.H. Freeman and Company.

Müller, M. (2002). *Statistikvorlesung Prof. M. Müller zum Thema Reliabilität*.Unpublished manuscript, Aarau.

Needham, J. (1997). Accuracy in workload measurement: a fact or fallacy? *Journal of Nursing Management, 5*, 83-87.

Noyes, B. (1994). Inter-rater reliability. Regaining credibility with your staff and financial officer while meeting JCAHO standards. *The Journal of Nursing Administration, 24*, 7-8.

O'Brien-Pallas, L. (1988). An analysis of the multiple approaches to measuring nursing workload. *Canadian journal of nursing administration, 1*, 8-11.

O'Brien-Pallas, L., Cockerill, R., & Leatt, P. (1992). Different systems, different costs? An examination of the comparability of workload measurement systems. *The Journal of Nursing Administration, 22*, 17-22.

O'Brien-Pallas, L., Giovannetti, P., Peereboom, E., & Marton, C. (1995). *Case Costing and Nursing Workload: Past, Present and Future*. Toronto: McMaster University Toronto.

O'Brien-Pallas, L., Irvine, D., Peereboom, E., & Murray, M. (1997). Measuring nursing workload: understanding the variability. *Nursing Economic$, 15*, 171-182.

O'Brien-Pallas, L., Irvine, D., Peereboom, E., Murray, M., Ho, R., Beed, J., & Young, J. (1994). *Factors that Influence Variability in Nursing Workload at the Hospital for Sick Children*. Toronto: McMaster University.

O'Brien-Pallas, L., Leatt, P., Deber, R., & Till, J. (1989). A comparison of workload estimates using three methods of patient classification. *Canadian journal of nursing administration, 2*, 16-23.

Ontario Guide to Case Costing [Online in Internet]. Retrieved 30.09.2002, 2002, from the World Wide Web: http://www.occp.com/

Pesut, D. J., & Herman, J. (1999). *Clinical Reasoning; The Art and Science of Critical and Creative Thinking*. Albany: Delmar.

Phillips, C. Y., Castorr, A., Prescott, P. A., & Soeken, K. (1992). Nursing intensity. Going beyond patient classification. *The Journal of Nursing Administration, 22*, 46-52.

Polit, D. F., & Hungler, B. P. (1999). *Nursing Research; Principles and Methods*. Philadelphia: Lippincott.

Poulson, E. (1987). A method for training and checking interrater agreement for a patient classification study. *Nursing Management , 18*, 72-74, 78, 80.

Prescott, P. A., & Phillips, C. Y. (1988). Gauging nursing intensity to bring costs to light. *Nursing & health care : official publication of the National League for Nursing, 9*, 17-22.

Prescott, P. A., Ryan, J. W., Soeken, K. L., Castorr, A. H., Thompson, K. O., & Phillips, C. Y. (1991). The Patient Intensity for Nursing Index: a validity assessment. *Research in nursing & health, 14*, 213-221.

Prescott, P. A., & Soeken, K. L. (1996a). Measuring nursing intensity in ambulatory care. Part I: Approaches to and uses of patient classification systems. *Nursing Economic$, 14*, 14-21, 33.

Prescott, P. A., & Soeken, K. L. (1996b). Measuring nursing intensity in ambulatory care. Part II: Developing and testing PINAC. *Nursing Economic$, 14*, 86-91, 116.

Prescott, P. A., Soeken, K. L., & Ryan, J. W. (1989). Measuring patient intensity. A reliability study. *Evaluation & the Health Professions , 12*, 255-269.

Procter, S. (1991). Subjectivity and objectivity in the measurement of nursing workload. *Journal of Clinical Nursing, 1992*, 123-129.

Procter, S., & Hunt, M. (1994). Using the Delphi survey technique to develop a professional definition of nursing for analysing nursing workload. *Journal of Advanced Nursing, 19*, 1003-1014.

Reitz, J. A. (1985a). Toward a comprehensive nursing intensity index: Part I, development. *Nursing Management , 16*, 21-30.

Reitz, J. A. (1985b). Toward a comprehensive Nursing Intensity Index: Part II, Testing. *Nursing Management , 16*, 31-42.

SAHO. (2003). Nursing Information System Saskatchewan (NISS) Overview (pp. 1-19). Saskatoon: Saskatchewan Association of Health Organizations (SAHO).

Santamaria, N., Daly, S., Addicott, R., & Clayton, L. (2001). The development, validity and reliability of the Hospital in the Home Dependency Scale (HDS). *Australian Journal of Advanced Nursing, 18*, 8-14.

Sarnecki, A. J., Haas, S., Stevens, K. A., & Willemsen, J. A. (1998). Design and implementation of a patient classification system for rehabilitation nursing. *The Journal of Nursing Administration, 28*, 35-43.

Schnetzler, R. (2002). Messungen in der Pflege - Realitäten und Visionen. *Managed Care*, 13-17.

Schulz, H. (2003). *Neuigkeiten* [Online in Internet]. Schulz, Helmut Pflegemanagementberatung. Retrieved 01.05.2003, 2003, from the World Wide Web: http://www.fim-pflegeplanung.de/whatsnew.htm

Sherman, J. J. (1990). Costing nursing care: a review. *Nursing Administration Quarterly, 14*, 11-17.

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care, 24*, 733-741.

Sovie, M. D. (1988). Variable costs of nursing care in hospitals. *Annual Review of Nursing Research, 6*, 131-150.

Tiesinga, L. J., Halfens, R. J., Algera-Osinga, J. T., & Hasman, A. (1994). The application of a factor evaluation system for community nursing in The Netherlands. *Journal of Nursing Management, 2*, 175-179.

Topf, M. (1986). Three estimates of interrater reliability for nominal data. *Nursing Research, 35*, 253-255.

Trivedi, V. M., & Hancock, W. M. (1975). Measurement of Nursing Work Load Using Head Nurses' Perceptions. *Nursing Research, 24*, 371-376.

Turner Stokes, L., Tonge, P., Nyein, K., Hunter, M., Nielson, S., & Robinson, I. (1998). The Northwick Park Dependency Score (NPDS): a measure of nursing dependency in rehabilitation. *Clinical Rehabilitation, 12*, 304-318.

Uebersax, J. (2002a, 20 July 2002). *Kappa Coefficients* [Online in Internet]. Retrieved 26.04.2003, 2003, from the World Wide Web: http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm

Uebersax, J. (2002b, 15 August 2002). *Raw Agreement Indices* [Online in Internet]. Retrieved 25.04.2003, 2003, from the World Wide Web: http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm

Unger, J. (1985). Building a classification system that works. *The Journal of Nursing Administration, 15*, 18-24.

Van Slyck, A. (1991). A systems approach to the management of nursing services--Part II: Patient classification system. *Nursing Management , 22*, 23-25.

Walker, L. O., & Avant, K. C. (1998). *Theoriebildung in der Pflege*. Wiesbaden: Ullstein Medical.

Waltz, C. F., Strickland, O. L., & Lenz, E. (1991). *Measurement in Nursing Research*. Philadelphia: Davis.

Willems, Y. (1992). *Methoden zur Pflegeaufwandmessung: eine kritische Bewertung.* Unpublished Diplomarbeit Kurs BO8, Kaderschule für die Krankenpflege, Aarau.

Williams, M. A. (1977). Quantification of direct nursing care activities. *The Journal of Nursing Administration, 7*, 15-18, 49-51.

WORKLOAD MEASUREMENT - THE GRASP® SYSTEM.: GRASP Systems Nurse Consulting International Ltd.