Dr. Ralf-Dieter Hilgers, Dr. Bernd Friedel, Prof. Dr. Günter Berghaus

Äquivalenztestung im Rahmen experimenteller Untersuchungen zur Fahrtüchtigkeit

1 Einleitung

Bei experimentellen Untersuchungen zur Fahrtüchtigkeit werden in der Regel Funktionsbereiche des Teilsystems Fahrer betrachtet. Mit Hilfe von Testverfahren werden diese Bereiche weiter operationalisiert. So kann etwa die visuelle Wahrnehmung mit Hilfe tachistokopischer Verfahren untersucht werden, wobei der Umfang sowie die Schnelligkeit und Genauigkeit der optischen Wahrnehmung bei kurzfristiger Reizexposition gemessen werden kann.

Basis der Interpretation sind im allgemeinen mehrere Testverfahren und innerhalb der Testverfahren die Placebo/Leerwert/Verum-Resultate mehrerer Meßgrößen, auch Endpoints genannt. Eine der grundlegenden Fragen ist nun, wie diese vielen Endpoints statistisch optimal bewertet werden sollen, d.h. mit welchen statistischen Methoden der simultane Vergleich mehrerer Meßwerte vorgenommen werden kann.

Im folgenden sollen Möglichkeiten vorgestellt und an einem Beispiel verglichen werden.

2 Art der Hypothesenprüfung

Zunächst gilt es, sehr deutlich zwischen zwei verschiedenen Strategien bei der Hypothesenbildung zu unterscheiden: Die Testung einer Differenz im Gegensatz zur Testung der Äquivalenz (Tabelle 1).

Bei der Testung einer Differenz zwischen Verum- und Placebo-Gruppe lautet die Null-Hypothese: Die Differenz ist nur zufällig von 0 unterschieden. Die Alternativhypothese lautet: Die Differenz ist > 0, d. h., die Gruppen unterscheiden sich signifikant. Die Fehler, die man bei einer derartigen Testung zu berücksichtigen hat, sind a, d.h. das Konsumentenrisiko (es wird ein Unterschied gefunden, obwohl in Wirklichkeit keiner besteht) und b, das Produzentenrisiko (es wird kein Unterschied gefunden, obwohl tatsächlich ein solcher vorhanden ist). Diese für klinische Studien adäquate Hypothesenbildung ist bisher fast ausschließlich auch im Bereich der Alkohol-, Medikamenten- und Drogentestung auf Fahrtüchtigkeit eingesetzt worden. Sie hat in diesem Forschungsbereich dann eine Berechtigung, wenn lediglich die Frage nach einer Differenz zwischen Leistungen interessiert.

Will man jedoch - und das ist ja letztlich Ziel der experimentellen Studien - Aussagen zur Fahrtüchtigkeit ableiten, hat die Testung der Hypothese einer Differenz entscheidende Nachteile. Zunächst muß eine statistisch signifikante Differenz nicht notwendig auch praktisch relevant sein. (Die Reduzierung der Reaktionszeit etwa um eine tausendstel Sekunde kann zwar statistisch signifikant sein, für die Fahrtüchtigkeit braucht sie jedoch keine Rolle zu spielen). Der zweite Nachteil ist die Abhängigkeit des Ergebnisses der Signifikanzprüfung von der Präzision der Testdurchführung. Bei großer Variabilität der Meßwerte in den Gruppen (etwa infolge eines schlechten Testverfahrens bzw. einer

Test auf Differenzen (z.B. klinische Prüfung)	
Δ : Differenz Verum zu Placebo-Gruppe	
Test: Nullhypothese	$H_0 : \Delta \leq 0$ kein Unterschied
Alternative	$H_1 : \Delta >$ Verum wirkt
Fehlerarten	
Fehler 1. Art α =	Wahrscheinlichkeit, Nullhypothese abzulehnen, obwohl sie zutrifft
=	Konsumentenrisiko: Verum wirkt, obwohl dies nicht zutrifft
Fehler 2. Art β =	Wahrscheinlichkeit, Nullhypothese anzunehmen, obwohl sie falsch ist
=	Produzentenrisiko: Verum wirkt nicht, obwohl dies nicht zutrifft

Tab. 1a: Hypothesenprüfung: Test auf Differenzen

Test auf Gleichwertigkeit (Äquivalenz)	
Δ : Differenz Verum zu Placebo	
δ : sachlich akzeptable Abweichung	
Test: Nullhypothese	$H_0 : \Delta \geq \delta$: Differenz ist „sachlich" bedeutsam
Alternativhypothese	$H_1 : \Delta < \delta$: Differenz ist nicht „bedeutsam"
Fehler α:	Verum wirkt nicht, obwohl dies nicht zutrifft
Fehler β:	Verum wirkt, obwohl dies nicht zutrifft

Tab. 1b: Hypothesenprüfung: Test auf Gleichwertigkeit (Quelle: WELLEK (1994), WESTLAKE (1972), WOLLMAR (1991))

„schlampigen" Arbeitsweise des Experimentators) wird im allgemeinen eine Differenz zwischen Verum und Placebo nicht signifikant sein.

Optimaler als die Testung auf Differenz ist für Fragen der Fahrtüchtigkeit daher die Testung auf Äquivalenz. Unter Äquivalenzprüfung wird folgendes verstanden: Bei einem Vergleich zwischen Verum- und Placebogruppe wird ein sachlich bedeutender Unterschied vor Versuchsbeginn fixiert. Die 0-Hypothese lautet dann: Der tatsächlich empirisch ermittelte Unterschied ist größer als dieser Effekt. Die Gegenhypothese besagt: Der tatsächlich beobachtete Unterschied ist kleiner. Unter dieser Art der Hypothesenprüfung ist der Fehler erster Art die Wahrscheinlichkeit, einen wirklich vorhandenen Unterschied zu übersehen. Diese Wahrscheinlichkeit klein zu halten bedeutet, ein mögliches Risiko für den Patienten, der ein Medikament einnimmt, zu minimieren. Der Fehler zweiter Art gibt die Wahrscheinlichkeit an, einen Unterschied aufgrund der Experimente zu behaupten, obwohl das Medikament keinen Einfluß auf die Fahrtüchtigkeit hat.

Vor dem Hintergrund dieser beiden verschiedenen Hypothesenstrategien ist zu fragen, mit welchen statistischen Methoden entweder die Differenz oder die Äquivalenz von Items simultan getestet werden kann.

3 Methoden der simultanen Testung

Wir werden vier Methoden vorstellen,

- den multiplen Vergleich
- die ANOVA-Varianzanalyse
- die Methode der multiplen Endpoints
- und die Akkumulationsstatistik,

von denen wir die Akkumulationsstatistik mit Äquivalenztestung unter bestimmten Bedingungen als optimal zur Beantwortung der Fragen der Fahrtüchtigkeit anhand der Messung mehrerer fahrrelevanter Merkmale ansehen.

Die Methoden sollen am Beispiel von 5 Items, die im Rahmen unserer Methadonsubstitutionsstudie (BERGHAUS et al. 1993) gemessen wurden, demonstriert werden. Hierbei handelt es sich um folgende Merkmale: periphere Wahrnehmung, Daueraufmerksamkeit, Tachistoskop, Tracking und mittlere Entscheidungszeit. Untersucht wurden 2 mal 13 Probanden (Methadon- und Kontrollgruppe).

3.1 Multiple Vergleiche, Alphaadjustierung

Die Tabelle 2 zeigt das Procedere: Führt man paarweise Tests auf Signifikanz der Differenz bei einem à priori gewählten α von 5 % durch, so ergibt sich lediglich für die periphere Wahrnehmung ein signifikanter Unterschied, während sich alle anderen Vergleiche als nicht signifikant herausstellen.

Bei simultaner Betrachtung aller Items muß eine Alphaadjustierung vorgenommen werden. Ein mögliches Vorgehen ist das Verfahren nach HOLM (1979). Hierbei werden die P-Werte der Größe nach sortiert. Das Signifikanzniveau wird durch die Ordnungszahl dividiert und dann werden mehrere Einzelvergleiche vorgenommen. Eine simultane Signifikanz liegt dann vor, wenn mindestens ein Test signifikant unterschiedlich ist. In unserem Beispiel liegt keine statistisch signifikante Beeinträchtigung unter Methadon bei simultanem Vergleich vor.

Ein Vorteil dieses Verfahrens liegt darin, daß man die individuellen Testresultate interpretieren kann. Nachteile sind u.a. die Konservativität, die Nichtberücksichtigung der Korrelationsstruktur, die Schwierigkeit der Interpretation bei gegensätzlichen Signifikanzen und schließlich Schwierigkeiten bei der Festlegung von Toleranzbereichen bei der Äquivalenztestung.

Merkmale	Methadon $\bar{x}$ (s)	Kontrolle $\bar{x}$ (s)	(geordnete) p-Werte (t-Test)	Korrigierter Signifikanzwert α
periphere Wahrnehmung (mittl. Reaktion gesamt)	116.46 (27.73)	93.38 (13.60)	0.0151	0.05 / 5
Daueraufmerksamkeit (Zahl der bearbeiteten Vorlagen)	452.08 (41.41)	499.23 (90.69)	0.1066	0.05 / 4
Tachistoskop (Bearbeitungszeit)	283.85 (69.61)	245.77 (74.85)	0.1919	0.05 / 3
Tracking (mittl. Abweichung von Spurmitte)	7.83 (2.62)	8.50 (1.52)	0.4376	0.05 / 2
mittlere Entscheidungszeit	401.62 (76.57)	407.92 (86.98)	0.8461	0.05 / 1

Tab. 2: Vergleich von 5 Merkmalen zwischen 2 Gruppen (beobachtet an jeweils 13 Probanden)
(Quelle: BERGHAUS et al. (1993), BAUER (1991), SHAFFER (1986), HOLM (1979), SIMES (1986), HOMMEL (1989))

3.2 Multivariates ANOVA

Dies ist ein in der experimentellen Fahrtüchtigkeitsliteratur häufig eingesetztes Verfahren, das die 0-Hypothese testet, daß Verum- und Kontrollgruppe in allen Items gleiche Mittelwerte aufweisen. Hier werden die 5 Items unseres Beispiels als ein 5-simensionaler Zielvektor aufgefaßt. Für unser Beispiel, berechnet anhand des Programms BMDP 3 D, ergab sich ein P von .07 (Tabelle 3) und somit keine Signifikanz.

Vorteil dieses Verfahrens ist, daß es auch unter schwacher Korrelationsstruktur aussagefähig bleibt. Ein wesentlicher Nachteil ist, daß fast nie die Voraussetzungen für den Einsatz der Methode, nämlich die multivariate Normalverteilung der Items, vorliegt. Außerdem ist für eine Äquivalenztestung die Definition von multivariaten Toleranzbereichen schwierig.

13 Fälle Methadon	
13 Fälle Kontrolle	
Nullhypothese:	Mittelwert für alle Variablen in beiden Gruppen gleich
F Wert	2.4323
p-Wert	0.07

Tab. 3: Multivariate Varianzanalyse
(Quelle: MORRISON (1967))

3.3 Multiple Endpoints

Dieses Verfahren zählt zu den neueren Methoden. Es besteht im 1. Schritt aus einer linearen Transformation der Merkmale in Einheiten der Standardabweichung um den Mittelwert 0 und im 2. Schritt aus einer Summation der Itemwerte pro Versuchsperson. Bei korrekter Anwendung des Verfahrens muß zusätzlich die Korrelationsstruktur zwischen den Items berücksichtigt werden. Bei der Annahme, diese Korrelationsstruktur sei vernachlässigbar, ergeben sich die in der Tabelle 4 zusammengestellten Parameter für die beiden Gruppen. Die Unterschiede sind bei Testung auf Differenz nach dem t-Test für unverbundene Stichproben nicht signifikant.

Dieses Verfahren führt auch unter moderaten Korrelationen zwischen den Items zu verwertbaren Resultaten. Entscheidender Nachteil ist jedoch u.a., daß diese Methode nicht auf das häufig eingesetzte cross-over-design anwendbar ist.

3.4 Akkumulationsstatistik

Bei diesem Verfahren müssen zunächst Gewichte definiert werden, die die Bedeutung der Items für die Fahrtüchtigkeit widerspiegeln. Dann wird für jeden Probanden die gewichtete Summe der Itemwerte gebildet. Schließlich wird ein Toleranzbereich festgelegt, wie beispielsweise eine Standardabwei-

5 Merkmale für je 13 Probanden in 2 Gruppen

- Berechnung des gemeinsamen Mittelwertes und der Standardabweichung für die 5 Merkmale

	Tracking	mittlere Entscheidungszeit	Tachistoskop	Daueraufmerksamkeit	periphere Wahrnehmung
$\bar{x}$	8.16	404.77	264.81	475.65	104.92
s	2.13	80.35	73.43	73.14	24.42

- Transformation der Originalwerte in Einheiten der Standardabweichung
- Berechnung der Summe dieser transformierten Werte pro Proband (unter Umständen gewichtet nach der Korrelationsstruktur)
- Berechnung T Test

Gruppe	N	$\bar{x}$	s
Methadon	13	0.2135	2.7358
Kontrolle	13	- 0.2135	2.0886

t = 0.4474 fg = 24 p = 0.6586

Tab. 4: Multiple Endpoints (Quelle: LEHMACHER (1991), O'BRIAN (1984), POCOCK (1987))

chung. Der Abschluß bildet der Test auf Äquivalenz in Form des t-Tests.

0-Hypothese ist hier, daß der Unterschied zwischen den Mittelwerten den Toleranzbereich übersteigt. Dementsprechend wird für Fahrtüchtigkeit entschieden, wenn der Test signifikant ausfällt. Die Tabelle 5 zeigt ein Beispiel für beliebig gewählte Gewichte. Im konkreten Fall muß die 0-Hypothese beibehalten werden, d.h. Methadonpatienten wären nicht fahrtüchtig, Äquivalenz liegt nicht vor.

4 Zusammenfassung

Wir halten die zuletzt vorgestellte Akkumulationsstatistik als eine optimale, für die Thematik adäquate Bewertung für mehrere simultan gemessene fahrrelevante Leistungen. Eine wesentliche Voraussetzung für ihren Einsatz ist allerdings, daß die Verkehrswissenschaften sich über Art, Umfang und Gewichtung der verschiedenen Komponenten der Leistungen, die für ein sicheres Fahren notwendig sind, einig sind. Im Rahmen einer internationalen Arbeitsgruppe der ICADTS werden Bemühungen in diese Richtung unternommen. Eine pragmatische Festlegung ausgewählter Merkmale mittels des Verfahrens der 'consensus conference' erscheint uns möglich und dringend.

- Definition von Gewichten entsprechend der Bedeutung der Merkmale für die Fahrtüchtigkeit z.B.

score = 0.41 x Tracking - 0.06 Entscheidungszeit
+ 0.52 Tachistoskop - 0.31 Daueraufmerksamkeit
- 0.67 periphere Wahrnehmung

Gruppe	N	$\bar{x}$	s
Methadon	13	- 91.46	39.04
Kontrolle	13	- 110.52	43.97

t-Test für Äquivalenzhypothese

akzeptable Abweichung: Δ zwischen den Mittelwerten beider Gruppen $\leq 1 \times s$

T = - 1.1689

kritischer Wert: 0.816

Tab. 5: Akkumulationsstatistik

5 Literatur

JONES, B., KENWARD, M.G. (1990): Design and analysis of crossover trials. Chapman and Hall, London

WESTLAKE, W.J. (1972): Use of confidende intervals in analysis of comparative bioavailability trials. J. Pharmac. Sci. 61, 1340-1341

WELLEK, S. (1994): Statistische Methoden zum Nachweis von Äquivalenz, Gustav Fischer, Stuttgart

VOLLMAR, J. ed. (1991): Biometrie in der chemisch-pharmazeutischen Industrie. Statistische Beurteilung der Bioäquivalenz sofort freisetzender Arzneiformen. Fischer, Stuttgart

BAUER, P. (1991): Multiple testing in clinical trials. Statistics in Medicine 10: 871-890

HOLM, S. (1979): A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70

HOMMEL, G. (1989): A comparison of two modified Bonferroni procedures. Biometrika 76: 624-625

SHAFFER, J.P. (1986): Modified sequentially rejective multiple test procedures. JASA 81: 826-831

SIMES, R.J. (1986): An improved Bonferroni procedure for multiple test of significance. Biometrica 73: 751-754

MORRISON, D.F. (1967): Multivariate Statistical Methods. McGraw-Hill, New York

LEHMACHER, W., WASSMER, G., REITMEIR, P. (1991): Procedures for two sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 47: 511-521

O'BRIAN, P.C. (1984): Procedures for comparing samples with multiple endpoints. Biometrics 40: 1079-1087

POCOCK, S.J., GELLER, N.L., TSIATIS, A.A. (1987): The analysis of multiple endpoints in clinical trials. Biometrics 43: 487-498

BERGHAUS, G., STAAK, M., GLAZINSKI, R., HÖHER, K., JOÓ, S., FRIEDEL, B. (1993): Complementary empirical study on the driver fitness of methadone substitution patients. In: Utzelmann H.-D., Berghaus, G., Kroj, G. (eds) Alcohol, Drugs and Traffic Safety T'92 Mensch - Fahrzeug - Umwelt, Bd. 29. TÜV Rheinland Verlag, Köln, S. 120-126

Anschriften der Verfasser

Dr. Ralf-Dieter Hilgers
Institut für Medizinische Statistik,
Informatik und Epidemiologie
der Universität zu Köln
Joseph-Stelzmann-Straße 9

D - 50931 Köln

Dr. Bernd Friedel
Bundesanstalt für Straßenwesen
Brüderstraße 53

D - 51427 Bergisch Gladbach

Prof. Dr. Günter Berghaus
Institut für Rechtsmedizin der
Universität zu Köln
Melatengürtel 60 - 62

D - 50823 Köln