ÜBERSICHTSARBEIT

Vom richtigen Umgang mit dem Crossover-Design in klinischen Studien

Teil 18 der Serie zur Bewertung wissenschaftlicher Publikationen

Stefan Wellek, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Viele klinische Studien werden nach dem sogenannten Crossover-(Überkreuzungs-)Design durchgeführt. Der wissenschaftliche Wert der Ergebnisse hängt entscheidend davon ab, dass bei der Planung und Auswertung gewisse Besonderheiten dieses Designs beachtet werden, die in standardmäßigen Parallelgruppen-Versuchen keine Rolle spielen.

Methoden: Darstellung der Grundprinzipien und der statistischen Methoden unter Bezugnahme auf statistische Lehrbücher und ausgewählte Originalliteratur.

Ergebnisse: Im einfachsten und häufigsten Fall werden in einem Crossover-Versuch zwei Behandlungen verglichen, die jedem rekrutierten Patienten zeitlich konsekutiv verabreicht werden, wobei die Reihenfolge der Verabreichung variiert wird. Hauptzweck des Designs ist es, sicherzustellen, dass Behandlungs- von Periodeneffekten sauber getrennt werden können. Hierzu müssen die Behandlungseffekte in beiden – per Randomisierung gebildeten – Sequenzgruppen separat berechnet werden. Der anschließende Test auf Behandlungsunterschiede lässt sich durchführen als unverbundener t-Test mit den intraindividuellen Differenzen zwischen den Ergebnissen aus beiden Versuchsperioden als den Einzelwerten. Voraussetzung ist dabei, dass keine sogenannte Carryover-(Überhang-)Effekte existieren, was üblicherweise in einem gesonderten Vorschalttest überprüft wird. Auf das Ersetzen des t-Tests durch nichtparametrische Tests sowie kompliziertere Designs mit mehr als zwei Versuchsperioden und/oder Behandlungen wird ebenfalls kurz eingegangen.

Schlussfolgerungen: Wenn bei der Auswertung von Crossover-Studien keine Auftrennung nach Sequenzgruppen erfolgt, sind die Ergebnisse verfälscht und von geringer wissenschaftlicher Aussagekraft. Eine weitere Voraussetzung für eine korrekte Auswertung solcher Studien ist, dass keine Überlagerung (Interaktion) von Behandlungs- mit Carryover-Effekten stattfindet. Falls sich die Annahme, dass solche Interaktionseffekte vernachlässigbar sind, nicht rechtfertigen lässt, muss sich die Evaluierung der Therapieeffekte auf eine Analyse der Daten aus der 1. Versuchsperiode beschränken. Allerdings ist auch dann die statistische Gültigkeit der Resultate nicht ohne weiteres gewährleistet.

▶ **Zitierweise**
Wellek S, Blettner M: On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2012; 109(15): 276–81. DOI: 10.3238/arztebl.2012.0276

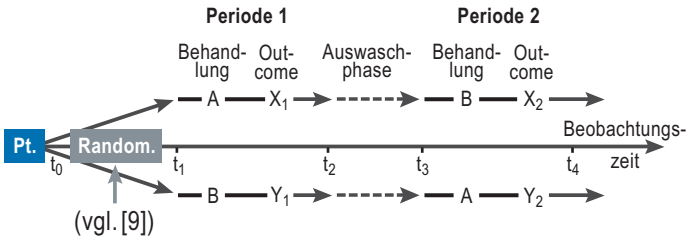
Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsklinikum Mainz: Prof. Dr. rer. nat. Wellek, Prof. Dr. rer. nat. Blettner

Das Crossover-Design hat als Prinzip der Planung wissenschaftlicher Versuche eine lange Geschichte (1, § 1.4) und bildet die Basis für eine große Zahl klinischer Studien, die alljährlich publiziert werden. Man findet dieses Design in fast allen klinischen Disziplinen, allerdings fällt eine deutliche Häufung in den „ZNS-Fächern“ Neurologie und Psychiatrie sowie auf dem Gebiet der Schmerztherapie auf. Ein Beispiel aus dem letzteren Bereich ist die häufig zitierte Studie zum Nachweis des analgetischen Effekts synthetischer Cannabinoide (2). Hierbei handelt es sich um einen klassischen Crossover-Versuch mit insgesamt 21 an chronischen neuropathischen Schmerzen leidenden Patienten. Jedem Patienten wurden in zwei konsekutiven Behandlungsperioden von je einer Woche Dauer täglich vier beziehungsweise acht äußerlich nicht unterscheidbare Kapseln verabreicht, die entweder Placebo oder Dimethylheptyl-THC-11-Carbonsäure (CT-3) enthielten. Als Hauptzielkriterium wurde die Änderung der Schmerzintensität am Ende jeder Behandlungsperiode auf einer visuellen Analogskala (VAS) erfasst.

Der wesentliche Unterschied zwischen einem Crossover- und einem herkömmlichen Parallelgruppen-Versuch liegt darin, dass jeder Proband beziehungsweise Patient als seine eigene Kontrolle dient. Somit stellt sich die Frage nach der Vergleichbarkeit von Versuchs- und Kontrollgruppe hinsichtlich konfundierender Variablen (wie zum Beispiel Lebensalter und Geschlecht) im Crossover-Design offensichtlich nicht. Vorteile bietet das Crossover-Design weiterhin hinsichtlich der Power des zur statistischen Absicherung eines Behandlungseffekts durchzuführenden Signifikanztests. Dies bedeutet, dass man unter gleich strikten Anforderungen an das Risiko eines Fehlers erster und zweiter Art mit geringeren Fallzahlen auskommt als in einem Parallelgruppen-Versuch.

Eine notwendige Voraussetzung dafür, dass diese Vorteile auch wirklich zum Tragen kommen, ist allerdings, dass bei der Planung und Auswertung einer solchen Studie einige spezifische Fallstricke vermieden werden. Bei der Planung muss garantiert sein, dass zwischen die beiden Behandlungsperioden, in denen der Patient die zu vergleichenden Behandlungen erhält, eine Auswaschphase geschaltet wird. Diese muss lang genug sein, damit keine Überhang-(„Carryover“)Effek-

GRAFIK



Schema eines Crossover-Versuchs: Pt. Patient; Random., Randomisation

te auftreten. Die Wirkung des ersten Medikamentes muss also vollständig abgebaut sein. Bei der Auswertung von Crossover-Studien wird häufig so verfahren, als handele es sich um einen einfachen Prä/Post-Vergleich. Dieses Vorgehen ist leider auch in angesehenen Zeitschriften immer wieder zu beobachten (3–8). Crossover-Studien, in denen die Auswertung mit dem verbundenen t-Test (oder einem anderen Verfahren für paarige Stichproben) vorgenommen worden ist, sind methodisch fehlerhaft und leisten keinen Beitrag zur evidenzbasierten Beurteilung der geprüften Behandlungen.

Leitfaden für die korrekte statistische Auswertung

Das formale Schema eines Crossover-Versuchs für den Vergleich von zwei Behandlungen A und B (im obigen Beispiel: A = Placebo, B = CT-3) ist in *Grafik 1* dargestellt. Die beiden Phasen, die der Patient während des Versuchs zu durchlaufen hat, werden üblicherweise als die beiden Versuchs-Perioden bezeichnet (10, S. 79). Die Wirksamkeit von A und B beurteilt man anhand der intraindividuellen Differenz zwischen den Werten, die man unter beiden Behandlungen für die Outcome-Variable erhalten hat. Der entscheidende Unterschied zwischen einem Crossover-Versuch und einer einfachen Studie, die zu Vergleichen zwischen verbundenen Stichproben (prä/post-Vergleich) führt, ist der folgende: Bei der Planung eines Crossover-Versuchs muss man davon ausgehen, dass es bei Patienten, die in Periode 1 Behandlung A und in Periode 2 Behandlung B erhalten (oder umgekehrt), aufgrund von Zeiteffekten auch dann systematische Unterschiede im Outcome geben kann, wenn A und B völlig identisch wirken (indem zum Beispiel beide Male dasselbe Medikament gegeben wird). Dieser Sachverhalt hat zur Konsequenz, dass bei der Planung und Analyse einer Crossover-Studie spezielle Maßnahmen erforderlich sind, um zu verhindern, dass es zu einer Vermengung (Confounding) (11, 12) zwischen Behandlungs- und Periodeneffekten kommt. Ein einfacher Grund dafür kann die Gewöhnung an die Studiensituation sein.

Hauptschritte für die konfirmatorische Daten-Analyse (Kasten 1 und 2)

Basis für die konfirmatorische Auswertung sind Vergleiche zwischen den Sequenzgruppen A-B und B-A, in die die Patienten zufällig eingeteilt worden sind.

- Die entscheidenden Messwerte für die Auswertung sind die intraindividuellen Differenzen zwischen den Outcome-Werten, die jeweils in den beiden Versuchsperioden gemessen werden. Für eine statistisch gültige Beurteilung der Behandlungseffekte ist ein unverbundener Test durchzuführen.
- Die Annahme, dass die Auswaschphase ausreichend lang angesetzt wurde, um Überhangeffekte auszuschließen, sollte in einem eigenen Vorschalttest überprüft werden. Hierzu werden die Summen der Messwerte aus beiden Perioden betrachtet und ein unverbundener Vergleich der Sequenzgruppen durchgeführt. Falls dieser Test zu einem statistisch signifikanten Ergebnis führt, besitzt der übliche Test auf Unterschiedlichkeit der Behandlungseffekte keine Aussagekraft.

Power- und Fallzahlberechnung, Effizienz

Wie für klinische Studien generell zu fordern ist (17), gehört auch zur Planung einer Crossover-Studie eine nachvollziehbare Fallzahlkalkulation, die von präzisen Vorgaben bezüglich der Trennschärfe (Power) des Tests der primär interessierenden Hypothese ausgeht. Im Falle des Crossover-Designs ist dies der Test auf Unterschiede zwischen den Behandlungseffekten. Bei der Planung wird generell vorausgesetzt, dass Carryover-Effekte aufgrund einer ausreichend langen Auswaschphase auszuschließen sind.

Power- und Fallzahlberechnung im Crossover-Design sind im Prinzip völlig identisch mit dem aus dem t-Test für unverbundene Stichproben bekannten Berechnungsverfahren (18). Der einzige Unterschied betrifft die Spezifikation der Annahmen, unter denen eine vorgegebene Power (zum Beispiel 80 %) erreicht werden soll *(Kasten 3a)*.

Eine wichtige Frage ist, ob das Crossover-Design im Vergleich zu einer herkömmlichen Zwei-Arm-Studie mit Messdaten aus nur einer Versuchsperiode effizienter ist. Gemeint ist dabei das Verhältnis der Stichprobenumfänge, die in beiden Designs benötigt werden, um unter sonst identischen Vorgaben und Bedingungen dieselbe Power zu erzielen.

Unter den üblichen statistischen Modellannahmen für die parametrische Analyse von Crossover-Studien (19) lässt sich die Frage mittels der in *Kasten 3b* dargestellten Näherungsbeziehung beantworten. Danach besitzt das Crossover-Design stets die höhere Effizienz. Da die Messfehler-Varianz im Allgemeinen eine geringere Größenordnung hat als die der interindividuellen Variabilität zuzuschreibende Varianzkomponente, ist der Unterschied sehr oft erheblich. Zum Beispiel benötigt man in einer Situation, in der die letztere doppelt so hoch ist wie die Messfehler-Varianz, rund sechsmal so viele Patienten, um im Parallelgruppen-Design dieselbe

KASTEN 1

Schritte für die konfirmatorische statistische Auswertung eines Crossover-Versuchs (1, § 2.3 ; 10, § 4.1)

Symbole:

- X_{1i} beziehungsweise X_{2i} = Messergebnis aus Periode 1 beziehungsweise 2 von Patient Nr. i aus Sequenzgruppe A-B
- Y_{1j} beziehungsweise Y_{2j} = Messergebnis aus Periode 1 beziehungsweise 2 von Patient Nr. j aus Sequenzgruppe B-A
- $C_i(X) = X_{1i} + X_{2i}$, $C_j(Y) = Y_{1j} + Y_{2j}$ [intraindividuelle Summen der Messergebnisse aus beiden Perioden]
- $D_i(X) = X_{1i} – X_{2i}$, $D_j(Y) = Y_{1j} – Y_{2j}$ [intraindividuelle Differenzen der Messergebnisse aus Periode 1 versus 2]
- m beziehungsweise n = Anzahl der Patienten in Sequenzgruppe A-B beziehungsweise B-A,
- N = m + n [Gesamtfallzahl]

Hinweis: Im Beispiel aus *Kasten 3* gilt:
m = 7, n = 6;
$X_{11} = 310$, $X_{21} = 270$, $C_1(X) = 310+270 = 580$, $D_1(X) = 310–270 = 40$;
$Y_{11} = 370$, $Y_{21} = 385$, $C_1(Y) = 370+385 = 755$, $D_1(Y) = 370–385 = –15$;
usw. für die übrigen Patienten.

1. Vorschalttest zur Überprüfung der Annahme zu vernachlässigender Carryover-Effekte
Wird durchgeführt wie ein „normaler" unverbundener t-Test (vgl. 13) mit $C_1(X)$, ..., $C_m(X)$ und $C_1(Y)$, ..., $C_n(Y)$ als den beiden Stichproben.
Die Prüfgröße hat man also zu berechnen nach der Formel

$$T = \sqrt{\frac{mn}{N}} \frac{\bar{C}(X) - \bar{C}(Y)}{\sqrt{(SQ_{CX} + SQ_{CY}) / (N-2)}},$$

mit $\bar{C}(X) = (C_1(X) + \ldots + C_m(X)) / m$,
$SQ_{CX} = (C_1(X) - \bar{C}(X))^2 + \ldots + (C_m(X) - \bar{C}(X))^2$
und analoger Berechnungsweise von $\bar{C}(Y)$ bzw. SQ_{CY}.

Der (2-seitige) p-Wert (vgl. 14) bestimmt sich dann wie immer im unverbundenen t-Test, nämlich als die Wahrscheinlichkeit, dass der Absolutbetrag einer (zentral) t-verteilten Größe mit N-2 Freiheitsgraden den errechneten absoluten Wert der Prüfgröße T überschreitet.

2. Test auf Unterschiedlichkeit der Behandlungseffekte
Der Test wird formal nach genau demselben Berechnungsschema durchgeführt wie der erste. Der einzige, inhaltlich allerdings entscheidende Unterschied besteht darin, dass die üblichen Formeln für den unverbundenen t-Test jetzt anzuwenden sind auf die intraindividuellen Differenzen $D_1(X)$, . . . , $D_m(X)$ und $D_1(Y)$, . . . , $D_n(Y)$.

Power zu garantieren wie im Crossover. Zu beachten ist aber, dass sich dieser Gewinn unter Kosteneffizienz-Gesichtspunkten dadurch reduziert, dass in einer Crossover-Studie bei gleicher Patientenzahl die doppelte Anzahl von Messungen durchzuführen ist. Außerdem erhöht sich der zeitliche Aufwand für die Durchführung aufgrund der Tatsache, dass jeder Patient zwei Versuchsperioden mit dazwischen geschalteter Auswaschphase zu durchlaufen hat.

Modifikationen und Verallgemeinerungen

Die oben beschriebenen konfirmatorischen Verfahren auf der Basis von unverbundenen t-Statistiken setzen voraus, dass die Messwerte (annähernd) normalverteilt sind. Nicht selten ist nur die schwächere Modellannahme realistisch, dass die zugehörigen Variablen eine stetige Verteilung von gemeinsamer, aber unbekannter Form besitzen, deren Mediane sich aus dem jeweiligen Behandlungs-, Perioden- und einem etwaigen Carryover-Effekt additiv zusammensetzen. Eine konfirmatorische Auswertungsstrategie, die auch unter diesen schwächeren Voraussetzungen gültige Ergebnisse liefert, besteht darin, dass jeweils anstelle eines unverbundenen t-Tests ein Wilcoxon-Rangsummen-Test durchgeführt wird (20). Für den Vorschalttest auf Vernachlässigbarkeit der Carryover-Effekte wird also mit den intraindividuellen Messwertsummen $C_1(X)$, ..., $C_m(X)$, $C_1(Y)$, ..., $C_n(Y)$ die Wilcoxon-Teststatistik berechnet (wie zum Beispiel in [13] beschrieben), und analog für den Test auf Unterschiedlichkeit der Behandlungseffekte.

KASTEN 2

Beispiel für die konfirmatorische statistische Auswertung eines Crossover-Versuchs (15, 16)

Studie:
Vergleich der bronchodilatatorischen Wirkung von inhaliertem Formoterol (A) und Salbutamol (B) auf den Peak Expiratory Flow (PEF) von Kindern mit Asthma bronchiale.

Daten:

Sequenzgruppe A-B

Pt.-Nr. (i)	X_{1i}	X_{2i}	$C_i(X)$	$D_i(X)$
1	310	270	580	40
2	310	260	570	50
3	370	300	670	70
4	410	390	800	20
5	250	210	460	40
6	380	350	730	30
7	330	365	695	–35

Für Tests benötigte arithmetische Mittel und Summen von Abweichungsquadraten:
$\bar{C}(X) = 643.57$, $\bar{D}(X) = 30.71$; $SQ_{CX} = 78435.71$, $SQ_{DX} = 6521.43$.

Sequenzgruppe B-A

Pt.-Nr. (j)	Y_{1j}	Y_{2j}	$C_j(Y)$	$D_j(Y)$
1	370	385	755	–15
2	310	400	710	–90
3	380	410	790	–30
4	290	320	610	–30
5	260	340	600	–80
6	90	220	310	–130

Für Tests benötigte arithmetische Mittel und Summen von Abweichungsquadraten:
$\bar{C}(Y) = 629.17$, $\bar{D}(Y) = -62.50$; $SQ_{CY} = 151320.83$, $SQ_{DY} = 9987.50$.

1. Vorschalttest zur Überprüfung der Annahme zu vernachlässigender Carryover-Effekte:

$$\text{Prüfgröße: } T = \sqrt{\frac{7 \times 6}{13}} \frac{643.57 - 629.17}{\sqrt{(78435.71 + 151320.83) / 11}} = 0.1791;$$

p-Wert: $p = 0.8611$.

2. Test auf Unterschiedlichkeit der Behandlungseffekte:

$$\text{Prüfgröße: } T = \sqrt{\frac{7 \times 6}{13}} \frac{30.71 - (-62.50)}{\sqrt{(6521.43 + 9987.50) / 11}} = 4.3247;$$

p-Wert: $p = 0.0012$.

3. Signifikanzentscheidungen: Signifikante Verbesserung des PEF unter Formoterol (A) im Vergleich zu Salbutamol (B); kein Hinweis auf relevante Carryover-Effekte.

KASTEN 3a

Für die Bestimmung der Effektstärke bei der Fallzahlplanung einer Crossover-Studie festzulegende Größen

1. Erwartete Differenz τ zwischen A und B bezüglich des Outcome-Maßes, unter Absehung von Perioden-effekten
2. Messmethodische Varianz σ_e^2, mit der zu rechnen wäre, wenn beim selben Patienten die Bestimmung des Outcome-Maßes unter identischen Bedingungen (gleiche Versuchsperiode und gleiche Behandlung) sehr oft wiederholt würde.
3. Die Effektstärke, die in die Formeln für Power- und Fallzahlen im unverbundenen t-Test einzusetzen ist, beträgt

$$(\mu_1 - \mu_2) / \sigma = \sqrt{2}\tau / \sigma_e$$

KASTEN 3b

Umrechnungsfaktor für die Effizienz des Crossover – relativ zum Parallelgruppen-Design

$$\frac{\sigma_e^2 + \sigma_s^2}{0.5 \times \sigma_e^2}$$

wobei σ_s^2 die interindividuelle (Englisch: between-subject variance) und σ_e^2 die intraindividuelle, messmethodische Varianz (Englisch: within-subject variance) bezeichnet.

Auf einer wesentlich anderen Ebene liegt eine Modifikation des Tests zum Vergleich der Behandlungseffekte, die in Zusammenhang mit Studien zum Nachweis der Bioäquivalenz zweier Formulierungen des gleichen Arzneimittels sehr häufig zur Anwendung gelangt. Dieser Test folgt einer grundsätzlich veränderten „statistischen Logik“, da die Alternativhypothese, die man anhand der Messdaten aus der Studie bestätigen will, im Falle des Bioäquivalenznachweises aussagt, dass es zwischen den beiden Behandlungen (Arzneimittelformulierungen) A und B keine wesentlichen Unterschiede gibt. Für eine Darstellung von Grundprinzipien und wichtigen speziellen Verfahren für das Testen auf Äquivalenz verweisen die Autoren neben der Originalliteratur (21) auf eine spätere Folge der Serie zur Bewertung wissenschaftlicher Publikationen.

Eine weitere wichtige, wenn auch in medizinischen Anwendungen vergleichsweise selten zu findende Modifikation betrifft die Ausdehnung des Versuchs auf mehr als zwei Messperioden. In einem solchen Mehrperioden-Crossover-Design braucht die Anzahl von Perioden nicht mit derjenigen von zu vergleichenden Behandlungen überein zu stimmen. Zum Beispiel wird für Bioäquivalenzstudien alternativ zum herkömmlichen Design mit zwei Perioden ein repliziertes Crossover-Design mit vier Perioden empfohlen, wobei sowohl A als auch B je zweimal wiederholt werden (22). Die Analyse von Mehrperioden-Crossover-Studien ist im Allgemeinen vergleichsweise kompliziert und erfordert spezielle Software für lineare Regressionsmodelle mit gemischten Effekten (1).

Diskussion

Das Crossover-Design ist als Versuchsschema für klinische und auch experimentelle Studien unverändert sehr populär, und in einer nicht unbeträchtlichen Zahl von Publikationen erscheint der Begriff bereits im Titel. Bei einem bedenklich hohen Anteil solcher Publikationen wird der Leser aber feststellen, dass die in dieser Arbeit dargestellten Anforderungen an eine statistisch sachgerechte Auswertung der Ergebnisse in keiner Weise erfüllt sind. Der häufigste Fehler besteht darin, dass die Aufgliederung in Sequenzgruppen unberücksichtigt bleibt, indem die Auswertung genau wie in einer Studie mit fester Behandlungsreihenfolge anhand eines verbundenen t- oder Wilcoxon-Tests vorgenommen wird. Ein solches Vorgehen stellt die Gültigkeit der Resultate einer Crossover-Studie grundsätzlich in Frage: Im Extremfall zeigt ein signifikantes Ergebnis dann lediglich an, dass es einen ausgeprägten Periodeneffekt gab, während die Wirksamkeit der Behandlungen als solche praktisch identisch war.

Ein weiterer Fallstrick, der in Zusammenhang mit Crossover-Studien unbedingt zu beachten ist, muss bereits in der Planungsphase abgefangen werden. Entscheidend ist hier, dass die zwischen die Behandlungsperioden der Studie einzuschiebende Auswaschphase von der zeitlichen Ausdehnung her ausreicht, um sicherzustellen, dass es keine in die nächste Periode hineinwirkenden Überhangeffekte einer Behandlung geben kann. Der Vorschalttest, durch den dies bei der nachmaligen konfirmatorischen Analyse der Studiendaten zu überprüfen ist, hat im Wesentlichen die Funktion, ein entsprechendes Defizit bei der Versuchsplanung aufzudecken. Die Frage, wie man verfahren sollte, wenn dieser Vorschalttest ein signifikantes Ergebnis liefert, ist anhand der statistischen Originalliteratur nicht abschließend zu beantworten: Lange Zeit konnte es als etablierte biometrische Praxis gelten, im Zweiperioden-Crossover nach einem signifikanten Test auf Carryover-Effekte die Studie mit den Daten aus der ersten Versuchsperiode wie einen gewöhnlichen Parallelgruppen-Versuch auszuwerten. In Routineanwendungen ist diese Vorgehensweise nach wie vor üblich, obwohl schon vor über 20 Jahren gezeigt worden ist, dass der zugehörige „nachgeschaltete“ unverbundene t-Test nicht mehr die gewohnten Eigenschaften hat und unter Umständen das Signifikanzniveau deutlich überschreitet, also antikonservativ werden kann (23).

Interessenkonflikt
Beide Autoren erklären, dass kein Interessenkonflikt besteht.

Manuskriptdaten
eingereicht: 12. 7. 2011, revidierte Fassung angenommen: 10. 11. 2011

LITERATUR

1. Jones B, Kenward MG: Design and analysis of cross-over trials. 2nd edition. Boca Raton: Chapman & Hall/CRC 2003.
2. Karst M, Salim K, Burstein S, Conrad I, Hoy L, Schneider U: Analgesic effect of the synthetic cannabinoid CT-3 on chronic neuropathic pain. A randomized controlled trial. JAMA 2003; 290: 1757–62.
3. Ganesan A, Crum-Cianflone N, Higgins J, et al.: High dose atorvastatin decreases cellular markers of immune activation without affecting HIV-1 RNA levels: results of a double-blind randomized placebo controlled clinical trial. J Infect Dis 2011; 203: 756–64.
4. Davis AR, Westhoff CL, Stanczyk FZ: Carbamazepine coadministration with an oral contraceptive: effects on steroid pharmacokinetics, ovulation, and bleeding. Epilepsia 2011; 52: 243–7.
5. Black KJ, Koller JM, Campbell MC, Gusnard DA, Bandak SI: Quantification of indirect pathway inhibition by the adenosine A2a antagonist SYN115 in Parkinson disease. J Neurosci 2010; 30: 16284–92.
6. Mellor DD, Sathyapalan T, Kilpatrick ES, Beckett S, Atkin SL: High-cocoa polyphenol-rich chocolate improves HDL cholesterol in Type 2 diabetes patients. Diabet Med 2010; 27: 1318–21.
7. Chung KA, Lobb BM, Nutt JG, Horak FB: Effects of a central cholinesterase inhibitor on reducing falls in Parkinson disease. Neurology 2010; 75: 1263–9.
8. Page TH, Turner JJ, Brown AC, et al.: Nonsteroidal anti-inflammatory drugs increase TNF production in rheumatoid synovial membrane cultures and whole blood. J Immunol 2010; 185: 3694–701.
9. Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M: Randomized controlled trials: part 17 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(39): 663–8.
10. Lehmacher W: Verlaufskurven und Crossover. Statistische Analyse von Verlaufskurven im Zwei-Stichproben-Vergleich und von Cross-over-Versuchen. In: Überla K, Reichertz PL, Victor N (eds.): Medizinische Informatik und Statistik, Vol 67. Berlin: Springer 1987.
11. Ressing M, Blettner M, Klug SJ: Data analysis of epidemiological studies: part 11 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(11): 187–92.
12. Sauerbrei W, Blettner M: Interpreting results in 2 x 2 tables: extensions and problems: part 9 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(48): 795–800.
13. du Prel JB, Röhrig B, Hommel G, Blettner M: Choosing statistical tests—part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(19): 343–8.
14. du Prel JB, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(19): 335–9.
15. Graff-Lonnevig V, Browaldh L: Twelve hours bronchodilating effect of inhaled formoterol in children with asthma: a double-blind cross-over study versus salbutamol. Clin Exp Allergy 1990; 20: 429–32.
16. Senn S: Crossover designs. In: Armitage P, Colton T (eds.): Encyclopedia of biostatistics, Volume 2 . Chichester: John Wiley & Sons 1998: 1033–49.
17. du Prel JB, Röhrig B, Blettner M: Critical appraisal of scientific articles—part 1 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(7): 100–5
18. Röhrig B, Prel JB du, Wachtlin D, Kwiecien R, Blettner M: Sample size calculation in clinical trials—part 13 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(31–32): 552–6.
19. Grizzle JE: The two-period change-over design and its use in clinical trials. Biometrics 1965; 21: 467–80.
20. Koch GG: The use of non-parametric methods in the statistical analysis of the two-period changeover design. Biometrics 1972; 28: 577–84.
21. Wellek S: Testing statistical hypotheses of equivalence and noninferiority. 2nd edition. Boca Raton: Chapman & Hall/CRC 2010.
22. Food and Drug Administration (FDA): Guidance for industry: Statistical approaches to establishing bioequivalence. Rockville, MD: Center for Drug Evaluation and Research (CDER) 2001.
23. Freeman P: The performance of the two-stage analysis of two treatment, two period crossover trials. Statistics in Medicine 1989; 8: 1421–32.

Anschrift für die Verfasser
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie
Epidemiologie u. Informatik der
Johannes Gutenberg-Universität
Obere Zahlbacher Straße 69
55131 Mainz
blettner@imbei.uni-mainz.de

SUMMARY

On the Proper Use of the Crossover Design in Clinical Trials: Part 18 of a Series on Evaluation of Scientific Publications

Background: Many clinical trials have a crossover design. Certain considerations that are relevant to the crossover design, but play no role in standard parallel-group trials, must receive adequate attention in trial planning and data analysis for the results to be of scientific value.

Methods: The authors present the basic statistical methods required for the analysis of crossover trials, referring to standard statistical texts.

Results: In the simplest and most common scenario, a crossover trial involves two treatments which are consecutively administered in each patient recruited in the study. The main purpose served by the design is to provide a basis for separating treatment effects from period effects. This is achieved via computing the treatment effects separately in two sequence groups formed via randomization. The differences between treatment effects can be assessed by means of a standard t-test for independent samples using the intra-individual differences between the outcomes in both periods as the raw data. The existence of carryover effects must be ruled out for this method to be valid. This assumption is usually checked using a pre-test, which is also described in this article. Finally, we briefly discuss the use of nonparametric tests instead of t-tests and more complicated designs with more than two test periods and/or treatments.

Conclusion: Crossover trials in which the results are not analyzed separately by sequence group are of limited, if any, scientific value. It is also essential to guard against carryover effects. Whenever ignoring such effects proves unjustified, the treatment effect must be analyzed solely via an analysis of the data obtained during the first trial period. Even the use of this restricted dataset yields results whose validity is not beyond question.

Zitierweise
Wellek S, Blettner M: On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2012; 109(15): 276–81. DOI: 10.3238/arztebl.2012.0276

The English version of this article is available online:
www.aerzteblatt-international.de