Prof. Dr. Rainer Koch Juli 1997, Februar 2008

Hinweise zur biometrischen Planung der notwendigen Anzahl Patienten in klinischen Studien oder Anzahl Versuchstiere im Fall a) des Vergleichs von prozentualen Häufigkeiten und b) des Vergleichs von Mittelwerten

Beispiele für eine Zielstellung einer klinischen Studie bzw. des Tierversuches:
Nachweis des Unterschiedes zwischen Behandlungsgruppe und Kontrollgruppe

a) bezüglich des Anteils Patienten bzw. Tiere mit einer bestimmten Reaktion (z.B. „Anteil Responder“, wobei Response sowohl ein erwünschtes als auch ein unerwünschtes Ereignis sein kann)
b) bezüglich des Erwartungswertes (= Mittelwert in der Population) eines quantitativen Merkmals.

Zur Zielstellung biometrischer Verfahren und zu Fragen der Ethik im Zusammenhang mit der Fallzahl:

Ziel biometrischer Verfahren ist die Verallgemeinerung der empirischen Beobachtungsergebnisse auf eine Population, d.h. der Schluß von der Stichprobe (den konkret untersuchten Patienten bzw. Tieren) auf eine (fiktive) Gesamtheit (z.B. aller prinzipiell möglichen zukünftigen Patienten bzw. Tiere) unter kontrollierten Fehlentscheidungsrisiken bzw. inferentiellen Gütekriterien.
Mit anderen Worten: Ziel ist die Abgrenzung des Versuchsergebnisses vom Zufall.
Hauptsächlich verwendet man statistische Tests (zur Prüfung von Hypothesen, z.B. für inter- und intraindividuelle Vergleiche) und statistische Schätzungen (für Effektmaße, z.B. relative Wirksamkeit, Anteil Responder usw.).

Experimentelle Studien ohne biometrische Planung sind möglich, z.B. wenn

- Vorversuche an einzelnen Patienten bzw. Tieren durchgeführt werden sollen, um neue Methoden oder Meßprinzipien zu erproben, ohne auf eine Population schließen zu wollen,
- im Rahmen eines Pilotversuches für die biometrische Planung eines Folgeversuches die erreichbaren Effekte und die Variabilität der Zielmerkmale gefunden werden sollen
- prinzipiell neue Forschungshypothesen zu entwickeln sind, die in einem Folgeversuch überprüft werden sollen.

In diesem Fall muß die notwendige Fallzahl anders begründet werden, und konfirmatorische (=„beweisende“) Signifikanzaussagen müssen vermieden werden.

Fehlschlußrisiken beim Schluß von den empirischen Versuchsergebnissen auf die Population („inferentielle Gütemaße“) im Fall, daß ein Unterschied zwischen Behandlungs- und Kontrollgruppe nachzuweisen ist (im Gegensatz dazu sind zum Nachweis der Äquivalenz Vorgehens- und Sprechweisen etwas anders):

α Risiko, fälschlich Unterschiede zu behaupten (= Signifikanzniveau)
gesellschaftlicher Konsens:
$\alpha = 0{,}05$ [am schwächsten] oder $\alpha = 0{,}01$ oder $\alpha = 0{,}001$ [am schärfsten]

β Risiko, wirkliche Unterschiede in der Studie nicht nachweisen zu können

$1-\beta$ statistische Power = Chance, mit der Studie den vermuteten Effekt nachweisen zu können (d.h. vom Zufall abgrenzen zu können)

Eine experimentelle Studie ist nur dann ethisch vertretbar,
wenn

a) im Fall statistischer Tests die statistische Power zur Verifizierung einer Forschungshypothese mindestens 80% beträgt (ein gesellschaftlicher Konsens);
b) im Fall der Schätzung von Zielkriterien die (fachwissenschaftlich zu begründende) maximal tolerierbare Unschärfe (= Breite des Konfidenzintervalls) nicht überschritten wird.

Andernfalls ist das Risiko unschlüssiger Ergebnisse größer als akzeptabel, und die klinische Studie bzw. der Tierversuch ist nicht ohne weiteres genehmigungsfähig.

Zur Planung der Anzahl notwendiger Patienten bzw. Tiere

(Beispielszenarien für den Vergleich und die Schätzung von Wahrscheinlichkeiten und Mittelwerten)

A Beispielszenarien für den Vergleich von Wahrscheinlichkeiten:

Der Vergleich von Wahrscheinlichkeiten für ein bestimmtes Zielereignis (z.B. Reaktion auf eine Noxe) zwischen Testbehandlung und Kontrollbehandlung wird anhand der beobachteten Prozentzahlen (= Schätzungen für die Wahrscheinlichkeiten in der Population) vorgenommen. Üblich ist z.B. der Nachweis der Überlegenheit einer der beiden Gruppen (Achtung: Äquivalenz ist auf diese Weise z.B. durch Nonsignifikanz nicht nachweisbar!)

- Vergleich einer Behandlungsgruppe mit einer Kontrollgruppe in einem Hauptzielkriterium,
- Inferentielle Güte: $\alpha = 0{,}05$ bei einseitiger Prüfung; Power mindestens 80%,
- Nachzuweisender Unterschied im Anteil Responder mindestens δ Prozentpunkte, wobei zusätzlich noch diejenige Prozentzahl der beiden Gruppen vorab geschätzt und angegeben werden muß, die näher an 50% liegt (falls das nicht möglich ist, nimmt man den ungünstigsten Fall 50% an). Es muß glaubhaft versichert werden, daß dieser im Experiment nachzuweisende Mindestunterschied δ erwartet wird und zur Beantwortung des Forschungsproblems ausreichend ist, d.h. die Begründung muß fachwissenschaftlich erfolgen (z.B. Literaturangaben, Voruntersuchungen)!

	δ = 80%Pkte.	δ = 40%Pkte.		δ = 20%Pkte.		δ = 10%Pkte.	
	Anteil Responder						
Gruppe 1	maximal 5%	maximal 5%	maximal 30%	maximal 5%	maximal 40%	maximal 5%	maximal 45%
Gruppe 2	mindestens 85%	mindestens 45%	mindestens 70%	mindestens 25%	mindestens 60%	mindestens 15%	mindestens 55%
Anzahl notwendiger Patienten bzw. Tiere pro Gruppe	4	14	19	39	77	111	309

Falls mehrere Gruppen zu vergleichen sind und/ oder mehrere Zielkriterien gleichzeitig und gleichgewichtig benutzt werden, erhöht sich die Fallzahl durch die notwendigen Testadjustierungen beträchtlich. Deshalb sollte man die Anzahl der statistischen Tests möglichst beschränken.

Sonderfall „100%“:
Es sei nachzuweisen, daß das betrachtete Zielereignis immer oder überhaupt nicht auftritt.
(Biometrisch: Vergleich einer Wahrscheinlichkeit gegen die Konstante 1 oder 0).
In diesem Fall erfolgt der Nachweis der statistischen Evidenz für die Forschungshypothese durch Angabe des Konfidenzintervalles, das dann genügend eng um die 100% (bzw. 0%) liegen muß.

Beispiele für den Nachweis „100%“
(Versuchsergebnis sei, daß alle untersuchten Patienten bzw. Tiere das betrachtete Zielereignis haben):

	Untere Grenze des 95%-Konfidenzintervalles			
	90%	95%	99%	99,5%
Anzahl notwendiger Patienten bzw. Tiere	29	59	297	592

Schreibweise für das Versuchsergebnis dann z.B.:
„Mit einer Irrtumswahrscheinlichkeit von 5% beträgt die Wahrscheinlichkeit für das Zielereignis mindestens 90% (bzw. wie in der Tabelle als untere Grenze des Konfidenzintervalles angegeben)“.

B Beispielszenarien für die Schätzung von Wahrscheinlichkeiten:

Die Schätzung für eine gesuchte Populationswahrscheinlichkeit erfolgt mit Hilfe des Prozentanteiles der Patienten bzw. Tiere, die das Zielereignis aufweisen, ergänzt um den statistischen „Unschärfebereich“.
(„Unschärfebereich“ = Konfidenzintervall = Intervall, das die gesuchte Populationswahrscheinlichkeit mit einer statistischen Sicherheit von z.B. 95% überdeckt)

Die maximal tolerierbare Breite δ des „Unschärfebereiches“ muß fachwissenschaftlich begründet werden! Aus ihr ergibt sich die notwendige Fallzahl.

Beispiele:

	Breite des 95%-Konfidenzintervalles			
	δ = 40%Pkte.	δ = 20%Pkte.	δ = 10%Pkte.	δ = 6%Pkte.
	1. Zeile: erwünschtes 95%-Konfidenzintervall 2. Zeile: Anzahl notwendiger Patienten bzw. Tiere			
Beispiel 1: 10% Patienten bzw. Tiere mit Zielereignis erwartet	1% ... 41% n = 10	4% ... 24% n = 54	6% ... 16% n = 153	7% ... 13% n = 442
Beispiel 2: 50% Patienten bzw. Tiere mit Zielereignis erwartet	30% ... 70% n = 20	40% ... 60% n = 82	45% ... 55% n = 290	47% ... 53% n = 784

C Beispielszenarien für den Vergleich von Mittelwerten:

Der Vergleich von Erwartungswerten eines quantitativen Merkmals zwischen verschiedenen Behandlungsgruppen wird anhand der beobachteten Mittelwerte (= Schätzungen für die Erwartungswerte in der Population) vorgenommen. Üblich ist z.B. der Nachweis der Überlegenheit einer von mehreren konkurrierenden Testbehandlungen gegenüber der Kontrollbehandlung (Achtung: Äquivalenz ist auf diese Weise z.B. durch Nonsignifikanz nicht nachweisbar!).

- Vergleich mehrerer Behandlungsgruppen (k Gruppen) mit einer Kontrollgruppe in einem Hauptzielkriterium (DUNNETT-Test),
- Inferentielle Güte: $\alpha = 0{,}05$ bei zweiseitiger Prüfung; Power mindestens 80%,
- Vorab geschätzte Variabilität des Hauptzielmerkmals: Standardabweichung = s (z.B. aus Literaturangaben oder Vorversuchen; falls total unbekannt, muß der ungünstigste Fall angenommen werden oder ein Pilotversuch durchgeführt werden oder auf die Angabe des relativen anstelle des absoluten Mindesteffektes ausgewichen werden; siehe nächsten Punkt).
- Nachzuweisender Unterschied im Erwartungswert absolut (d.h. Mindesteffekt in der Maßeinheit des Hauptzielmerkmals) mindestens δ. (Für die Planung der Fallzahl genügt der relative Mindesteffekt δ/s). Es muß glaubhaft versichert werden, daß dieser Effekt erwartet und im Sinne des Studienzieles als ausreichend angesehen wird, d.h. die Begründung muß fachwissenschaftlich erfolgen.

Beispiele (angenäherte Normalverteilung und gleiche Varianzen vorausgesetzt):

Anzahl k Behandlungs-gruppen	Anzahl notwendiger Patienten bzw. Tiere		
	relativer Mindesteffekt 80%	relativer Mindesteffekt 100%	relativer Mindesteffekt 120%
1	in Behandlungs- und Kontrollgruppe je 26, zusammen 52	in Behandlungs- und Kontrollgruppe je 17, zusammen 34	in Behandlungs- und Kontrollgruppe je 12, zusammen 24
2	in jeder Behandlungsgruppe 26, in der Kontrollgruppe 37, zusammen 89	in jeder Behandlungsgruppe 17, in der Kontrollgruppe 24, zusammen 58	in jeder Behandlungsgruppe 12, in der Kontrollgruppe 17, zusammen 41
3	in jeder Behandlungsgruppe 27, in der Kontrollgruppe 43, zusammen 124	in jeder Behandlungsgruppe 17, in der Kontrollgruppe 29, zusammen 80	in jeder Behandlungsgruppe 12, in der Kontrollgruppe 21, zusammen 57
4	in jeder Behandlungsgruppe 27, in der Kontrollgruppe 50, zusammen 158	in jeder Behandlungsgruppe 17, in der Kontrollgruppe 34, zusammen 102	in jeder Behandlungsgruppe 12, in der Kontrollgruppe 24, zusammen 72

D Beispielszenarien für die Schätzung der Differenz zweier Mittelwerte

Die Schätzung für die Differenz zweier Mittelwerte ist eine Maßzahl für den Unterschied beider zu vergleichender Gruppen bezüglich des Erwartungswertes und wird als Hauptzielkriterium z.B. für die Überlegenheit einer Testbehandlung gegenüber einer Kontrollbehandlung bei einem quantitativen Merkmal benutzt. Diese Schätzung muß durch Angabe des statistischen „Unschärfebereiches" ergänzt werden, um die Validität dieser Maßzahl einschätzbar zu machen.
(„Unschärfebereich" = Konfidenzintervall = Intervall, das die gesuchte Differenz mit einer statistischen Sicherheit von z.B. 95% überdeckt)

Die maximal tolerierbare Breite δ des „Unschärfebereiches" muß fachwissenschaftlich begründet werden. Aus dem Verhältnis δ/s dieser Breite δ zur Variabilität s des Merkmals (s = vorab zu schätzende Standardabweichung) ergibt sich die notwendige Fallzahl.

Beispiele (angenäherte Normalverteilung und gleiche Varianzen vorausgesetzt):

	relative Breite des 95%-Konfidenzintervalles				
	δ/s =60%	δ/s =80%	δ/s =100%	δ/s =120%	δ/s =140%
notwendige Anzahl Patienten bzw.Tiere	88	51	33	24	18

Literatur z.B.:

MACHIN/ CAMPBELL 1987: Statistical Tables for the Design of Clinical Trials. Blackwell Publ.
RASCH u.a. 1996: Verfahrensbibliothek Versuchsplanung. Oldenbourg Verlag
GÄBLER 1996: Biometrische Methodik pharmakologischer Tierexperimente. G.Fischer Verlag