Number-Needed-to-Treat, Berechnung von Varianz und Punktschätzern sowie Eigenschaften und Darstellung einer nicht unproblematischen Relevanz-Kenngröße

Dietrich Knoerzer
Martin Burke
Dieter Hauschke
ALTANA Pharma
Byk-Gulden-Str.2
Konstanz
Dietrich.knoerzer@altanapharma.com

Zusammenfassung

Die Kenngröße Number-Needed-to-Treat (NNT) beschreibt die Anzahl der Patienten, die einer gegebenen Behandlung zu unterziehen sind, bis sich ein Patient verändert (idealerweise verbessert). Mathematisch lässt sie sich als Reziprok aus der Variable 'Absolute Risk Reduction' (ARR) herleiten. Da die Zielgröße NNT anders als das ARR viel suggestiver interpretiert werden kann, erfreut sie sich in der Literatur zunehmender Beliebtheit. Für die Berechnung von Punkt- und Varianzschätzern für ARR bzw. NNT muß die Orginalskalierung der Variable und das Studiendesign berücksichtigt werden. Bei klassifizierten Variablen und Parallelgruppendesign stehen zwei Berechnungsvarianten zur Verfügung.
Durch Transformation auf die NNT-Skala entstehen statistisch unverwünschte Eigenschaften, die ggf. eine Interpretation dann stark erschweren können. In diesem Beitrag werden die Berechnungsvarianten und die Eigenschaften zusammengestellt und Darstellungsvorschläge mittels SAS gegeben.

Schlüsselworte: Number-Needed-to-Treat, Absolute Risk Reduction, Studiendesign, Skalierung, Punktschätzer, Varianzschätzer

1 Einleitung

Beim Vergleich zweier Gruppen hinsichtlich einer binären Variable gibt es verschiedene beschreibende Maße, z. B. Odds Ratio, Absolute Risk Reduction (ARR) sowie Number-Needed-to-Treat (NNT). Die folgenden Ausführungen beziehen sich auf Beispiele aus der klinischen Forschung, die Aussagen gelten analog aber auch für andere Forschungsgebiete. Die Kenngröße NNT beschreibt im Bereich der klinischen Forschung die Anzahl der Patienten, die einer gegebenen Behandlung zu unterziehen

sind, bis sich ein Patient verändert (idealerweise verbessert). Mathematisch lässt sie sich als Reziprok aus der Variable ARR herleiten. Die Zielgröße NNT ist für den Fachwissenschaftler, hier den Mediziner, intuitiv verständlicher als OR und ARR, sie kann zudem sehr suggestiv interpretiert werden, und erfreut sich deshalb in der medizinischen Literatur zunehmender Beliebtheit.
Als Folge der zunehmenden Beliebtheit wurde dieses Maß auch auf stetige Variablen übertragen (bedingt vorherige Klassifikation) und wird bei sehr unterschiedlichen Studiendesigns eingesetzt, wie z. B. Parallelgruppen- und Cross-over Studien.
Ziel dieses Beitrages ist nicht eine vertiefende Untersuchung der statistischen Eigenschaften (siehe z. B. [1],[8]), er soll vielmehr eine Art praktischer Hilfestellung im täglichen Gebrauch der angewandten Forschung liefern.

2 Berechnung des Punktschätzers

2.1 Allgemeines

Die Berechnung des Punktschätzers (PE) hängt ab von (i) der Skalierung der Variablen und (ii) dem Studiendesign. Daraus ergeben sich 4 Fälle:

		Studien Design	
		Cross-over	**Parallelgruppe**
Skalierung	**Binär**	Kap. 2.2	Kap. 2.3
	Stetig und klassifziert	Kap. 2.4	Kap. 2.5

Der Punktschätzer für ARR und damit auch der NNT kann numerisch positiv oder negativ sein, letzteres wenn der Effekt in der Behandlungsgruppe kleiner als in der Kontrollgruppe ist.
Für die Interpretation bei einem negativen NNT Punktschätzer hat [1] eine genauere Spezifikation vorgeschlagen. Eine Unterscheidung in NNTB (Number-needed-to-treat for one patient to benefit, d.h. die mittlere Untersuchungseinheit weist unter Behandlung ein verbessertes Ergebnis auf) bei positivem Punktschätzer und NNTH (Number-needed-to-treat for on patient to be harmed, d.h. Untersuchungseinheit hat unter Behandlung ein verschlechtertes Ergebnis) bei negativem Punktschätzer.

2.2 Binäre Variable und Cross-over Design

In dieser Konstellation sind reversible Ereignisse notwendig, d.h. Ereignisse, die eine Rückkehr des Patienten in den ursprünglichen Zustand erlauben, also nicht, als Extrem: Tod. Als Beispiel für ein solches reversibles Ereignis seien Exazerbationen (i.e. Verschlechterungen) von Asthmatikern bei unterschiedlichen Behandlungen in den Perioden genannt.
Es seien p_{ij} die Realisationen der wahren Risiken (Wahrscheinlichkeiten) π_{ij} unter Behandlung bzw. Kontrolle. Damit ergibt sich für die Stichprobe die Berechnung des Punktschätzers $\hat{NNT} = \frac{1}{(p_{21} - p_{12})}$ als das Reziprok des $\hat{ARR} = p_{21} - p_{12}$ basierend auf einer Kontingenztafel der Anteile (s.auch [14])

		Behandlung	
		Ereignis	Kein Ereignis
Kontrolle	Ereignis	p_{11}	p_{12}
	Kein Ereignis	p_{21}	p_{22}

Im Hinblick auf die Berechnung der Varianz des Punktschätzers und die Berechnung des Konfidenzintervalls wird auf Kap. 3.1 und 3.2 verwiesen.

2.3 Binäre Variable und Parallelgruppendesign

In dieser Kombination von Studiendesign und Skalierung der Variable können auch nicht reversible Ereignisse untersucht werden z. B. die Untersuchung von Todesfällen nach zwei in den Gruppen unterschiedlichen Behandlungen auf (chirurgischen Eingriff vs. Chemotherapie in der Onkologie).
Es seien p_B und p_K die Realisationen der wahren Risiken (Wahrscheinlichkeiten) π_B und π_K in der Behandlungs- bzw. Kontrollgruppe. Damit ergibt sich für die Stichprobe die $\hat{ARR} = p_B - p_K$ und $\hat{NNT} = \frac{1}{\hat{ARR}} = \frac{1}{p_B - p_K}$.

Die Anteile der Stichprobe ergeben sich für die jeweiligen Behandlungen des Patienten zu

$$p_i = \frac{e_i}{n_i}$$

p=Anteil
n=Anzahl Patienten/Gruppe
e=Anzahl Patienten mit Ereignis
i=Behandlung, Kontrolle

Zur Berechnung von Varianz und Konfidenzintervall siehe Kap. 3.1 und 3.2.
Die Teststatistik wird analog Kap. 3.3 gebildet, es wird ein exakter Fisher Test gerechnet, bei großem n die asymptotische Variante. Notabene: Alle Beobachtungen fließen in die Berechnung mit ein.

2.4 Klassifizierte stetige Variablen und Cross-over Design

Der Punktschätzer kann in diesem Fall ohne Verteilungsannahmen geschätzt werden. Dafür wird eine intra-individuelle (within) Differenz (wegen der unterschiedlichen Sequenzen eine Cross-over Differenz, d.h. jeweils Behandlung - Kontrolle) berechnet, die dann eine vorab festgelegte - klinisch relevante – Grenze über- oder unterschreitet. Diese Klassifikation führt direkt zu den Anteilen von profitierenden und nicht-profitierenden Patienten. Der Anteil profitierender Patienten (überhalb (bzw. unterhalb) der Grenze,) entspricht:

$$\hat{ARR} = \frac{n_{wd\_pos}}{N}$$

$wd\_pos = Behandlung - Kontrolle > \delta$ und δ : klinisch relevante Grenze, Patientenweise berechnet

Dieser Berechnung kann in einem ersten Schritt eine Post – Prä-Wert Differenzbildung innerhalb jeder Periode vorangehen.
Führt die Klassifizierung zu einer Dichotomisierung, so stellt diese Teststrategie in erster Linie einen Powerverlust gegenüber den Orginaldaten durch den Informationsverlust durch Dichotomisierung bzw. Klassifizierung dar. Der Power-Verlust ist in diesem Fall bedingt durch die nicht vollständig ausgenutzte Information der kontinuierlichen Variablen ([11],[7]). Diese durch die Klassifikation bedingte Reduktion der Power kann zwischen 2 und 40% liegen ([15],[7]).
Zur Berechnung von Varianz und Konfidenzintervall siehe Kap. 3.1 und 3.2.

2.5 Klassifizierte stetige Variablen und Parallelgruppendesign

Für diese Konstellation von Studiendesign und Skalierung der Variablen gibt es 2 Berechnungsmöglichkeiten, die sich darin unterscheiden, wieviele und welche Patienten bei der Analyse berücksichtigt werden.
Die **erste Berechnungsvariante** entspricht einer Dichotomisierung und damit dem Vorgehen bei einer binären Variable (siehe Kap. 2.3), d.h. nur Patienten mit einem Ergebnis über (bzw. unter) einer der beiden Grenzen werden als Ereignis gewertet, der Rest wird als ‚Kein Ereignis' gewertet.
Die **zweite Berechnungsvariante** berücksichtigt alle Kategorien indem die Anteile in den jeweiligen Klassen innerhalb der Gruppen als marginale Verteilungen aufgefasst werden, die zur Erstellung einer Kontingenztafel verwendet werden.

Dabei entspricht der Anteil in den einzelnen Gruppen $p_{.j} = \frac{n_{class}}{n_{Behandlung}}$ für die Behandlungs- und entsprechend $p_{i.} = \frac{n_{class}}{n_{Kontrolle}}$ für die Kontrollgruppe. Die Anteile in den einzelnen Klassen ergeben über alle Klassen hinweg die marginale Verteilung in der jeweiligen Gruppe. Unter Annahme der Unabhängigkeit der Ergebnisse in den Gruppen können über einen Produkt-Multinomial-Ansatz die Zellhäufigkeiten anhand der marginalen Verteilung berechnet werden. Durch Multiplikation des jeweiligen marginalen Anteils ($p_{ij} = p_{i.} * p_{.j}$) ergibt sich die empirische Zellhäufigkeit. Daraus ergibt sich im Fall von 3 (2) Klassen eine 3x3 (2x2) Kontingenztafel, z. B.:
Hierbei ergibt sich $\hat{ARR} = \left((p_{21} + p_{31} + p_{32}) - (p_{12} + p_{13} + p_{23})\right)$ und NNT als Reziprok damit als $\hat{NNT} = \frac{1}{\left((p_{21} + p_{31} + p_{32}) - (p_{12} + p_{13} + p_{23})\right)}$. Diese Variante berücksichtigt somit alle Zellen außerhalb der Hauptdiagonalen, d.h. der Anteil an Patienten mit der Information ‚Ereignis' ist ein anderer als bei der ersten Variante, bei dem $p_{.1} - p_{1.}$, d.h. $\frac{n_{Behand\_++}}{n_{Behand}} - \frac{n_{Kontrol\_++}}{n_{Kontrol}}$ berechnet werden.
In dieser zweiten Variante werden die Patienten, mit keinem Ereignis in beiden Perioden, bzw. solche mit einem Ereignis in beiden Perioden als nicht-informativ angesehen und nicht in die Berechnung einbezogen. Diese Anteile sind jedoch nicht tatsächlich beobachtet, sie folgen aus den Annahmen zu den Marginal-Verteilungen.
Die von [4] vorgeschlagene Berechnung des AQLQ, eines Fragebogens zur Lebensqualität von Asthmatikern entspricht dem vorgestellten zweiten Ansatz. Sie beruht auf zwei symmetrischen Klassengrenzen, die von [5] vorgestellt wurden.

		Behandlung: Verbessert (=positives Ereignis)	Behandlung: Unverändert (= kein Ereignis)	Behandlung: Verschlechtert (=negatives Ereignis)	
Kontrolle	Verbessert (=positives Ereignis)	p_{11}	p_{12}	p_{13}	$p_{1.} = \frac{n_{Kontrol\_++}}{n_{Kontrol}}$
	Unverändert (= kein Ereignis)	p_{21}	p_{22}	p_{23}	$p_{2.} = \frac{n_{Kontrol\_oo}}{n_{Kontrol}}$
	Verschlechtert (=negatives Ereignis)	p_{31}	p_{32}	p_{33}	$p_{3.} = \frac{n_{Kontrol\_--}}{n_{Kontrol}}$
		$p_{.1} = \frac{n_{Behand\_++}}{n_{Behand}}$	$p_{.2} = \frac{n_{Behand\_oo}}{n_{Behand}}$	$p_{.3} = \frac{n_{Behand\_--}}{n_{Behand}}$	

Da die zweite Variante eine – allerdings unbekannte - Korrelation (ρ) innerhalb eines Patienten voraussetzt, unterscheidet sich die Berechnung der Varianz von der ersten Variante (s. Kap. 3.1) und der Fallzahl.
Welcher der beiden Ansätze die größere Power aufweist, ist Gegenstand aktueller Untersuchungen, im ersten Fall werden mehr (i.e. alle) Patienten berücksichtigt, wobei einige wenig Information tragen, im zweiten Ansatz sind weniger, aber ausschließlich informative Patienten berücksichtigt. Mit der Bildung der Kontingenztafel sind zwei Aspekte verbunden: (i) Je höher die Anteile auf der Hauptdiagonalen, desto geringer die tatsächliche Fallzahl für die Berechnung, im Extrem wird nur ein kleiner Teil der Stichprobe tatsächlich zur Berechnung verwendet. (ii) Die Zellbesetzungen sind berechnet, nicht beobachtet (s.o.), damit hängen die Ergebnisse von der Gültigkeit der zur Konstruktion der Kontingenztafel notwendigen Annahmen ab.
Für mehr als 2 Klassengrenzen erfolgt die Berechnung analog. Im Falle des zweiten Ansatzes sind alle Zellen außerhalb der Hauptdiagonalen Grundlage der Berechnung.

3 Varianz- und Intervall-Schätzung, Teststatistik

3.1 Varianzberechnung

Analog des Vorgehens beim Punktschätzer wird die Varianz für das ARR berechnet und zur Berechnung des Konfidenzintervalls verwendet (Kap. 3.1 und 3.2). In einem zweiten Schritt werden die Ergebnisse auf die NNT-Skala transformiert. Dabei muß das Studiendesign, die Skalierung der Variable und die gewählte Variante zur Bestimmung der Anteile für die Berechnung der Varianz berücksichtigt werden:

Binär, Cross-over Design
Die Berechnung der Varianz des Punktschätzers von ARR nutzt beim Cross-over Design die zugrundeliegende Multinomialverteilung. Das Verfahren macht sich die asymptotische Annäherung an die Normalverteilung bei steigendem n zunutze (s. auch [14]. Sie berechnet sich zu:

$$\hat{\sigma}^2_{ARR,bin,CO} = \frac{(p_{21} + p_{12} - (p_{21} - p_{12})^2)}{N}$$

Binär, Parallelgruppendesign
Der Punktschätzer des ARR ist die Differenz zweier binomialverteilter Variablen, damit ergibt sich die Varianz als:

$$\hat{\sigma}^2_{ARR,bin,PG} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Stetig, Cross-Over Design
Die Varianz des Punktschätzers (letzterer konnte ohne Verteilungsannahmen geschätzt werden, s. Kap. 3.1) ergibt sich aus der Annahme einer Binomialverteilung für den Anteil von Patienten, deren intra-individuelle Differenz die - klinisch relevante – Grenze übersteigt. Sie berechnet sich und somit zu:

$$\hat{\sigma}^2_{ARR,stet,CO} = \frac{p_{wd\_pos}(1-p_{wd\_pos})}{N}$$

Stetig, Parallelgruppendesign, Berechnungsvariante: Anteilsdifferenzen
Die Berechnung des Punktschätzers nimmt eine dichotome Variable an, entsprechend erfolgt die Berechnung analog zum Fall einer binären Variable im Parallelgruppendesign über die Differenz zweier binomialverteilter Variablen.

Stetig, Parallelgruppendesign, Berechnungsvariante: Kontingenz-Tafel
Im Falle von klassifizierten Variablen, die im Rahmen einer Parallelgruppenstudie erhoben und mittels einer Kontingenztafel ausgewertet werden, berechnet sich die Varianz für den vereinfachten Fall von $n_1=n_2=n$ und $\sigma_1 = \sigma_2$ wie folgt [14]:

$$\hat{\sigma}^2_{ARR,stet,PG,KT} = \frac{1}{(1-\rho)}\phi^2\left(\frac{\mu_2-\mu_1-c}{\sigma\sqrt{2(1-\rho)}}\right) * \frac{\left(1+\frac{(\mu_2-\mu_1-c)^2}{8\sigma_{Gruppe}{}^2}\right)}{n}$$

Ansonsten nach [14] (Formel 13, S. 3953). Da in beiden Fällen die intra-individuelle Korrelation (ρ) unbekannt ist, sollten im Idealfall zur Sensitivitätsprüfung mehrere Berechnungen mit unterschiedlichem ρ erfolgen.

3.2 Intervallschätzung (Konfidenzintervall)

Analog dem Punktschätzer werden die Konfidenzgrenzen durch die Bildung des Reziproks der Konfidenzgrenzen von ARR berechnet, so dass gilt ARR: (LL_{ARR},UL_{ARR}) die entsprechenden Grenzen für NNT=1/ARR: ((1/ UL_{ARR}), (1/ LL_{ARR})) werden.
Zur Berechnung der Konfidenzintervalle stehen mehrere Möglichkeiten zur Verfügung, z. B. die Wald-Methode ($p_1 - p_2 \pm z_{1-\alpha/2} * SE_{(p1-p2)}$, mit SE=Standardfehler) oder die Wilson-Score-Methode ([9],[10],[2],[3]). Die Schwäche der auf Normalapproximation basierenden Wald-Methode (schlechte Überdeckungseigenschaften bei kleinen Stichproben, unbalancierten Designs und Anteile nahe 0 bzw. 1) sind bekannt und für abhängige Stichproben, wie z. B. bei Cross-over-Studien noch gravierender [10]. Trotzdem werden sie vielfach eingesetzt, basierend auf der Annahme, dass die NNT die guten asymptotischen Eigenschaften bei ARR im Falle kleiner Stichproben erbt [8]. Konfidenzintervall basierend auf der Wilson-Score Methode haben weniger Schwächen, werden aber seltener verwendet. Hinsichtlich des Vergleichs der verschiedenen Methoden sei auf [2],[3] verwiesen.

3.3 Teststatistik

Die Teststatistik wird üblicherweise über die Analyse einer Kontingenztafel berechnet und den jeweiligen Tests auf Assoziation. Dabei handelt es sich, unabhängig vom Design meist um den exakten Fisher Test.
Beim cross-over design ist die Kontingenztafel etwas anders aufgebaut, mit den design-üblichen Annahmen: (i) kein carry-over effect, (ii) keine patient-by-treatment Interaktion [6]. Die Kontingenztafel hat die folgende Form:

	Ereignis / kein Ereignis	Kein Ereignis / Ereignis
Behandlung	n_{12}	n_{13}
Kontrolle	n_{22}	n_{23}

Es ist zu beachten, dass die Kombinationen ‚Ereignis / Ereignis' bzw. ‚Kein Ereignis / Kein Ereignis' in der letztgenannten Kontingenztafel nicht aufgeführt sind. Für die Berechnung der Teststatistik werden alle Patienten mit einem Ereignis in nur einer Periode berücksichtigt, d.h. bei einem 2x2 Cross-over Design, unter nur einer von beiden Behandlungen. Bei der Darstellung des Ergebnisses sollte ein Hinweis auf die Zahl der nicht-berücksichtigten Patienten aufgeführt werden.
Ein asymptotischer Test für eine cross-over Studie, z. B. bei großen Fallzahlen, ist z. B. der Mainland-Gart Test [6].

4 Interpretation

4.1 Punktschätzer und Konfidenzintervall

Da ARR eine unimodale Verteilung aufweist, gilt dies auch für den Reziprokwert NNT [8]. Die Autoren führen auch weitergehende Untersuchungen zu Verzerrung und Variabilität durch. Die Bestimmung der genauen Verteilung ist allerdings schwierig, da für ARR = 0 eine Sprungstelle in der Verteilung von NNT ist. Dies bedeutet anschaulich, dass bei Gleichheit des Ergebnisses in beiden Gruppen, das Resultat für den Punktschätzer ∞ ist.
Gelten die Regularitätsbedingungen (z. B. Stetigkeitsaxiom) so ist die Transformation kein Problem. Ist diese Bedingung verletzt, können die Konfidenzintervalle nicht mehr intuitiv interpretiert werden (siehe Beispiel im Kap. 5.2). In diesem Fall scheint der PE nicht mehr im Konfidenzintervall zu liegen. Allerdings besteht in diesem Fall das Konfidenzintervall aus der Union von zwei Teilintervallen ([1],[8]), so dass gilt $\left(-\infty, \frac{1}{LL_{ARR}}\right] \cup \left[\frac{1}{UL_{ARR}}, \infty\right)$.

Bezogen auf die unterschiedlichen Situationen lässt sich konstatieren:

- Bei aktiv kontrollierten Studien wird der o.g. Fall relativ regelmäßig eintreten, da bereits die Guidelines eine geringere Wirksamkeit unter bestimmten Bedingungen zulassen (non-inferiority acceptance limit). Im Rahmen der Hypothese wird akzeptiert, dass die Zielvariable um einen bestimmten Betrag δ schlechter sein darf, ohne dass von einer klinisch relevanten Unterlegenheit gesprochen werden muß. Damit werden regelmässig Punktschätzer im Bereich von NNTH das Ergebnis sein, mit zwei Konsequenzen: (i) Der NNT-Schätzer basiert auf einer nicht-relevanten Unterlegenheit und (ii) das Konfidenzintervall zerfällt in die oben beschreibenen zwei Teilintervalle. Die Vermittlung und Interpretation gegenüber den Fachwissenschaftlern dürfte sich schwierig gestalten.
- Bei einer Meta-Analyse, d.h. bei einer Kombination mehrerer Studien kann ein ‚overall' Punktschätzer nicht direkt durch Mittelung bestimmt werden, hierfür muß zuerst der ‚overall' Schätzer für ARR berechnet werden, und dann erneut auf die NNT-Skala transformiert werden [8].
- Im Fall von klassifizierten Variablen und bei Berechnung über eine Kontingenztafel gilt: Je größer der Anteil der Patienten, die sich nicht verändern, desto mehr basieren die Ergebnisse auf einem kleinen Prozentsatz von ‚Respondern'.

5 Fazit

Hinsichtlich der statistischen Eigenschaften ist NNT dem ARR unterlegen, letztlich müssen die Ergebnisse auf der ARR-Skala berechnet und in die NNT-Skala transformiert werden. Die suggestivere Interpretierbarkeit der Ergebnisse von sowohl Punktschätzer wie auch Konfidenzintervall bei NNT ist nur unter bestimmten Bedingungen gegeben. Werden diese verletzt, ist eine Interpretation deutlich erschwert und nicht mehr intuitiv. Für die Interpretation bzw. Präsentation sollte ein zweites Maß mit beschrieben werden (s. auch [8]). Hier bietet sich ebenfalls die ARR an, da es (i) ohnehin zur Berechnung von NNT benötigt wird und (ii) gegenüber relativen Größen (relative risk reduction) den Vorteil aufweist, ohne die Annahme gleicher Risiken in den Vergleichsgruppen auszukommen.

Sollte die Zielgröße stetig sein, so kommen zusätzlich noch alle Nachteile zum tragen, die sich ganz allgemein durch die Klassifikation ergeben. Auch wenn man der auf Klassifizierung beruhenden Responder-Analyse nicht so grundsätzlich ablehnend gegenüber steht, wie dies z. B. [12] tut, sollte NNT (aber in diesem Fall auch ARR) nicht die erste Wahl der Analyse ein. Im Falle der vorherigen Klassifikation ist meist die Validität der Relevanz-Grenze(n) nach erfolgter Berechnung der NNT, wegen deren suggestiver Interpretation bei gleichzeitig komplexer Berechnung nicht mehr

Teil der Interpretation und Diskussion des Ergebnisses. Es wird dann oft vergessen, dass die Begründung für die Wahl der Klassengrenzen fachlich (hier: klinisch) fundiert sein sollte, d.h. sich an fachlichen Relevanzkriterien genügen. Vielfach fehlt eine solche Rationale und die Grenze wird willkürlich gewählt [12]. Selbst im günstigen Fall einer fachlichen Überprüfung beruht die Wahl zumindest anfangs auf sehr geringen Fallzahlen. Für den schon angesprochenen Fragebogen (AQLQ) beruhte die Definition von zwei symetrischen Klassengrenzen, auf einer Studie mit 39 Patienten [5]. Die tatsächlich resultierenden Grenzen differieren je nach Teilbereich des Fragebogens um 20-60% (bis zu 20%) von den postulierten von 0.5 (-0.5) score-Punkten. Die dadurch erreichte Vereinfachung darf nicht dazu führen, dass diese Grenzen nicht hinterfragt werden.
Insgesamt lässt sich konstatieren, dass NNT ein nur sehr bedingt taugliches Maß zur Responder-Analyse ist, sowohl von den statistischen Eigenschaften, wie auch der Berechnung etc. ist ARR zu präferieren.
Werden seitens der Fachwissenschaftler dennoch NNT Auswertungen gewünscht, so ist neben der korrekten Berechnung von Punktschätzer, Varianz und Konfidenzintervall (s.o.) in den folgenden Kapiteln eine Möglichkeit für (i) die Umsetzung in SAS-Programmen und (ii) Möglichkeiten der praktischen Darstellung wiedergegeben.

6 Umsetzung und Darstellung

6.1 Umsetzung in SAS

Die Umsetzung in SAS bildet die in Kap. 2.2. bis 2.5. beschriebenen Fälle ab. Für die Berechnung wurden ein SAS-Program NNT.sas mit 6 Makroprogrammen geschrieben. Die Parametrisierung erfolgt durch Übergabe von SAS - Macro-Variablen, die inhaltlich Studiendesigntyp, Skalierung der Variable, Berechnungsvarianten für Konfidenzgrenzen und Punktschätzer festlegen. Auch Informationen zu einer möglichen Klassifikation (z. B. Klassengrenzen). Weitere Steuerparameter betreffen Parameter, Niveau, Gruppenbezeichnungen und Eingangsdataset.

In die Programme wurden bestehende, anerkannte und publizierte Teillösungen integriert (z. B. die Berechnung der Wilson-Score Konfidenzintervalle durch [2],[3].
Vorgehen: Klassifizierung, Berechnung der ARR und NNT sowie der Konfidenzintervalle basierend auf Wald- und Wilson-Score Konfidenzintervalle. Je nach Berechnungsvariante erfolgt die Angabe der nicht berücksichtigten Patient. Die Graphische

Aufarbeitung von NNT bzw. ARR Punktschätzer und Konfidenzintervall in Anlehnung an [1].

6.2 Darstellung von Ergebnissen

Im Folgenden wird eine Darstellung für einen aufwändigen Fall dargestellt: Parallelgruppenvergleich mit kontinuierlicher Ausgangsvariablen und einer Klassifikation mit 2 Grenzen, damit 3 Klassen. Zur Berechnung wurde eine Kontingenztafel verwendet, wobei alle Zellen mit Ausnahme der Hauptdiagonalen berücksichtigt werden (Kap. 2.5, Variante 2). Als realer Ausgangspunkt dient der Fragebogen zur Lebensqualität von Asthmatikern, der auf einem einfachen Mittelwert aus den numerischen Antworten auf 32 Fragen basiert, wobei die Einzelwerte in den Antworten zwischen 0 und 7 liegen können. Basierend auf den Relevanzüberlegungen ergeben sich symetrische und studiendesign-unabhängige Klassifizierungsgrenzen, die durch [5] publiziert wurden. Diese Grenzen gelten sowohl für die intra-individuelle als auch für die inter-individuelle Differenz. Durch die 2 Grenzen ergeben sich 3 Klassen: Verbesserung AQLQ score Differenz $\geq$ 0.5 score-Punkte, Unverändert (- 0.5<diff<0.5) und verschlechtert (diff $\leq$ -0.5).
Die Anteile im dargestellten Beispiel sind willkürlich gewählt und spiegeln eine nicht signifikante Unterlegenheit der Behandlung wieder, wie dies z. B. bei einer Nichtunterlegenheitsstudie auftreten könnte. Damit können die Schwierigkeiten bei der Darstellung der Konfidenzintervalle einfacher aufgezeigt werden.
Erstellt werden 2 Dateien, die sowohl Teile der Berechnung nachvollziehbar machen (Abb. 1), sowie eine Graphik (Abb. 2) die eine unterstützende optische Interpretation erlaubt.

PE + Confidence Intervals for Activity alpha=0.05 (two-sided)

n not used for calculation: 14 (35.5%)

	Absolute Risk Reduction		Number Needed to Treat		
Point estimate	-0.105		-9.524		(contingency table)
Confidence Limit	-0.393	0.183	-2.545	5.468	(Wald)
Confidence Limit	-0.365	0.176	-2.742	5.697	(Wilson Score)

Contingency table of subjects with classified score variation for Activity
Design= parallel Scale= continuous (3 classes)

	Treatment			
	Deteriorated	Unchanged	Improved	Total
Control				
Deteriorated	0.045	0.060	0.045	0.150
Unchanged	0.165	0.220	0.165	0.550
Improved	0.090	0.120	0.090	0.300
Total	0.300	0.400	0.300	1.000

Das Beispiel ist so gewählt, dass die Regularitätsbedingungen verletzt werden und Punktschätzer und Konfidenzintervall für NNT nicht mehr intuitiv verstanden werden. U.a. für diesen Fall liefert die graphische Darstellung eine Hilfestellung zur Interpretation.

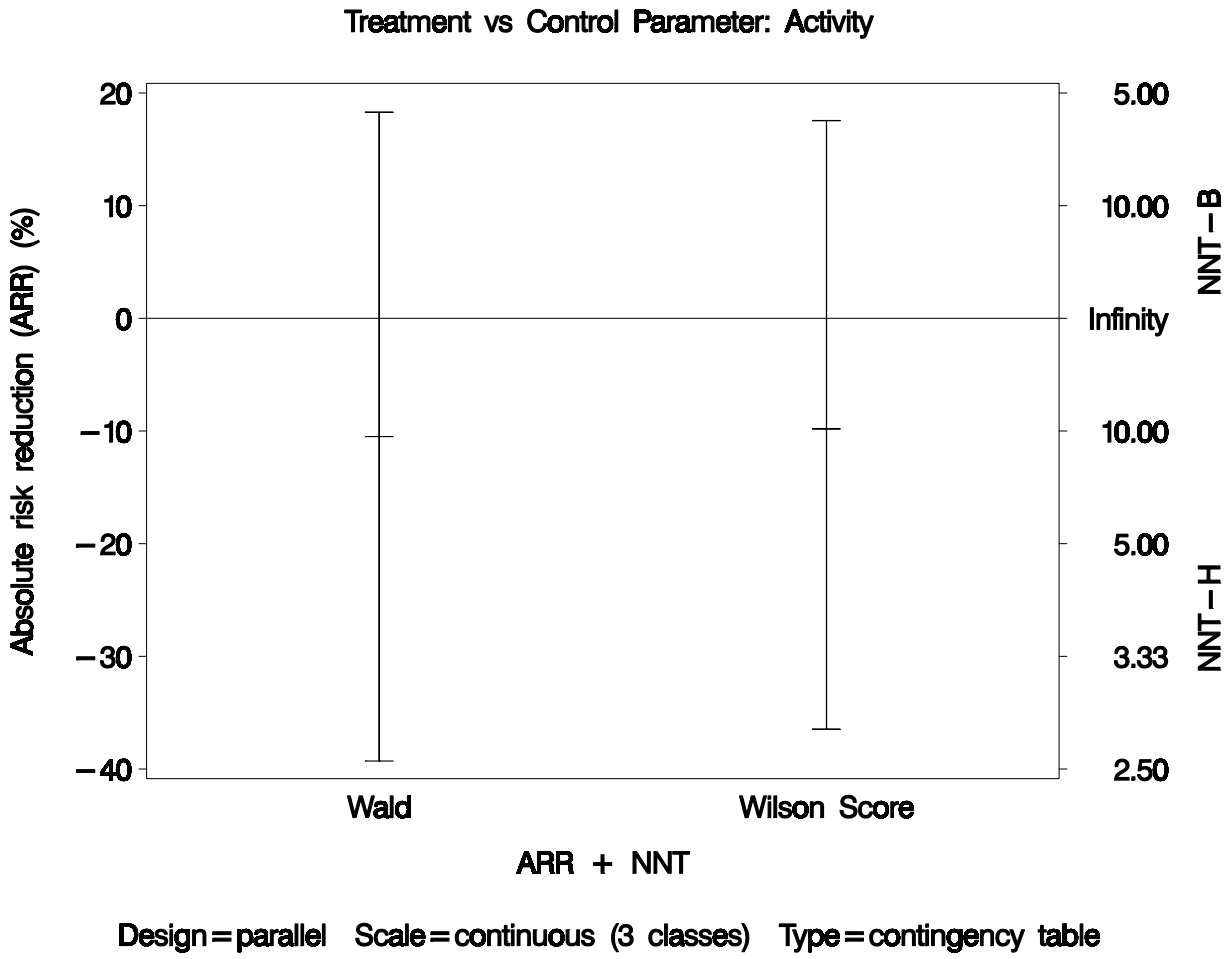

Abb.2: Diese Darstellung erleichtert im vorliegenden Fall z. B. die Vorstellung von der Union zweier Teilintervalle.

Literatur

[1] Altman D G (1998): Confidence intervals for the number needed to treat, BMJ, 317: 1309-12.

[2] Bender, R. (2000): Improving the calculation of confidence intervals for the number needed to treat, in: Hasman et al. (eds.) Medical Infobahn for Europe, IOS Press, 29 – 32 S.

[3] Bender, R (2001): Calculating confidence intervals for the number needed to treat, Control Clin Trials 22: 102-110.

[4] Guyatt, G.H., Juniper, E.F., Walter, S.D., Griffith, L.E. & Goldstein, R.S. (1998): Interpreting treatment effects in randomised trials, BMJ 316: 690-693.

[5] Juniper, E.F., Guyatt, G.H., Willan, A. & Griffith, L.E. (1994): Determining a minimal important change in a disease-specific quality of life questionnaire, J. Clin. Epidemiol. 47: 81-87.

[6] Jones, B. & Kenward, BJ (2003): Design and análisis of cross-over trials, 2nd ed., Monographs on Statistics and Applied Probability 98, Chapman & Hall, Boca Raton, 382 S.

[7] Kieser, M., Röhmel, J. & Friede, T. (2004): Power and sample size determination when assessing the clinical relevance of trial results by ‚responder analyses', Statist. Med. 23: 3287 – 3305.

[8] Lesaffre, E &Pledger, G: A note on the number needed to treat. Control Clin Trials. 1999 Oct;20(5):439-47

[9] Newcombe, RG (1998a): Interval estimation for the difference between independent proportions. A comparative evaluation of eleven methods, Statist. Med. 17: 873-890.

[10] Newcombe (1998b): Improved confidence intervals for the difference between binomial proportions based on paired data, Statist. Med. 17: 2635-2650.

[11] Senn, SJ (1997): Statistical issues in drug development, Wiley, Chichester 423 S.

[12] Senn, SJ (2003): Disappointing dichotomies, Pharmaceut. Statist 2: 239-240.

[13] Walter, SD & Irwig, L (2001): Estimating the number needed to treat (NNT) index when the data are subject to error.Stat Med. 2001 Mar 30;20(6):893-906.

[14] Walter, SD (2001): Number needed to treat (NNT): estimation of a measure of clinical benefit. Stat Med. 2001 Dec 30;20(24):3947-62

[15] Whitehead, J (1993): Sample size calculations for ordered categorical data, Statist. Med. 12: 2257-2271.