OPEN ACCESS

Originalarbeit

Untersuchung von Methoden zur Überprüfbarkeit von Ergebnissen von Studienpopulationen auf Teilpopulationen

Analysis of methods for the transferability of results from study populations to subpopulations

Lars Beckmann[1]
Ulrich Grouven[1]
Meinhard Kieser[2]
Wiebke Sieben[1]
Guido Skipka[1]
Ralf Bender[1]

1 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln, Deutschland

2 Institut für Medizinische Biometrie und Informatik, Ruprecht-Karls-Universität, Heidelberg, Deutschland

Abstract

Background: When assessing the benefit of an intervention, the study population (SP) may consist of a relevant target population (ZP) and a non-relevant population (nZP). We consider the situation that a significant treatment effect is observed only in SP but not in ZP leading to the question if and how the effect in SP may be used for conclusions about the effect in ZP.
Methods: We assessed three test procedures: the first increases the level of significance α_{ZP} for ZP (elevation rule, ER). The second procedure involves a permutation-based test for a qualitative interaction between ZP and nZP (extension rule, EWR). The third one is a modification of the EWR, which takes the relation between ZP and nZP into account.
In a simulation study, we compared the empirical type 1 error and power for all three test procedures.
Results: EWR unacceptably exceeds the significance level for some simulated parameter constellations (median 5.8%, maximum 15.9%). The modified version of EWR has a lower empirical type 1 error (median 5.5%, maximum 10.2%). But EWR has no advantages with respect to the empirical power and type 1 error compared to ER with an increased significance level of α_{ZP}=15%.
Conclusion: ER, with an increased significance level of α_{ZP}=15%, is the appropriate procedure with respect to the empirical power, when accepting a slightly increased type 1 error (median 6.1%, maximum 10.9% over all simulated scenarios).

Keywords: simulation study, transferability, subpopulation

Zusammenfassung

Hintergrund: In Nutzenbewertungen kann der Fall auftreten, dass sich die Studienpopulation (SP) aus einer relevanten Zielpopulation (ZP) und Nicht-ZP (nZP) zusammensetzt und ein nicht statistisch signifikanter Behandlungseffekt in ZP und ein statistisch signifikanter Behandlungseffekt in SP vorliegt. Es stellt sich hier die Frage unter welchen Umständen und mit welcher Methodik das Ergebnis in SP auf ZP übertragen werden kann.
Methoden: Wir haben drei Testprozeduren untersucht: eine Anhebung des Signifikanzniveaus α_{ZP} für ZP (Anhebungsregel, AHR), eine Testprozedur, die auf einem permutationsbasierten Test auf qualitative Interaktion zwischen ZP und nZP beruht (Erweiterungsregel, EWR) sowie eine Modifikation derselben. Die Testprozeduren wurden in einer Simulationsstudie bzgl. des empirischen Fehlers 1. Art und der empirischen Power verglichen.
Ergebnisse: Die EWR zeigte für einzelne Datenkonstellationen eine nicht akzeptable Niveauüberschreitung (Median 5,8%, Maximum 15,9%). Die modifizierte EWR unter Berücksichtigung der Relation der Stichprobengrößen in ZP und nZP führte zwar zu einer Reduktion des empirischen Fehlers 1. Art (Median 5,5%, Maximum 10,2%). Ein Vergleich

bezüglich empirischer Power und Fehler 1. Art mit der AHR mit einer Erhöhung des Signifikanzniveaus auf α_{ZP}=15% ließ jedoch insgesamt keine Vorteile erkennen.
Schlussfolgerung: Bei Inkaufnahme einer geringen Niveauüberschreitung (Median 6,1%, Maximum 10,9% in den untersuchten Datenkonstellationen) stellt die AHR mit bedingter Erhöhung des Signifikanzniveaus auf α_{ZP}=15% unter Berücksichtigung des Fehlers 1. Art und der Power das geeignetste Verfahren dar.

Schlüsselwörter: Simulationsstudie, Übertragbarkeit, Teilpopulation

Hintergrund

Nutzenbewertungen des IQWiG haben zum Ziel, Nutzen und Schaden einer Prüfintervention im Vergleich zu einer Kontrollbehandlung zu bewerten. Es kann der Fall auftreten, dass für die Untersuchung einer konkreten Fragestellung lediglich eine Teilpopulation (TP) der gesamten Studienpopulation (SP) relevant ist.
Eine besondere Situation liegt vor, wenn sich die SP aus der für die untersuchte Fragestellung relevanten Zielpopulation (ZP) und Nicht-ZP (nZP) zusammensetzt und eine Datenkonstellation wie in Abbildung 1 dargestellt vorliegt, d.h.:

- ein nicht statistisch signifikanter Behandlungseffekt in der ZP
- ein gleichgerichteter Behandlungseffekt in der nZP
- ein statistisch signifikanter Behandlungseffekt in der SP
- ein nicht statistisch signifikanter Interaktionstest (p_{int}≥5%)
- nicht zu unpräzise Effektschätzung in der ZP im Vergleich zur nZP

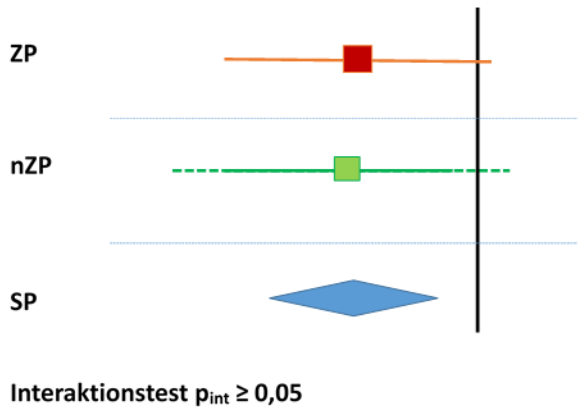

Abbildung 1: Notwendige Datenkonstellation für die Übertragung der Ergebnisse für die Studienpopulation (SP) auf die Zielpopulation (ZP).
Dargestellt ist die Effektschätzung in der jeweiligen Population mit zugehörigem 95%-KI, vertikaler Strich stellt Nulleffekt für das betrachtete Effektmaß dar.

Es stellt sich die Frage, ob der nicht statistisch signifikante Effekt in der ZP eine Folge zu geringer Power ist und unter welchen Umständen das Ergebnis der SP herangezogen werden kann. Ein nicht statistisch signifikanter Interaktionstest zu α=5% allein ist nicht hinreichend, um eine Aussage im Sinne der Gleichheit von Effekten abzuleiten und Aussagen zu einer ZP durch Heranziehen der Ergebnisse der gesamten SP zu treffen. So kann es trotz eines nicht statistisch signifikanten Interaktionstests zu Situationen kommen, in denen zwischen TPen relevant unterschiedliche Effekte geschätzt werden. Dies bedeutet, dass eine qualitative Interaktion zwischen der interessierenden ZP und der nZP mit ausreichender Sicherheit ausgeschlossen werden muss, um das Ergebnis der SP auf die ZP übertragen zu können [1].
Eine mögliche Vorgehensweise bieten die Erweiterungsregel (EWR), die als Test auf eine qualitative Interaktion verstanden werden kann, sowie die Anhebungsregel (AHR), bei der das Signifikanzniveau für den Test in der ZP angehoben wird. Die Anwendung von mehrstufigen Testprozeduren, die die EWR oder die AHR enthalten, führen konstruktionsbedingt zu einer Niveauüberschreitung für den Test auf einen Effekt in der ZP. Folgende Fragestellungen sollen untersucht werden:

- **Frage 1: Signifikanzniveau.** Es soll die Stärke der Niveauüberschreitung quantitativ untersucht werden. Ziel ist es, einfache Anforderungen an die Parameter(konstellationen) zu formulieren, sodass eine Testprozedur mit entsprechend modifizierten Bedingungen mit akzeptabler Niveauüberschreitung angewendet werden kann.
- **Frage 2: Power.** Sofern eine Formulierung der Anforderungen wie unter Punkt 1 beschrieben gelingt, soll der Powergewinn durch die Anwendung der modifizierten Testprozedur untersucht werden.
- **Frage 3:** Es soll ein Vergleich der alternativen Testprozeduren hinsichtlich Fehler 1. Art und Power durchgeführt werden.

Methodik & Daten

Hypothese

Folgendes Testproblem wird betrachtet:
H_0: $\theta_{ZP}=0$ vs. $\theta_{ZP}\neq 0$, mit θ_{ZP} wahrer Effekt in ZP

Testprozedur

- Schritt 1: Es wird zweiseitig getestet, ob für ZP eine statistisch signifikante Effektschätzung zum Niveau α=5% vorliegt.
  Falls ja: H_0 wird abgelehnt.
  Falls nein: Führe Schritt 2 durch.
- Schritt 2: Es wird zweiseitig getestet, ob für SP eine statistisch signifikante Effektschätzung zum Niveau α=5% vorliegt.
  Falls ja: Führe Schritt 3 durch.
  Falls nein: H_0 wird nicht abgelehnt.
- Schritt 3: Es wird geprüft, ob die Effektschätzungen von ZP und nZP dieselbe Effektrichtung haben.
  Falls ja: Führe Schritt 4 durch.
  Falls nein: H_0 wird nicht abgelehnt.
- Schritt 4: Es wird getestet, ob zwischen ZP und nZP eine statistisch signifikante Interaktion zum Niveau α=5% vorliegt.
  Falls ja: H_0 wird nicht abgelehnt.
  Falls nein: Führe Schritt 5 durch.
- Schritt 5: Wird dieser Schritt erreicht, so liegen hinreichend homogene Effektschätzungen für ZP und nZP mit derselben Effektrichtung vor, und der Effekt in der SP ist statistisch signifikant von Null verschieden. Unter diesen Voraussetzungen können weitere statistische Tests bzgl. der Hypothesen durchgeführt werden.

Erweiterungsregel (EWR)

Die Erweiterungsregel untersucht, wie wahrscheinlich das beobachtete Ergebnis ist, wenn in Wahrheit kein Effekt in der ZP vorliegt. Dabei werden die Effektschätzungen in den Populationen SP, ZP und nZP berücksichtigt, sowie die Heterogenität zwischen ZP und nZP. Somit kann die EWR auch als Test auf eine qualitative Interaktion verstanden werden. Jedoch ist zu beachten, dass hierbei in einer Population, ZP, auch der Nulleffekt berücksichtigt wird. Damit geht die betrachtete Situation über die bekannte Situation hinaus, in denen bei einer qualitativen Interaktion von zwei vom Nulleffekt verschiedenen und nicht gleichgerichteten Effekten ausgegangen wird [2].
Die EWR beinhaltet die Simulation eines empirischen p-Wertes. Als Effektmaß wird die standardisierte Mittelwertdifferenz SMD (=Cohen's d, θ) betrachtet.

Für beobachtete Werte $[\theta_{ZP}^{beob}, SE_{ZP}^{beob}]$ und $[\theta_{nZP}^{beob}, SE_{nZP}^{beob}]$ in den Teilpopulationen ZP und nZP (mit $n_{i,ZP}$ und $n_{i,nZP}$ als Fallzahlen der zwei Gruppen in ZP bzw. nZP) werden die folgenden Schritte n_{rep}-mal durchlaufen:

1. Zufälliges Ziehen von

a.

$$m_{1,ZP}^{rand} \sim N\left(0{,}1 \,/\, \sqrt{n_{1,ZP}}\right),$$

$$S_{1,ZP}^{rand} \sim \sqrt{rand(X^2_{n_{1,ZP}-1})} \,/\, (n_{1,ZP}-1)$$

und

$$m_{2,ZP}^{rand} \sim N\left(0{,}1 \,/\, \sqrt{n_{2,ZP}}\right),$$

$$S_{2,ZP}^{rand} \sim \sqrt{rand(X^2_{n_{2,ZP}-1})} \,/\, (n_{2,ZP}-1)$$

für ZP

b.

$$m_{1,nZP}^{rand} \sim N\left(\theta_{nZP}^{beob}, 1 \,/\, \sqrt{n_{1,nZP}}\right),$$

$$S_{1,nZP}^{rand} \sim \sqrt{rand(X^2_{n_{1,nZP}-1})} \,/\, (n_{1,nZP}-1)$$

und

$$m_{2,nZP}^{rand} \sim N\left(0{,}1 \,/\, \sqrt{n_{2,nZP}}\right),$$

$$S_{2,nZP}^{rand} \sim \sqrt{rand(X^2_{n_{2,nZP}-1})} \,/\, (n_{2,nZP}-1)$$

für nZP

$rand(X_k^2)$ bezeichnet dabei eine Zufallszahl aus einer Chi-Quadrat-Verteilung mit *k* Freiheitsgraden. Aus den Angaben kann in beiden Populationen die SMD mit zugehörigem Standardfehler geschätzt werden:

c.

$$\theta_{ZP}^{rand} = \frac{m_{1,ZP}^{rand} - m_{2,ZP}^{rand}}{S_{pool,ZP}^{rand}}$$

mit

$$S_{pool,ZP}^{rand} = \sqrt{\frac{(n_{1,ZP}-1)\left(S_{1,ZP}^{rand}\right)^2 - (n_{2,ZP}-1)\left(S_{2,ZP}^{rand}\right)^2}{n_{1,ZP}+n_{2,ZP}-2}}$$

und

$$SE\left(\theta_{ZP}^{rand}\right) = \sqrt{\frac{(n_{1,ZP}+n_{2,ZP})}{n_{1,ZP}n_{2,ZP}} + \frac{\theta_{ZP}^{rand}}{2(n_{1,ZP}+n_{2,ZP})-4}}$$

d.

$$\theta_{nZP}^{rand} = \frac{m_{1,ZP}^{rand} - m_{2,nZP}^{rand}}{S_{pool,nZP}^{rand}}$$

mit

$$S_{pool,nZP}^{rand} = \sqrt{\frac{(n_{1,nZP}-1)\left(S_{1,nZP}^{rand}\right)^2 - (n_{2,nZP}-1)\left(S_{2,nZP}^{rand}\right)^2}{n_{1,nZP}+n_{2,nZP}-2}}$$

und

$$SE\left(\theta_{nZP}^{rand}\right) = \sqrt{\frac{(n_{1,nZP}+n_{2,nZP})}{n_{1,nZP}n_{2,nZP}} + \frac{\theta_{nZP}^{rand}}{2(n_{1,nZP}+n_{2,nZP})-4}}$$

2. Durchführen eines Interaktionstests basierend auf $[\theta_{ZP}^{rand}, SE_{ZP}^{rand}]$ und $[\theta_{nZP}^{rand}, SE_{nZP}^{rand}]$ mit Ergebnis P_{int}^{rand}, p-Wert des Q-Tests auf Homogenität.

3. Überprüfung:
   (i) $\theta_{ZP}^{rand} \leq \theta_{ZP}^{beob}$
   und
   (ii) $P_{int}^{rand} \geq P_{int}^{beob}$

Tabelle 1: Szenarien für die Simulationsuntersuchungen

Parameter	Beschreibung	Werte für die Simulation	Anzahl Werte
Stichprobengröße N_{nZP}	Stichprobengröße pro Therapiearm in nZP	50/100/200/500/750/1.000	6
N_{ZP}/N_{nZP}	Stichprobengröße in ZP im Verhältnis zu nZP	0,2/0,33/0,5/0,75/1,0/1,5/2,0/3,0/5,0	9
θ_{nZP}	Wahrer Effekt in nZP	0/–0,1/–0,2/–0,3/–0,4/–0,5/–0,6/–0,7/–0,8/–0,9/–1,0	11
Fehler 1. Art			
θ_{ZP}	Wahrer Effekt in ZP	0	1
Power			
θ_{ZP}	Wahrer Effekt in ZP	–0,1/–0,2/–0,3/–0,4/–0,5/–0,6/–0,7/–0,8/–0,9/–1,0	10
n_{sim}	Anzahl Replikationen pro Szenarium		10.000

Ein empirischer p-Wert ergibt sich aus der Anzahl an Replikationen, in denen die beiden Bedingungen unter 3. erfüllt sind, geteilt durch die Gesamtzahl (n_{rep}) an Replikationen. Als Signifikanzniveau wird α=2,5% gewählt. Die Anzahl an Replikationen beträgt n_{rep}=100.000. Ist der empirische p-Wert kleiner als 2,5%, so wird das Ergebnis der Gesamtpopulation SP auf die jeweilige ZP übertragen, d.h. es wird geschlossen, dass der Behandlungseffekt auch in der Zielpopulation signifikant vom Nulleffekt verschieden ist.

Das vorgestellte Verfahren kann mit entsprechenden Verteilungsananahmen auf weitere Effektmaße wie das relative Risiko, das Odds Ratio oder das Hazard Ratio angewendet werden.

Anhebungsregel (AHR): Testprozedur mit bedingter Erhöhung des Signifikanzniveaus

Durchführung der Schritte 1 bis 4, in Schritt 5 wird erneut ein zweiseitiger Test auf einen Effekt in der ZP mit erhöhtem Signifikanzniveau α_{ZP}>5% durchgeführt.

Standardprozedur (A_5)

Um darstellen zu können, welche Vor- und Nachteile mit den genannten Testprozeduren EWR und AHR einhergehen, wird als Referenz das Standardvorgehen A_5, d.h. ein Test auf einen von Null verschiedenen Effekt in der ZP mit einem Signifikanzniveau von 5%, in den Vergleich der Testprozeduren mit einbezogen.

Simulationsstudie

Im Rahmen von Simulationsuntersuchungen werden empirischer Fehler 1. Art und Power der Testprozeduren untersucht. Gegenstand der im Folgenden beschriebenen Simulationsuntersuchungen ist die Anwendung der gesamten Testprozedur. Davon abzugrenzen ist die Simulation des empirischen p-Wertes im Rahmen der EWR, die Teil der Methodik der EWR ist.

Tabelle 1 zeigt die geplanten Szenarien für die Simulationsuntersuchungen. Der Wertebereich der untersuchten Simulationsparameter wurde so gewählt, dass praxisrelevante Szenarien abgebildet sind. Jedes Szenario wird für die Untersuchung des empirischen Fehlers 1. Art und der empirischen Power 10.000 mal simuliert. Als Effektmaß wird die standardisierte Mittelwertdifferenz (Cohen's d) verwendet.

Ergebnisse

Empirischer Fehler 1. Art

Für die Untersuchung des Fehlers 1. Art wurden insgesamt 594 Szenarien simuliert. Die Anzahl der Replikationen je Szenario betrug 10.000, von denen zufällig ausgewählt 6.667 als Trainingsdaten und die übrigen 3.333 Szenarien als Testdaten verwendet wurden.

Über alle Szenarien betrachtet ist der empirische Fehler 1. Art auf den Trainingsdaten in 5,56% der Szenarien größer als 10% (Abbildung 2). Auch wenn Mittelwert und Median des Fehlers 1. Art mit 6,31% und 5,70% leicht erhöht sind, gibt die Häufigkeit einer großen Niveauüberschreitung Anlass, den Einsatz der EWR auf solche Szenarien zu beschränken, in denen nicht oder nur sehr selten mit einem Fehler 1. Art von mehr als 10% zu rechnen ist.

Das 97,5%-Quantil der Verteilung des empirischen Fehlers 1. Art der Testprozedur mit EWR bei einem Signifikanzniveau von 5% für den Interaktionstest ist 12,0%; d.h., in 2,5% der simulierten Szenarien ist ein empirischer Fehler 1. Art größer als dieser Wert zu erwarten.

Die Relation der Stichprobengrößen in ZP und nZP erweist sich als ein einfacher Ansatz, um Szenarien zu identifizieren, die nur in seltenen Fällen einen empirischen Fehler 1. Art größer als 10% aufweisen.

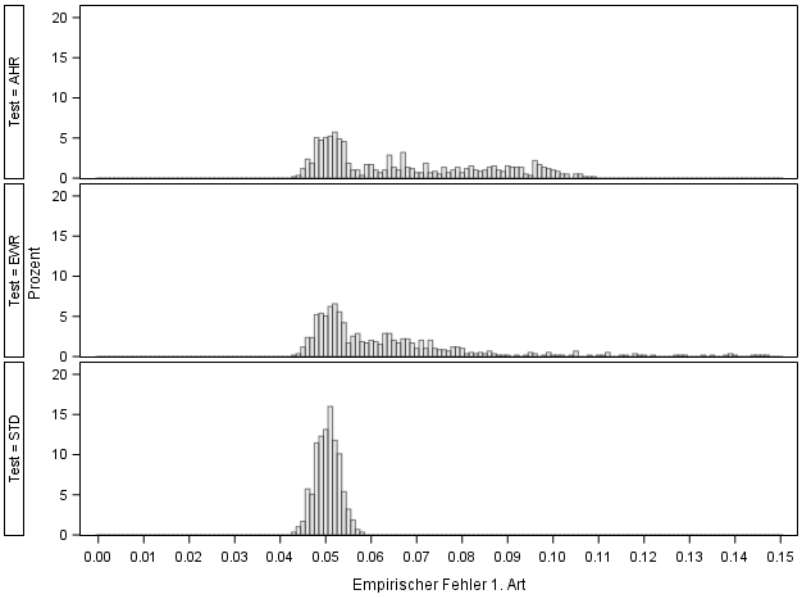

Abbildung 2: Verteilung des empirischen Fehlers 1. Art der Testprozeduren über alle Szenarien (Trainingsdatensatz)

Tabelle 2: 97,5%-Quantil der Verteilung des empirischen Fehlers 1. Art der Testprozedur mit EWR für verschiedene Cut-offs für die Relation der Stichprobengrößen (Trainingsdaten)

Cut-off	0,2	**0,33**	0,5	0,75	1,0	1,5	2,0	3,0	5,0
97,5%-Quantil (%)	12,03	10,14	8,50	7,95	7,45	7,09	6,88	6,63	6,154

Tabelle 3: 97,5%-Quantil der Verteilung des empirischen Fehlers 1. Art der Testprozedur A_{Bed15} für verschiedene Signifikanzniveaus des Tests auf einen von Null verschiedenen Effekt in der ZP (Trainingsdaten)

Signifikanzniveau	0,05	0,09	0,10	0,11	0,12	0,13	0,14	0,145	**0,15**
97,5%-Quantil (%)	5,58	7,23	7,74	8,23	8,70	9,19	9,66	9,93	10,15

Die Hinzunahme weiterer Parameter brachte keine bedeutsame Verbesserung der Identifikation von Szenarien mit häufiger erhöhtem empirischen Fehler 1. Art.

Mit fallendem Wert der Relation der Stichprobengrößen ist mit einem zu häufig deutlich erhöhten empirischen Fehler 1. Art zu rechnen, sodass die EWR dann nicht mehr angewendet werden sollte. Es kann ein Cut off so bestimmt werden, dass folgendes gilt: Beschränkt man die Anwendung der EWR auf Szenarien, in denen die Relation der Stichprobengrößen größer gleich dem Cut-off ist, so haben weniger als 2,5% der Szenarien einen empirischen Fehler 1. Art von über 10%. Aus Tabelle 2 kann für verschiedene Cut-offs entnommen werden, wie hoch der empirische Fehler 1. Art für die 2,5% mit dem größten empirischen Fehler 1. Art mindestens ist (97,5%-Quantile der Verteilung der simulierten Fehler 1. Art der Szenarien). Beschränkt man die Anwendung der EWR auf Szenarien, in denen die Relation der Stichprobengrößen ≥0,33 ist, so haben weniger als 2,5% der Szenarien einen empirischen Fehler 1. Art von über 10%. Bei einem Cut-off von 0,2 hätten mehr als 2,5% der Szenarien einen empirischen Fehler 1. Art von über 10%. Hieraus ergibt sich die Testprozedur $EWR_{0,33}$: zusätzlich zu den unter den Schritten 1 bis 4 genannten Bedingungen wird 0,33 als Cut-off für die Relation der Stichprobengrößen n_{ZP}/n_{nZP} als weitere Voraussetzung für die Anwendung der EWR gewählt.

Analog zum Vorgehen bei der EWR wird für die Anwendung der AHR die Erhöhung des Signifikanzniveaus so festgelegt, dass auch für diese Testprozedur das 97,5%-Quantil der Verteilung des Fehlers 1. Art für die Trainingsdaten kleiner als 10% ist. Aus Tabelle 3 kann entnommen werden, dass dies bei einem Niveau von knapp unter 15% erfüllt ist. Das Niveau für den Test auf einen Effekt in der ZP innerhalb dieser Testprozedur wird daher auf 15% festgesetzt. Für die Testdaten ergibt sich für AHR_{15} ein 97,5%-Quantil von 10,23% für die empirische Verteilung des Fehlers 1. Art.

Tabelle 4 fasst die Simulationsergebnisse zum empirischen Fehler 1. Art zusammen. Die für den Trainingsdatensatz ermittelten Werte werden für den Testdatensatz bestätigt.

Vergleich der Testprozeduren bzgl. der empirischen Power

Die Größenordnung des Fehlers 1. Art der Testprozedur mit EWR (ohne zusätzliche Bedingungen) erwies sich in den Simulationsuntersuchungen als inakzeptabel hoch. Im Folgenden wird diese Testprozedur daher nicht weiter

Tabelle 4: Ergebnisse für den empirischen Fehler 1. Art (%) für die untersuchten Testprozeduren

Testprozedur	Daten	Mittelwert	Median	97,5%-Quantil	Maximum
$EWR_{0.33}$	Trainingsdaten	5,97	5,40	10,14	12,03
	Testdaten	5,99	5,52	10,17	12,45
$AHR_{0,15}$	Trainingsdaten	6,69	6,19	10,15	10,90
	Testdaten	6,70	6,09	10,23	10,92
A_5	Trainingsdaten	5,04	5,04	5,58	5,85
	Testdaten	5,04	5,04	5,85	6,18

Tabelle 5: Mittlere empirische Power für die Testprozeduren in Abhängigkeit von der Power der Standardprozedur A_5

		Mediane empirische Power			Differenzen der Power zu A_5 pro Szenarium (%-Punkte)					
					Median		90%-Quantil		Maximum	
Szenarien mit Power x von A_5	Anteil Szenarien (%)	A_5	$EWR_{0,33}$	AHR_{15}	$EWR_{0,33}$	AHR_{15}	$EWR_{0,33}$	AHR_{15}	$EWR_{0,33}$	AHR_{15}
Alle	100	100	100	100	<0,1	<0,1	5,1	10,6	22,0	22,7
x=100	48,9	100	100	100	0	0	0	0	0	0
90≤x<100	20,5	99,2	99,4	99,5	<0,1	<0,1	1,4	2,2	5,2	5,9
80≤x<90	4,1	85,1	87,6	89,5	1,5	4,2	6,1	9,0	10,5	11,6
70≤x<80	3,4	77,4	78,4	80,4	1,7	4,7	9,1	12,9	15,0	15,6
60≤x<70	3,3	66,0	69,6	72,0	2,7	6,4	11,7	16,3	19,2	19,0
50≤x<60	2,8	55,4	58,7	65,7	3,1	11,8	13,1	19,1	19,4	20,8
40≤x<50	3,2	43,1	48,1	52,2	0,5	9,3	12,4	20,4	21,4	22,7
30≤x<40	2,1	35,3	39,2	44,6	2,9	10,1	15,8	20,4	22,0	21,8
20≤x<30	4,1	25,1	29,1	34,2	3,0	9,1	14,6	18,8	21,7	20,3
10≤x<20	4,7	14,1	17,1	21,4	1,9	7,4	9,6	14,9	16,8	18,2
x<10	2,8	7,6	10,0	14,1	2,1	7,1	7,7	10,4	12,8	12,5

betrachtet. Die folgenden Vergleiche beziehen sich auf die Testprozeduren, $EWR_{0.33}$, AHR_{15} und A_5.

Die mittlere empirische Power unterscheidet sich zwischen den Testprozeduren über alle 5.940 Szenarien kaum und liegt bei 82,9% für A_5, bei 84,1% für $EWR_{0,33}$ und bei 85,3% für AHR_{15}. Um Unterschiede bezüglich der empirischen Power näher zu untersuchen, wurden pro Szenarium die Differenzen in der Power von $EWR_{0,33}$ und AHR_{15} im Vergleich zu A_5 betrachtet (Tabelle 5). Dabei sind deutliche Powergewinne in Szenarien zu beobachten, in denen die Standardprozedur A5 eine geringe Power hat. Es ergeben sich Powergewinne von $EWR_{0,33}$ im Median bis 3,1 Prozentpunkten und maximal von 22,0 Prozentpunkten. Für AHR15 ergeben sich im Median Powergewinne bis 11,8 Prozentpunkte und maximal 22,7 Prozentpunkte. Die 90%-Quantile ($EWR_{0,33}$: 1,4 bis 15,8 Prozentpunkte; AHR_{15}: 2,2 bis 20,4 Prozentpunkte) zeigen, dass deutliche Powergewinne nicht auf einzelne Szenarien zurückzuführen sind. Insgesamt erwies sich die Testprozedur AHR_{15} als diejenige mit dem höchsten Powergewinn. Die Testprozedur $EWR_{0.33}$ zeigt bezüglich der empirischen Power keine Vorteile, die ihren Einsatz trotz des erhöhten Rechenaufwands rechtfertigen. In der Abwägung von empirischem Fehler 1. Art, empirischer Power sowie Praktikabilität erweist sich die Anhebungsregel AHR_{15} als das geeignetste Verfahren, insbesondere in Situationen mit zu erwartender niedriger Power.

Diskussion

Ausgangspunkt für die vorliegenden Untersuchungen war die Tatsache, dass in Nutzenbewertungen der Fall auftreten kann, dass für die Untersuchung konkreter Fragestellungen lediglich eine Teilpopulation aus einer vorliegenden Studienpopulation relevant ist. Die Auswertung der TP kann zu einer reduzierten Power zur Aufdeckung eines vorhandenen Behandlungseffekts führen. Es stellt sich die Frage, ob und unter welchen Umständen es gerechtfertigt ist, die gesamte SP für eine Aussage zur relevanten TP heranzuziehen. Für die Situation, dass eine spezifische Datenkonstellation vorliegt, wurde die EWR definiert mit dem Ziel, einen relevanten Powergewinn bei Inkaufnahme einer moderaten Niveauüberschreitung zu erzielen.

Die Untersuchung des Fehlers 1. Art bei Anwendung der Testprozedur mit EWR zeigte für einzelne Parameterkonstellationen eine nicht akzeptable Niveauüberschreitung. Es wurde eine modifizierte Testprozedur basierend auf der Relation der Stichprobengrößen $EWR_{0.33}$ definiert.

Die Anwendung der $EWR_{0.33}$ wurde mit einer Testprozedur mit bedingter Erhöhung des Signifikanzniveaus (AHR_{15}) sowie mit dem Standardvorgehen A_5 hinsichtlich empirischem Fehler 1. Art und empirischer Power verglichen.

Deutliche Powergewinne lassen sich in den Szenarien erreichen, in denen A_5 eine geringe Power aufweist.

Insgesamt zeigte die Testprozedur $EWR_{0.33}$ weder bezüglich der empirischen Power noch bezüglich des empirischen Fehlers 1. Art Vorteile gegenüber der alternativen Testprozedur, die ihren Einsatz in Anbetracht des erhöhten Rechenaufwands rechtfertigen würden. In der Abwä-

gung von Einbußen beim empirischen Fehler 1. Art, Zugewinn bei der empirischen Power sowie Praktikabilität erweist sich die Anhebungsregel AHR_{15} insgesamt als das Verfahren der Wahl.

Limitationen der Untersuchungen

Die Abhängigkeit der Ergebnisse von den gewählten Szenarien stellt eine grundsätzliche Limitation von Simulationsuntersuchungen, so auch der vorliegenden, dar. Insbesondere die Tatsache, dass die Ergebnisse extremerer Szenarien mit der gleichen Gewichtung versehen wurden wie die in der Praxis üblicherweise auftretenden, schränkt möglicherweise die Übertragbarkeit der Simulationsergebnisse ein. Um diesem Problem zu begegnen, wurde durch die Wahl geeigneter Parameterwerte versucht, unrealistische Szenarien von vorneherein auszuschließen. Nur eine Gewichtung der Szenarien gemäß ihrer zu erwartenden Auftrittswahrscheinlichkeit hätte dieses Problem tatsächlich lösen können. Es hätte hierzu bekannt sein müssen, welche Parameterkonstellationen in der Realität (z.B. in den Bewertungen des IQWiG) wie häufig auftreten.

Schlussfolgerung

Die Testprozedur mit EWR zur Ableitung von Nutzenaussagen für die Zielpopulation unter Berücksichtigung der gesamten Studienpopulation zeigte für einzelne Datenkonstellationen eine nicht akzeptable Niveauüberschreitung. Eine modifizierte Testprozedur unter Berücksichtigung der Relation der Stichprobengrößen in ZP und nZP führte zwar zu einer Reduktion des empirischen Fehlers 1. Art. Ein Vergleich mit alternativen, einfacheren Testprozeduren AHR_{15} bezüglich der empirischen Power und des Fehlers 1. Art ließ jedoch insgesamt keine Vorteile erkennen. Unter Berücksichtigung des Fehlers 1. Art, der Power sowie des Rechenaufwands liefert die Anhebungsregel AHR_{15} die besten Ergebnisse. Die Anwendung der Methode erfordert die Abwägung zwischen Inkaufnahme eines erhöhten Fehlers 1. Art und erzielbarem Powergewinn.

Anmerkung

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Literatur

1. Grouven U, Beckmann L, Bender R, Lange S. Kriterien zur Überprüfbarkeit der Anwendung von Studienergebnissen [Präsentation]. In: IQWiG im Dialog; 2013 Jun 21; Köln. Köln: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG); 2013. Available from: https://www.iqwig.de/download/13-06-21_IQWiG_im_Dialog_Ulrich_Grouven_Kriterien_zur_Ueberpruefung_der_Anwendbarkeit_von_Studienergebnissen.pdf
2. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985 Jun;41(2):361-72. DOI: 10.2307/2530862

Korrespondenzadresse:
Lars Beckmann
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Im Mediapark 8, 50670 Köln, Deutschland
Lars.beckmann@iqwig.de

Bitte zitieren als
Beckmann L, Grouven U, Kieser M, Sieben W, Skipka G, Bender R. Untersuchung von Methoden zur Überprüfbarkeit von Ergebnissen von Studienpopulationen auf Teilpopulationen. GMS Med Inform Biom Epidemiol. 2018;14(2):Doc11.
DOI: 10.3205/mibe000189, URN: urn:nbn:de:0183-mibe0001896

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/mibe/2018-14/mibe000189.shtml

Veröffentlicht: *30.08.2018*