Kurzfassung

Im Rahmen des Generalauftrags wurde das Thema Untersuchung der statistischen Eigenschaften von Verfahren zur Übertragbarkeit von Studienergebnissen auf Teilpopulationen bearbeitet.

Hintergrund

Nutzenbewertungen des IQWiG haben zum Ziel, Nutzen und Schaden einer Prüfintervention im Vergleich zu einer Kontrollbehandlung zu bewerten. Es kann die Situation entstehen, dass für die Untersuchung einer konkreten Fragestellung in einer vorliegenden Studie 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 Nichtzielpopulation (nZP) zusammensetzt und ein nicht statistisch signifikanter Behandlungseffekt in der ZP, ein gleichgerichteter Behandlungseffekt in der nZP, ein statistisch signifikanter Behandlungseffekt in der SP sowie ein nicht statistisch signifikanter Interaktionstest ($p_{ia} \geq 0{,}05$) vorliegt. Zudem soll die Effektschätzung in der ZP nicht zu unpräzise sein im Vergleich zur nZP. Es stellt sich hier die Frage, unter welchen Umständen das Ergebnis der SP herangezogen werden kann. Eine mögliche Vorgehensweise bietet die sogenannte Erweiterungsregel (EWR).

Fragestellung

Die Anwendung einer mehrstufigen Testprozedur, die die EWR enthält, führt konstruktionsbedingt zu einer Niveauüberschreitung für den Test auf einen Effekt in der ZP. Folgende Fragestellungen sollen untersucht werden:

1) **Signifikanzniveau:** Es soll die Stärke der Niveauüberschreitung quantitativ untersucht werden. Es soll analysiert werden, wie stark einzelne Parameter in welchen Konstellationen die Niveauüberschreitung beeinflussen. Insbesondere sollen die Konstellationen identifiziert werden, die zu einem empirischen Fehler 1. Art von 10 % oder mehr führen. Ziel ist es, einfache Anforderungen an die Parameter(-konstellationen) zu formulieren, sodass eine Testprozedur mit entsprechend modifizierten Bedingungen hinsichtlich der Anwendung der EWR mit einer festgelegten maximalen Niveauüberschreitung angewendet werden kann.
2) **Power:** Sofern eine Formulierung der Anforderungen wie unter Punkt 1 beschrieben gelingt, soll der Powergewinn durch die Anwendung der modifizierten Testprozedur untersucht werden.
3) Es soll ein Vergleich der Testprozedur mit EWR mit weiteren alternativen Testprozeduren hinsichtlich Fehler 1. Art und Power durchgeführt werden.

Methoden

Folgendes Testproblem wird betrachtet:

H_0: $\theta_{ZP} = 0$ vs. H_1: $\theta_{ZP} \neq 0$,

wobei θ_{ZP} der wahre Effekt in der ZP ist. Das Effektmaß θ ist Cohen's d. Um die Nullhypothese H_0 zu testen, wird eine Testprozedur angewendet, die aus mehreren Schritten besteht. Die Anzahl der Schritte variiert je nach Konstellation der Ergebnisse. Die Testprozedur wird als Testprozedur mit EWR bezeichnet.

Die Testprozedur mit EWR wird mit alternativen Testprozeduren bezüglich des empirischen Fehlers 1. Art und der empirischen Power verglichen. Hierzu wird eine Simulationsuntersuchung durchgeführt. Der Wertebereich der untersuchten Simulationsparameter wurde so gewählt, dass praxisrelevante Szenarien abgebildet sind.

Ergebnisse

Simulationsuntersuchung

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 6667 als Trainingsdaten und die übrigen 3333 Szenarien als Testdaten verwendet wurden.

Die Anwendung der Testprozedur mit EWR führt konstruktionsbedingt zu einer Erhöhung des Fehlers 1. Art. Über alle Szenarien betrachtet war der empirische Fehler 1. Art auf den Trainingsdaten in 5,56 % der Szenarien größer als 10 %. Das 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art bei einem Signifikanzniveau von 5 % für den Interaktionstest war 12,0 %, d. h., in 2,5 % der simulierten Szenarien ist ein empirischer Fehler 1. Art größer als dieser Wert zu erwarten.

Alternative Testprozeduren

Modifizierte Testprozedur mit EWR ($EWR_{0.33}$)

Die Testprozedur mit EWR wurde optimiert, indem der Fehler 1. Art durch die Berücksichtigung des Verhältnisses der Stichprobengrößen in ZP und nZP (N_{ZP} / N_{nZP}) mit Cut-off 0,33 reduziert wurde.

Testprozedur mit bedingter Erhöhung des Signifikanzniveaus (AHR_{15})

Diese Testprozedur wird als Anhebungsregel bezeichnet und entspricht im Wesentlichen der Testprozedur mit EWR. Es wird jedoch statt der EWR ein Test auf einen Effekt in der ZP mit einem erhöhten Signifikanzniveau von 15 % angewendet.

Testprozedur mit alternativen Bedingungen (PInt und $PInt_{pmax}$)

Diese Testprozedur überträgt die Signifikanzaussage der SP auf die ZP, falls der p-Wert des Tests auf einen von 0 verschiedenen Effekt in der ZP kleiner ist als der p-Wert des Tests auf Heterogenität zwischen ZP und nZP. Diese Testprozedur wird mit PInt bezeichnet. Um den Fehler 1. Art zu reduzieren, wurde eine Modifikation der Testprozedur PInt untersucht, bei der der p-Wert für den Test in der ZP auf ein maximales Niveau beschränkt wird. Die Testprozedur wird mit $PInt_{pmax}$ bezeichnet, wobei sich für p_{max} ein Wert von 15 % ergab.

Test auf Effekt in der ZP zum Standard-Signifikanzniveau (A_5)

Um darstellen zu können, welche Vor- und Nachteile mit den genannten Testprozeduren einhergehen, wird auch das Standardvorgehen (also ein Test auf einen von 0 verschiedenen Effekt in der ZP mit einem Signifikanzniveau von 5 %) in den Vergleich der Testprozeduren mit einbezogen. Im Folgenden wird die so definierte Testprozedur mit A_5 bezeichnet.

Vergleich der Testprozeduren bzgl. Fehlers 1. Art und Power

Die folgende Tabelle 1 fasst die Ergebnisse zum empirischen Fehler 1. Art und zur empirischen Power der untersuchten Testprozeduren zusammen.

Tabelle 1: Empirischer Fehler 1. Art [%] (Testdaten) und mediane empirische Power [%] für die untersuchten Testprozeduren

Test-prozedur	**Fehler 1. Art**		**Mediane Power**		
	Median	**Maximum**	**alle Szenarien**	**ohne Szenarien mit 100 % Power in allen TP**	**Szenarien mit Power < 60 % in A_5**
$EWR_{0.33}$	5,52	12,45	100	80,8	30,7
AHR_{15}	6,09	10,92	100	84,3	35,5
$PInt_{15}$	5,91	10,89	100	83,5	35,1
A_5	5,04	6,18	100	78,5	26,3
AHR: Anhebungsregel; EWR: Erweiterungsregel					

Der mediane empirische Fehler 1. Art lag bei der Standard-Testprozedur A_5 erwartungsgemäß bei etwa 5 %, die übrigen Testprozeduren zeigten keine ausgeprägten Unterschiede und lagen mit dem Median im Bereich von 5,5 bis 6,1 %. Das Maximum des Fehlers 1. Art der Testprozeduren erreichte Werte von bis zu knapp über 12 %, während das Maximum bei Standardvorgehen (A_5) 6,2 % betrug.

Die Testprozedur AHR_{15} erzielt insgesamt eine leicht höhere empirische Power als die anderen Testprozeduren. Um Unterschiede bezüglich der empirischen Power näher zu untersuchen, wurden die Differenzen der empirischen Power der Testprozeduren $EWR_{0.33}$ und AHR_{15} zur Standardprozedur A_5 pro Szenarium in Abhängigkeit von der Power der Standardprozedur A_5 betrachtet (Tabelle 9). Deutlich höhere Powergewinne sind in Szenarien zu beobachten, in denen die Standardprozedur A_5 eine geringe Power hat. Für einzelne Szenarien zeigt sich ein Powergewinn von über 20 % im Vergleich zur Standardprozedur A_5. Insgesamt erwies sich die Testprozedur AHR_{15} als diejenige mit dem höchsten Powergewinn.

Fazit

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 bezüglich empirischer Power und Fehler 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.