IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

IQWiG-Berichte – Nr. 638

Untersuchung der statistischen Eigenschaften von Verfahren zur Übertragbarkeit von Studienergebnissen auf Teilpopulationen

Arbeitspapier

Auftrag: GA18-01
Version: 1.0
Stand: 20.06.2018

Impressum

Herausgeber:
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema:
Untersuchung der statistischen Eigenschaften von Verfahren zur Übertragbarkeit von Studienergebnissen auf Teilpopulationen

Auftraggeber:
Bearbeitung im Rahmen des Generalauftrags

Interne Auftragsnummer:
GA18-01

Anschrift des Herausgebers:
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Im Mediapark 8
50670 Köln

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
E-Mail: berichte@iqwig.de
Internet: www.iqwig.de

ISSN: 1864-2500

Für die Inhalte des Berichts ist allein das IQWiG verantwortlich.

Mitarbeiter des IQWiG

- Lars Beckmann
- Ralf Bender
- Ulrich Grouven
- Wiebke Sieben
- Guido Skipka

Schlagwörter: Simulationsstudie, Übertragbarkeit, Teilpopulation, Nutzenbewertung

Keywords: simulation study, transferability, subpopulation, benefit assessment

Inhaltsverzeichnis

Seite

Tabellenverzeichnis

Seite

Abbildungsverzeichnis

Seite

Abkürzungsverzeichnis

Abkürzung	**Bedeutung**
AHR	Anhebungsregel
EWR	Erweiterungsregel
G-BA	Gemeinsamer Bundesausschuss
HR	Hazard Ratio
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
MD	Mittelwertdifferenz
nZP	Nichtzielpopulation
p_{ia}	p-Wert des Interaktionstests zwischen ZP und nZP
RR	relatives Risiko
SE	Standard Error (Standardfehler)
SMD	standardisierte Mittelwertdifferenz
SP	Studienpopulation
TP	Teilpopulation
ZP	Zielpopulation

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

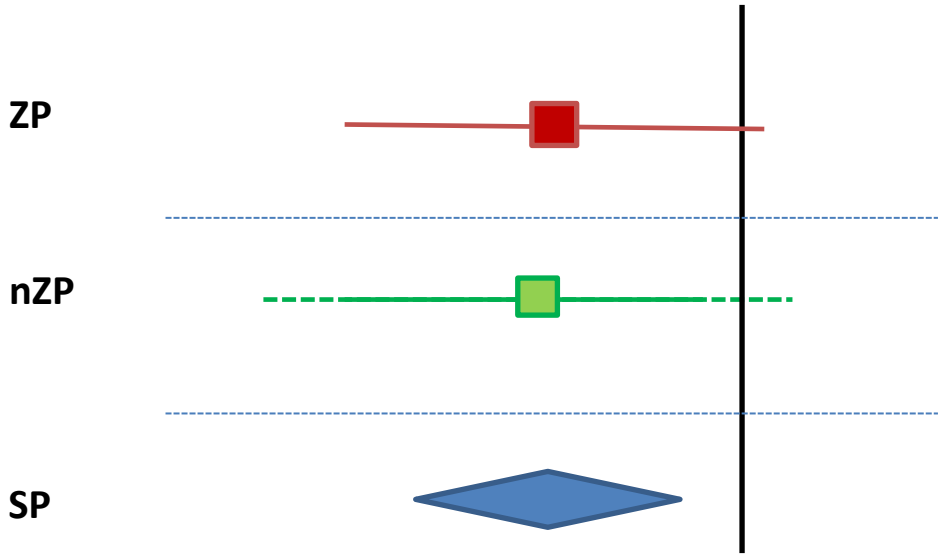
Berücksichtigung der Relation der Stichprobengrößen in ZP und nZP führte zwar zu einer Reduktion des empirischen Fehlers 1. Art, ein Vergleich mit alternativen, einfacheren Testprozeduren bezüglich empirischer Power und Fehler 1. Art ließ jedoch insgesamt keine Vorteile erkennen. Unter Berücksichtigung 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.

1 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 sich eine Datenkonstellation wie in Abbildung 1 dargestellt ergibt, d. h., es liegen 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 ($p_{ia} \geq 0{,}05$) vor. Zudem soll die Effektschätzung in der ZP nicht zu unpräzise sein im Vergleich zur nZP. Es stellt sich hier 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 $\alpha = 0{,}05$ allein ist nicht hinreichend, um hieraus eine Aussage im Sinne der Gleichheit von Effekten abzuleiten und damit zu begründen, dass Aussagen zu einer ZP durch Heranziehen der Ergebnisse der gesamten SP getroffen werden. So kann es trotz eines nicht statistisch signifikanten Interaktionstests zu Situationen kommen, in denen zwischen TPs relevant unterschiedliche Effekte geschätzt werden. Dies bedeutet, dass eine qualitative Interaktion zwischen der interessierenden ZP und der nZP, d. h. kein oder gegenläufiger Effekt in der Punktschätzung zwischen ZP und nZP, mit ausreichender Sicherheit ausgeschlossen werden muss, um das Ergebnis der SP auf die ZP übertragen zu können [1].

Eine mögliche Vorgehensweise bietet die sogenannte Erweiterungsregel (EWR), die in Abschnitt 4.2 und in Anhang A näher beschrieben wird.



HR: Hazard Ratio; KI: Konfidenzintervall; MD: Mittelwertdifferenz; nZP: Nichtzielpopulation; RR: relatives Risiko; SMD: standardisierte Mittelwertdifferenz (Hedges' g); SP: Studienpopulation (= ZP + nZP); ZP: Zielpopulation

Abbildung 1: Notwendige Datenkonstellation für die Anwendung der untersuchten Testprozeduren. Dargestellt ist das betrachtete Effektmaß in der jeweiligen Population (z. B. RR, HR, MD oder SMD mit zugehörigem 95 %-KI, vertikaler Strich stellt Nulleffekt dar, d. h. RR, HR = 1 bzw. MD, SMD = 0).

2 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.

3 Projektverlauf

Die vorliegende Untersuchung wurde im Rahmen des Generalauftrags bearbeitet. Bearbeitungsbeginn war der 11.04.2016. Dieser Bericht wurde an den G-BA übermittelt und 4 Wochen später auf der Website des IQWiG veröffentlicht.

4 Methoden

4.1 Hypothese und Testprozedur

Hypothese

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. Konkret ist die Testprozedur folgendermaßen definiert:

Testprozedur

Schritt 1: Es wird 2-seitig getestet, ob für die ZP ein statistisch signifikanter Effekt zum Niveau 0,05 vorliegt.

- Falls ja: H_0 wird abgelehnt und die Testprozedur stoppt.
- Falls nein: Führe Schritt 2 durch.

Schritt 2: Es wird 2-seitig getestet, ob für die SP ein statistisch signifikanter Effekt zum Niveau 0,05 vorliegt.

- Falls ja: Führe Schritt 3 durch.
- Falls nein: H_0 wird nicht abgelehnt und die Testprozedur stoppt.

Schritt 3: Es wird geprüft, ob die Effektschätzungen von ZP und nZP dieselbe Effektrichtung aufweisen.

- Falls ja: Führe Schritt 4 durch.
- Falls nein: H_0 wird nicht abgelehnt und die Testprozedur stoppt.

Schritt 4: Es wird getestet, ob zwischen ZP und nZP eine statistisch signifikante Interaktion zum Niveau 0,05 vorliegt.

- Falls ja: H_0 wird nicht abgelehnt und die Testprozedur stoppt.
- 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 0 verschieden. Unter der Annahme, dass der wahre Effekt in der ZP gleich 0 ist (H_0) und der wahre Effekt der nZP seiner beobachteten Schätzung entspricht, wird die Wahrscheinlichkeit dafür bestimmt, mindestens so homogene Ergebnisse zwischen ZP und

nZP (gemessen an der Q-Statistik) und mindestens einen so großen Effekt für die ZP zu erhalten wie beobachtet. Diese Wahrscheinlichkeit wird mithilfe der weiter unten beschriebenen EWR geschätzt und resultiert in einem empirischen p-Wert. Es wird geprüft, ob diese Wahrscheinlichkeit kleiner als 0,025 ist (entspricht dem üblichen Signifikanzniveau eines 1-seitigen Tests).

- Falls ja: H_0 wird abgelehnt.
- Falls nein: H_0 wird nicht abgelehnt.

Die beschriebene Testprozedur wird im Folgenden als Testprozedur mit EWR bezeichnet.

4.2 Erweiterungsregel (EWR)

Die EWR beinhaltet die Simulation eines empirischen p-Wertes. Ist der empirische p-Wert kleiner als 0,025, 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 genaue Vorgehen bei der Anwendung der EWR wird in Anhang A dargestellt.

4.3 Alternative Testprozeduren

Um die Eignung der in Abschnitt 4.1 und 4.2 beschriebenen Testprozedur mit EWR zu beurteilen, wird bezüglich des empirischen Fehlers 1. Art und der empirischen Power ein Vergleich mit alternativen Testprozeduren durchgeführt:

- Modifizierte Testprozedur mit EWR (siehe Abschnitt 5.2.1)
- Testprozedur mit bedingter Erhöhung des Signifikanzniveaus:

  Durchführung der Schritte 1 bis 4 der in Abschnitt 4.1 beschriebenen Testprozedur, jedoch in Schritt 5 keine Anwendung der EWR, sondern Test auf einen Effekt in der ZP mit erhöhtem Signifikanzniveau (siehe Abschnitt 5.2.2)
- Alternative Testprozedur mit und ohne Modifikation (siehe Abschnitt 5.2.3). Das Niveau für die letztgenannten Testprozeduren wird jeweils auf einen Wert angehoben, sodass sich ein Fehler 1. Art ergibt, der nur selten über 10 % liegt (siehe Abschnitt 5.1).
- 2-seitiger Test auf Effekt in der ZP zum Signifikanzniveau $\alpha = 0{,}05$ (Standardvorgehen, siehe Abschnitt 5.2.4)

4.4 Simulationsuntersuchungen

Im Rahmen von Simulationsuntersuchungen werden empirische Fehler 1. Art und Power der Anwendung der im Abschnitt 4.3 beschriebenen Testprozeduren untersucht. Es ist zu beachten, dass der Gegenstand der im Folgenden beschriebenen Simulationsuntersuchungen die Anwendung der gesamten Testprozedur ist. Davon abzugrenzen ist die Simulation des

empirischen p-Wertes im Rahmen der EWR, die Teil der Methodik der EWR ist (siehe Anhang A).

Tabelle 2 zeigt die geplanten Szenarien für die Simulationsuntersuchungen. Als Effektmaß wird die standardisierte Mittelwertdifferenz SMD (Cohen's d) verwendet. 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. Das Vorgehen zur Erzeugung der simulierten Daten für die betrachteten Szenarien entspricht den Schritten 1 bis 3 der im Abschnitt 4.2 beschriebenen Simulationen für die Erweiterungsregel.

Tabelle 2: Szenarien für die Simulationsuntersuchungen

Parameter	**Beschreibung**	**Werte für Simulation**	**Anzahl Werte**
Stichproben-größe N_{nZP}	Stichprobengröße pro Therapiearm in nZP bzw. ZP	50 / 100 / 200 / 500 / 750 / 1000	6
N_{ZP} / N_{nZP}	Verhältnis Stichprobengröße in ZP zu Stichprobengröße in nZP	0,2 / 0,33 / 0,5 / 0,75 / 1,0 / 1,5 / 2,0 / 3,0 / 5,0	9
θ_{nZP}	wahrer Effekt (SMD) in nZP	0 / −0,1 / −0,2 / −0,3 / −0,4 / −0,5 / −0,6 / −0,7 / −0,8 / −0,9 / −1,0	11
n_{rep}	Anzahl Replikationen für Simulation des empirischen p-Werts bei der EWR		100 000
Fehler 1. Art			
θ_{ZP}	wahrer Effekt (SMD) in ZP	0	1
n_{sim}	Anzahl Replikationen pro Szenario		10 000
Power			
θ_{ZP}	wahrer Effekt (SMD) 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 Szenario		10 000

EWR: Erweiterungsregel; nZP: Nichtzielpopulation; SMD: standardisierte Mittelwertdifferenz; ZP: Zielpopulation

5 Ergebnisse

5.1 Simulationsuntersuchungen der Testprozedur mit EWR bzgl. 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 6667 als Trainingsdaten und die übrigen 3333 Replikationen 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 % (Abbildung 2). Auch wenn Mittelwert und Median des Fehlers 1. Art mit 6,31 % und 5,70 % im Vergleich zum Standardvorgehen mit einem Test zum Niveau von 5 % nur leicht erhöht sind, gibt die Häufigkeit einer großen Niveauüberschreitung Anlass, den Einsatz der EWR auf solche Szenarien beschränken zu wollen, in denen nicht (oder nur sehr selten) mit einem Fehler 1. Art von mehr als 10 % zu rechnen ist.

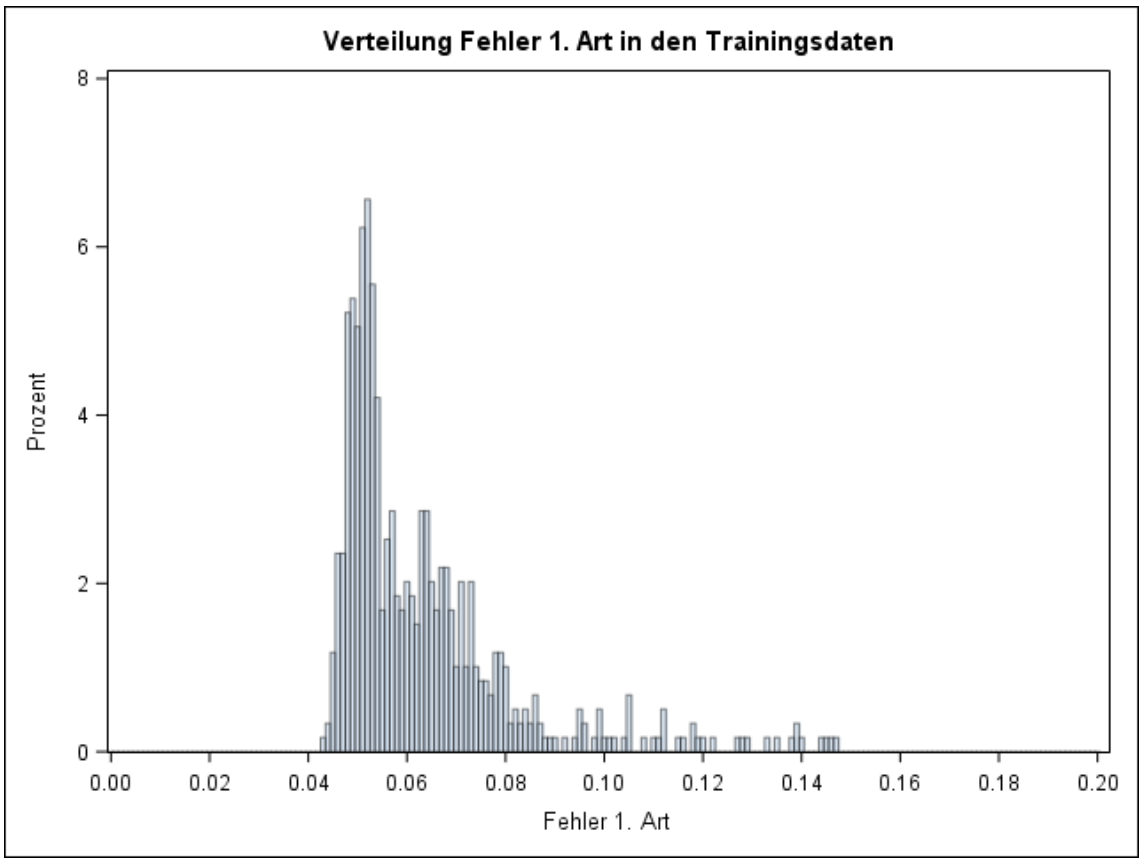


Abbildung 2: Verteilung des empirischen Fehlers 1. Art für die EWR über alle Szenarien (Trainingsdatensatz)

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.

Eine Erhöhung des Signifikanzniveaus des Interaktionstests auf Werte von 20, 30, 40 und 50 % lieferte 97,5 %-Quantile von vergleichbarer Größenordnung. Aus diesem Grund werden in diesem Arbeitspapier nur noch die Ergebnisse zum üblichen Signifikanzniveau von 5 % präsentiert.

Nachfolgend wird untersucht, ob einfache Anforderungen an Parameter(-konstellationen) gestellt werden können, sodass der empirische Fehler 1. Art der Testprozedur mit EWR nur in Ausnahmefällen über 10 % liegt. Konkret wird nach einer einfachen Bedingung an 1 oder 2 der Simulationsparameter gesucht, sodass das 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art kleiner als 10 % ist. Auf diese Weise kann erreicht werden, dass nur in 2,5 % der simulierten Szenarien ein empirischer Fehler 1. Art von mehr als 10 % zu erwarten ist.

5.2 Definition und Herleitung alternativer Testprozeduren

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

Untersuchung des Zusammenhangs zwischen Parameter(-konstellationen) und Niveauüberschreitung

Mithilfe grafischer Darstellungen des Fehlers 1. Art der Testprozedur mit EWR wird der Zusammenhang zwischen dem empirischen Fehler 1. Art und den Parametern

- Effekt in der nZP (θ_{nZP}),
- Stichprobengröße in ZP und nZP und
- Relation der Stichprobengröße in ZP und nZP (N_{ZP} / N_{nZP})

untersucht (siehe Abbildung 3 bis Abbildung 8 in Anhang B).

Die Relation der Stichprobengrößen in ZP und nZP erweist sich als einfache und praktikable Lösung, um Szenarien zu identifizieren, die nur in seltenen Fällen einen empirischen Fehler 1. Art größer als 10 % aufweisen (siehe Abbildung 8 in Anhang B).

Die Hinzunahme einer weiteren Variablen brachte keine bedeutsame Verbesserung der Identifikation von Szenarien mit häufiger erhöhtem empirischen Fehler 1. Art.

Festlegung des Cut-offs für die Relation der Stichprobengrößen

Mit fallendem Wert der Relation der Stichprobengrößen ist mit einem zunehmend häufig deutlich erhöhten empirischen Fehler 1. Art zu rechnen, sodass die EWR dann nicht mehr angewendet werden sollte. Es ist ein Cut-off so zu bestimmen, 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 3 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 größer gleich 0,33 ist, so haben weniger als 2,5 % der Szenarien einen empirischen Fehler 1. Art von über etwa 10 %. Bei einem Cut-off von 0,2 hätten mehr als 2,5 % der Szenarien einen empirischen Fehler 1. Art von über 12 %; daher wird 0,33 als Cut-off für die Einschränkung der Anwendung der EWR gewählt.

Tabelle 3: 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art der Testprozedur mit EWR für verschiedene Cut-offs für die Relation der Stichprobengrößen (Trainingsdaten)

Szenarien mit Relation der Stichprobengröße ≥	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	6,15

Für die Testdaten bestätigt sich, dass der Anteil der Szenarien mit einem empirischen Fehler 1. Art über 10 % angemessen beschränkt werden kann, wenn die EWR nur dann angewendet wird, wenn die Relation der Stichprobengrößen größer oder gleich 0,33 ist. Hier beträgt das 97,5 %-Quantil der Verteilung des Fehlers 1. Art 10,17 %.

Zusammenfassend beinhaltet die modifizierte Testprozedur mit EWR folgende Bedingungen für die Anwendung der EWR:

- kein statistisch signifikanter Effekt in ZP,
- statistisch signifikanter Effekt in SP,
- ZP und nZP sind homogen ($p_{int} \geq 0{,}05$) und
- Relation der Stichprobengrößen: $N_{ZP} \,/\, N_{nZP} \;\geq 0{,}33$.

Im Folgenden wird diese Testprozedur mit $EWR_{0.33}$ bezeichnet.

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

Als weitere Alternative wird eine Testprozedur betrachtet, die die Anwendung der Schritte 1 bis 4 der Testprozedur wie in Abschnitt 4.1 beschrieben beinhaltet, jedoch in Schritt 5 der Testprozedur nicht die EWR anwendet, sondern stattdessen einen Test auf einen Effekt in der ZP mit einem erhöhten Signifikanzniveau. Diese Testprozedur wird als Anhebungsregel bezeichnet und im Folgenden mit AHR_{15} abgekürzt. Die Erhöhung des Signifikanzniveaus wird dabei so festgelegt, dass auch für diese Testprozedur das 97,5 %-Quantil der Verteilung des Fehlers 1. Art für die Trainingsdaten kleiner ist als 10 %. Nachfolgender Tabelle 4 kann entnommen werden, dass dies bei einem Niveau von etwa 15,0 % 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: 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art der Testprozedur AHR_{15} für verschiedene Signifikanzniveaus des Tests auf einen von 0 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

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

Es wurde außerdem eine weitere Testprozedur untersucht, die die Signifikanzaussage der SP auf die ZP überträgt, wenn folgende Bedingungen gelten:

- Der Effekt in ZP ist nicht statistisch signifikant,
- der Effekt in SP ist statistisch signifikant,
- die Effekte in ZP und nZP sind gleichgerichtet und
- der p-Wert des Tests auf einen von 0 verschiedenen Effekt in der ZP ist kleiner als der p-Wert des Interaktionstests (p_{ia}) zwischen ZP und nZP.

Diese Testprozedur wird mit PInt bezeichnet.

Für die Testprozedur PInt ergibt sich ein durchschnittlicher Fehler 1. Art von 8,25 %. Das 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art liegt bei 21,49 %. Diese Testprozedur ist damit aufgrund des hohen Fehlers 1. Art ohne weitere Modifizierungen nicht geeignet. Es wurde daher eine Modifikation der Testprozedur PInt untersucht, bei der der p-Wert für den Test in der ZP auf ein maximales Niveau

$$p_{neu} = \min (p_{ia}, p_{max})$$

beschränkt wird. p_{max} wird wiederum so festgelegt, dass das 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art maximal 10 % beträgt (Tabelle 5). Die Testprozedur wird mit $PInt_{pmax}$ bezeichnet.

Tabelle 5: 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art der Testprozedur $PInt_{pmax}$ für verschiedene Signifikanzniveaus des Tests auf einen von 0 verschiedenen Effekt in der ZP (Trainingsdaten)

Signifikanzniveau p_{max}	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	7,71	8,20	8,68	9,07	9,84	9,84	10,03

Es ergibt sich für die Testprozedur $PInt_{pmax}$ ein 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art von 10,03 %, wenn p_{max} auf 15 % gesetzt wird. Für die Testdaten ergibt sich für $PInt_{pmax}$ ein 97,5 %-Quantil von 10,05 % für die Verteilung des empirischen

Fehlers 1. Art. Diese Testprozedur kommt damit für einen Powervergleich infrage und wurde weiter untersucht.

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

Um darstellen zu können, welche Vor- und Nachteile mit den 3 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 folgenden Vergleich der Testprozeduren mit einbezogen. Im Folgenden wird die so definierte Testprozedur mit A_5 bezeichnet.

5.2.5 Zusammenfassung der Ergebnisse zur Untersuchung des Fehlers 1. Art

Die folgende Tabelle 6 fasst die Simulationsergebnisse zum empirischen Fehler 1. Art zusammen. Die für den Trainingsdatensatz ermittelten Werte werden für den Testdatensatz bestätigt.

Tabelle 6: Empirischer 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_{15}	Trainingsdaten	6,69	6,19	10,15	10,90
	Testdaten	6,70	6,09	10,23	10,92
$PInt_{15}$	Trainingsdaten	6,52	5,95	10,03	10,74
	Testdaten	6,53	5,91	10,05	10,89
A_5	Trainingsdaten	5,04	5,04	5,58	5,85
	Testdaten	5,04	5,04	5,85	6,18
AHR: Anhebungsregel; EWR: Erweiterungsregel					

Eine grafische Darstellung der Verteilung des empirischen Fehlers 1. Art der untersuchten Testprozeduren für Trainings- und Testdatensatz findet sich in Abbildung 9 in Anhang B.).

5.3 Vergleich der Testprozeduren bzgl. der Power

Die Größenordnung des Fehlers 1. Art der Testprozedur mit EWR (ohne zusätzliche Bedingungen) sowie der Testprozedur PInt erwies sich in den Simulationsuntersuchungen (Abschnitt 5.1 und 5.2.3) als inakzeptabel hoch. Im Folgenden werden diese Testprozeduren daher nicht weiter betrachtet. Die folgenden Vergleiche beziehen sich auf die Testprozeduren, $EWR_{0.33}$, AHR_{15}, $PInt_{15}$ und A_5. Für alle Szenarien wurden jeweils 10 000 Replikationen durchgeführt.

Die auf Basis der Simulationen berechnete empirische Power unterscheidet sich zwischen den Testprozeduren über alle Szenarien betrachtet kaum. Nimmt man keine Änderung am

Standardvorgehen vor und testet in der Zielpopulation zu einem Niveau von 5 %, ob der Effekt von 0 verschieden ist (A_5), so ergibt sich eine durchschnittliche empirische Power von 82,9 %. Um Unterschiede bezüglich der empirischen Power näher zu untersuchen, wurden diejenigen 49 % der Szenarien, in denen alle Testprozeduren eine empirische Power von 100 % aufwiesen, in allen folgenden Auswertungen nicht weiter betrachtet, da diese Szenarien keine Information zum Unterschied der Testprozeduren liefern. Abbildung 3 zeigt die Verteilung der empirischen Power ohne diese Szenarien.

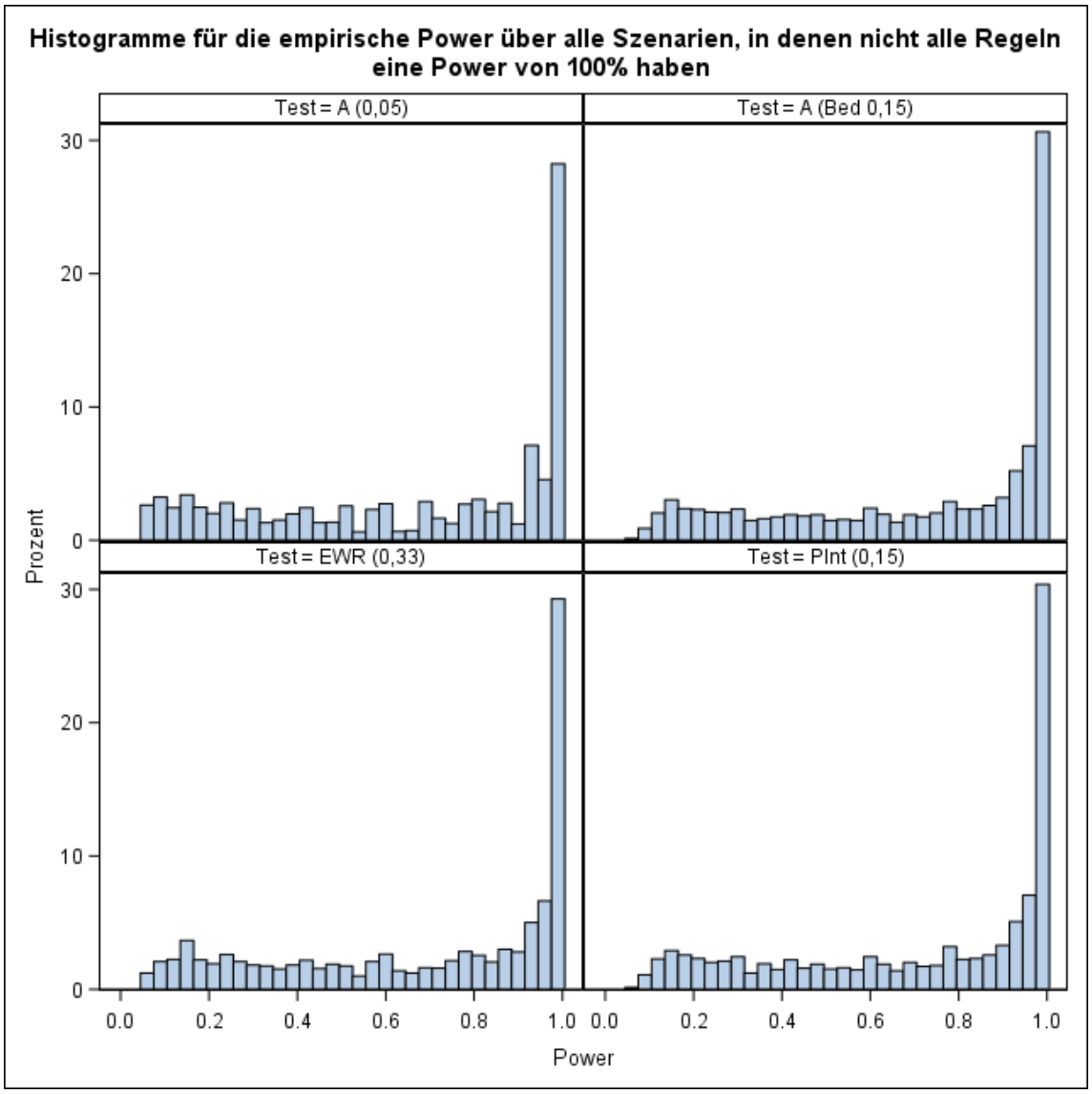
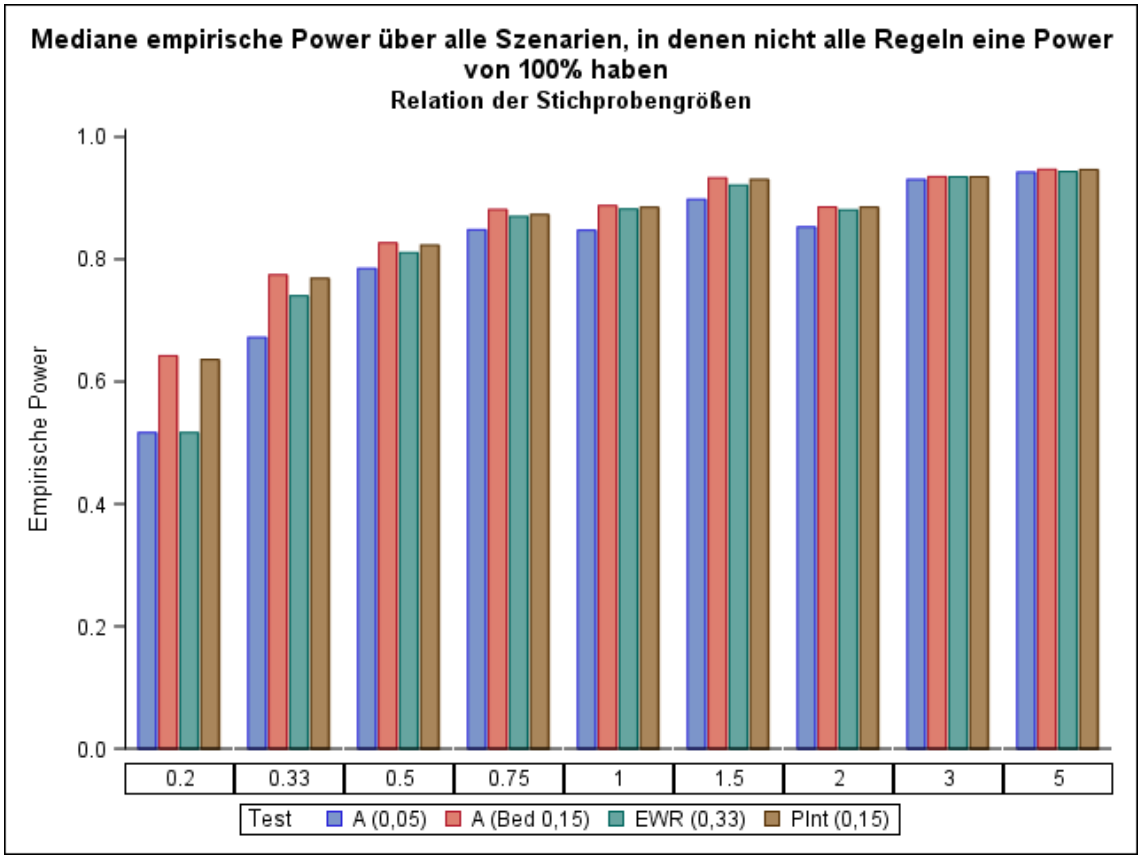


Abbildung 3: Empirische Power der betrachteten Testprozeduren ohne die Szenarien, in denen alle Testprozeduren eine empirische Power von 100 % aufweisen



AHR: Anhebungsregel; EWR: Erweiterungsregel

Abbildung 4: Empirische Power für verschiedene Relationen der Stichprobengrößen N_{ZP} / N_{nZP}

Die medianen Werte der empirischen Power für die eingeschränkten Szenarien betragen 80,8 % für die Testprozedur $EWR_{0.33}$, 83,5 % für die Testprozedur $PInt_{15}$ und 84,4 % für Testprozedur AHR_{15}. Der größte Powergewinn im Vergleich zur Standardprozedur A_5 mit einer medianen empirischen Power von 78,5 % ergibt sich mit 5,9 Prozentpunkten für die Testprozedur AHR_{15}. Die empirische Power der Testprozeduren $EWR_{0.33}$, AHR_{15} und $PInt_{15}$ unterscheidet sich für Relationen der Stichprobengrößen von ZP zu nZP von größer oder gleich 0,33 wenig. Insgesamt erzielt die Testprozedur AHR_{15} eine leicht höhere empirische Power als die anderen Testprozeduren.

Die Ergebnisse zur empirischen Power sind in folgender Tabelle 7 zusammengefasst:

Tabelle 7: Empirische Power [%] für die untersuchten Testprozeduren

	alle Szenarien		**Szenarien, in denen nicht alle Test-prozeduren eine Power von 100 % haben**	
Testprozedur	**Mittelwert**	**Median**	**Mittelwert**	**Median**
$EWR_{0.33}$	84,1	100	68,9	80,8
AHR_{15}	85,3	100	71,2	84,4
$PInt_{15}$	85,1	100	70,9	83,5
A_5	82,9	100	66,5	78,5
AHR: Anhebungsregel; EWR: Erweiterungsregel				

Auswahl der Testprozedur mit den günstigsten Eigenschaften

Über alle Szenarien betrachtet fällt der Powergewinn im Vergleich zum Standardvorgehen nur gering aus. Ein deutlicherer Powergewinn ergibt sich in den Szenarien, bei denen die Power der Standardtestprozedur A_5 geringer ist. Tabelle 8 und Abbildung 10 in Anhang B zeigen, dass sich bei verschiedenen oberen Grenzwerten für die mediane Power der Standardprozedur A_5 ein konsistentes Bild für den Powergewinn der betrachteten Testprozeduren ergibt. Die Testprozedur $EWR_{0.33}$ zeigt durchweg eine niedrigere mediane empirische Power als die Testprozedur AHR_{15}.

Tabelle 8: Empirische Power [%] für die untersuchten Testprozeduren für Szenarien mit geringer Power bei A5

Szenarien mit Power A_5 < C	**Mediane Power**		
	C = 60 %	**C = 50 %**	**C = 40 %**
Anzahl (%)	1174 (19,8)	1005 (16,9)	816 (13,7)
Testprozedur			
$EWR_{0.33}$	30,7	26,4	22,8
AHR_{15}	35,5	30,4	26,3
$PInt_{15}$	35,1	29,8	25,7
A_5	26,3	22,8	17,8
AHR: Anhebungsregel; EWR: Erweiterungsregel			

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). Dabei sind deutliche Powergewinne in Szenarien zu beobachten, in denen die Standardprozedur A_5 eine geringe Power hat. Für einzelne Szenarien zeigt sich ein Powergewinn von über 20 %. Insgesamt erwies sich die Testprozedur AHR_{15} als diejenige mit dem höchsten Powergewinn.

Tabelle 9: Verteilung der Differenzen der empirischen Power der Testprozeduren $EWR_{0.33}$ und AHR_{15} zur Standardprozedur A5 pro Szenarium

					Differenzen der Power zu A_5 pro Szenarium [%-Punkte]					
		Mediane empirische Power [%]			**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	< 0,1	< 0,1	< 0,1	< 0,1	< 0,1	< 0,1
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

Die folgende Tabelle 10 fasst die Ergebnisse zum empirischen Fehler 1. Art und zur empirischen Power der untersuchten Testprozeduren zusammen.

Tabelle 10: Empirischer Fehler 1. Art [%] (Testdaten) und empirische Power [%] für die untersuchten Testprozeduren

Test-prozedur	**Fehler 1. Art**		**Mediane Power**		
	Median	**Maximum**	**alle Szenarien**	**ohne Szenarien mit 100 % Power in allen Testprozeduren**	**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					

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 (Rechenaufwand) erweist sich die Anhebungsregel (Testprozedur AHR_{15}) als das geeignetste Verfahren.

6 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 führt naturgemäß zu einer reduzierten Power zur Aufdeckung eines vorhandenen Behandlungseffekts. 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 Szenarien (Parameterkonstellationen) eine nicht akzeptable Niveauüberschreitung. Es wurde daher untersucht, ob einfache Anforderungen an Parameter(-konstellationen) gestellt werden können, sodass der empirische Fehler 1. Art der Testprozedur mit EWR reduziert wird und nur in Ausnahmefällen über 10 % liegt. Konkret wurde nach einer einfachen Bedingung an 1 oder 2 der untersuchten Parameter gesucht, sodass das 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art kleiner als 10 % ist. Die Relation der Stichprobengrößen erwies sich als einfache und praktikable Lösung. Darauf basierend wurde eine modifizierte Testprozedur mit EWR ($EWR_{0.33}$) definiert.

In den vorliegenden Auswertungen wurde als Grenze für das 97,5 %-Quantil der Verteilung des empirischen Fehlers 1. Art ein Wert von 10 % gewählt. Dieser Wert wurde pragmatisch festgelegt. Abbildung 11 zeigt, dass der Zusammenhang zwischen Grenzwert und mittlerer bzw. medianer empirischer Power einer monotonen Kurve folgt. Die Wahl einer strengeren Grenze für den Fehler 1. Art von z. B. 8 % führt zu einer alternativen Testprozedur mit einem Signifikanzniveau von 10 % für den Test auf einen Effekt in der ZP (AHR_{10}). Abbildung 10 zeigt, dass die Testprozedur AHR_{10} im Vergleich zu $EWR_{0.33}$ – trotz niedrigerem Fehler 1. Art – eine leicht höhere Power aufweist. Die Wahl der konkreten Testprozedur ist abhängig von der (subjektiven) Abwägung von akzeptablem Fehler 1. Art und möglichem Powergewinn.

Die Anwendung der $EWR_{0.33}$ wurde mit 2 alternativen Testprozeduren mit bedingter Erhöhung des Signifikanzniveaus (AHR_{15} und $PInt_{15}$) sowie mit dem Standardvorgehen A_5 hinsichtlich empirischen Fehlers 1. Art und empirischer Power verglichen.

Bei Betrachtung der Szenarien, in denen nicht alle Testprozeduren eine empirische Power von 100 % aufweisen, ließ sich ein medianer Powergewinn von 2,3 Prozentpunkten ($EWR_{0.33}$), 5,0 Prozentpunkten ($PInt_{15}$) und 5,9 Prozentpunkten (AHR_{15}) im Vergleich zum Standardvorgehen A_5 erreichen. 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 zu A_5. Insgesamt erwies sich die Testprozedur AHR_{15} als diejenige mit dem höchsten Powergewinn.

Zusammenfassend zeigte die Testprozedur $EWR_{0.33}$ weder bezüglich der empirischen Power noch bezüglich des empirischen Fehlers 1. Art Vorteile gegenüber den alternativen Testprozeduren, die ihren Einsatz in Anbetracht des erhöhten Rechenaufwands rechtfertigen würden. In der Abwägung von Einbußen beim empirischen Fehler 1. Art, Zugewinn bei der empirischen Power sowie Praktikabilität erweist sich die Anhebungsregel (Testprozedur AHR_{15}) als das Verfahren der Wahl.

Als Effektmaß im Rahmen der vorliegenden Untersuchungen wurde Cohen's d gewählt. Die vorgestellten Verfahren können mit entsprechenden Verteilungsananahmen auch auf weitere Effektmaße wie das relative Risiko, das Odds Ratio oder das Hazard Ratio angewendet werden.

Limitationen der Untersuchungen

Die Niveauüberschreitung der Testprozedur AHR_{15} ist anpassbar durch die Wahl des Niveaus, zu dem bei erfüllten Einstiegsbedingungen auf einen Effekt in der ZP getestet wird. Wird, wie hier untersucht, das Niveau 15 % gewählt, so ist dies auch der maximal mögliche erwartete empirische Fehler 1. Art der Testprozedur. Durch die Auswahl der Szenarien der Simulation betrug der empirische Fehler 1. Art im Testdatensatz im Median nur 6,09 %. Wären andere Szenarien gewählt worden, hätte sich hier ein anderer Wert ergeben können. 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 im Vorfeld versucht, unrealistische Szenarien von vorneherein auszuschließen. Nur eine Gewichtung der Szenarien gemäß ihren zu erwartenden Auftrittswahrscheinlichkeiten hätte dieses Problem tatsächlich lösen können. Es hätte hierzu bekannt sein müssen, welche Parameterkonstellationen in der Realität (d. h. in den Bewertungen des IQWiG) wie häufig auftreten. Diese Verteilung der Simulationsparameter ist jedoch unbekannt und kann auch nicht mit ausreichender Sicherheit geschätzt werden.

7 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.

8 Literatur

1. Grouven U, Beckmann L, Bender R, Lange S. Kriterien zur Überprüfbarkeit der Anwendung von Studienergebnissen [online]. In: IQWiG im Dialog 2013: Bedeutung der Zulassung für die Nutzenbewertung; 21.06.2013; Köln, Deutschland. [Zugriff: 15.05.2018]. URL: https://www.iqwig.de/download/13-06-21_IQWiG_im_Dialog_Ulrich_Grouven_Kriterien_zur_Ueberpruefung_der_Anwendbarkeit_von_Studienergebnissen.pdf.

Anhang A – Beschreibung der Anwendung der EWR

Die EWR beinhaltet die Simulation eines empirischen p-Wertes. Das genaue Vorgehen der Anwendung der EWR wird im Folgenden dargestellt. Als Effektmaß wird die standardisierte Mittelwertdifferenz SMD (Cohen's d, θ) betrachtet. Im Folgenden wird davon ausgegangen, dass die geschätzten SMD kleiner als 0 sind.

Es wird ausgegangen von individuellen Patientendaten X_{ij}, i = 1, 2; j = 1,…,n_i mit

$$X_{ij} \sim N(\mu_i, \delta^2) \text{ und } \theta = \frac{\mu_1 - \mu_2}{\delta},$$

wobei $N(\mu, \delta^2)$ die Normalverteilung mit Erwartungswert μ und Varianz δ^2 bezeichnet.

Eine Schätzung der SMD θ und zugehörigem Standardfehler ist gegeben durch

$$\hat{\theta} = \frac{m_1 - m_2}{S_{pool}} \text{ mit } S_{pool} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \text{ und}$$

$$SE(\hat{\theta}) = \sqrt{\frac{(n_1+n_2)}{n_1 n_2} + \frac{\hat{\theta}^2}{2(n_1+n_2)-4}}$$

wobei m_1 und m_2 die geschätzten Mittelwerte in den Therapiearmen einer Studienpopulation und S_1^2 und S_2^2 die entsprechenden Varianzen sind.

Es gilt

$$\hat{\theta}\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sim T_{n_1+n_2-2;ncp} \text{ mit } ncp = \theta\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

$$\hat{\theta} \overset{approx}{\sim} N\left(\theta, \frac{n_1+n_2}{n_1 n_2} + \frac{\theta^2}{2(n_1+n_2)-4}\right),$$

wobei $T_{n_1+n_2-2;ncp}$ die nicht zentrale *t*-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden und Nichtzentralitätsparameter *ncp* bezeichnet.

Für beobachtete Werte [θ_{ZP}^{beob}, SE_{ZP}^{beob}] und [θ_{nZP}^{beob}, SE_{nZP}^{beob}] in den Teilpopulationen ZP und nZP (mit $n_{i,ZP}$ und $n_{i,nZP}$ als Fallzahlen der 2 Gruppen in ZP bzw. nZP) werden, unter den in Kapitel 1 aufgeführten Bedingungen, 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.

2) Aus den Angaben kann in beiden Populationen die SMD mit zugehörigem Standardfehler geschätzt werden:

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

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

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

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

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

4) Überprüfung:

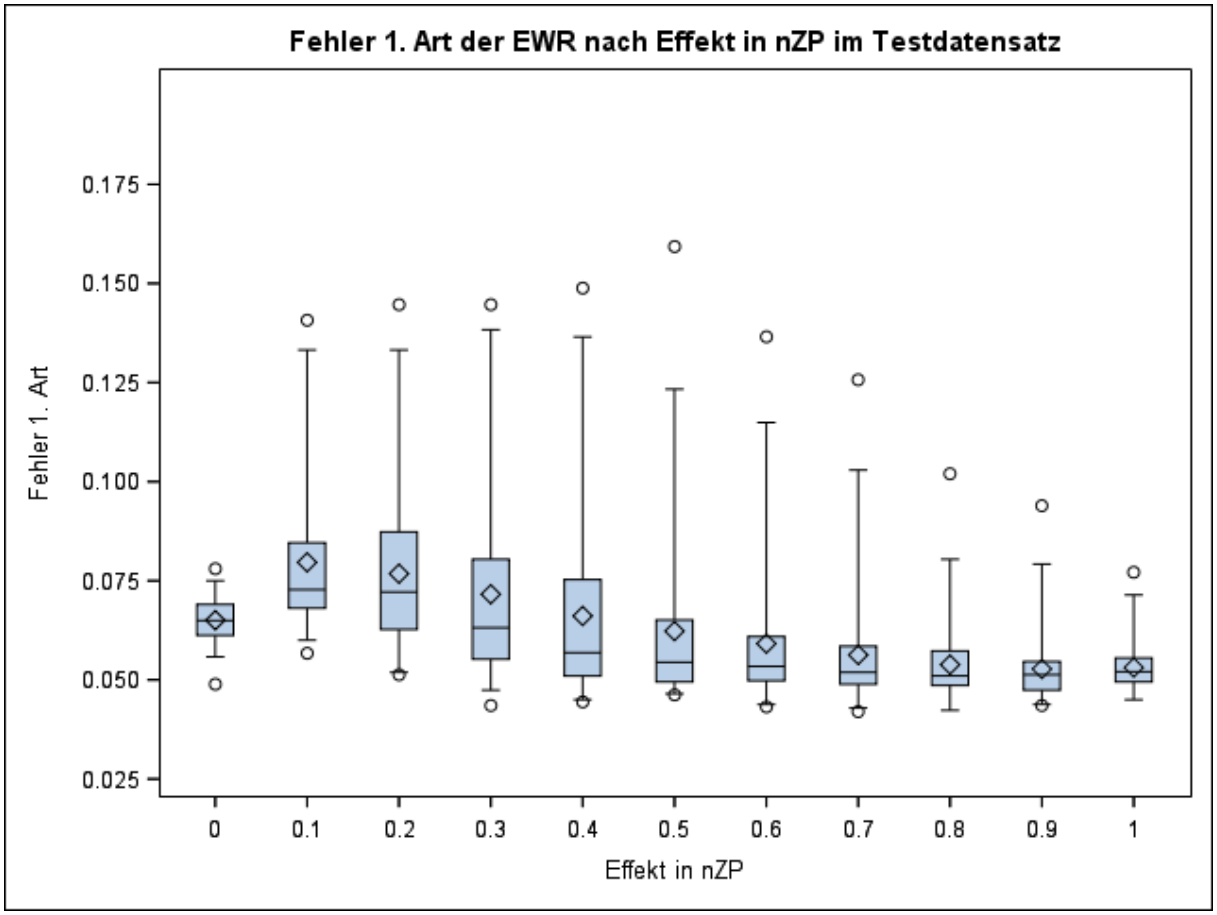
   $\theta_{ZP}^{rand} \leq \theta_{ZP}^{beob}$

   und

   $P_{int}^{rand} \geq P_{int}^{beob}$

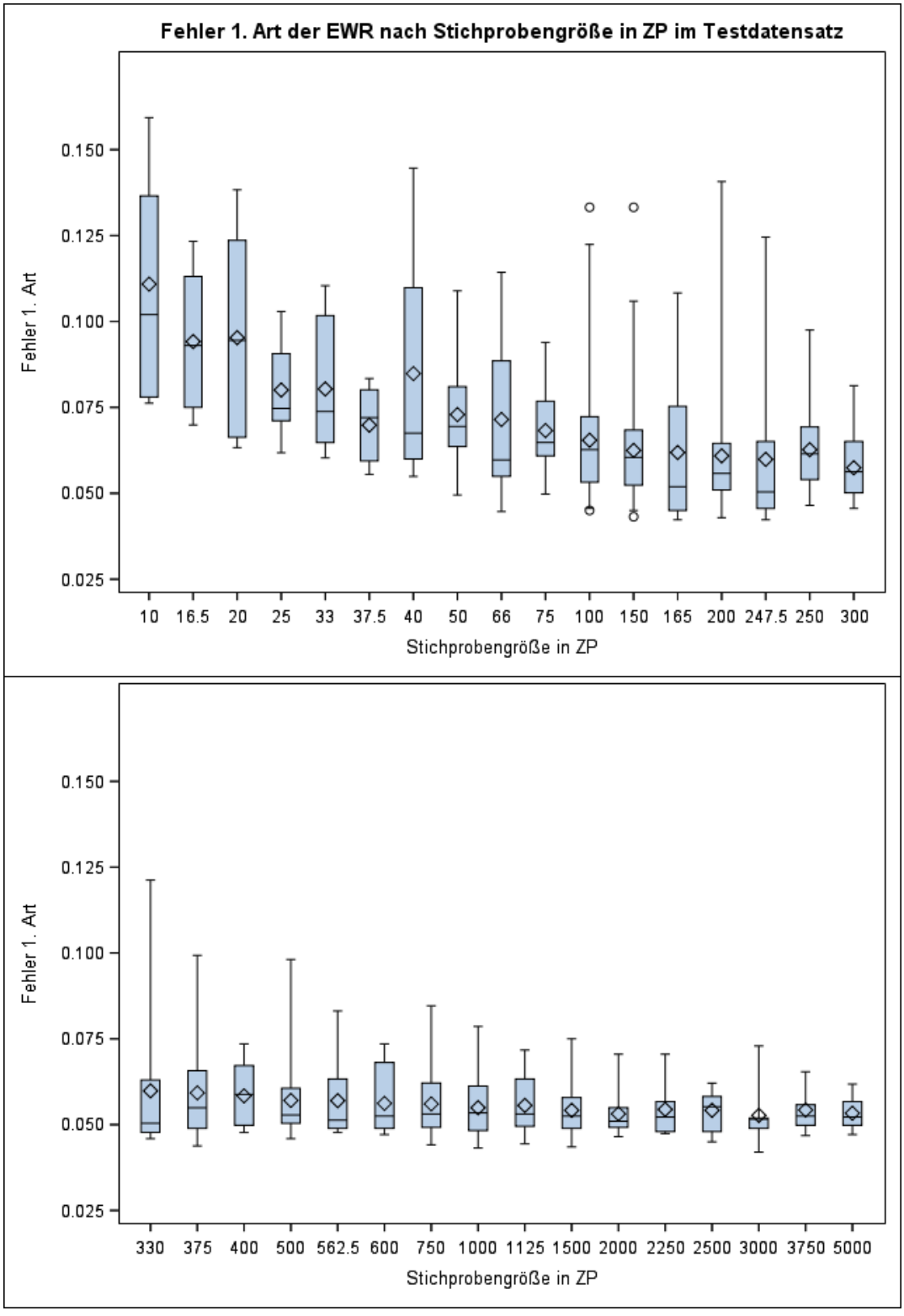
Ein empirischer p-Wert ergibt sich aus der Anzahl an Replikationen, in denen die beiden Bedingungen unter 4) erfüllt sind, geteilt durch die Gesamtzahl (n_{rep}) an Replikationen. Als Signifikanzniveau wird $\alpha = 0{,}025$ gewählt. Die Anzahl an Replikationen beträgt $n_{rep} = 100\,000$. Ist der empirische p-Wert kleiner als 0,025, 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.

Anhang B – Simulationsergebnisse



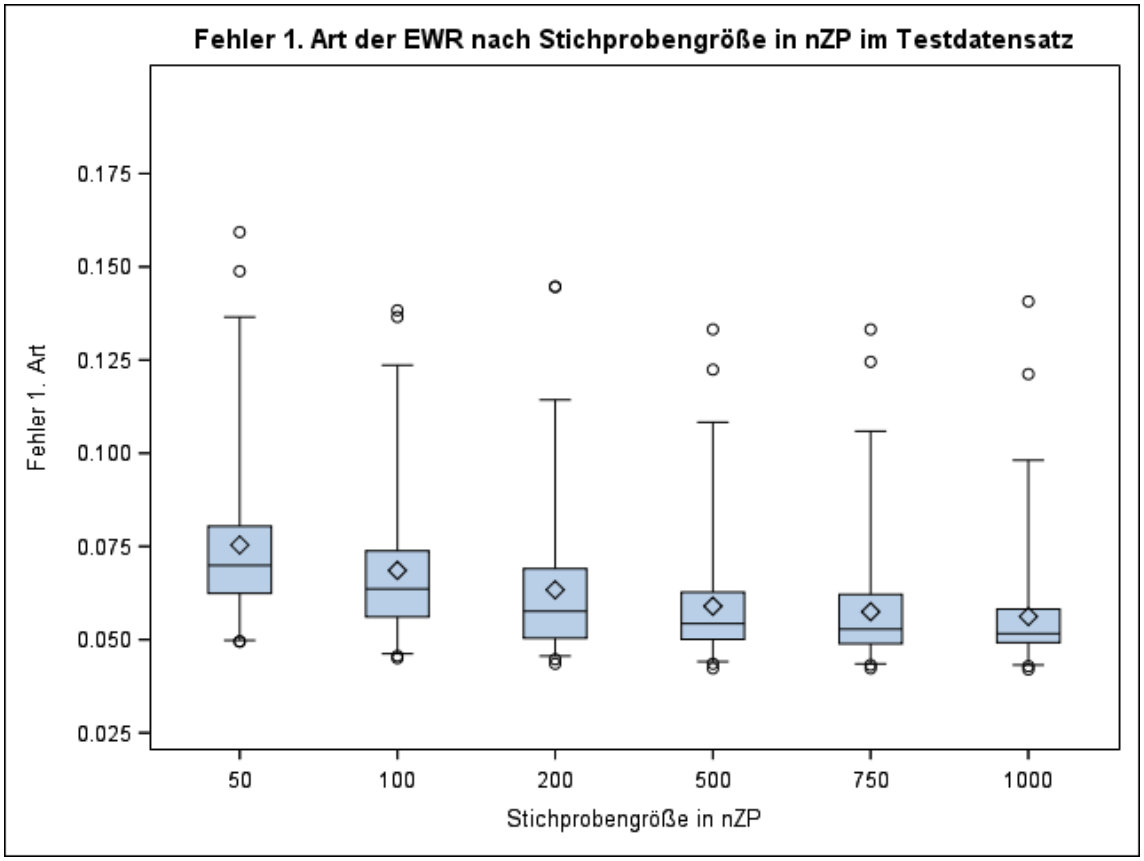
EWR: Erweiterungsregel; nZP: Nichtzielpopulation

Abbildung 5: Empirischer Fehler 1. Art der Testprozedur mit EWR in Abhängigkeit vom Effekt in der nZP. Die Boxen entsprechen 0,25 bzw. 0,75 Perzentilen, die Whisker 0,025 bzw. 0,975 Perzentilen, die Kreise bezeichnen Ausreißer jenseits der genannten Perzentile.



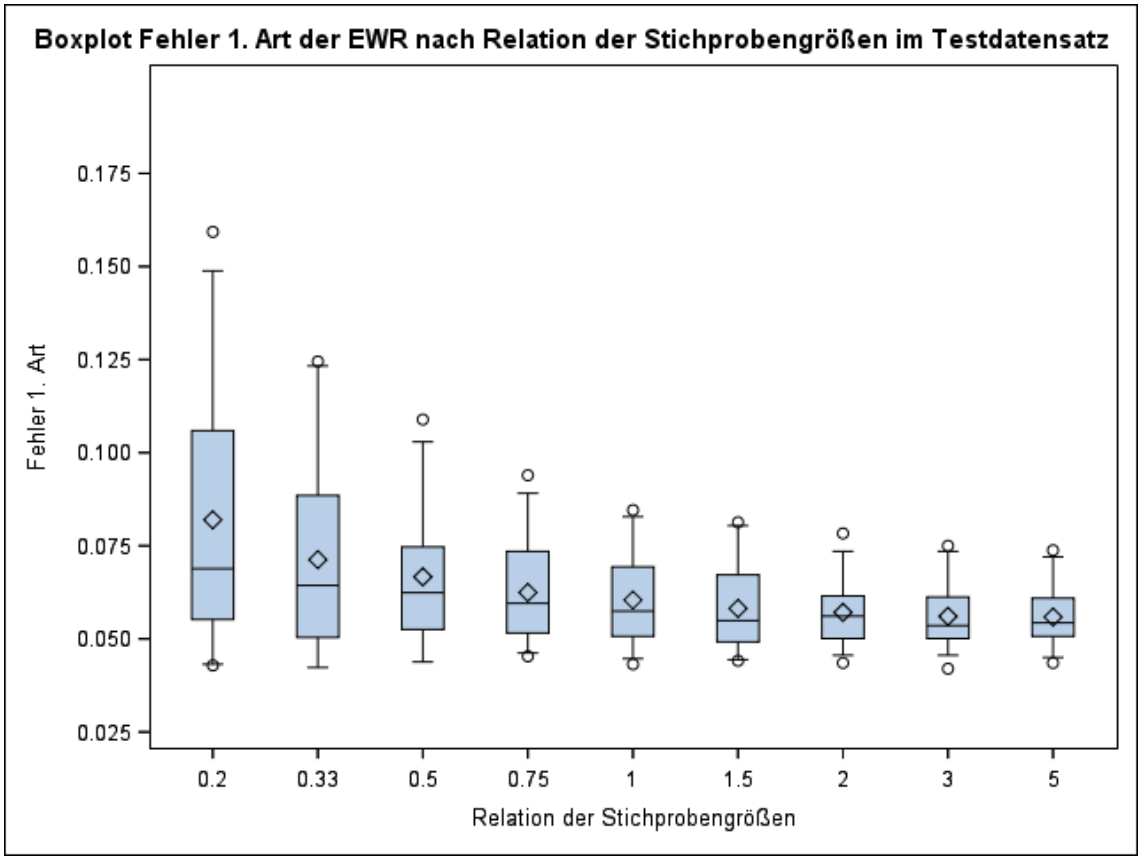
EWR: Erweiterungsregel; ZP: Zielpopulation

Abbildung 6: Empirischer Fehler 1. Art der Testprozedur mit EWR in Abhängigkeit von der Stichprobengröße in der ZP. Die Boxen entsprechen 0,25 bzw. 0,75 Perzentilen, die Whisker 0,025 bzw. 0,975 Perzentilen, die Kreise bezeichnen Ausreißer jenseits der genannten Perzentile.



EWR: Erweiterungsregel; nZP: Nichtzielpopulation

Abbildung 7: Empirischer Fehler 1. Art der Testprozedur mit EWR in Abhängigkeit von der Stichprobengröße in der nZP. Die Boxen entsprechen 0,25 bzw. 0,75 Perzentilen, die Whisker 0,025 bzw. 0,975 Perzentilen, die Kreise bezeichnen Ausreißer jenseits der genannten Perzentile.



EWR: Erweiterungsregel

Abbildung 8: Empirischer Fehler 1. Art der Testprozedur mit EWR in Abhängigkeit von der Relation der Stichprobengröße in ZP und nZP (N_{ZP} / N_{nZP}). Die Boxen entsprechen 0,25 bzw. 0,75 Perzentilen, die Whisker 0,025 bzw. 0,975 Perzentilen, die Kreise bezeichnen Ausreißer jenseits der genannten Perzentile.

Trainingsdaten **Testdaten**

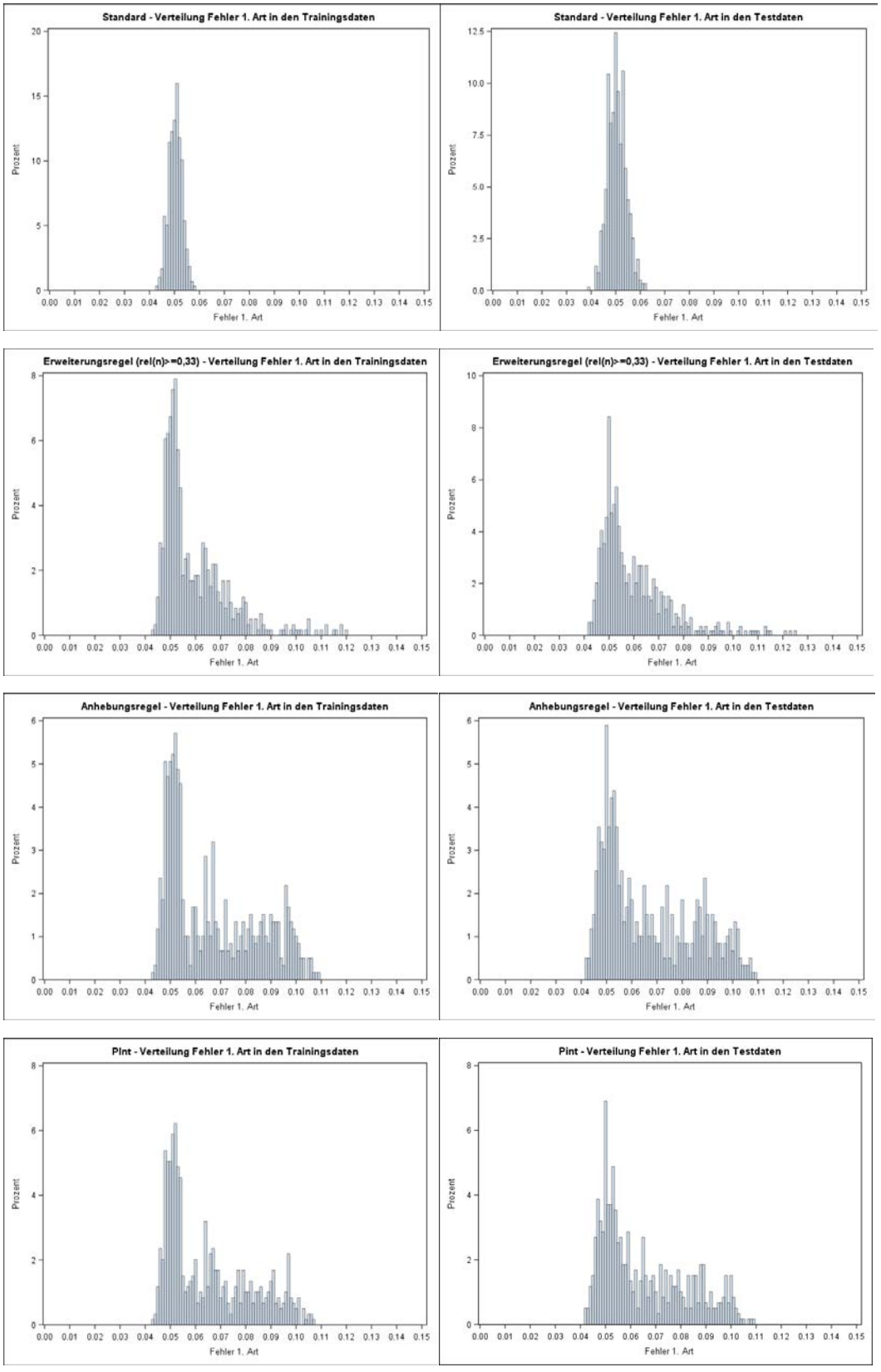
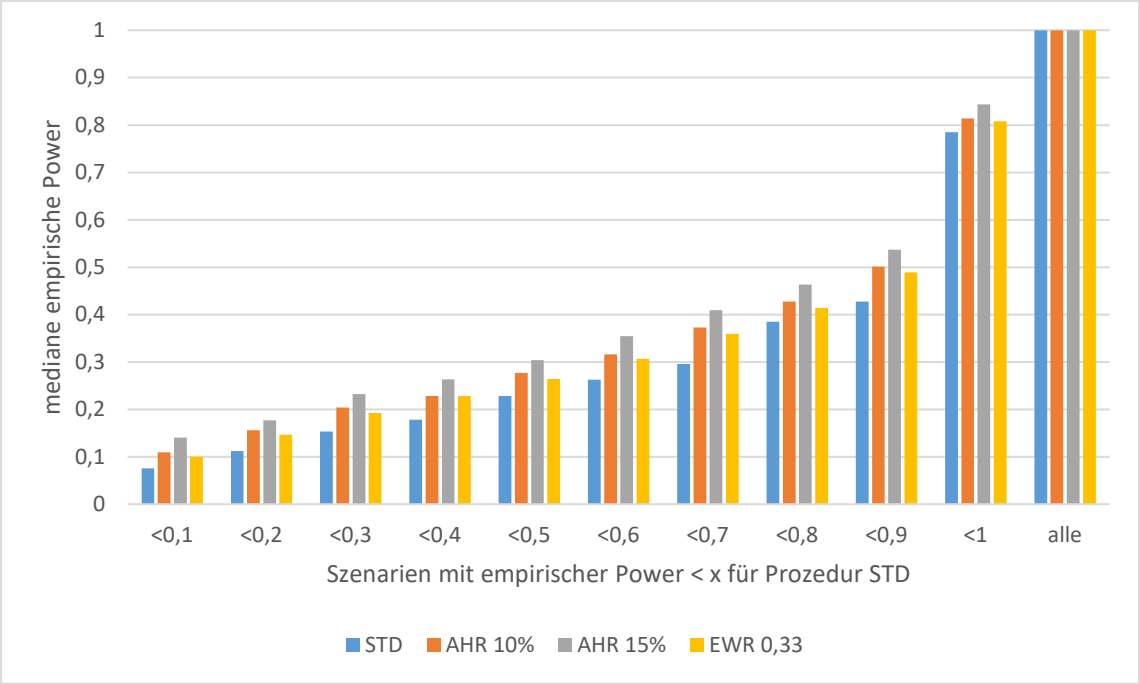
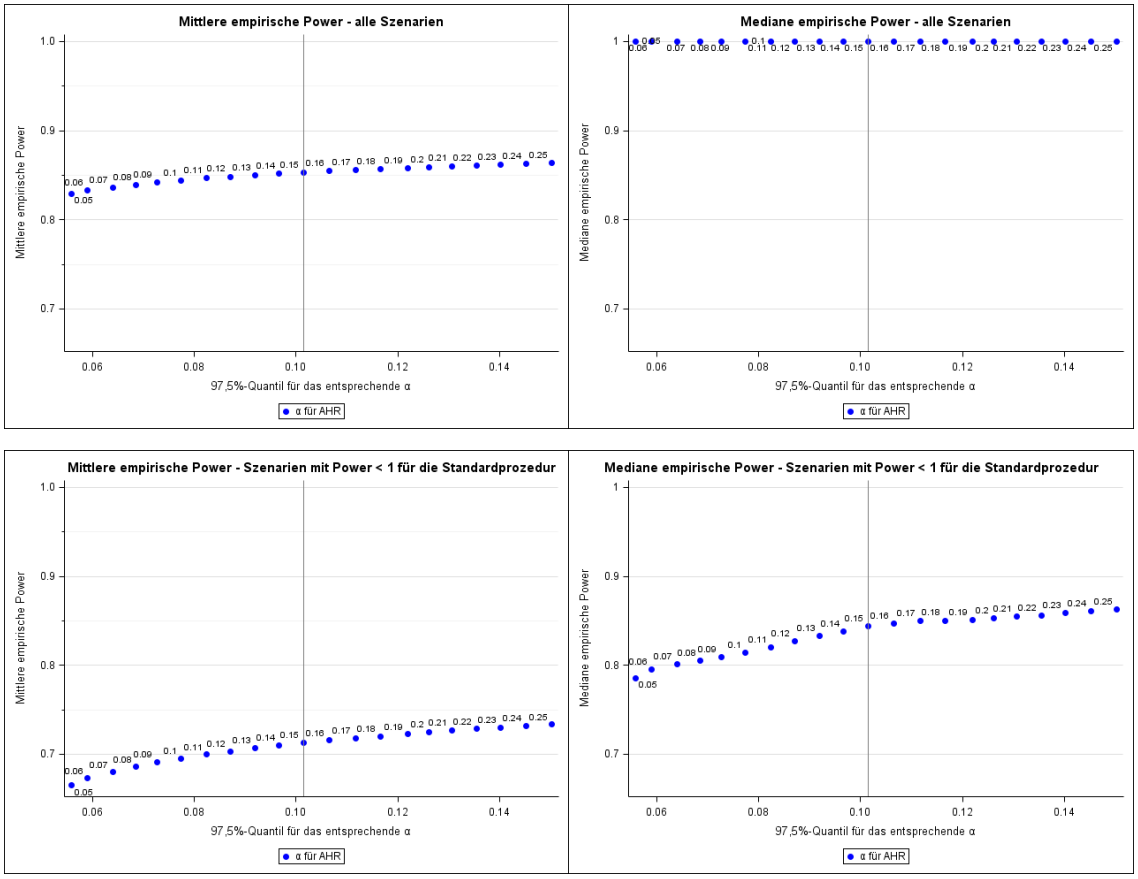


Abbildung 9: Verteilung des empirischer Fehlers 1. Art der untersuchten Testprozeduren für Trainings- und Testdatensatz



AHR: Anhebungsregel; EWR: Erweiterungsregel; STD: Standardtestprozedur A_5

Abbildung 10: Mediane empirische Power der untersuchten Testprozeduren für verschiedene Obergrenzen der empirischen Power der Standardtestprozedur A_5



AHR: Anhebungsregel

Abbildung 11: Mittlere und mediane empirische Power mit Signifikanzniveau α für die Testprozedur AHR_{α} in Abhängigkeit von der festgelegten Grenze für das 97,5 %-Quantil