OPEN ACCESS

Originalarbeit

Surrogatvalidierung durch Korrelation und Surrogate Threshold Effect – Ergebnisse von Simulationsstudien

Validation of surrogates by correlation and surrogate threshold effect – Results of simulation studies

Johanna Gillhaus[1]
Ralf Goertz[2]
Ulli Jeratsch[3]
Friedhelm Leverkus[1]

1 Pfizer Deutschland GmbH, Berlin, Deutschland

2 AMS Advanced Medical Services GmbH, Mannheim, Deutschland

3 AMS Advanced Medical Services GmbH, München, Deutschland

Abstract

Background: Progression-free survival (PFS) is often used instead of the patient-relevant endpoint overall survival (OS) in cancer clinical trials. In order for PFS to be accepted as a patient-relevant outcome within the benefit assessment of pharmaceuticals in accordance with the German Social Code, Book Five (SGB V), section 35a, it has to be validated as a surrogate endpoint for OS in the relevant indication. As part of a rapid report the Institute for Quality and Efficiency in Health Care (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen – IQWiG) presented methods for surrogate endpoints validation and recommendations for correlation-based procedures. These methods include the evaluation of the certainty of conclusion of study results and the correlation between estimates of surrogate outcome and patient-relevant outcome on trial-level. The correlation is estimated by sample Pearson correlation coefficient r or coefficient of determination R^2 and respective confidence interval (CI). Requirements for surrogate validation are a high correlation and a high certainty of conclusion of the study results. In case of medium correlation IQWiG methods propose applying the concept of surrogate threshold effect (STE) to determine thresholds for the estimate of the surrogate endpoint.

Methods: In simulation studies we investigate the requirements for a successful surrogate validation when applying a correlation-based approach. Simulation parameters are the estimates of the surrogate and the patient-relevant outcome, the correlation between them, the number of patients and the number of studies. We analyzed different scenarios in order to figure out parameters contributing to high correlation. Furthermore, we investigate requirements of the STE method, allowing conclusions on patient-relevant endpoints by means of surrogate endpoints. Finally, in consideration of IQWiG methods we analyze the challenges of surrogate validation in practical use.

Results: Both, simulations of the surrogate validation using correlation-based procedure as well as an analytical derivation show low statistical power despite a medium-sized number of studies and a high true correlation. The power for n=5 studies and correlation ρ=0.9 is below 6%. A very high true correlation of ρ=0.95 in at least n=25 studies would be required in order to preserve a power of 80%, however this scenario is considered implausible in practice. Further simulations investigating the power of the method of STE showed that only one fifth of the considered scenarios have power above 80%. However, these scenarios included parameter constellations with impractical values regarding number of studies, number of patients and effect estimate of OS. The correlation parameter ρ as well as the parameter of the estimate of PFS barely have an impact on the power of the STE procedure.

Conclusion: Our simulations show that in practical use it is quite unlikely to fulfill the condition of high correlation as defined in the rapid report of IQWiG, proposing the lower limit of confidence interval to be crucial. Despite setting the true correlation in the model to a high value, statistical power will be quite small as long as the number of studies remains

low or medium which is a realistic assumption in validation of surrogate endpoints within the framework of early benefit assessment. Besides, recommendation to involve certainty of studies in the analysis remains problematic. On closer inspection of the density function of sample correlation coefficient and assuming a given true correlation we can conclude that sample correlation does not depend on the variance of the single estimates but only on sample size (representing the number of studies in the model). Therefore, patient number does not have an impact on the confidence interval of the correlation whether using weight vectors for studies or not. Application of the STE concept according to the requirements described in the rapid report appears to be rather complicated as well. We propose an alternative solution of comparing the value of STE with point estimate of the surrogate endpoint instead of its lower level of confidence interval showing low α-errors in realistic scenarios.

Keywords: validation of surrogates, correlation, surrogate threshold effect, progression-free survival, benefit assessment

Zusammenfassung

Hintergrund: In onkologischen Studien wird oftmals statt des patientenrelevanten Endpunkts Gesamtüberleben (overall survival, OS) der Endpunkt progressionsfreies Überleben (progression-free survival, PFS) erfasst. Für eine Anerkennung von PFS als patientenrelevant im Verfahren der Nutzenbewertung nach § 35a SGB V gilt es, dieses als Surrogatendpunkt für OS in der betrachteten Indikation zu validieren. Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) hat im Rahmen eines Rapid Report Methoden zur Validierung von Surrogatendpunkten dargestellt und Empfehlungen zur Verwendung von korrelationsbasierten Verfahren ausgesprochen. In diesen Methoden werden die Einschätzung der Aussagesicherheit der Studienergebnisse und der Zusammenhang zwischen den Effektschätzern des Surrogats und des patientenrelevanten Endpunkts auf Studienebene einbezogen. Der Zusammenhang wird mit dem Korrelationskoeffizienten r bzw. dem Bestimmtheitsmaß R^2 und entsprechendem Konfidenzintervall (KI) gemessen. Für den Nachweis der Validität des Surrogats müssen eine hohe Korrelation sowie eine hohe Aussagesicherheit der Studienergebnisse vorliegen. Im Falle einer mittleren Korrelation kann das Konzept des Surrogate Threshold Effects (STE) zur Festlegung von Schwellenwerten für den Effektschätzer des Surrogatendpunkts angewandt werden.
Methoden: In Simulationsstudien wird nun untersucht, welche Bedingungen für eine erfolgreiche Surrogatvalidierung mit korrelationsbasierten Verfahren erfüllt sein müssen. Variierende Parameter sind die Effektschätzer des Surrogats und des patientenrelevanten Endpunkts, die Korrelation zwischen den Effektschätzern, die Patientenanzahl sowie die Anzahl der Studien. Es wird analysiert, in welchen Szenarien der Nachweis einer hohen Korrelation gelingt und falls nicht, welche Voraussetzungen vorliegen müssen, dass anhand des Surrogats unter Einbeziehen des STE-Konzepts noch Schlüsse auf den patientenrelevanten Endpunkt gezogen werden können. Die Herausforderungen der vom IQWiG präferierten Methodik zur Surrogatvalidierung in der Praxis werden analysiert.
Ergebnisse: Die Simulation der Surrogatvalidierung über das korrelationsbasierte Verfahren sowie die analytische Herleitung der Power zeigen, dass diese bei moderater Studienanzahl und starker zugrundeliegender wahrer Korrelation dennoch sehr gering ist. Die Power liegt für n=5 Studien und Korrelation ρ=0,9 unter 6%. Es wäre eine sehr hohe Korrelation von ρ=0,95 in mindestens n=25 Studien erforderlich, um eine Power von 80% zu erhalten. Dieses Szenario ist in der Realität al-

lerdings als unplausibel anzusehen. In der Simulation zur Anwendung des STE-Konzepts lag die Power nur bei etwa ein Fünftel der betrachteten Szenarien über 80%. Dabei handelte es sich jedoch um Szenarien, in denen die Parameterkonstellationen aus hoher Studien- und Patientenanzahl und großem Effekt des OS eher unrealistisch sind. Der Parameter der Korrelation ρ zwischen den Effektschätzern der Studien wirkt sich ebenso wie verschiedene Effekte des PFS kaum auf die Power des STE-Verfahrens aus.

Schlussfolgerung: Die durchgeführten Simulationen zeigen, dass die im Rapid Report beschriebene Methodik, wonach die untere Grenze des Konfidenzintervalls ausschlaggebend für eine hohe Korrelation bei der Surrogatvalidierung sein soll, eine in der Praxis kaum zu überwindende Hürde darstellt. Bei gering bis moderat angenommener Studienanzahl - wie es für eine Validierung von Surrogatendpunkten im Rahmen der frühen Nutzenbewertung realistisch erscheint - ist die Power selbst bei hoher, wahrer Korrelation äußerst gering. Problematisch erscheint weiterhin die Empfehlung, die Aussagekraft der Studien in die Analyse mit einzubeziehen, auch wenn dies prinzipiell gerechtfertigt erscheint. Bei Betrachtung der Definition des Korrelationskoeffizienten und dessen Dichtefunktion wird zudem klar, dass die empirische Korrelation unter Annahme einer festen wahren Korrelation gar nicht von der Varianz der Einzelschätzer, sondern nur von der Anzahl der Wertepaare abhängt. Die Patientenanzahl hat somit keine Auswirkung auf das Konfidenzintervall der Korrelation. Dies gilt ebenso, wenn Modelle mit Gewichtung der Studien verwendet werden. Die Anwendung des STE-Konzeptes gemäß der im Rapid Report beschriebenen Methodik erscheint ebenfalls schwierig. Ein Vergleich des STE mit dem Punktschätzer des Surrogatendpunkts wäre eine Alternative, die in realistischen Szenarien geringe α-Fehler zeigte.

Schlüsselwörter: Surrogatvalidierung, Korrelation, Surrogate Threshold Effect, progressionsfreies Überleben, Nutzenbewertung

Hintergrund

Im Rahmen der frühen Nutzenbewertung nach § 35a SGB V des Arzneimittelmarktneuordnungsgesetzes (AMNOG) werden Entscheidungen auf Basis patientenrelevanter Endpunkte getroffen [1]. Dazu zählen laut Verfahrensordnung Mortalität, Morbidität und gesundheitsbezogene Lebensqualität [1]. Falls keine wahren Endpunkte vorgelegt werden, können anstelle dieser unter Umständen auch Surrogatparameter akzeptiert werden. Diese Aussage findet sich auch in den ICH E9 Guidelines [2]. Surrogatparameter sind im Gegensatz zu wahren Endpunkten oftmals einfacher, schneller und kostengünstiger zu erheben [3]. In der Literatur wird ein Surrogatparameter oftmals als ein Biomarker verstanden, der den klinischen Endpunkt ersetzen soll, mit der Absicht den wahren Endpunkt vorherzusagen [4]. Oftmals wird missverstanden, dass Biomarker per se validierte Surrogate sind [5]. Im Rahmen der frühen Nutzenbewertung wird ein Surrogatendpunkt nur dann akzeptiert, wenn dieser im betrachteten Zusammenhang als valide gesehen werden kann, oder daraus hinreichend präzise Aussagen zum patientenrelevanten Endpunkt gefolgert werden können (Anlage II.6: Modul 4 [1]). Hier wird der oftmals in onkologischen Studien gemessene Endpunkt PFS in der Regel bisher nicht als patientenrelevanter Endpunkt für OS akzeptiert.

Der Gemeinsame Bundesausschuss (G-BA) hat das IQWiG mit der Fragestellung der Validierung von Surrogaten beauftragt, das 2011 den Rapid Report „Aussagekraft von Surrogatparametern in der Onkologie“ veröffentlichte [6]. Darin wurden in Teilziel 1 („Darstellung und Bewertung methodischer Verfahren“) sowohl aktuelle wissenschaftliche Methoden vorgestellt als auch eine Empfehlung zur Validierung von Surrogatendpunkten abgegeben. Zur Validierung empfiehlt das IQWiG das aktuell primär befürwortete korrelationsbasierte Verfahren zur Schätzung von Korrelationsmaßen auf Studien- und individueller Ebene. Andere Validierungsmethoden müssten ausreichend begründet werden. Damit wird eine deutliche Präferenz für die Benutzung der Korrelation ausgesprochen. Auch beim korrelationsbasierten Verfahren ist anzumerken, dass das bloße Aufzeigen einer Korrelation zwischen Surrogatendpunkt und klinischem Endpunkt nicht genügt. Für eine erfolgreiche Validierung soll „vorzugsweise eine Meta-Analyse von mehreren randomisierten Studien mit ausreichender Ergebnissicherheit“ eingesetzt werden [6]. Dabei muss die Eingrenzung sowohl auf das Indikationsgebiet als auch auf die Intervention berücksichtigt werden. Des Weiteren wird angemerkt, dass kein „universell anzuwendendes Maß noch eine allgemein beste Schätz-

methode noch eine allgemein akzeptierte Grenze, deren Überschreitung den Nachweis der Validität bedeuten würde“ existiere [6].

Da das Aufzeigen einer Korrelation zwischen Surrogatendpunkt und klinischem Endpunkt auf Basis einer einzelnen Studie nicht ausreiche, entwickelte das IQWiG zur Durchführung einer Validierung einen zweistufigen Algorithmus. Während im ersten Schritt die für die Validierung herangezogenen Studien bezüglich ihrer Aussagekraft bewertet werden, wird im zweiten Schritt die Validität des Surrogats beurteilt, welche maßgeblich durch das Ergebnis der Korrelation zwischen Surrogatendpunkt und patientenrelevantem Endpunkt bestimmt wird.

Nach der im Rapid Report des IQWiG erläuterten Methodik wird die Stärke der Korrelation auf dem 95%-Konfidenzintervall des Korrelationskoeffizienten zwischen den Effektschätzern des Surrogatendpunkts und des patientenrelevanten Endpunkts in die Kategorien hoch, mittel und niedrig klassifiziert. Hohe Korrelation besteht dann, wenn das Konfidenzintervall vollständig über 0,85 liegt. In Verbindung mit hoher Aussagesicherheit der Ergebnisse ist dies Voraussetzung für den Nachweis einer Validität des Surrogats. Wenn das Konfidenzintervall vollständig unter 0,7 liegt, wird von niedriger Korrelation gesprochen und es ist keine Aussage über die Validität des Surrogats möglich. In allen anderen Fällen, in denen die eben genannten Grenzwerte nicht komplett über- bzw. unterschritten werden, liegt mittlere Korrelation vor. Die Validität des Surrogats ist dann unklar.

In Situationen, in denen keine hohe Korrelation vorliegt, können durch Anwendung des STE Konzeptes bei hinreichend großen Effekten für das Surrogat noch Aussagen bezüglich patientenrelevanter Endpunkte getroffen werden. Buyse und Burzykowski [5] beschreiben den STE in einem Modell mit patientenindividuellen Überlebenszeiten in Surrogatendpunkt und wahrem Endpunkt. Im Rahmen der Nutzenbewertung erscheint dieser Ansatz eher ungeeignet, da der pharmazeutische Unternehmer (pU) die dafür benötigten Daten in der Regel nur für seine eigene Studie – in der das Surrogat validiert werden soll – zur Verfügung hat, nicht aber für diejenigen Studien, dessen Daten für die Surrogatvalidierung herangezogen werden sollen. Auch das IQWiG sieht von einem Ansatz mit individuellen Patientendaten im Hinblick auf eine Nutzenbewertung ab. Es wird der oben erwähnte meta-analytische Ansatz auf Studienebene verfolgt, bei dem zunächst die Korrelation der Behandlungseffekte auf Surrogat- und patientenrelevantem Endpunkt berechnet wird. Sollte die Korrelation im mittleren Bereich liegen, bestünde nach Methodik des Rapid Reports immer noch die Möglichkeit eine Schlussfolgerung für einen Effekt bezüglich des patientenrelevanten Endpunkts unter Verwendung des STE-Konzepts zu ziehen.

Im Folgenden werden am besonderen Beispiel der onkologischen Endpunkte PFS und OS Simulationsstudien vorgestellt, die sich mit den Voraussetzungen auseinander setzen, die für eine erfolgreiche Surrogatvalidierung laut aktueller vorgeschlagener wissenschaftlicher Methodik gemäß Rapid Report [6] vorliegen müssen. Die erste Simulation basiert auf dem „einfachen“ Korrelationskoeffizienten nach Bravais-Pearson, eine zweite Simulation bezieht sich auf den STE zur Festlegung von Schwellenwerten für den Effektschätzer des Surrogatendpunkts. Das Design und die Ergebnisse der Simulationen werden jeweils präsentiert und diskutiert. Hierbei wird insbesondere auf die Umsetzung in der Realität mit Hinblick auf die Onkologie eingegangen. Abschließend werden mögliche Ideen für alternative Methoden zur Surrogatvalidierung genannt.

Methoden

Die Durchführung der Simulation sowie der Erstellung aller Grafiken erfolgt mit der Statistik-Software R [7]. Insbesondere werden die Pakete MASS, und metafor [8] verwendet.

Simulation 1: Surrogatvalidierung über korrelationsbasierte Verfahren

Geplante Vorgehensweise

Ziel dieser ersten Simulationsstudie ist es, die Power des Tests zur Surrogatvalidierung zu ermitteln, die unter den Vorgaben des IQWiG zum Nachweis einer hohen Korrelation vorliegt. Es ist also zu bestimmen, wie oft das 95%-KI des empirischen Korrelationskoeffizienten *r* zwischen den Effektschätzern des patientenrelevanten Endpunktes OS und des Surrogatendpunktes PFS vollständig oberhalb von 0,85 liegt. Wie in der Biometrie üblich betrachtet man als Effektschätzer für OS und PFS Hazard Ratios (HR), die hier logarithmiert werden. Die Simulation wird mit der Statistik-Software R [7] durchgeführt. Dazu werden folgenden Annahmen gemacht:

a) Es liegen für eine Validierung *n* Studien vor mit $n \in \{5, 10, 20\}$.

b) Für jede der *n* Studien wird angenommen, dass diese mit jeweils *N* Patienten durchgeführt wurde, die in zwei gleich große Studienarme (Verum und Kontrolle) randomisiert wurden mit $N \in \{100, 200, 500\}$.

c) Für jede der *n* Studien wird für das OS ein Hazard Ratio simuliert, mit $\ln(HR_{OS}) \sim N(\mu_{OS}, \sigma_{OS}^2)$.

d) Für jede der *n* Studien wird für das PFS ein HR simuliert, mit $\ln(HR_{PFS}) \sim N(\mu_{PFS}, \sigma_{PFS}^2)$, wobei eine wahre Korrelation von ρ=0,9 zwischen HR_{OS} und HR_{PFS} zugrunde gelegt wird.

e) Die Varianzen σ_{OS}^2 und σ_{PFS}^2 ergeben sich aus den Ereigniszahlen der jeweiligen Studie nach der Formel

$$\sigma^2 = \frac{1}{N_V} + \frac{1}{N_C} \quad (1)$$

wobei N_v und N_c die erwartete Anzahl der Ereignisse im Verum- bzw. im Kontrollarm bezeichnen (Gleichungen 3 u. 6, [9]). Es wird von einer Todesrate von 80% und einer Progressionsrate von 90% im Beobachtungszeitraum ausgegangen.

f) Jedes der durch die Kombination der Parameter in a) und b) entstehenden neun Szenarien wird 10.000 Mal simuliert und dabei der empirische Korrelationskoeffizient *r* mit 95% KI zwischen den *HR* über die *n* Studien hinweg ermittelt. Die relative Anzahl der Fälle, in denen die untere Grenze des Konfidenzintervalls größer als 0,85 ist, ergibt die Power des Validierungstests.

Das Konfidenzintervall der Korrelation wird der üblichen Vorgehensweise entsprechend über die Fisher-z-Transformation [10], die eine Areatangens-Hyperbolicus-Transformation (artanh) ist, und deren Inverse, also eine Tangens-Hyperbolicus-Transformation (tanh), bestimmt:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad (2)$$

$$r = \frac{e^{2z}-1}{e^{2z}+1} \quad (3)$$

Durch diese Transformation erhält man ein approximativ normalverteiltes Maß *z* mit Erwartungswert $E(z) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ und Standardabweichung $\sigma = \frac{1}{\sqrt{n-3}}$, wobei *n* die Anzahl der zur Bestimmung des Korrelationskoeffizienten benutzten Paare ist. Das 95%-KI von *z* ergibt sich dann wie gewohnt aus

$$z \pm \frac{\Phi^{-1}(0{,}975)}{\sqrt{n-3}} \quad (4)$$

mit Φ^{-1} als der Inversen der Verteilungsfunktion der Standardnormalverteilung. Durch die Rücktransformation gemäß (3) kommt man zum Konfidenzintervall für *r*.

Die Korrelation der zu generierenden Zufallsvariablen HR_{OS} und HR_{PFS} wird über die Zerlegung der gewünschten Kovarianz-Matrix $\Sigma = LL^T$ realisiert. Die Einträge von Σ ergeben sich aus den Vorgaben. Auf der Diagonale finden sich die Varianzen für HR_{OS} und HR_{PFS}. Die Kovarianz ergibt sich aus der Umstellung der Definition des Korrelationskoeffizienten

$$\rho = \frac{\mathrm{cov}(HR_{OS}, HR_{PFS})}{\sigma_{OS}\sigma_{PFS}} \quad (5)$$

und der Forderung, dass die wahre Korrelation ρ einen festen Wert hat.

Für die Bestimmung der Matrix *L* bieten sich zum Beispiel die Cholesky-Zerlegung und die Eigenwert-Zerlegung an. Für diese Simulation wird letztere verwendet, da diese in der R-Funktion mvrnorm(n, μ, Σ) aus dem Paket MASS implementiert ist [11]. Diese Funktion liefert *n* multivariat normalverteilte Zufallszahlen mit zentraler Tendenz μ und Kovarianzmatrix Σ.

Tatsächliches Vorgehen

Unabhängigkeit der Korrelation von der Varianz der Einzelschätzer

Die geplante Vorgehensweise erweist sich in einem Punkt als undurchführbar. Durch die Einführung des Parameters Patientenanzahl *N* soll dem Einfluss der Aussagesicherheit der Studien Rechnung getragen werden. Dies geschieht in der Annahme, dass große Studien auch eine höhere Aussagekraft bezüglich der Schätzer HR_{OS} und HR_{PFS} haben. Dies ist tatsächlich der Fall, denn mit steigender Patientenzahl nimmt auch die Anzahl der erwarteten Ereignisse zu, was nach Gleichung (1) zu niedrigeren Varianzen führt. Die Schätzer haben somit ein kleineres Konfidenzintervall. Die Variabilität der Schätzer hat jedoch *keinen* Einfluss auf die Korrelation der beiden Maße, da die Korrelation gerade die an der Varianz der Einzelschätzer normierte Kovarianz darstellt und eine Veränderung der Varianz eines Schätzers exakt durch die Kovarianz beider Schätzer kompensiert wird. Durch die Annahme einer wahren Korrelation zwischen den beiden Maßen bleibt die Kopplung also vollständig erhalten.

Die Patientenanzahl hat zudem keine Auswirkung auf das Konfidenzintervall der Korrelation, denn dieses ist, wie der Term (4) zeigt, einzig und allein vom Korrelationskoeffizienten *r* selbst und der Studienanzahl *n* abhängig. Insbesondere ist also das Konfidenzintervall für ein gegebenes *r* vollständig durch die Anzahl der Studien *n* determiniert. Diese Eigenschaft ist kein Nebenprodukt der Fisher-z-Transformation, denn auch die exakte Verteilung des empirischen Korrelationskoeffizienten (und damit dessen Varianz) zeigt nur eine Abhängigkeit von ρ und *n*. Diese an sich bekannte und triviale Tatsache erscheint bei oberflächlicher Betrachtung in diesem Zusammenhang dennoch zunächst kontraintuitiv. Geringere Patientenanzahl beeinflusst nur die Lage der Maße, nicht jedoch ihre Korrelation.

Es ergeben sich auch keine anderen Ergebnisse, wenn Studien unterschiedlich gewichtet würden oder metaanalytische Verfahren angewandt werden. Unter der Voraussetzung, dass die wahre Korrelation ρ für alle diese Studien dieselbe ist, führt eine Gewichtung im Einzelfall zu anderer empirischer Korrelation *r*, der Erwartungswert für dieses *r* ändert sich jedoch nicht. Lediglich die Verteilung von *r* mag eine andere sein. Genauso wenig führt eine Gewichtung zu anderen Konfidenzintervallen, da für dessen Bestimmung außer *r* auch dann nur *n* ausschlaggebend ist. Beispielhaft wird dies im folgenden Abschnitt gezeigt.

Angenommen, die wahre Korrelation ist selbst kein fester Wert sondern eine Zufallsvariable, ähnlich einem Modell mit zufälligen Effekten. Dann stellt sich die Frage, ob eine Gewichtung bessere Schätzungen der Korrelation erlaubt. Sei zum Beispiel S_0 die Menge der „genauen" Studien mit vielen Patienten, für die die Korrelation auf ρ=0,82 festgelegt ist, sowie die Menge S_1 der weniger genauen Studien mit nur einem Fünftel der Patientenzahl und ρ=0,778. Damit wäre der über die Fisher-z-Skala gebildete Mittelwert wieder 0,8. (Allein diese Richtungsentscheidung einer höheren Korrelation für S_0 ist willkürlich, denn warum sollten gerade die *genauen* Studien aus einer Grundgesamtheit kommen, der eine *höhere* Korrelation zugrunde liegt.) Eine Simulation (*n*=1.000) dieses Szenarios mit jeweils fünf Studien aus S_0 und S_1 ergab im Mittel Korrelationen von 0,781 für die ungewichtete und 0,786 für die mit inverser Varianz gewichtete Korrelation, so dass man von einer marginalen Verbesserung der Schätzungen ausgehen könnte. Allerdings war der mittlere

Schätzer 0,800 für den Fall, dass die Studien aus S_i mit dem Wert für die Studien aus S_{1-i} gewichtet wurden. In diesem Beispiel hat also die (falsche) geringere Gewichtung der „hoch-korrelierten" Studien zu einer im Mittel höheren Korrelationsschätzung geführt, als die eigentlich korrekte Gewichtung. Es ist daher nicht davon auszugehen, dass eine wie auch immer geartete Gewichtung prinzipiell zu besseren Resultaten führt.

Simulation und analytische Untersuchung der Power

Die dargestellten Zusammenhänge zeigen, dass eine Variation der Patientenanzahl *N* keinen Einfluss auf die Ergebnisse haben würde. Die Simulation dieses Parameters wird daher fallen gelassen und stattdessen die Power der Surrogatvalidierung für Studienzahlen n=5,...,100 durchgeführt. Dabei wird zusätzlich zu der in Punkt d) geplanten wahren Korrelation von ρ=0,9 auch von noch höheren Werten für die Korrelation von ρ=0,95 und ρ=0,97 und jeweils derselben Untergrenze von ρ_u=0,85 ausgegangen. Ansonsten wird nach dem oben beschriebenen Schema vorgegangen.
Weiterhin kann nun auch eine analytische Untersuchung der Power erfolgen. Dazu betrachtet man zu jedem n die Verteilungsdichte des Fisher-z-transformierten Korrelationskoeffizienten für ein festes wahres ρ. In Abbildung 1 ist die Dichtefunktion exemplarisch für n=20 und ρ=0,9 dargestellt (schwarze Linie). Bei jeder Validierung ergibt sich ein transformierter Korrelationskoeffizient z=artanh(r), wobei in diesem Beispiel r=0,92 ist. Die Bestimmung des Konfidenzintervalls für dieses z erfolgt dann über (4). Die dabei benutzte Dichtefunktion ist in roter Farbe dargestellt. Liegt die untere Grenze (im Beispiel nach Rücktransformation etwa 0,805) über dem transformierten Wert von ρ_u=0,85, gilt das Surrogat als validiert.

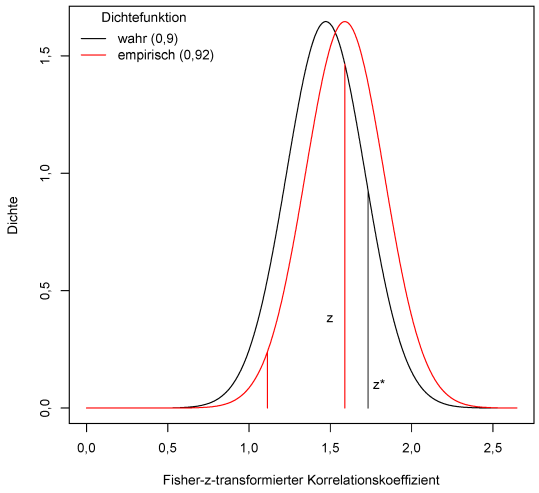

Abbildung 1: Schwarze Linie: Dichtefunktion des Fisher-z-transformierten Korrelationskoeffizienten mit n=20 Studien und wahrer Korrelation ρ=0,9. Rote Linie: Dichtefunktion zur Bestimmung des Konfidenzintervalls bei einer empirischen Korrelation r=0,92 (mit unterer Konfidenzintervallgrenze). Die Fläche unter der schwarzen Kurve rechts von z* ergibt die Power.

Da bei festem n die untere Intervallgrenze streng monoton mit z steigt, gibt es ein minimales z^*, so dass für alle $z>z^*$ die Validierung gelingt. Für das Beispiel erhält man z^*=artanh(0,9392)≈1,73. Integriert man daher die Dichtefunktion mit diesem z^* als unterer Grenze ergibt sich unmittelbar die Power Π für die betrachteten Parameter n, ρ und ρ_u:

$$\Pi(n,\rho,\rho_u) = \int_{z^*}^{\infty} \frac{1}{\sqrt{\frac{2\pi}{n-3}}} e^{\frac{n-3}{2}(z-\text{artanh}(\rho))^2} dz$$

$$= 1 - \Phi\left(z^*; \text{artanh}(\rho); \frac{1}{\sqrt{n-3}}\right) \quad (6)$$

Es bleibt also nur die Bestimmung von z^*. Auch diese kann analytisch erfolgen, es muss lediglich zur unteren Grenze die halbe Länge des Konfidenzintervalls hinzugefügt werden:

$$z^* = \text{artanh}(\rho_u) + \frac{\Phi^{-1}(0{,}975; 0{,}1)}{\sqrt{n-3}} \quad (7)$$

Simulation 2: Konzept des Surrogate Threshold Effects (STE)

Das Ziel der zweiten Simulationsstudie ist es, zu untersuchen, unter welchen Bedingungen die Anwendung des STE-Ansatzes zum Erfolg führen könnte. Wie in der ersten Simulation werden verschiedene Szenarien betrachtet, bei denen die Parameter der zur Verfügung stehenden Studienlage variiert werden. Im Vergleich zu Punkt a) und b) in Simulation 1 werden hier $n \in \{5, 10\}$ Studien mit jeweils drei verschiedenen Fallzahlen $N \in \{200, 500, 1.000\}$ betrachtet. Im Unterschied zur Korrelation ρ, die als skaleninvariantes Maß nicht von den Erwartungswerten der Variablen abhängt, müssen für die Überprüfung des STE-Konzepts auch Erwartungswerte für die Effektschätzer des OS und PFS betrachtet werden. Analog zu Punkt c), d) und e) in Simulation 1 wird das Tupel der logarithmierten Effektschätzer HR_{OS} und HR_{PFS} als bivariat normalverteilte Zufallsgröße mit Kovarianzmatrix Σ modelliert. Für die Erwartungswerte μ_{OS} und μ_{PFS} werden jeweils die Werte 0,5, 0,7 und 0,8 angenommen und für die wahre Korrelation die Werte ρ=0,85 und ρ=0,9 zugrunde gelegt. Zu den n Studien werden das HR_{OS} und das HR_{PFS} einer weiteren Studie mit denselben Parametern N, μ_{OS} und μ_{PFS} generiert. Diese soll die eigene Studie des pU darstellen, anhand der er überprüfen will, ob der darin vorliegende Effekt des PFS den STE unterschreitet. Diese wird im Folgenden als „pU-Studie" bezeichnet.
Durch die Kombination aller Ausprägungen der Parameter n, N, HR_{OS}, HR_{PFS} und ρ ergeben sich 108 unterschiedliche Szenarien. Jedes dieser Szenarien wird 10.000 Mal simuliert, wobei in jedem Iterationsschritt zuerst überprüft wird, ob die empirische Korrelation r zwischen den Studien tatsächlich im mittleren Bereich liegt, und eine Überprüfung anhand des STE überhaupt eingesetzt werden muss. Dies ist laut der im Rapid Report des IQWiG beschriebenen Methodik genau dann der Fall, wenn das 95%-KI von r zwischen HR_{OS} und HR_{PFS} über n Studien weder vollständig oberhalb von 0,85 liegt, was hoher

Korrelation entspräche, noch vollständig unterhalb von 0,7, was niedrige Korrelation darstellt. Wenn keine mittlere Korrelation vorliegt, werden neue Zufallszahlen gezogen. Anschließend wird über eine Meta-Regression mit zufälligen Effekten – mithilfe der R-Funktion rma.uni aus dem metafor-Paket – ein Konfidenzband (auf dem Signifikanzniveau α=0,05) für HR_{OS} bestimmt. Der STE ergibt sich als der minimale Wert, den HR_{PFS} annehmen darf, sodass das HR_{OS} gerade noch statistisch signifikant ist, d. h. die obere Grenze des Konfidenzintervalls für das HR_{OS} gerade noch unterhalb von 1 liegt. Die Power des Tests ergibt sich schließlich als relative Häufigkeit der Fälle, bei denen die obere Grenze des Konfidenzintervalls des HR_{PFS} aus der „pU-Studie“ unterhalb des STE liegt.
Abbildung 2 zeigt ein Beispiel eines simulierten Szenarios mit n=5 Studien und N=500 Patienten sowie jeweils 0,7 als Erwartungswert für HR_{OS} und HR_{PFS} und wahrer Korrelation ρ=0,85. Die simulierten Werte ergeben eine empirische Korrelation r [95% KI]=0,92 [0,19; 0,99]. Somit liegt nach der im Rapid Report geschilderten Methodik mittlere Korrelation vor, und das Überprüfen des STE kommt zur Anwendung. Das Dreieck zeigt die Koordinaten HR_{OS} und HR_{PFS} der „pU-Studie“ und die waagerechte Linie das entsprechende 95% KI des HR_{PFS} ([0,57; 0,86]). Der STE liegt hier bei 0,8092, d. h. der in der „pU-Studie“ geschätzte Effekt auf dem Surrogat ist nicht groß genug, um mit ausreichender Sicherheit von einem signifikanten Effekt auf dem patientenrelevanten Endpunkt auszugehen, denn das gesamte Konfidenzintervall des HR_{PFS} müsste dafür kleiner als der STE sein.

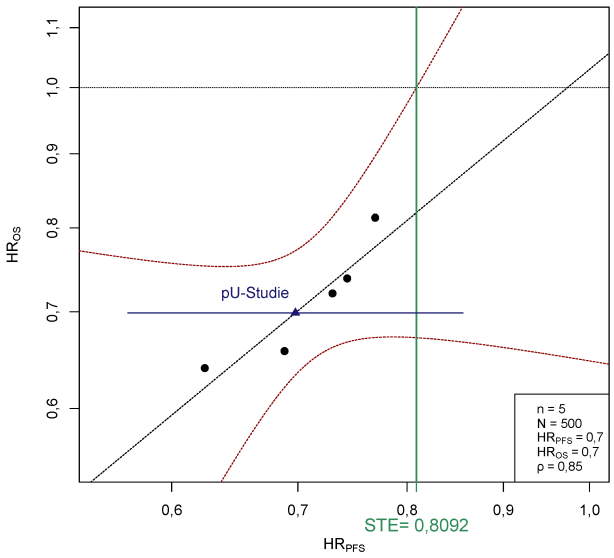

Abbildung 2: Simuliertes Szenario zur Anwendung des STE-Ansatzes am Beispiel von n=5 Studien mit je N=500 Patienten sowie jeweils 0,7 als Erwartungswert für HR_{os} und HR_{PFS} und wahrer Korrelation ρ=0,85.

Ergebnisse

Simulation 1

In Abbildung 3 sind die Ergebnisse sowohl für die Simulation als auch die analytische Herleitung der Power der Surrogatvalidierung dargestellt. Die Diskrepanzen erklären sich aus der Tatsache, dass durch die Fisher-z-Transformation nur eine approximative Normalisierung des Korrelationskoeffizienten erreicht wird und der zurücktransformierte Mittelwert einen positiven Bias aufweist [12]. Das bedeutet im Fall der Simulation, dass man im Mittel etwas höhere Untergrenzen für das Konfidenzintervall erhält, als es die analytische Herangehensweise erwarten lassen würde. Daher finden wir in letzterer etwas niedrigere Power für alle n.

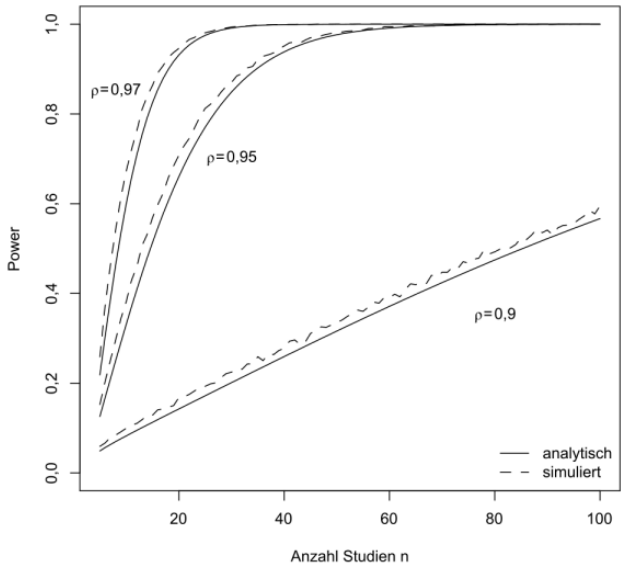

Abbildung 3: Ergebnisse der Simulation und der analytischen Herleitung der Power

Die Power der Surrogatvalidierung ist bei einer wahren Korrelation von ρ=0,9 insgesamt sehr gering. Der in der Statistik oft geforderte Mindestwert der Power von 80% wird hier nicht annähernd erreicht, die selbst bei n=100 Studien unter 60% liegt. Selbst für ρ=0,95 sind n=25 Studien erforderlich, um eine Power von 80% zu erhalten. Dass auch eine Gewichtung von Studien nach der Anzahl der Patienten oder eine meta-analytische Zusammenfassung keine Verbesserung der Resultate bringt, zeigt eine weitere Simulation, bei der für ρ=0,9 und einer Untergrenze des Konfidenzintervalls von ρ_u=0,85 von n=20 Studien ausgegangen wird, die eine Hälfte mit 100 Patienten, die andere mit 500. Der Erwartungswert des HR_{OS} wird auf 0,8, die Todesrate auf 80% und das Verhältnis von Tod zu Progression auf 0,9 gesetzt. Es werden wiederum 10.000 Durchgänge simuliert und einerseits ein gewichtetes lineares Modell mit Patientenzahl N als Gewicht, andererseits eine Meta-Analyse für das HR_{OS} mit Random Effects Modell und HR_{PFS} als Moderator gerechnet. In beiden Fällen wird als „Korrelation“ die Wurzel des Be-

stimmtheitsmaßes R^2 verwendet. Im zweiten Fall wird dieses über

$$R^2 = 1 - \left(\frac{L_0}{L_{PFS}}\right)^{\frac{2}{n}} \qquad (8)$$

bestimmt, wobei L_0 die Likelihood ohne bzw. L_{PFS} die Likelihood mit Verwendung des HR_{PFS} als Moderatorvariable darstellen. Während in der Simulation des Modells mit Gewichtung der Studien (bzw. Effektschätzer) die Power mit einem Wert von 0,7026 praktisch mit dem Wert, der sich mit ungewichteten Studien ergab (0,7081), übereinstimmt, ist das Ergebnis für die Meta-Analyse gleich 0. Der Maximalwert für die untere Grenze des Konfidenzintervalls lag bei 0,8191, also immer noch deutlich unter der geforderten Grenze ρ_u=0,85.

In Modul 4 des Nutzendossiers zu Dabrafenib [13] hatte der pU den Versuch einer Surrogatvalidierung des PFS durch Verwendung korrelationsbasierter Verfahren unternommen. Darin wurde die Berechnung der Korrelation vermutlich über (8) vorgenommen, zumindest konnte so das *r* aus Tabelle 4-123 (Seite 282, Modul 4 des Nutzendossiers zu Dabrafenib) reproduziert werden. Auch in der Meta-Analyse im Nutzendossier zu Dabrafenib liegt das *r* deutlich niedriger als das der Analyse mit Gewichtung nach Patientenzahl. Dies hat seine Ursache darin, dass der Zusammenhang $r = \sqrt{R^2}$ nur im Falle linearer Regression gilt und daher *r* und R^2 in anderen Fällen gar nicht vergleichbar sind. Auch scheint eine Berechnung des Konfidenzintervalls für R^2 über (4) daher fraglich.

Simulation 2

Die Simulation zur Anwendung des STE-Konzepts ergibt, dass zwischen den Effektschätzern HR_{OS} und HR_{PFS}, unabhängig von der Wahl der Parameter *n*, *N*, μ_{OS}, μ_{PFS} und ρ, überwiegend mittlere Korrelation vorliegt: der Anteil der Simulationsdurchgänge, in denen sich die empirische Korrelation im mittleren Bereich befand, liegt in jedem der 108 Szenarien bei mindestens 89%. Obwohl mit ρ=0,85 bzw. 0,9 sehr starke Korrelationen in das Modell gelegt werden, muss eine Überprüfung anhand des STE (wegen nicht vorliegender hoher Korrelation) tatsächlich in den meisten Fällen eingesetzt werden. Dies bestätigt nochmals die Ergebnisse aus Simulation 1, die verdeutlicht, dass hohe Korrelation in der Praxis kaum zu erreichen ist.

Die Ergebnisse der Power aller Szenarien aus der Simulation zur Anwendung des STE-Konzepts sind in Abbildung 4 veranschaulicht. Der Großteil der Szenarien zeigt dabei niedrige Power: nur bei 21 der 108 Szenarien liegt die Power über 80%. Unter diesen 21 befinden sich 18 Szenarien, bei denen der Erwartungswert des HR_{OS} bei 0,5 liegt, 15 Szenarien bei denen die Studienanzahl bei *n*=10 liegt, und 15 Szenarien bei denen die Patientenanzahl bei *N*=1.000 liegt. Szenarien, in denen der Erwartungswert des HR_{OS} mit 0,8 angenommen wird, erreichen maximal eine Power von 47%. Es müsste demzufolge ein extrem starker Behandlungseffekt zum OS (HR_{OS}≈0,5) in den für die Validierung herangezogenen Studien vorliegen, um anhand der Ergebnisse des Surrogatendpunkts PFS aus der „pU-Studie“ noch Schlussfolgerungen auf das OS ziehen zu können.

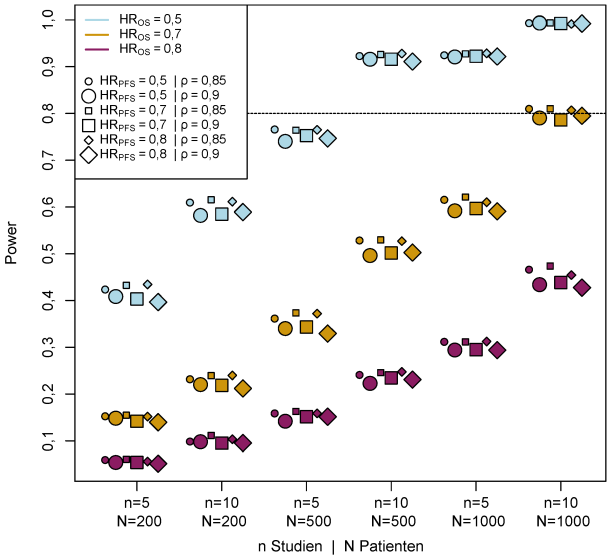

Abbildung 4: Darstellung der Power der 108 simulierten Szenarien, die sich durch die Kombination der Parameter Studiengröße *n*, Patientenanzahl *N*, Erwartungswerte der Effektschätzer HR_{OS}, HR_{PFS} und wahrer Korrelation ρ ergeben.

Außerdem verdeutlicht die Übersicht, dass die Power mit sinkender Studien- und Patientenanzahl abnimmt. Dies ist leicht mit Hilfe des Beispiels in Abbildung 2 zu erklären. Werden *n* und *N* kleiner, vergrößert das die Varianz der beiden Effektschätzer HR_{OS} und HR_{PFS}. Dadurch wird das Konfidenzband breiter und dessen Obergrenze schneidet die horizontale Linie an der Stelle HR_{OS}=1 bereits bei kleineren Werten des HR_{PFS}. Es bedarf dann eines stärkeren Surrogateffektes in der zu testenden Studie um den nun weiter links liegenden STE unterschreiten zu können. Wie oben erwähnt, sinkt die Power rapide, wenn in den Studien geringe Effekte auf dem OS (also näher an 1 liegende Werte des HR_{OS}) vorliegen. In der Grafik des Beispiels in Abbildung 2 würde dies zu einer Verschiebung der Studienpunkte und des Konfidenzbandes nach oben führen, dessen Obergrenze die horizontale Linie an der Stelle HR_{OS}=1 dadurch weiter links schneiden würde. Die „pU-Studie“ unterschreitet den STE dann mit geringerer Wahrscheinlichkeit.

Des Weiteren zeigt sich, dass verschiedene Ausprägungen des HR_{PFS} zu keiner Änderung der Power führen. Verändert man den Erwartungswert des HR_{PFS} der Studien, so würde sich auch der STE in gleichem Maße verändern – im Beispiel in Abbildung 2 würde das eine waagerechte Verschiebung der Studienpunkte bedeuten. Interessant ist zudem, dass sich der Parameter ρ kaum auf die Power des STE-Verfahrens auswirkt. Die Aussagen können anhand eines linearen Modells mit Power als Responsevariable und Simulationsparametern als erklärende Variablen unterstützt werden: Studienanzahl *n* (p<0,001), Patientenanzahl *N* (p<0,001), wahre Korrelation ρ (p=0,225), logarithmiertes HR_{OS} (p<0,001) und logarithmiertes HR_{PFS} (p=0,947).

Diskussion

Diskussion der Empfehlungen des Rapid Report

In onkologischen Studien wird oftmals statt des patientenrelevanten Endpunkts Gesamtüberleben der Endpunkt progressionsfreies Überleben erfasst. Für eine Anerkennung von PFS als patientenrelevant im Verfahren zur frühen Nutzenbewertung gilt es, dieses als Surrogatendpunkt für OS in der betrachteten Indikation zu validieren. Das IQWiG hat im Rahmen eines Rapid Report Methoden zur Validierung von Surrogatendpunkten dargestellt und Empfehlungen zur Verwendung von korrelationsbasierten Verfahren ausgesprochen.

In der hier vorliegenden Arbeit wurde mithilfe zweier Simulationsstudien untersucht, inwiefern diese Vorgaben in der Realität am Beispiel der Onkologie umgesetzt werden können. In der ersten Simulationsstudie wurde die Validierung anhand des Korrelationskoeffizienten und in der zweiten Simulation das Konzept des STE zur Überprüfung der Schwellenwerte für den Effektschätzer des Surrogatendpunkts untersucht.

Gemäß der im Rapid Report des IQWiG erläuterten Methodik muss neben der Einschätzung der Aussagesicherheit ein gleichgerichteter Zusammenhang, gemessen durch die Korrelation zwischen den Effektschätzern des Surrogats und des patientenrelevanten Endpunkts vorliegen, um die Validität für das Surrogat auf Studienebene nachzuweisen. Der Korrelationskoeffizient nach Bravais-Pearson stellt bei $\rho=1$ den perfekten linearen Zusammenhang dar, während $\rho=0$ für völlige Unkorreliertheit steht. Mit dem Wert von ρ steigt die Korrelation, jedoch gibt es keine allgemein gültige Definition, in welchen Wertebereichen diese hoch oder niedrig wäre. Im Rapid Report werden einige Stellen aus der biometrischen Fachliteratur zitiert, in denen die jeweiligen Autoren verschiedene Vorschläge von Schwellenwerten zur Korrelation bzw. zum Bestimmtheitsmaß R^2 angeben, ab denen ein hoher Zusammenhang bzw. eine gute statistische Validität vorliegt (S. 71 u. 108, [6]). Basierend auf diesen Angaben werden für die Korrelation die Kategorien „hoch“, „mittel“ und „niedrig“ hinsichtlich der Einstufung der Validität des Surrogats festgelegt. Hohe Korrelation liegt vor, wenn die untere Konfidenzintervallgrenze des empirischen Korrelationskoeffizienten r mindestens den Wert 0,85 (bzw. bei Bestimmtheitsmaßen $R^2 \geq 0{,}72$) annehme. Weiter wird von niedriger Korrelation gesprochen, wenn die obere Konfidenzintervallgrenze von r kleiner 0,7 (bzw. beim Bestimmtheitsmaß $R^2<0{,}49$) ist. Der geforderte Schwellenwert von 0,85 entstamme daher, dass bei einem wahrem Wert von $\rho=0{,}9$ die empirische Korrelation selbst bei genauen Schätzungen unvermeidlich den Wert 0,9 unterschreite. Jedoch wird nicht hergeleitet, wie der Wert 0,85 ermittelt wurde. Dieser Wert stellt lediglich eine Abschätzung dar, denn mithilfe der Fisher-z-Transformation und in Abhängigkeit der Studienanzahl n lässt sich die Untergrenze des 95%-KI von r analytisch bestimmen.

Abbildung 5 veranschaulicht, dass bei $n=10$ Studien mit $r=0{,}9$ die Untergrenze des 95%-KI für r bei 0,62 liegt. Da dies den Wert 0,85 unterschreitet, könnte selbst bei diesem hohen Punktschätzer nur eine mittlere Korrelation konstatiert werden. Ginge man von $r=0{,}9$ und einer hohen Studienanzahl von $n=50$ aus – was äußerst unrealistisch wäre, da sich alle Studien auf dieselbe Indikation bzw. denselben Schweregrad sowie dieselbe Intervention eingrenzen müssten, um hohe Aussagesicherheit zu gewährleisten – läge die Untergrenze des 95%-KI für r bei 0,83, und somit immer noch unterhalb des für hohe Korrelation geforderten Schwellenwerts. Wenn man prüfen will, wie hohe Korrelation überhaupt zu erreichen ist, d. h. einen Schwellenwert von 0,85 für die Untergrenze des 95%-KI für r vorgibt, müsste bei einer Studienanzahl von $n=10$ ein Punktschätzer der Korrelation von mindestens 0,9638 gemessen werden. Dieser Wert erscheint extrem hoch, wenn man bedenkt, dass $r=1$ einen perfekten linearen Zusammenhang darstellt, der in realen Daten quasi nicht beobachtet wird. Es stellt sich außerdem die Frage, warum von einem zweiseitigen 95% Konfidenzintervall ausgegangen werden muss. Die zu verwerfende Hypothese lautet schließlich $\rho<\rho_{relevant}$. Bei einem Punktschätzer von $r=0{,}95$ überschritte man mit einseitiger Testung die Grenze 0,85 schon bei $n=12$ Studien, während es $n\geq 15$ bedarf, um dies mit zweiseitiger Testung zu erreichen.

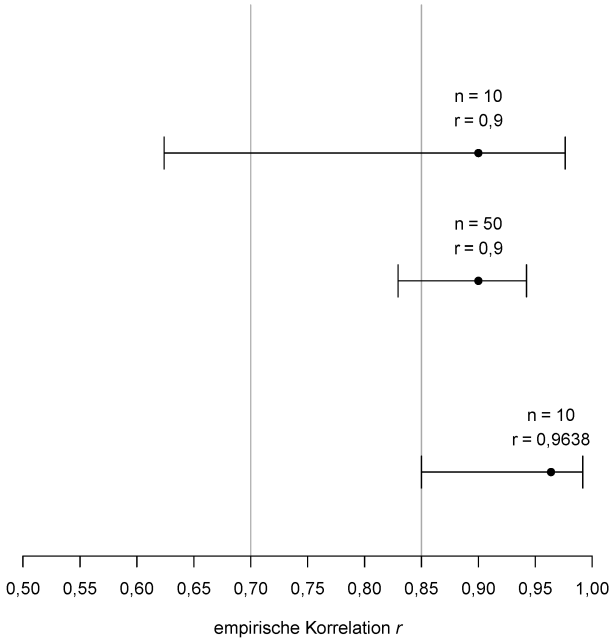

Abbildung 5: Konfidenzintervalle für verschiedene Werte der empirischen Korrelation r und der Studienanzahl n

Die Simulation der Surrogatvalidierung über das korrelationsbasierte Verfahren sowie die analytische Herleitung der Power zeigten, dass diese bei moderater Studienanzahl und starker zugrundeliegender wahrer Korrelation dennoch sehr gering ist. Selbst bei $\rho=0{,}9$ in $n=100$ Studien liegt die Power unter 60% und es wäre $\rho=0{,}95$ in $n=25$ Studien erforderlich, um eine Power von 80% zu erhalten.

Während also der Nachweis für eine hohe Korrelation kaum zu erbringen sein dürfte, ist es ganz einfach, eine mittlere Korrelation nachzuweisen, wenn nur wenige Studien zur Verfügung stehen. Denn selbst bei unkorrelierten Daten (ρ=0) aus n=5 Studien ist die Obergrenze des 95%-KI mit einer Wahrscheinlichkeit von 76,8% größer als 0,7. Für n=10 sind es immerhin noch fast 18%.
Die Beziehung $r^2=R^2$ zwischen Korrelation und Bestimmtheitsmaß gilt nur im Fall des linearen (einfachen) Regressionsmodells. Im Allgemeinen kann aber nicht aus Bestimmtheitsmaßen spezieller Modelle (vgl. Buyse und Burzykowski [3]) auf Korrelation rückgeschlossen werden. Im Rapid Report wird aus den Schwellenwerten der Korrelation offensichtlich auf die Schwellenwerte des Bestimmtheitsmaßes ($r^2=0{,}85^2=0{,}7225=R^2$) geschlossen (S. 71, [6]). Das würde eine Präferenz der Anwendung des linearen Modells in der Surrogatvalidierung bedeuten. In der Nutzenbewertung zu Dabrafenib, in der der pU den Versuch einer Surrogatvalidierung auf Basis des linearen Modells unternommen hatte, wird dieser Ansatz wiederum kritisiert (S. 62, [14]). Es wird im Weiteren noch ausgeführt, dass – abgesehen von der Unsicherheit des linearen Modells – die in Modul 4 vorgelegten Ergebnisse zum Schätzer der Korrelation ohnehin nicht zum Erfolg geführt hätten. Darin präsentierte der pU verschiedene Korrelationsanalysen für die Validierung des Surrogats, unter denen das Modell mit dem stärksten Effekt eine Korrelation zwischen logarithmierten HR_{OS} und HR_{PFS} von r [95%-KI]=0,95 [0,83; 0,99] ergab. Damit liegt gemäß dem im Rapid Report festgelegtem Schwellenwert von 0,85 keine hohe Korrelation vor. Das Beispiel verdeutlicht noch einmal, dass der angegebene Grenzwert von 0,85 (als Mindestmaß für hohe Korrelation) ziemlich konservativ ist. Die Fachliteratur gibt zwar zahllose Vorschläge für Einstufungen des Grades der Korrelation, doch beziehen sich diese auf den Punktschätzer und werden nicht als Untergrenze eines 95%-KI verstanden. Beispielsweise bezeichnet Cohen [15] Korrelation von 0,5 bereits als hoch. Burzykowski [16] erläuterte, dass es schwierig sei, einen Schwellenwert für das Validierungsmaß R^2 anzugeben, da dieses keine intuitive Skala besäße, und zudem abhängig von der Variabilität des Behandlungseffekts des klinischen Endpunktes sei.
Sollte die Validierung des Surrogatendpunkts über das korrelationsbasierte Verfahren nicht gelingen, besteht nach Methodik des Rapid Report noch die Möglichkeit unter Verwendung des STE-Konzepts eine Schlussfolgerung für einen Effekt bezüglich des patientenrelevanten Endpunkts zu ziehen. Die Ergebnisse der zweiten Simulationsstudie zeigen, dass jedoch auch dieses Konzept an realen Daten nur sehr schwer erfolgreich durchgeführt werden kann. Lediglich diejenigen Szenarien, bei denen eine Kombination aus außergewöhnlich hohen Effekten des wahren Endpunkts (HR_{OS}=0,5), hoher Studienanzahl (n=10) sowie hoher Patientenanzahl (N=1.000) angenommen wurde, können eine Power von über 80% erreichen. Für plausiblere Werte der Effekte (HR_{OS}=0,8) liegt die Power bei einer Patientenanzahl N=1.000 maximal bei 47%, bei N=500 höchstens bei 25%.

Beurteilung der angewendeten Simulationsmodelle

Die in beiden Simulationsstudien vorliegenden Szenarien versuchen der Realität entsprechende Beispiele abzubilden. Durch Variationen der Parameter Studienanzahl, Patientenanzahl, Stärke der Korrelation und der Effektgrößen ist ein starker Praxisbezug gegeben. Ein großer Teil der Auswahl der simulierten Parameterwerte charakterisiert allerdings optimistische Szenarien, die in der Wirklichkeit nicht bzw. nur sehr selten vorzufinden sind. So wäre eine für die Validierung herangezogene Studienanzahl n=10 eher unrealistisch. Dies gilt auch für die hoch gewählten Werte der zugrundeliegenden Korrelation. Nichtsdestotrotz zeigen diese – teils optimistisch gestalteten – Szenarien, dass eine Surrogatvalidierung in der Praxis selbst unter optimalen Bedingungen sowohl anhand des korrelationsbasierten Verfahrens als auch anhand des STE-Konzepts eine große Herausforderung ist. Freilich ist anzumerken, dass alle Simulationen auf Basis aggregierter Daten liefen und keine patientenindividuellen Daten verwendeten. Wie eingangs erwähnt, ist die Verfügbarkeit patientenindividueller Daten in der Praxis jedoch im Allgemeinen nicht gegeben. Dieser Aspekt beeinflusst auch nicht das Ziel dieser Arbeit, welcher in der Überprüfung der Anwendbarkeit der Methodik anhand praktischer Beispiele bestand. Abschließend wird nun eine mögliche Alternative zur STE-Methodik vorgestellt.

Möglicher Alternativvorschlag für Surrogatvalidierung anhand STE-Konzept

Die Surrogatvalidierung über das STE erscheint grundsätzlich als geeignete Methode. Allerdings sind die methodischen Vorgaben im Rapid Report sehr streng. Wir möchten im Folgenden von der Forderung absehen, in der das ganze 95%-KI des Surrogateffekts HR_{PFS} unterhalb des STE liegen muss. Durch dieses Kriterium ist es, wie Simulation 2 zeigte, äußerst schwierig einen tatsächlich vorhandenen Effekt auch nachzuweisen. Die Rationale hinter dieser zweifachen Anwendung von Konfidenzintervallen ist die Unsicherheit, mit der beide Schätzer behaftet sind. Durch diese Vorgehensweise sinkt der α-Fehler – zu Lasten der Power – allerdings deutlich unter die 5%-Marke.
In einer abschließenden Simulation untersuchen wir, wie groß der α-Fehler ist, wenn statt der oberen Konfidenzintervallgrenze des HR_{PFS} nur der Punktschätzer mit dem STE verglichen wird. Mit anderen Worten wird die Fehlerwahrscheinlichkeit je Szenario angegeben, mit der aus dem vorliegenden Effekt des PFS auf einen Effekt des OS geschlossen wird, obwohl letzterer nicht vorliegt. Für den Erwartungswert des HR_{PFS} wird 0,7 gewählt, der Erwartungswert des HR_{OS} wird auf 1 gesetzt, es gibt also einen Effekt der Intervention auf das PFS nicht aber auf das OS. Variiert werden wieder die Patientenzahl N pro Studie (200, 500 und 1.000), die Anzahl der Studien n (5 und 10) sowie die wahre Korrelation ρ (0; 0,25; 0,5).

Parameterkonstellationen mit höheren Werten von ρ sind zwar mathematisch gesehen möglich, in der Realität allerdings völlig unplausibel. Inhaltlich entspräche das einem starken Zusammenhang zwischen den beiden Hazard Ratios bei gleichzeitig großer Effektivität der Behandlung hinsichtlich des PFS und völligem Fehlen eines Effekts auf das OS.

Der α-Fehler ergibt sich als Anteil derjenigen von jeweils 10.000 Durchgängen, bei denen der Punktschätzer des HR_{PFS} kleiner als das STE ist. Dies entspricht also einer einseitigen Testung. Wie in Abbildung 6 zu erkennen ist, liegt der α-Fehler für ρ=0 erwartungsgemäß bei etwa 2,5% (=0,05/2). Weiterhin zeigt sich, dass die Patientenanzahl N keinen Einfluss hat, was ebenfalls den Erwartungen entspricht. Mit zunehmender Korrelation und gleichzeitig höherer Studienanzahl n steigt auch der α-Fehler. Für n=10 Studien mit je N=1.000 Patienten und ρ=0,5 liegt er bei etwa 7,6%.

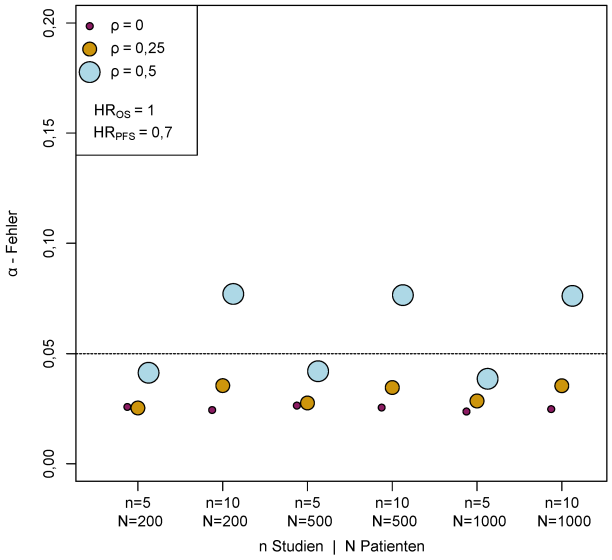

Abbildung 6: Darstellung des α-Fehlers der simulierten Szenarien, die sich durch die Kombination der Parameter Studiengröße *n*, Patientenanzahl *N* und wahrer Korrelation *ρ* ergeben. Die Erwartungswerte der Effektschätzer HR_{OS} bzw. HR_{PFS} hatten die Werte 1 bzw. 0,7.

Bei Durchführung von Simulation 2 wird eine wesentlich höhere Power in allen Szenarien beobachtet, wenn anstelle der oberen Konfidenzintervallgrenze des HR_{PFS} nur der Punktschätzer mit dem STE verglichen wird (Abbildung 7). Eine Power von über 80% wird somit schon für Szenarien erreicht, die zuvor lediglich nur die 20%-Grenze überschritten haben (vgl. Abbildung 4). Nichtsdestotrotz stellen die der Simulation zugrundeliegenden Annahmen, wie beispielsweise eine wahre Korrelation von 0,85 bzw. 0,9, weiterhin eine praxisferne Herausforderung dar.

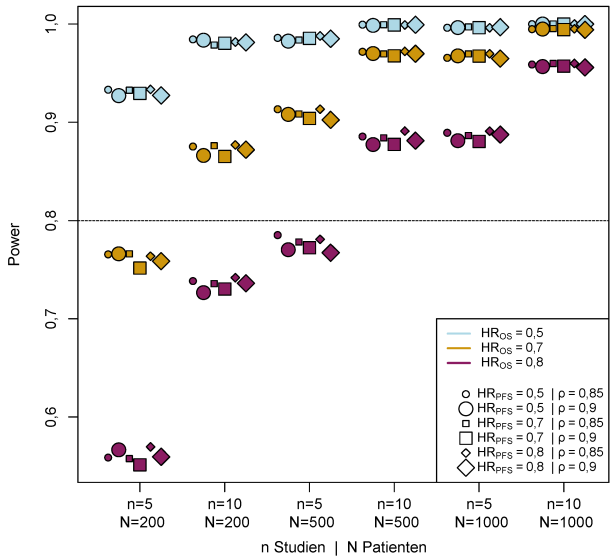

Abbildung 7: Durchführung von Simulation 2, in der anstelle der oberen Konfidenzintervallgrenze des HR_{PFS} nur der Punktschätzer mit dem STE verglichen wird.

Schlussfolgerung

Die durchgeführten Simulationen zeigten, dass die im Rapid Report beschriebene Methodik, wonach die untere Grenze des Konfidenzintervalls ausschlaggebend für eine hohe Korrelation bei der Surrogatvalidierung sein soll, eine in der Praxis kaum zu überwindende Hürde darstellt. Bei gering bis moderat angenommener Studienanzahl – wie es für eine Validierung von Surrogatendpunkten im Rahmen der frühen Nutzenbewertung realistisch erscheint – ist die Power selbst bei hoher, wahrer Korrelation äußerst gering. Problematisch erscheint weiterhin die Empfehlung, die Aussagekraft der Studien in die Analyse mit einzubeziehen, auch wenn dies prinzipiell gerechtfertigt erscheint. Bei Betrachtung der Definition des Korrelationskoeffizienten und dessen Dichtefunktion wird zudem klar, dass die empirische Korrelation unter Annahme einer festen wahren Korrelation gar nicht von der Varianz der Einzelschätzer, sondern nur von der Anzahl der Wertepaare abhängt. Die Patientenanzahl hat somit keine Auswirkung auf das Konfidenzintervall der Korrelation. Dies gilt ebenso, wenn Modelle mit Gewichtung der Studien verwendet werden. Die Anwendung des STE-Konzeptes gemäß der im Rapid Report beschriebene Methodik erscheint ebenfalls schwierig. Ein Vergleich des STE mit dem Punktschätzer des Surrogatendpunkts wäre eine Alternative, die in realistischen Szenarien geringe α-Fehler zeigte. Gegebenenfalls kann in weiteren Simulationen überprüft werden, ob eine vorgeschalteter Test (z. B. ob der ein-σ-Bereich der Verteilung des HR_{OS} unter 1 liegt) den α-Fehler auch in diesen unrealistischen Szenarien kontrolliert.

Anmerkung

Interessenkonflikte

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

Literatur

1. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses (in Kraft getreten am 16. April 2015). 2015. Verfügbar unter: https://www.g-ba.de/downloads/62-492-1002/VerfO_2014-12-18_iK-2015-04-16.pdf
2. International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use. ICH Harmonised Tripartite Guideline. Statistical Principles For Clinical Trials E9 (Current Step 4 Version). 1998. Verfügbar unter: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf
3. Burzykowski T, Molenberghs G, Buyse M, eds. The Evaluation of Surrogate Endpoints. New York: Springer; 2005. (Statistics for Biology and Health). DOI: 10.1007/b138566
4. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar;69(3):89-95. DOI: 10.1067/mcp.2001.113989
5. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat. 2006 Jul-Sep;5(3):173-86. DOI: 10.1002/pst.207
6. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. IQWiG-Berichte – Jahr: 2011 Nr. 80: Aussagekraft von Surrogatendpunkten in der Onkologie. Rapid Report: A10-05, Version 1.1 vom 21.11.2011. 2011. Verfügbar unter: https://www.iqwig.de/download/A10-05_Rapid_Report_Version_1-1_Surrogatendpunkte_in_der_Onkologie.pdf
7. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. Verfügbar unter: http://www.R-project.org/
8. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. J Stat Softw. 2010;36(3):1-48. Verfügbar unter: https://www.jstatsoft.org/article/view/v036i03/v36i03.pdf
9. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007 Jun 7;8:16. DOI: 10.1186/1745-6215-8-16
10. Fisher RA. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron. 1921;1:3-32. Verfügbar unter: http://hdl.handle.net/2440/15169
11. Ripley BD. Stochastic Simulation. Wiley; 1987.
12. Gorsuch RL, Lehmann CS. Correlation Coefficients: Mean Bias and Confidence Interval Distortions. J Methods Meas Soc Sci. 2010;1(2):52-65. DOI: 10.2458/azu_jmmss_v1i2_gorsuch
13. GlaxoSmithKline GmbH & Co. KG. Dossier zur Nutzenbewertung gemäß §35a SGB V. Dabrafenib (Tafinlar®) – Modul 4 A Melanom. Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen. 2013. Verfügbar unter: https://www.g-ba.de/downloads/92-975-391/2013-09-20_Modul4_Dabrafenib.pdf
14. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. IQWiG-Bericht - Nr. 203: Dabrafenib - Nutzenbewertung gemäß § 35a SGB V. Dossierbewertung: A13-35, Version 1.0 vom 23.12.2013. 2013. Verfügbar unter: https://www.g-ba.de/downloads/92-975-393/2013-12-23_A13-35_Dabrafenib_Nutzenbewertung-35a-SGB-V.pdf
15. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1977.
16. Burzykowski T. What Threshold for the validity measure? In: BfArM Workshop; Bonn; 2012.

Korrespondenzadresse:
Johanna Gillhaus
Pfizer Deutschland GmbH, Linkstraße 10, 10785 Berlin, Deutschland
Johanna.Gillhaus@pfizer.com

Bitte zitieren als
Gillhaus J, Goertz R, Jeratsch U, Leverkus F. Surrogatvalidierung durch Korrelation und Surrogate Threshold Effect – Ergebnisse von Simulationsstudien. GMS Med Inform Biom Epidemiol. 2017;13(1):Doc01.
DOI: 10.3205/mibe000168, URN: urn:nbn:de:0183-mibe0001683

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/mibe/2017-13/mibe000168.shtml

Veröffentlicht: *11.01.2017*