



Methodological problems in the method used by IQWiG within early benefit assessment of new pharmaceuticals in Germany

Matthias Herpers¹ · Charalabos-Markos Dintsios² 

Received: 7 November 2017 / Accepted: 16 April 2018 / Published online: 25 April 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Background The decision matrix applied by the Institute for Quality and Efficiency in Health Care (IQWiG) for the quantification of added benefit within the early benefit assessment of new pharmaceuticals in Germany with its nine fields is quite complex and could be simplified. Furthermore, the method used by IQWiG is subject to manifold criticism: (1) it is implicitly weighting endpoints differently in its assessments favoring overall survival and, thereby, drug interventions in fatal diseases, (2) it is assuming that two pivotal trials are available when assessing the dossiers submitted by the pharmaceutical manufacturers, leading to far-reaching implications with respect to the quantification of added benefit, and, (3) it is basing the evaluation primarily on dichotomous endpoints and consequently leading to an information loss of usable evidence.

Objective To investigate if criticism is justified and to propose methodological adaptations.

Methods Analysis of the available dossiers up to the end of 2016 using statistical tests and multinomial logistic regression and simulations.

Results It was shown that due to power losses, the method does not ensure that results are statistically valid and outcomes of the early benefit assessment may be compromised, though evidence on favoring overall survival remains unclear. Modifications, however, of the IQWiG method are possible to address the identified problems.

Conclusion By converging with the approach of approval authorities for confirmatory endpoints, the decision matrix could be simplified and the analysis method could be improved, to put the results on a more valid statistical basis.

Keywords IQWiG · Early benefit assessment · Benefit quantification · Dichotomization · Pivotal trials · Statistical validity

JEL Classification I18

Introduction

Within the early benefit assessment, the Institute for Quality and Efficiency in Health Care (IQWiG) uses a 3×3 matrix, to quantify a potential added benefit of a new drug. For the endpoints analyzed in the dossier, three different categories

are used and if an added benefit can be observed, this benefit is categorized in three different levels[1–3]. This scheme is well-documented in the IQWiG method paper[4] and has also been published in relevant literature[1]. Nevertheless, the IQWiG approach is unique and can hardly be compared to the established methods applied by other Health Technology Assessment (HTA) agencies.

IQWiG assumes that the international regulatory standard of two pivotal phase III studies is met, and that these two studies are pooled for the early benefit assessment. Based on the pooled data set, IQWiG has derived upper bounds of the confidence limits for each of the three types of endpoints and each of the three levels of added benefit, resulting in the above-mentioned 3×3 matrix used in the evaluation.

This matrix is designed, such that an added benefit for the endpoint ‘overall mortality’ can be demonstrated easily. Also, any prolongation of survival has per default at least

Matthias Herpers and Charalabos-Markos Dintsios contributed equally to the work.

✉ Charalabos-Markos Dintsios
dintsios@hhu.de

¹ ClinStat GmbH, Cologne, Germany

² Institute for Health Services Research and Health Economics, Medical Faculty, Heinrich-Heine University Düsseldorf, Gebäude 12.49, Moorenstraße 5, 40225 Düsseldorf, Germany

a minor added benefit, whereas there is a lower boundary for all other endpoints in place, below which an additional benefit will not be concluded, even if a difference between the new therapy and the reference therapy was proven.

Points of criticism

As every HTA body, IQWiG has to apply a standardized method and when designing such a method, has to stick to the law [5, 6]. There are, however, some points of critique with regard to the method introduced by IQWiG:

1. By favoring ‘overall mortality’, drugs also that have mortality as primary endpoint in the phase III studies may be preferred over drugs whose primary endpoint is not overall survival, even if the drug may be a therapeutic breakthrough in the respective indication. Such a drug may, nevertheless, not be granted a major added benefit, simply because the endpoint, in which success in this specific indication is measured, falls in a less favored IQWiG category [2, 5].
2. Despite the regulatory norm of two pivotal phase III studies, the reality currently seen in drug approval is that more and more drugs are approved with only one pivotal study. IQWiG, however, uses its evaluation matrix for such dossiers as well, thereby, applying confidence bounds calculated for at least twice the number of patients, making it harder to stay below the critical upper limit for dossiers with less patients than anticipated [2, 5, 7]. This phenomenon is even reinforced, when only a small subgroup of the entire study population becomes relevant for IQWiG’s assessment [8, 9].
3. IQWiG always evaluated a dichotomous endpoint and if the original endpoint was not measured in a dichotomous manner, IQWiG will dichotomize such an endpoint. This, however, is accompanied with a loss of information and usually results in a loss of precision (i.e., a wider confidence interval) and could lead to the situation that an added benefit, which can be demonstrated in the original endpoint, is either of lower effect size or even disappears in the dichotomized endpoint [2].

These points are dealt with in the present article.

The IQWiG method in detail

In a first step, every endpoint is transformed to a 2×2 matrix, with the four fields being the product of a response category (yes/no) and the treatment category (new/reference). Based on this matrix, relative risks or hazard ratios, respectively, are calculated. The governing law, the Legislative Decree on the benefit assessment of pharmaceuticals (in German Arzneimittel-Nutzenbewertungsverordnung:

AM-NutzenV), requires that different levels of added benefit are derived based on the effect size, and IQWiG interprets the Decree in a way that it presupposes to consider not only the magnitude of the effect, but the outcome itself (e.g., mortality, adverse events, health-related quality of life etc.), as well [10]. IQWiG has decided to use the upper bound of the 95% confidence intervals to measure the effect size. For ratios, this means that the closer the upper bound of the 95% CI is to 1, the lesser the effect. Since two pivotal studies are assumed to be available and every single study is designed and powered to show a statistically significant effect, the pooled analysis has to be adjusted to the higher number of patients, to derive meaningful confidence intervals. Based on the 2×2 matrix, IQWiG assumes that the dichotomized endpoint is tested by a chi-squared test. IQWiG sets a relative risk of 0.5 to represent a relevant effect and then applies the respective formulas to derive corresponding confidence bounds for a major added benefit [1, 11].

IQWiG interprets the Legislative Decree such that different types of endpoints must be evaluated differently, as well [1, 10]. This is how IQWiG gets to the 3×3 evaluation matrix it does apply in the early benefit assessment (see Table 1). The confidence bound calculated for two studies and a relative risk of 0.5 is used as reference for a major added benefit, demonstrated in overall mortality. To fill the other fields of the 3×3 evaluation matrix, IQWiG divides the range of relative risks in six parts and assigns each of the six values to one field in their matrix. The assumed relative risk increases with decreasing added benefit, and compared to the risk for a given level of added benefit in the overall mortality endpoint category, the other endpoint categories will get a lower relative risk (i.e., it will be harder to demonstrate the same level of added benefit in other categories compared to mortality). Based on the given relative risks and the assumed sample size, the necessary confidence bounds for all fields of the matrix are derived.

This approach is based on two assumptions that in reality may be violated:

1. Data from two studies are available and can be pooled for the benefit assessment.
2. Endpoints are dichotomous.

Furthermore, the relative risks and the confidence bounds derived from them are quite arbitrarily chosen by IQWiG. Only the starting relative risk of 0.5 for a major benefit in survival is justified by IQWiG, citing one single reference [11]. All other relative risks were chosen by IQWiG in “pragmatic manner” to be a multiple of 1/6 of the interval 0–1 [1].

Table 1 IQWiG's outcome-specific decision matrix

IQWiG approach	Outcome category			
	All-cause mortality	Serious (or severe) symptoms (or late complications) and adverse events	Health-related quality of life	Non-serious (or non-severe) symptoms (or late complications) and adverse events
Benefit extent category				
Major sustained and great improvement in the therapy-relevant benefit, which has not previously been achieved vs the appropriate comparator therapy	Major increase in survival time CI ₁ : 0.85 (RR ₁ = 0.50)	Long-term freedom or extensive avoidance CI ₁ : 0.75 (RR ₁ = 0.17) Risk ≥ 5% ^b	<i>Major improvement^a</i> CI ₁ : 0.75 (RR ₁ = 0.17) Risk ≥ 5% ^b	<i>Not applicable</i>
Considerable marked improvement in the therapy-relevant benefit, which has not previously been achieved vs the appropriate comparator therapy	Moderate increase in survival time CI ₁ : 0.95 (RR ₁ = 0.83)	Alleviation or relevant avoidance CI ₁ : 0.90 (RR ₁ = 0.67)	<i>Important improvement^a</i> CI ₁ : 0.90 (RR ₁ = 0.67)	Important avoidance CI ₁ : 0.80 (RR ₁ = 0.33)
Minor moderate and not only marginal improvement in the therapy-relevant benefit, which has not previously been achieved vs the appropriate comparator therapy	<i>Any increase in survival time</i> CI ₁ : 1.00	<i>Any reduction</i> CI ₁ : 1.00	<i>Relevant improvement^a</i> CI ₁ : 1.00	Relevant avoidance CI ₁ : 0.90 (RR ₁ = 0.67)

Amendments to the Legislative Decree on the benefit assessment of pharmaceuticals (AM-NutzenV) in italics

CI₁ threshold for the upper limit of the 95% confidence interval, RR₁ true relative risk

^aProvided that a validated instrument with a validated response criterion is used

^bFor at least one of two compared groups

Does the IQWiG method prefer indications?

As stated above, the IQWiG method prefers the endpoint mortality by making it easiest to get an added benefit approved in this category. One might expect that this leads to a preference of indications that have survival as the primary endpoint, irrespectively of the effect size. To test this hypothesis, all available dossiers were downloaded from the IQWiG homepage [12] and transferred to a database. The dossiers were downloaded on 11-DEC-2016. Dossiers not published at that date, were not considered for the analysis. Filtering the published dossiers to:

Department	Drug assessment
Report type	Dossier assessments
Therapeutic area	All
Year	All

resulted 174 available dossiers.

To test the hypothesis, the indication of each dossier was recorded in the database as well as a high-level indication (e.g., all oncological indications were grouped together). The 174 dossiers were about 121 substances in 23 high-level indications. For 38 substances, more than one dossier was available, either because that substance has more than one indication or after a temporarily granted benefit an update of the dossier with additional data had to be re-submitted to IQWiG. A large portion of dossiers (102 of 174) could not be analyzed, because there was no demonstrated added benefit in any endpoint or because IQWiG deemed the dossier unsuitable. The main reason of IQWiG to reject a dossier was the inclusion of unsuited studies (44 of 102). In 34 of these, there was no study available at all and in the other 10 cases, the submitted studies were not accepted by IQWiG. In 27 dossiers, the reference treatment (appropriate comparative therapy) chosen by the sponsor was not the reference treatment set by the Federal Joint Committee (in German Gemeinsamer Bundesausschuss: G-BA). IQWiG evaluated all endpoints according to its 3 × 3 matrix. Therefore, the analysis of the remaining dossiers could be performed on an endpoint level. In the 72 dossiers, eligible for analysis, 568 endpoints were assessed. Of these, 403 were risk ratios, hazard ratios, odds ratios or were transformed to such for the early benefit assessment. These 403 endpoints were used for analysis. The fact that 165 endpoints were other measures, e.g., mean differences, Hedges g , is quite remarkable, given that IQWiG actually requires a dichotomous or dichotomized endpoint for the dossier.

If the hypothesis is true that IQWiG's method prefers indications, for which the primary endpoint is measured as survival, this should result in a preference of oncological

or cardiovascular indications, as survival as a primary endpoint will be more commonly used in these indications than in the others.

Methods

To investigate the question, if the IQWiG method prefers some indications over others, all 403 evaluable endpoints and their respective evaluation by IQWiG were analyzed. To compare the different types of endpoints, homogeneity between the four types of endpoints would be necessary. This is tested by comparing the four mean estimates of the endpoint types by an F test. The underlying hypothesis for the F test is:

H_0 : All group means are equal

versus

H_1 : At least one pair of means is different.

The endpoints are rates and rates usually are distributed log-normal. Therefore, log-transformed values were used for the F test. The F test was also used to compare the mean upper confidence interval bounds of the four types of endpoints.

If the F Test shows differences between the mean estimates or the mean upper confidence limit, the analysis of the IQWiG evaluations has to take this non-homogeneity into account. This would be done by statistical modeling. The variable to be explained, added benefit, is ordinal with more than two factors. This means that the appropriate statistical model for this type of variable is the multinomial logistic regression. The model equation would be:

added benefit = indication + upper confidence limit.

If the F test shows no difference between the endpoint types, the model would be reduced to:

added benefit = indication.

The logistic regression models chances (odds) and the calculated estimators from the model can be interpreted as odds ratios after transformation using the exponential function. With the target variable being qualitative, a reference must be given, to model the odds ratios for the other categories compared to the reference. It is assumed that the IQWiG method favors fatal diseases. Therefore, malignant tumors were set as reference for the analysis. Results of the logistic regression are, therefore, the chances of an indication to get a higher added benefit than a malignant tumor, if necessary adjusted for differences in the upper confidence limits. This approach uses all available endpoints and one can object that even if IQWiG has not specified how different benefit signals will be combined to come to an overall judgement for a compound, it can safely be assumed that only the strongest

signal will in the end be the used one to set the overall added benefit, provided no signal of harm was detected. Using all endpoints, therefore, would be a biased analysis, as endpoints not relevant for the overall added benefit are included as well. For sensitivity reasons, the analysis will be repeated with the highest rated endpoint only. In case that endpoints of different types have the same added benefit, and that this is also the highest rated added benefit, the following order will be used to decide which endpoint is included in the sensitivity analysis:

Mortality → Morbidity → Side Effects/AE → Health-related quality of life

This order was chosen, as it emulates the IQWiG decision matrix. The IQWiG matrix differentiates between serious adverse events (SAE) and non-serious adverse events (AE), which we do not in our analysis to keep it more comprehensible. The IQWiG matrix values equally SAE and health-related quality of life (HRQoL). However, in most cases where some data on HRQoL were presented, no added benefit could be derived [13]. Furthermore, for explorative HRQoL endpoints, no agreement between market authorization and early benefit assessment is ascertainable [14]. Thus, we decided to rank our category *side effects* higher than HRQoL.

If not mentioned otherwise, analyses were conducted using the software SAS Version 9.4 (SAS Institute, Cary, North Carolina).

Results

Table 2 shows the summary statistics for the risk ratios and the upper confidence limits. The one observation with

“severe symptoms” comes from dossier A15-29 Edoxaban and was used there as such by IQWiG, without assigning it to one of the four categories. This observation is, therefore, excluded from further analyses. The result of the *F* test shows that the various endpoint categories differ and consequently the upper confidence bound needs to be included in the final model as explanatory variable.

Results of the logistic regression (Table 3) show that the odds for comparison of malignant tumors with the other indications are in favor of tumors as well as against

them. Remarkable is the estimated odds for metabolic disorders, which is favored over tumors. This may probably be due to the fact that most substances in that class are for treatment of diabetes mellitus and most of the dossiers were rejected by IQWiG due to an inappropriate comparator or an inappropriate implementation of the control intervention. The comparator given by the Federal Joint Committee (Sulfonylurea), however, does not represent the current standard of care and, therefore, any drug that actually provides accepted comparative study data for mortality vs that treatment has quite high chances to prove a major added benefit [i.e., for the combination of sitagliptin with metformin compared with sulfonylurea (glipizide) plus metformin]. However, all confidence intervals for the odds (except for lung diseases) include the neutral value 1, which means that indication does not have a statistically significant influence on the assigned added benefit. This also means that the underlying hypothesis that oncological indications are favored by the IQWiG methodology, could not be proven.

Table 2 Summary statistics for the risk ratios and upper confidence limits (all endpoints)

Variable	Mortality	Morbidity	Side effects/AE	HRQoL	Severe symptoms	<i>p</i> value
Risk ratio						
<i>N</i>	74	129	163	36	1	
Mean	0.76	0.68	0.46	0.69	0.89	< 0.0001
Std	0.53	0.46	0.23	0.21		
Min	0.14	0.03	0.03	0.33	0.89	
Med	0.73	0.60	0.47	0.67	0.89	
Max	4.31	3.10	1.14	1.47	0.89	
Upper CI bound						
<i>N</i>	74	129	163	36	1	
Mean	1.75	1.51	0.71	0.94	0.96	< 0.0001
Std	4.26	2.93	0.27	0.29		
Min	0.54	0.13	0.16	0.72	0.96	
Med	0.93	0.87	0.74	0.89	0.96	
Max	36.38	29.48	2.37	2.26	0.96	

Table 3 Results of the logistic regression (all endpoints)

Comparison	Odds ratio		
	Lower CI	Estimate	Upper CI
Eye diseases vs malignant tumor	0.187	1.346	9.663
Vascular diseases vs malignant tumor	0.299	0.580	1.125
Cardiac diseases vs malignant tumor	0.722	2.295	7.297
Immunosuppressive agents vs malignant tumor	0.277	0.999	3.606
Infections vs malignant tumor	0.700	0.868	10.772
Blood diseases vs malignant tumor	0.132	2.051	31.914
Pulmonary diseases vs malignant tumor	0.127	0.350	0.963
Nervous system disorders vs malignant tumor	0.558	1.789	5.734
Metabolic disorder vs malignant tumor	0.499	1.215	2.958
Viruses vs malignant tumor	0.438	0.800	1.462
Benign tumor vs malignant tumor	0.100	0.346	12.154

Table 4 Summary statistics for the risk ratios and upper confidence limits (highest rated endpoints)

Variable	Mortality	Morbidity	Side effects/AE	<i>p</i> value
Risk ratio				
<i>N</i>	30	21	21	
Mean	0.59	0.46	0.38	0.0045
Std	0.18	0.25	0.19	
Min	0.22	0.03	0.10	
Med	0.59	0.44	0.42	
Max	1.02	1.06	0.64	
Upper CI bound				
<i>N</i>	30	21	21	
Mean	0.91	0.83	0.62	0.0023
Std	0.33	0.41	0.20	
Min	0.54	0.22	0.24	
Med	0.82	0.81	0.67	
Max	2.20	2.31	0.91	

The baseline status of the sensitivity analysis, which included the highest available endpoint per drug only, shows similar results as the main analysis. The mean upper bounds of the confidence intervals are different for the types of endpoints and have to be included as corrective factor in the analysis (Table 4).

When looking at the frequencies of added benefit per indication, one problem of this analysis can be observed: some combinations have no counts and the remaining classes have very few counts (Table 5). This results in failure of the numeric algorithms to provide estimates from the logistic regression.

In summary, the sensitivity analysis did not have sufficient observations to provide estimates for the statistical analysis and the main analysis, which included all endpoints, did not support the hypothesis that the IQWiG methodology favors indications with survival endpoints, like malignant tumors. For a conclusive analysis of this question, more data

Table 5 Frequency count of added benefit by indication

	<i>N</i>	Added benefit			
		Major <i>N</i> (%)	Considerable <i>N</i> (%)	Minor <i>N</i> (%)	None <i>N</i> (%)
Eye diseases	1	1 (100.00)	0 (0.00)	0 (0.00)	0 (0.00)
Vascular diseases	5	0 (0.00)	4 (80.00)	1 (20.00)	0 (0.00)
Cardiac diseases	2	0 (0.00)	2 (100.00)	0 (0.00)	0 (0.00)
Immunosuppressive agents	2	0 (0.00)	1 (50.00)	0 (0.00)	1 (50.00)
Infections	1	0 (0.00)	0 (0.00)	0 (0.00)	1 (100.00)
Blood diseases	1	0 (0.00)	1 (100.00)	0 (0.00)	0 (0.00)
Pulmonary diseases	6	0 (0.00)	2 (33.33)	3 (50.00)	1 (16.67)
Nervous system disorders	3	0 (0.00)	2 (66.67)	1 (33.33)	0 (0.00)
Metabolic disorders	5	3 (60.00)	1 (20.00)	0 (0.00)	1 (20.00)
Viruses	8	2 (25.00)	5 (62.50)	0 (0.00)	1 (12.50)
Benign tumor	1	1 (100.00)	0 (0.00)	0 (0.00)	0 (0.00)
Malignant tumor	37	25 (67.57)	8 (21.62)	1 (2.70)	3 (8.11)

is needed and with additional benefit dossiers being published constantly, this analysis may be repeated in the future.

Problems inherent to the IQWiG methodology

Since the early benefit assessment starts immediately after market approval of a drug, the dossiers are based on the phase III studies also used for market authorization of the drug. This causes some problems, because the approach taken by IQWiG to answer the question of an added benefit is different from the approach taken by an approval authority, like EMA or FDA. For approval, the main focus is to prove efficacy of a new drug. Phase III studies are, therefore, planned for usually one or at most two primary endpoint(s) that are accepted as valid endpoints for efficacy in the indication of interest. Approval of the drug is tied to the outcome of the associated statistical test:

H_0 : The drug does not work

versus

H_1 : The drug does work.

It does not really matter with which statistical method this hypothesis is tested for a specific indication. Important is the fact, that the whole study is planned to test this one hypothesis. When planning a study, three parameters have to be kept in mind:

- Type I error (α)
- Type II error (β)
- Clinically relevant difference in the primary endpoint (Δ)

The clinically relevant difference is usually agreed upon with the approval authority beforehand and is, therefore, set. The type I error denotes the error when the statistical test rejects the null hypothesis incorrectly. This can be interpreted as the regulatory risk, as it would mean that a non-working drug would get approved. The approval authorities have an interest to keep this error as low as possible, and the traditional agreement for approval studies is $\alpha = 0.05$. For approval of a drug, usually two independent studies have to show the effectiveness of a drug for this given type I error of 5%. Exceptions are, however, possible, especially if the observed p value is very low, a deviation from the two-study requirement is possible [15, 16]. Nowadays, the exception has become the standard especially for pharmaceuticals with oncological indications and a lot of approvals are based on only one phase II or III pivotal study and only the results of the phase II or III study/studies are then incorporated in the value dossier [17]. This is a further problem of the IQWiG method, which is covered

further on. For the current problem, only the fact that through regulatory requirement, the type I error given as well is relevant. Therefore, the only variable that can be changed when planning a phase III study is the type II error. This is the error of not rejecting the null hypothesis, although it should have been rejected. This can be seen as the sponsors' risk, because it would mean, that a working drug is not identified as such. Given the two variables (α , Δ), it holds, that the lesser the type II error, the higher the required sample size for the study. Balancing economic and ethical aspects of a study, a range of 10–30% for the type II error has been proven to be feasible. If looking at the complementary probability ($1 - \beta$), instead of β , this represents the chance that a working drug is identified as such. This is called *power*. Given the above-mentioned error rates, typical phase III studies have a power of 70–90%, with a power of 80% being the desired target in planning. In summary, the sample size in a phase III study was planned for the clinically relevant difference in the indication, $\alpha = 0.05$ and $\beta \cong 0.2$. If there is more than one endpoint for a given indication, adjustments have to be made (see below), but the typical case is one endpoint. IQWiG, however, does look at all endpoints of a study and this is an inherent problem, because for the given sample sizes, the values of α and β in this endpoint are unknown. As the approval authorities, IQWiG sets $\alpha = 0.05$, which still leaves β as completely unknown for the given endpoint. The given sample size, therefore, can be too high or too low. Too low a sample size is the case that can be understood most easily. If the sample size is too low, the type II error increases, because α and Δ are set. With increasing type II error, the reliability of the statistical test decreases. A statistical test can be seen as an arbitrary tool for decision making, obviously sophisticated enough, to protect against errors. But if the power falls below 50%, the decision to reject the null hypothesis or not, would be better taken by a coin toss, which has a power of 50%, instead of the statistical test. Too low a sample size, leads sooner or later to the point at which the statistical test loses its validity. Too high a sample size seems to be less problematic, because it would increase the power of the test and, therefore, the validity. To understand that too high a power, does also have adverse effects, one has to look at the clinically relevant difference (Δ). Δ is the limit that must be achieved from a medical perspective, at least to make a difference in treatment of patients. When planning a study, this is usually given. If the drug does not reach the pre-specified Δ , the null hypothesis will not be rejected, or at least there is a high probability for it, to be not rejected. If the power of a test increases, it does so for the "real" Δ , but it does so for all other potential values of Δ as well, even for the values considered to be too low, from the medical treatment perspective. In other words, if the power of a test is too high, the chances increase that the null hypothesis is rejected for a Δ that is not considered clinically relevant, but too low. Too high a power, therefore, can actually increase the probability

to reject the null hypothesis for a drug, with insufficient effect. One can now object that IQWiG uses confidence intervals and not tests, but both are interrelated. 95% confidence interval actually means that there is a 5% risk of having a true effect outside the confidence interval. So, the 95% represent $1 - \alpha$ of a corresponding test. The decision, to correctly reject the null hypothesis, would be equivalent to the confidence interval not including the neutral value (e.g., 1 for rates), and the power of a test, is corresponding to the probability of the confidence interval, excluding the neutral value correctly. In this sense, it does not matter, if the decision is based on a test or on a confidence interval and, therefore, the term of statistical tests (e.g., power) is used in this text also for decisions, based on confidence intervals.

The studies that are submitted to IQWiG have not been planned for this kind of evaluation, and the fact that IQWiG analyzes all available endpoints is, therefore, a problem, because it is completely unknown, which type II error is associated with the various endpoints. And, this in conclusion means that it is unknown, with what level of validity the results are associated. Given the process of the early benefit assessment, it is almost impossible to submit a dossier with studies designed for the benefit analysis, although this would be the gold standard. Having to accept the sub-optimal situation as it is now, the problem of the unknown β is difficult to solve. The most meaningful approach would be that IQWiG would require a power calculation for all endpoints, given the sample size of the included studies, and include only endpoints with a reasonable power (e.g., 70–90%) in the early benefit assessment.

As mentioned above, having more than one endpoint for decision making is a problem of its own. If more than one test is conducted, there are two types of type I error:

1. The type I error of each individual test. This is called *comparisonwise type I error* and is denoted by α_c .
2. The overall risk of having at least one test in which the type I error is committed. This is called the *experimentwise type I error*, denoted by α_e .

If n independent tests are conducted and the null hypothesis would be true for all tests, then $(1 - \alpha_c)$ would be the probability of a correct test decision for each individual test. The probability of x type I error in all n tests is the binomial distributed and given as:

$$P(x) = \frac{n!}{x!(n-x)!} \alpha_c^x (1 - \alpha_c)^{n-x} \quad (x = 1, \dots, n).$$

The probability to conduct no type I error is then:

$$P(x = 0) = (1 - \alpha_c)^n.$$

And, the probability of at least one type I error:

$$P(x \geq 1) = 1 - P(x = 0) = (1 - \alpha_c)^n = \alpha_e.$$

If $n = 5$, then the probability of conducting at least one type I error in the 5 tests is 0.226, which is considerably higher than the individual type I error of $\alpha_c = 0.05$.

There are multiple methods, to adjust for multiple tests and keep the experimentwise error at 0.05 [18–23], but taken the high number of endpoints in an IQWiG dossier, even these methods would be driven to their limits. Since IQWiG does look at the overall picture as presented in the dossier, it may be acceptable, if a few of the endpoints, showing an added benefit, do so erroneously. That there is some kind of added benefit would not be doubted.

In summary, the problem of the completely unknown power is an issue that is caused by the IQWiG approach, and which casts a doubt on whatever is seen in a dossier, unless the added benefit is seen in the primary endpoint, for which the studies were planned.

Number of available studies

Another issue with the method used by IQWiG is that the critical limits for the risk ratio and the confidence interval are calculated, based on the assumption that two studies are available and part of the dossier. As stated earlier, this is not reflecting the regulatory reality, and more and more approvals are based on only one study. A second problem is that IQWiG requires dichotomous variables and, if necessary, variables have to be transformed to a 0–1 scheme for the dossier. The consequences of these two assumptions are investigated in this section.

Two studies are assumed but not available

Table 6 shows that in nearly 3 out of 4 dossiers, submitted to IQWiG, only one study was included. Since IQWiG does not adjust the critical limits to this fact, this is a problem. In [1] IQWiG describes detailed, how it derives the critical bounds, based on a chi-squared test and two studies. The critical upper confidence bound of 0.85 for the reference category is calculated, based on two studies powered at 80%. But what happens, if only one study is available and, therefore, the number of patients is only half the assumed size? Furthermore, even in cases with more

Table 6 Number of studies in the available dossiers

# studies	Frequency	
	Absolute	Relative (%)
1	52	72.22
2	15	20.83
3	5	6.94

The number of studies refers to the dossier in general and not to specific patient subpopulations being evaluated

than one study, a meta-analysis might not be feasible due to heterogeneity or different patient groups. Even worse with regard to potential power losses, becomes the situation, where only a subgroup of the entire study population is included in IQWiG's assessment.

Methods

To investigate the consequences, the necessary sample size for a chi-squared test for different values of p_1 and a relative risk of 0.5 with a power of 80% is calculated. Then, this number is cut by half, and the resulting power is calculated.

Results

Table 7 shows that with only half the needed sample size, the power falls to values around 50%. This means, that only because of the fact that one instead of two studies is forming the dossier, the power falls to a critical value, which puts the statistical test on one level with a coin toss. This applies to the primary endpoint only, as only for this endpoint it is known that the study was designed to achieve 80% power. For all other endpoints, the previously mentioned problem that the power is not known is just worsened, because whatever power may be there in case of two studies, it is dramatically reduced, if only one study is available. The calculations assume that the original study had a dichotomous endpoint. If that is not the case, dichotomization will increase the problem of loss of power even further, as we will see below. Furthermore, this problem worsens, if only a subpopulation instead of the entire labeled population is used for the early benefit assessment.

Table 7 Power calculations for chi-squared test with only half the needed sample size

Risk p_1	Relative risk	Alpha	Power
0.1	0.5	0.05	0.507
0.2	0.5	0.05	0.504
0.3	0.5	0.05	0.503
0.4	0.5	0.05	0.507
0.5	0.5	0.05	0.503
0.6	0.5	0.05	0.498
0.7	0.5	0.05	0.483
0.8	0.5	0.05	0.480
0.9	0.5	0.05	0.483

Endpoint dichotomization

IQWiG dichotomizes endpoints, to fit study endpoints into the evaluation matrix. For endpoints not measured on 0–1 categories, that means a loss of information. The consequences of this loss are investigated in this section.

Methods

Parkinson's disease is chosen as example. The severity of the disease is measured by the Unified Parkinson's Disease Rating Scale (UPDRS). The UPDRS is calculated by answers to 42 questions, and the range of the UPDRS score is 0–199 [24]. Due to the wide range, this scale can be assumed to be nearly continuous and the distribution is a quasi-normal. So, the UPDRS score is tested, using the common t test. A reduction of at least 8 points in the UPDRS scale is considered to be a clinically relevant difference [24]. In clinical studies, a baseline score of around 31 points was observed, with a standard deviation of around 12 points [25, 26]. With these data, the needed sample size for a power of 80% can be calculated. With this sample size, a data set of normally distributed data is then created, having a reduction of 8 points in one group and 0 in the other. For dichotomization, a second variable is calculated, based on the randomly generated data, resulting in 1 for response if the difference is at least 8 points and 0 otherwise. For the dichotomized variable, the chi-squared test is then calculated. By means of Monte Carlo simulation, this is repeated 5000 times to estimate the power of the chi-squared test on the dichotomized variable. Based on this real life example, the simulations are repeated with other, artificial, values for differences of the mean and standard deviation, to get a feeling for the range of loss of power.

Results

Using SAS PROC POWER, a needed sample size of $N=37$ per group was calculated for a difference of at least 8 points between the groups, a baseline value of 31 points and a standard deviation of 12 points. Monte Carlo simulations were performed also for the t test, to confirm that 5000 runs are sufficient. The simulated power for the t test with 5000 iterations was 80.18%, which is a very good result, given the theoretical power of 80%, and shows that 5000 runs are sufficient for reliable results.

The same 5000 runs, using the dichotomized variable and a chi-squared test instead of the original variable and the t test, resulted in a power of 58.22%. In this example, dichotomization resulted in a loss of power of over 20%.

Using means differences of 1–5 points and a standard deviation of 5 and 8 points, respectively, these results were

confirmed. Dichotomization from normally distributed data leads to a loss in power of approximately 20%, with results having a variability of only a few percentage points.

Discussion

Interpretation of the results

The IQWiG methodology is subject to various criticisms. Our analyses investigated some of this critique. The first point was that due to the different types of endpoints and different critical bounds, indications measuring their success in survival, could be favored by IQWiG's methodology. Analysis of all dossiers by means of logistic regression could not verify this hypothesis; however, due to the limited number of dossiers (74), the results of this analysis are not valid enough, to answer the question definitively. The analysis will have to be repeated in the future, when a meaningful number of dossiers can be included.

The fact, that IQWiG includes all available endpoints, leads, without any adjustment for multiple testing, to an unknown power and an unknown number of positive results, that may happen to be the outcome of a high experimentwise type I error rate only. Both issues are obvious, and at least for regulatory approval of new drugs adjustment for multiplicity is required [27]. With multiple methods for adjustment available, it is surprising, that IQWiG ignores this problem completely.

Another problem is that IQWiG still assumes that two pivotal studies form the dossier, despite their own data, showing that this is not the case nowadays. Lacking the second study, we could show, that the power could drop down to the critical limit of 50% or, if the power was below 80% in the study, it could fall even below 50% in the IQWiG analysis, making a coin toss more reliable than the statistical test. It is more than surprising, that IQWiG does not adjust its decision boundaries to the number of studies at all, and stays in a virtual reality of the two studies gold standard.

The last point of critique was that IQWiG dichotomizes all endpoints, not originally available as 0–1 variable. For data originally normally distributed and analyzed by the t test, it was shown that this decreases the power of the test by approximately 20%.

In summary, IQWiG is facing a major problem with the power of the analyses it is performing on the endpoints. For most endpoints, the power is unknown, but due to methodological issues, it can safely be assumed that the power will be low anyway. It is, therefore, reasonable to investigate how this problem can be overcome and if this is feasible at all.

Modification of the IQWiG method

Unequal treatment of endpoints and a problem with reliability of the results, due to an unknown but presumably low power, are the two problems of the IQWiG method. Looking at the text of the Legal Decree, and in contrast to the interpretation of the IQWiG, we do not see the strict necessity for different types of endpoints. But, similar treatment of all endpoints should remain an option, when modifying the IQWiG methodology.

The problem of an unknown power, however, can be addressed and corrected. The current method has already included a mechanism to protect against too high a power. If the power is too high, this would result in a clinically not relevant difference (i.e., a higher than assumed risk ratio) but an upper confidence bound below the critical value. If IQWiG would decide, if added benefit can be established or not, not only on the upper confidence interval bound, but also on the estimated risk ratio, a too high power would result in failure to meet both criteria. And, since IQWiG sets the risk ratios to calculate the critical values for the upper bounds, it can easily use both values in the decision.

However, too low power seems to be the more prominent problem, as shown in the previous part. As indicated, it may be useful to include only endpoints with at least a minimal power in the analysis. This would ensure that each single endpoint analysis has certain validity. It would still not adjust for multiplicity, but the more signals for added benefit are seen, the less important the single endpoint will become. Multiplicity will only be a problem, if the final decision is based on very few, or in an extreme example on only one endpoint. Including only endpoints with a minimal power, would reduce the number of endpoints and to a certain degree also the chances to get an additional benefit granted, but the decision itself would come with a higher reliability. Furthermore, pharmaceutical manufacturers could include HTA endpoints in their phase III study planning, to adjust for such a modified methodology, and ensure that a sufficient number of endpoints meet the requirement of minimal power.

As already indicated in the power analysis earlier, IQWiG should also use multiple matrices for evaluation, depending on the number of studies included. The formulas are available and easy to implement. It is quite surprising that this is still not implemented, given the loss of power observed.

Dichotomization proves to be the trickiest problem of the IQWiG methodology. A form of standardization and also simplification is necessary, if IQWiG wants to follow a similar process for each dossier. The alternative would be an extensive consulting and evaluation process, similar to the process when planning the submission of a new drug application (NDA). For a HTA evaluation, a process similar to a NDA, however, is simply too time consuming, too expensive

and too complex. Having a method applied that is simple and transparent, is therefore necessary, to uphold a national HTA process. Despite its downside, dichotomization, therefore, seems to be the only way for the early benefit assessment. In summary, a modified IQWiG procedure could look like this:

- Non-dichotomous endpoints will be dichotomized
- Only (dichotomized) endpoints with a minimal power are included. Since 50% is the critical value, a relevant distance to that value has to be maintained. Given that 80% is the reasonable target in clinical studies, a lower level of 70% is suggested, but is obviously open for discussion. Important is that a lower level for inclusion is set at all.
- Currently, side effects are analyzed singularly. This causes a lot of signals, partly confusing and makes it hard to come to an overall conclusion, regarding harm or benefit of the new drug. It is proposed, to analyze them comprehensively using Cox regression with adjustment for repeating events and/or competing events. Serious and non-serious side effects should be analyzed separately. Since this would increase the power of the analysis, a lower type I error level should be applied, e.g., 1% instead of 5%.
- The categories for endpoints are dropped. The critical values of the current category “mortality” are applied to every endpoint.
- Risk ratio and upper confidence interval bound are used, to evaluate the added benefit. Only if both are met, the respective benefit is granted. If there is a mismatch, the lower category of benefit is granted only.
- The critical values for risk ratio and upper confidence interval bound are adjusted for the number of studies.
- If only endpoints with a minimal power, i.e., an acceptable statistical validity, are included in the analysis, the highest observed benefit category could be taken as the overall level of benefit granted for the drug, provided there is no evidence of a harmful effect.

This algorithm is simple, transparent and could be considered in phase III planning by the pharmaceutical manufacturer. The benefit, compared to the current process, is that only endpoints with a proven reliability are included, and that the overall assessment therefore can be trusted more. The downside is that less endpoints will be considered in evaluation and, therefore, a potential added benefit may be missed or that the decision relies on very few, or even just one endpoint. Since IQWiG does not decide upon the added benefit and the price of a new drug, the overall process, however, can adjust for this downside. To protect against too few endpoints forming the dossier, the pharmaceutical manufacturers can include the most important HTA endpoints in their phase III planning. Since the process would be published and everything is consequently known, this is

easily achievable, although it may mean a higher number of patients have to be included in the studies and, thus, increasing costs. If the overall decision is based on only few or even just one endpoint, the Federal Joint Committee (G-BA) may grant a restricted added benefit (as it can already and increasingly does so), and require the applicant to deliver more data for a second assessment [28].

Overall, it is possible with few adjustments to modify the IQWiG methodology, such that most of the points of critique can be addressed, and that the overall process of the early benefit assessment in Germany with different stakeholders involved, is not touched. There has been also a proposal to include in IQWiG’s methodology patients’ preferences for an explicit endpoint weighting [2, 29, 30], following an approach discussed for the regulatory benefit-risk assessment [31] as well as for HTA purposes [32]. In the past, patient preferences with regard to endpoints in the treatment of major depression and hepatitis C had been elicited on behalf of IQWiG [33, 34]. The main objection, put forward by IQWiG’s director, referred to the risk of inconsistencies while handling interventions in different indications [2]. Obviously, the proposed modification in the present paper is not the only one possible, but it stays as close as possible to the current method.

Limitations

It should be mentioned that IQWiG is not the decision maker in the German early benefit assessment, but the final decision (appraisal) is taken by the Federal Joint Committee (G-BA), the German self-administrative body of physicians, hospitals and health insurance funds. The Federal Joint Committee considers IQWiG’s assessments, but may rely on other evidence (e.g., results of currently ongoing studies) as well. We did focus in this work only on the methodology of the IQWiG and how this may be improved, not taking the overall process into account. Figure 1 shows the differences between IQWiG’s assessments and Federal Joint Committee’s appraisals on a subgroup level with regard to the extent of added benefit for all the early benefit assessments, orphan drugs excluded, up to 2016 following [35]. To explore why the assessment of the IQWiG and the final decision by the Federal Joint Committee differ, is a separate question also worth investigating.

Conclusion

The method used by IQWiG within early benefit assessment of new pharmaceuticals in Germany, is subject to manifold criticism. In general, the method does not ensure that results are statistically valid, and outcomes of the early benefit assessment may be compromised, due to power losses.

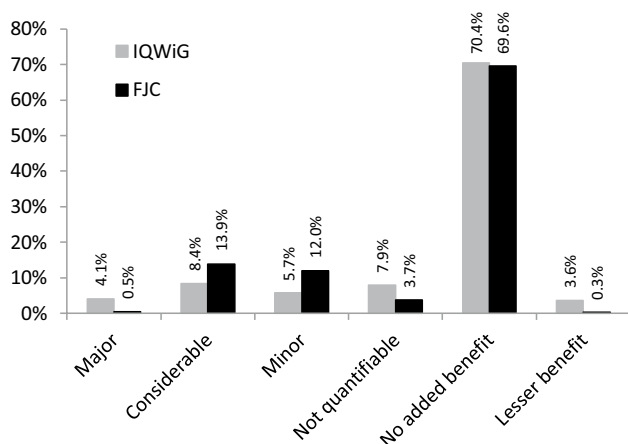


Fig. 1 Added benefit in IQWiG assessments and Federal Joint Committee's appraisals on a sub-population basis (orphan drugs excluded) up to end of 2016

By converging with the approach of approval authorities for confirmatory endpoints, the analysis method could be improved, to put the results on a more valid statistical basis.

Funding The authors did not receive any financial support for the preparation of the manuscript.

Compliance with ethical standards

Conflict of interest MH is working for ClinStat GmbH, Cologne, Germany, a service provider offering statistical support to different pharmaceutical companies. Next to his academic affiliation CMD is employed by Bayer Vital GmbH, Leverkusen, Germany.

References

1. Skipka, G., Wieseler, B., Kaiser, T., Thomas, S., Bender, R., Windeler, J., Lange, S.: Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biom J. Biometrische Zeitschrift*. **58**(1), 43–58 (2016). <https://doi.org/10.1002/bimj.201300274>
2. IQWiG.: Documentation and evaluation of comments on the Draft of the General Methods 4.1. Institute for Quality and Efficiency in Health Care (2013). https://www.iqwig.de/download/Dokumentation_und_Wuerdigung_der_Stellungnahmen_IQWiG_Methoden_4-1.pdf. Accessed 6 Nov 2017
3. AMNOG.: Act to Reorganize the Pharmaceuticals' Market in the SHI System (Arzneimittelneuordnungsgesetz-AMNOG) (2010). http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBI#_bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl110s2262.pdf%27%5D__1431036981006. Accessed 6 Nov 2017
4. IQWiG.: General Methods. Version 4.2 of 22 April 2015 (2015). https://www.iqwig.de/download/IQWiG_General_Methods_Version_4-2_no_longer_valid.pdf. Accessed 6 Nov 2017
5. Vach, W.: Quantifying the additional clinical benefit of new medicines: little-considerable-significant-6 remarks from a biometrician's point of view]. *Gesundheitswesen* (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany)). **76**(11), 757–762 (2014). <https://doi.org/10.1055/s-0033-1363682>
6. Röhmel, J.: Gutachten zum Vorschlag des IQWiG zur Bewertung des Ausmaßes des Zusatznutzens im Rahmen der Nutzenbewertung von Arzneimitteln nach § 35a SGB V, Bremen (2012)
7. Witte, J., Greiner, W.: Problembeobachtungen der Quantifizierung des Zusatznutzens im Rahmen der frühen Arzneimittelnutzenbewertung. *Gesundh ökon Qual manag*. **18**(05), 226–234 (2013). <https://doi.org/10.1055/s-0033-1350539>
8. Ruof, J., Dintsios, C.-M., Schwartz, F.W.: Questioning patient subgroups for benefit assessment: challenging the German Gemeinsamer Bundesausschuss approach. *Value Health*. **17**(4), 307–309 (2014). <https://doi.org/10.1016/j.jval.2014.05.001>
9. Rasch, A., Dintsios, C.-M.: Subgruppen in der frühen Nutzenbewertung von Arzneimitteln: eine methodische Bestandsaufnahme. *Zeitschrift für Evidenz, Fortbildung und Qualität. im Gesundheitswesen*. **109**(1), 69–78 (2015). <https://doi.org/10.1016/j.zefq.2015.01.001>
10. AM-NutzenV.: Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a Absatz 1 SGB V für Erstattungsvereinbarungen nach § 130b SGB V. https://www.bgbl.de/xaver/bgbl/start.xav#_bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl110s2324.pdf%27%5D__1523126802497 (2010). Accessed 09 Apr 2018
11. Djulbegovic, B., Kumar, A., Soares, H.P., Hozo, I., Bepler, G., Clarke, M., Bennett, C.L.: Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. *Arch. Internal Med*. **168**(6), 632–642 (2008). <https://doi.org/10.1001/archinte.168.6.632>
12. IQWiG.: IQWiG reports. <https://www.iqwig.de/en/projects-results/publications/iqwig-reports.1071.html>. Accessed 09 Apr 2018
13. Bender, H., Dintsios, C.M.: [Health-related quality of life in the context of early benefit assessment of drugs according to section sign 35a of the German social code book V: a challenging endpoint for all the involved stakeholders]. *Gesundheitswesen* (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany)). **80**(2), 132–143 (2018). <https://doi.org/10.1055/s-0042-116433>
14. Niehaus, I., Dintsios, C.M.: Confirmatory versus explorative endpoint analysis: decision-making on the basis of evidence available from market authorization and early benefit assessment for oncology drugs. *Health Policy (Amst. Neth.)* (2018). <https://doi.org/10.1016/j.healthpol.2018.03.017>
15. FDA.: Guidance for industry: providing clinical evidence of effectiveness for human drugs and biological products. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER) (1998)
16. EMA.: 11.14. CPMP/EWP/2330/99: points to consider on validity and interpretation of meta-analyses, and one pivotal study (2001). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf. Accessed 6 Nov 2017
17. Lange, S.: Challenges for clinical trials in oncology within the scope of early benefit assessment of drugs. *Z Evid Fortbild Qual Gesundhwes*. **109**(6), 417–430 (2015). <https://doi.org/10.1016/j.zefq.2015.08.003>
18. Marcus, R., Peritz, E., Gabriel, K.R.: On Closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. **63**(3), 655–660 (1976). <https://doi.org/10.2307/2335748>
19. Sonnemann, E.: Kombination unabhängiger tests. In: Vollmer, J. (ed.) *Biometrie in der chemisch-pharmazeutischen Industrie*, 4, Stand und Perspektiven. Gustav-Fischer, Stuttgart (1991)
20. Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**(2), 65–70 (1979)

21. Budde, M., Bauer, P.: Multiple test procedures in clinical dose finding studies. *J. Am. Stat. Assoc.* **84**(407), 792–796 (1989). <https://doi.org/10.1080/01621459.1989.10478839>
22. Horn, M., Volland, R.: *Multiple Tests und Auswahlverfahren*. Spektrum Akademischer Verlag, Heidelberg (1999)
23. Wahl, M.: *Proposal of a Two Stage Design for Dose Finding Studies*. University of Dortmund, Dortmund (2001)
24. Erklärung der UPDRS auf parkinson/web. http://www.parkinson-web.de/content/was_ist_parkinson/beurteilungskalen/updrs_skala/index_ger.html. Accessed 09 April 2018
25. EMA.: *EPAR Rotigotine (2006)*. http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Scientific_Discussion/human/000626/WC500026394.pdf. Accessed 6 Nov 2017
26. CADTH.: *Common drug review—clinical review report—Rotigotine (2016)*. https://www.cadth.ca/sites/default/files/cdr/clinical/SR0344_Neupro_CL_Report.pdf. Accessed 6 Nov 2017
27. ICH E9 Expert Working Group.: *Statistical principles for clinical trials: ICH harmonized tripartite guideline*. *Stat Med* **18**, 1905–1942 (1999)
28. Ruof, J., Staab, T., Dintsios, C.M., Schroter, J., Schwartz, F.W.: Comparison of post-authorisation measures from regulatory authorities with additional evidence requirements from the HTA body in Germany—are additional data requirements by the Federal Joint Committee justified? *Health Econ Rev.* **6**(1), 46 (2016). <https://doi.org/10.1186/s13561-016-0124-4>
29. Dintsios, C.M.: Nutzen-Schaden-Abwägung zur Herleitung eines Gesamtnutzens im Rahmen der frühen Nutzenbewertung: Kritische Würdigung und alternative Lösungsansätze. 4–14 (2013). <https://doi.org/10.3205/13dkvf040>. <http://www.egms.de/static/en/meetings/dkvvf2013/13dkvf040.shtml>. Accessed 6 Nov 2017
30. Dintsios, C.M.: Patientenpräferenzbasierte Nutzen-Schaden-Abwägung für die frühe Nutzenbewertung von Arzneimitteln: ein Erweiterungsvorschlag. 17 (2017). <https://doi.org/10.3205/17ebm066>. <http://www.egms.de/en/meetings/ebm2017/17ebm066.shtml>. Accessed 6 Nov 2017
31. Egbrink, M.O., IJzerman M.: The value of quantitative patient preferences in regulatory benefit-risk assessment. *J Market Access Health Policy* **2** (2014). <https://doi.org/10.3402/jmahp.v2.22761>
32. Muhlbacher, A.C., Johnson, F.R.: Giving patients a meaningful voice in european health technology assessments: the role of health preference research. *Patient.* **10**(4), 527–530 (2017). <https://doi.org/10.1007/s40271-017-0249-5>
33. Hummel, M.J., Volz, F., van Manen, J.G., Danner, M., Dintsios, C.M., IJzerman, M.J., Gerber, A.: Using the analytic hierarchy process to elicit patient preferences: prioritizing multiple outcome measures of antidepressant drug treatment. *Patient.* **5**(4), 225–237 (2012). <https://doi.org/10.2165/11635240-000000000-00000>
34. Muhlbacher, A.C., Bridges, J.F., Bethge, S., Dintsios, C.M., Schwalm, A., Gerber-Grote, A., Nubling, M.: Preferences for antiviral therapy of chronic hepatitis C: a discrete choice experiment. *Eur. J. Health. Econ. HEPAC Health Econ. Prev. Care.* **18**(2), 155–165 (2017). <https://doi.org/10.1007/s10198-016-0763-8>
35. Ruof, J., Schwartz, F.W., Schulenburg, J.M., Dintsios, C.M.: Early benefit assessment (EBA) in Germany: analysing decisions 18 months after introducing the new AMNOG legislation. *Eur. J. Health Econ. HEPAC Health Econ. Prev. Care.* **15**(6), 577–589 (2014). <https://doi.org/10.1007/s10198-013-0495-y>