# Implementation of AMNOG: An industry perspective

**Friedhelm Leverkus***,[1] and **Christy Chuang-Stein**[2]

[1] Pfizer Deutschland GmbH, Linkstr. 10, 10785, Berlin, Germany
[2] Chuang-Stein Consulting, 5857 Stoney Brook Road, Kalamazoo, MI 49009, USA

Received 31 October 2013; revised 13 June 2015; accepted 18 June 2015

In 2010, the Federal Parliament (Bundestag) of Germany passed a new law (Arzneimittelmarktneuordnungsgesetz, AMNOG) on the regulation of medicinal products that applies to all pharmaceutical products with active ingredients that are launched beginning January 1, 2011. The law describes the process to determine the price at which an approved new product will be reimbursed by the statutory health insurance system. The process consists of two phases. The first phase assesses the additional benefit of the new product versus an appropriate comparator (zweckmäßige Vergleichstherapie, zVT). The second phase involves price negotiation. Focusing on the first phase, this paper investigates requirements of benefit assessment of a new product under this law with special attention on the methods applied by the German authorities on issues such as the choice of the comparator, patient relevant endpoints, subgroup analyses, extent of benefit, determination of net benefit, primary and secondary endpoints, and uncertainty of the additional benefit. We propose alternative approaches to address the requirements in some cases and invite other researchers to help develop solutions in other cases.

*Keywords:* Additional benefit; AMNOG; Comparator; Early benefit assessment; Endpoint; Net benefit; Subgroup.

## 1 Introduction

The number of healthcare systems that conduct health technology assessment (HTA) has steadily increased over the last decade. Cost-effectiveness evaluation is a key component of the HTA in some countries such as UK, Canada, and Australia. In these countries, the incremental quality-adjusted life years (QALY) attributable to the use of a new health technology is evaluated against the incremental cost (e.g. National Institute for Health and Care Excellence, 2013). The resulting incremental cost-effectiveness ratio (ICER) is then compared to a threshold to decide if a new health technology is cost-effective. The threshold changes over time based on multiple considerations including inflation.

The QALY-based approach was judged to be flawed by some researchers because it assumes a multiplicative model for the calculation of the QALY values (Duru et al., 2002; Beresniak et al., 2012). Instead of using ICER, the German and French decision makers for pharmaceutical product reimbursement place their primary focus on determining the clinical benefit from clinical studies. After a clinical benefit has been ascertained, the price is then determined through a complex negotiation process between the manufacturer and the decision makers.

The Federal Parliament (Bundestag) of Germany passed a new law in 2010, called AMNOG (Arzneimittelmarktneuordnungsgesetz), on reimbursement decisions of health technology. The still relatively new law applies to all pharmaceutical products with a new active ingredient that have been launched beginning January 1, 2011. The law was enforced with the political goal of saving the sick

*Corresponding author: e-mail: Friedhelm.leverkus@pfizer.com

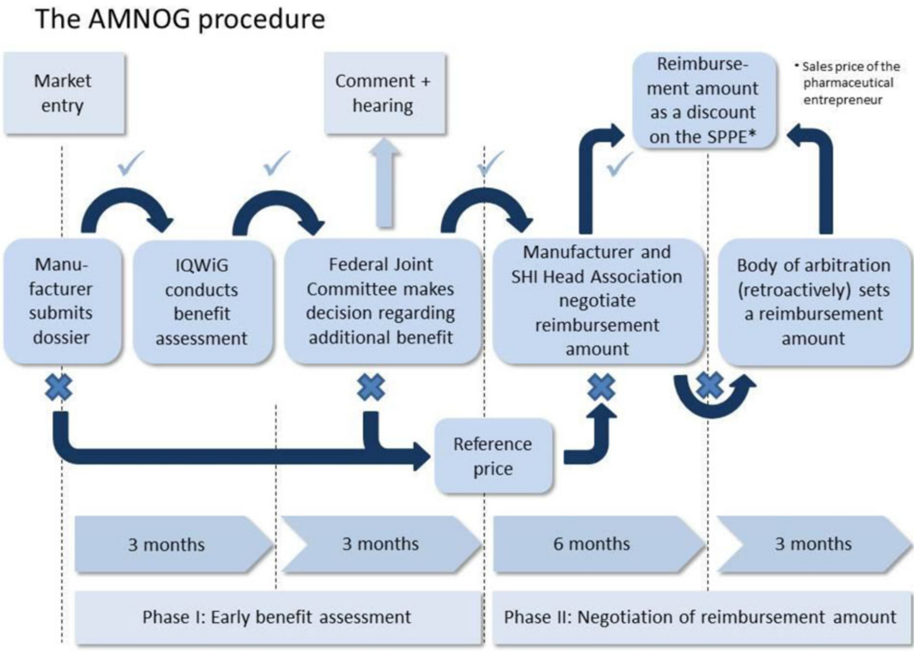
The AMNOG procedure
Market entry
Comment + hearing
Reimburse-ment amount as a discount on the SPPE*
* Sales price of the pharmaceutical entrepreneur
Manu-facturer submits dossier
IQWiG conducts benefit assessment
Federal Joint Committee makes decision regarding additional benefit
Manufacturer and SHI Head Association negotiate reimbursement amount
Body of arbitration (retroactively) sets a reimbursement amount
Reference price
3 months
3 months
6 months
3 months
Phase I: Early benefit assessment
Phase II: Negotiation of reimbursement amount

**Figure 1** The AMNOG process, effective as of 1 January 2011 (vfa, 2012).

funds 2.2 billion € per year (Federal Ministry of Health, 2013). The process to determine reimbursement decisions under the new law consists of two phases and is described in Fig. 1. At the product launch, a manufacturer needs to submit a Benefit Dossier intended to show that the new product has additional benefit against a comparator (zweckmäßige Vergleichstherapie, zVT) determined by the Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA). The G-BA is the highest self-governing decision-making body of physicians, dentists, hospitals, and health insurance funds in Germany. It issues directives for the benefit catalog of the statutory health insurance funds and specifies measures for quality assurance in inpatient and outpatient areas of the healthcare system. Additional benefit needs to be demonstrated in patient relevant endpoints such as mortality and morbidity. The G-BA retains the Institute for Quality and Efficiency in Healthcare (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, or IQWiG) to conduct benefit assessment.

IQWiG is a professionally independent and scientific institute that investigates the benefits and harms of medical interventions and is funded by contributions from insurers. It should be noted that the G-BA and IQWiG are responsible for different regulations. G-BA, through the Rules of Procedure it issued, regulates the methodological requirements for the scientific assessment of the benefit. The Rules of Procedure also contain regulations on the independence and on possible conflicts of interest, as well as on hearing procedures for directives (http://www.english.g-ba.de/legalmandate/rules/). The basis for IQWiG's assessment is the General Methods paper (IQWiG, 2013) issued by the Institute. The General Methods paper describes the procedures IQWiG follows in its benefit assessment. The steps the Institute chooses to assess a specific medical intervention depend primarily on the research question posed and the available scientific evidence.

After receiving IQWiG's recommendation, the G-BA has up to 3 months to decide on the benefit of the new product relative to the chosen comparator. The G-BA will classify the benefit into one of six categories: major additional benefit, considerable additional benefit, minor additional benefit, additional benefit not quantifiable, no additional benefit, and benefit smaller than that of the comparator. While the G-BA takes IQWiG's recommendation into consideration, the G-BA is not obligated to follow IQWiG's recommendation exactly. After the G-BA assessment a manufacturer has the option to withdraw the product without having a price ever discussed in Germany.

Price negotiations take place during the second phase. If the new product is deemed to have no additional benefit, it will go into a fixed reference price group. If a fixed reference price group does not exist, the maximum price at which the new product can be reimbursed is the price of the chosen comparator. On the other hand, if the G-BA judges the new product to have an additional benefit, the manufacturer will be invited to engage in price negotiations with the National Association of the Statutory Health Insurance (Spitzenverband der gesetzlichen Krankenkassen, GKV-SV). Product price, as a discount on the sales price set by the manufacturer, has to be negotiated within 6 months. The negotiated price becomes effective 12 months after the Benefit Dossier submission (or product launch). During these 12 months, the manufacturer can set the price for the product. If the negotiations do not yield an agreement, either side can appeal to an Arbitration Board (according to § 130b SGB V). The Board will determine the product price and the price will apply retroactively. A manufacturer can decide to withdraw their product from the German market after the Board's decision, but the new price would be available for cross-referencing and could impact price negotiations in other countries. The two-phase process shows the importance of demonstrating the additional benefit of a new product.

There are exceptions to the two-phase process. For example, an assessment of the additional benefit for an orphan drug will take place only if the annual sales are expected to exceed EUR 50 million. In this case, the G-BA will conduct the assessment itself based on the Benefit Dossier submitted by the manufacturer. This is illustrated by the path from "Manufacturer submits dossier" to "Reference price" and to "G-BA makes decision regarding additional benefit" in Fig. 1. According to Section 12 in Chapter 5 of the G-BA Rules of Procedure (G-BA, 2011a), certain parts of the Benefit Dossier are not required in the case of an orphan drug.

The primary objective of this paper is to offer statisticians working in the pharmaceutical industry who are planning Phase III trials an overview of the requirements of the G-BA and IQWiG in conducting benefit assessments. In Section 2, we discuss, from an industry perspective, requirements, and issues associated with IQWiG's and G-BA's assessment of additional benefit. We focus on seven areas: the choice of the comparator, patient relevant endpoints, subgroup analyses, extent of benefit, determination of net benefit, primary and secondary endpoints, and certainty of the additional benefit. We contrast the approaches taken by the G-BA and IQWiG in their benefit assessment with those of the European Medicines Agency (EMA), which makes marketing authorization decisions. For principles underlying regulatory decisions, we reference primarily the International Conference on Harmonization (ICH) E9 (Statistical Principles for Clinical Trials, 1998) document. Since 1998, ICH E9 has been the primary reference for statistical principles governing clinical trials in support of marketing applications in Europe, Japan, and the United States. For approaches behind G-BA/IQWiG's decisions, we reference IQWiG's General Methods paper, German Social Code Book, and G-BA's Rules of Procedure.

Even though the most current German version of the G-BA Rules of Procedure is G-BA (2014), we reference the English translation of a previous version in this paper so that more readers are able to understand the contents of the document. In general, we reference the English translation of a document if a translation exists. It should be understood that only the original German document is legally binding.

We recognize that the legal requirements and the tasks for G-BA/IQWiG and EMA differ, leading to potential differences in approaches and methods adopted by different decision bodies. In this paper, we identify these differences and propose ways to address these differences. By submitting our proposals, we are also inviting other statisticians who design and analyze clinical trials to join us in developing solutions to minimize the impact of the identified differences.

## 2 Assessment of additional benefit

When assessing a new product, a Benefit Dossier has to answer four questions broadly (G-BA, 2011a). They are:

(i) Is an additional benefit over the comparator determined by G-BA proven?
(ii) Are there special patient groups with an additional benefit?
(iii) How large is the benefit?
(iv) How certain are we about the conclusions?

In this section, we will discuss seven issues on how benefit assessments are conducted, based on our understanding of how the new law is implemented. We discuss these issues from an industry perspective.

### 2.1 Choice of the comparator

From a manufacturer's perspective, the way G-BA chooses the comparator for the benefit assessment is one of the most contentious points. The comparator is determined by the criteria in Table 1 extracted from Section 6 of the English version of G-BA's Rules of Procedure (G-BA, 2011a).

Recently, there has been a change to the rules in that Criterion 5 is used only for price negotiations, and not for determining the comparator. Under the revised rules, a manufacturer could choose the comparator if several comparators are allowed.

In general, the G-BA's decision on the comparator is driven by label and medical guidelines. The decision could also depend on a patient's disease state prior to initiating a new treatment. For example, when deciding on the comparator for patients with locally advanced or metastatic nonsmall cell lung cancer, the G-BA may choose a different comparator for cancer patients with different ECOG (Eastern Cooperative Oncology Group) performance status (PS). In a recent assessment, the G-BA determined that the comparator should be chemotherapy (e.g. docetaxel or pemetrexed) for patients with ECOG PS 0–1 and PS 2. Since patients with more advanced ECOG PS 2, or PS 3 or 4, cannot receive chemotherapy due to poor general health, the G-BA considered the comparator for these patients to be the best supportive care. So, if a manufacturer conducted studies that compared a new treatment with chemotherapy in patients with ECOG PS 0–2, the G-BA will consider data related to patients with ECOG PS 3–4 as missing and that there is no proven additional benefit from the new treatment in the PS 3–4 population. This will be taken into consideration during price negotiations. This is different from obtaining marketing authorization. Regulators do not necessarily restrict approval to the subgroup of PS 0–2 even though the registration studies were conducted only in these patients as long as there

**Table 1** G-BA criteria for determining the appropriate comparators (zVT) (G-BA, 2011a).

| | |
|---|---|
| 1 | Insofar as a medical therapeutic indication is considered as the comparator, the pharmaceutical must be authorized for the therapeutic indication. |
| 2 | Insofar as a nonmedicinal treatment is considered as the comparator, this must be deliverable within the framework of the statutory health insurance. |
| 3 | Medical therapeutic indications or nonpharmaceutical treatments are preferred as comparator, whose patient-relevant benefit has already been determined by the Federal Joint Committee. |
| 4 | The comparator should belong to the appropriate therapy in the therapeutic indication according to the generally accepted state of medical knowledge. |
| 5 | If there are several alternatives, the more economic therapy is selected, preferably a therapy, for which there is a reference price. (This criterion applies to price negotiations only.) |

is no biological reason to believe that the new product may not benefit more advanced patients. An example for this difference is the decisions about Crizotinib. (See European Public Assessment Report for Xalkori (2012) and G-BA Assessment (2013))

It is not unusual for the G-BA to choose a comparator that differs from that included in the Phase III trials used to support regulatory approval even if the comparator was the result of consultation with the EMA. When this happens, the Benefit Dossier will not include a head-to-head direct comparison between the new product and the comparator chosen by the G-BA since the manufacturer's Benefit Dossier is typically based on pivotal registration studies. This creates a significant challenge for the manufacturer. While a manufacturer can estimate the extent of additional benefit of a new substance over the comparator using indirect comparisons (IC), ICs bear a high potential for bias. This is reflected in the assessment of the probability of an additional benefit. This and the role of ICs in the benefit assessment by IQWiG are further discussed in Section 2.7.

In order to meet the requirements of the G-BA and the EMA on the selection of a suitable comparator, we encourage the adoption of a parallel consultation process where input can be sought from all relevant parties concurrently. When opinions differ, we hope efforts such as seeking medical expert opinions could be made to resolve the differences among the various decision-making bodies.

### 2.2 Patient-relevant endpoints

Regulation for Pharmaceutical Benefit Assessments ("AM-NutzenV") in Germany (http://www.gesetze-im-internet.de/am-nutzenv/__5.html) requires that pharmaceutical manufacturers demonstrate the added value of their products through a benefit assessment based on patient-relevant endpoints.

IQWiG describes patient relevant endpoints in their General Methods paper (IQWiG, 2013) as follows: "… "patient-relevant" refers to how a patient feels, functions, or survives. Consideration is given here to both the intentional and unintentional effects of the intervention that in particular allow an assessment of the impact on the following patient-relevant outcomes to determine the changes related to disease and treatment: (1) mortality, (2) morbidity (symptoms and complications), (3) health-related quality of life."

Safety endpoints such as adverse events may also be relevant. Safety laboratory parameters are generally not accepted as patient-relevant endpoints in a benefit assessment except when a change in a safety parameter represents a serious adverse event that has serious consequences for patients.

If a surrogate endpoint is used, its surrogacy for the clinical endpoint needs to be validated. IQWiG describes methods and procedures to validate a surrogate endpoint (IQWiG Reports, 2011a). The methods, based on the work of Buyse and Piedbois (2008, 2010), Lassere (2008), and Buyse et al. (2015), examine the correlation between the alleged surrogate endpoint and the clinical endpoint, both within studies and across studies. The overall correlation is derived from a meta-analysis of randomized clinical trials and needs to be high for an endpoint to qualify as a surrogate. For high correlation, IQWiG requires the lower bound of the confidence interval for the correlation coefficient to be greater than 0.85. Based on this criterion, IQWiG (2011a) concluded that there is not an adequate investigation of the validity of progression-free survival (PFS) as a surrogate endpoint for overall survival (OS) in breast cancer, colon cancer, and renal cell carcinoma. Furthermore, PFS as defined by the RECIST criteria (Eisenhauer et al., 2009) is based on radiographic assessment and is, therefore, not seen as a patient-relevant endpoint by IQWiG.

By comparison, regulators have accepted PFS as a primary endpoint on many occasions (European Medicines Agency, 2013). On the occasions when PFS is accepted as the primary endpoint, marketing authorization is typically conditioned on the new treatment not decreasing the OS when compared with the control in the study. In other words, a new treatment with a benefit on PFS and a noninferior benefit on OS (relative to a control) could be approved by regulators. When presented with the same data, IQWiG may conclude that additional benefit could not be proven.

Composite endpoints that have been traditionally used as the primary endpoints in many large registration trials can now be questioned by IQWiG. For example, when assessing the benefit of a new anti-coagulant (IQWiG, 2012a), IQWiG did not accept asymptomatic embolism as a component of a composite endpoint even though the study specified the composite endpoint of stroke and systemic embolism (both symptomatic and asymptomatic) as the primary endpoint.

Redefining a composite endpoint after the study was completed is problematic from a manufacturer's perspective. For one thing, removing a component in a composite endpoint could reduce the power of the study. Many composite endpoints are analyzed by time-to-event methods that compare time to the first occurrence of the composite endpoint between two treatment groups. Under the proportional hazard assumption, the semi-parametric log hazard ratio estimate has an asymptotic variance that is a decreasing function of the number of events in the two treatment groups. Therefore, reducing the number of events in the groups by removing a component from the composite endpoint leads to a larger variance for the log hazard ratio estimate and a wider confidence interval for the log hazard ratio. This could translate to a lower power to detect a hazard ratio that is less than one at a fixed significance level.

Because of the above, we do not support changing the definition of a primary endpoint after a study has been completed. If the G-BA insists on removing a component from a composite endpoint, we suggest the use of a lower confidence level when constructing a confidence interval. A lower confidence level could mitigate the increase in variance due to a reduction in the number of events in the two groups. The reduction in the confidence level should reflect the total number of events removed. The advantage of this approach is that, even with a lower confidence level, the point estimate for the additional benefit still needs to be of a certain magnitude for the confidence interval to meet the efficacy requirements.

### 2.3 Subgroup analysis

ICH E9 states that in most cases subgroup analyses are exploratory and should be clearly identified as such. Subgroup analyses are conducted to explore, rather than confirm, the uniformity of an observed overall treatment effect. When subgroup analysis is conducted to verify an anticipated difference in treatment effect, it should be part of the planned confirmatory analysis.

AMNOG requires subgroup analyses for age, gender and other possible effect modifiers. The latter includes stratification variable(s) in a stratified randomization. To assess the impact of an effect modifier on the treatment effect, IQWiG (2013, Section 7.3.8.(C)) requires an interaction test. If the $p$-value from the interaction test is $> 0.20$, IQWiG will conclude "no evidence for interaction." If the $p$-value is between 0.05 and 0.20, IQWiG will conclude a "hint of an interaction." In this case, results from the pooled data and results observed in the subgroups need to be reported. If the $p$-value is $< 0.05$, this will be interpreted as a "proof of interaction." When this happens, data may not be pooled and results need to be reported for each subgroup individually.

In addition to effect modifiers, the G-BA can require subgroup analyses that result from dividing (slicing) the study population into subpopulations. The G-BA may request comparing a new intervention to different comparators in the subpopulations.

Demonstrating additional benefit in subgroups in the fashion required by the G-BA (see Section 2.4) without adequate upfront sample size planning can be problematic for a manufacturer. This challenge can then be further exacerbated if the G-BA requires different comparators to demonstrate additional benefit in different subpopulations.

Bender and Lange (2001) cautioned that a difference in the observed treatment effects between subgroups could be completely due to sampling variability and pointed out the danger of selecting subgroups based on the observed treatment effect to draw statistical inference. Paget et al. (2011) wrote about subgroup analysis of clinical effectiveness in support of HTA. Paget et al. recommended a set of considerations for conducting subgroup analyses of clinical effectiveness. These considerations include: definition of subgroups, identification of differences in treatment effects in subgroups, multiplicity associated with subgroup analyses, sensitivity analysis to support subgroup findings, replication of subgroup results, source of subgroup evidence, presentation, and reporting of subgroup analyses. Many of these are acknowledged in IQWiG's General Methods paper (IQWiG, 2013).

We support the position of ICH E9 and the considerations advocated by Paget et al. (2011). While we understand the desire to conduct subgroup analysis in support of the search for personalized medicine, we are concerned that subgroup analyses conducted in the manner by IQWiG and G-BA could lead to unintended consequences. In January 2014, the EMA released a draft guideline on the investigation of subgroups in confirmatory clinical trials for public consultation (EMA, 2014). We encourage collaborations between regulators, HTA decision makers, and academic researchers on this subject so that health care decisions based on subgroup results could be made in a science-based and consistent manner.

### 2.4 Extent of benefit

The G-BA classifies benefit into one of six categories: major additional benefit, considerable additional benefit, minor additional benefit, additional benefit not quantifiable, no additional benefit, and benefit smaller than comparator (G-BA, 2011a, 2011b). Table 2 describes the basis for the classification.

AMNOG itself is vague on the quantitative definition of an additional benefit. Since IQWiG is obligated to give the G-BA a recommendation on the extent of additional benefit, IQWiG developed criteria to quantify the benefit (IQWiG, 2013). IQWiG's rationale is described briefly as below.

IQWiG has categorized patient-relevant endpoints into survival, severe symptoms and side effects, quality of life, and nonsevere symptoms and side effects. In Djulbegovic et al. (2008), a relative risk (RR) of 0.5 for mortality was observed and the new treatment was considered a breakthrough. IQWiG translated this to mean "considerable or major benefit" concerning survival (IQWiG, 2013). Assuming a survival endpoint with a true RR of 0.5 and two well-powered studies (trials of the size like PLATO (IQWiG, 2011b) or TRITON (Antman et al., 2008)), IQWiG calculated a 95% CI for the RR in a meta-analysis combing data from two studies. They found the upper limit of the CI close to 0.85. Based on this insight, IQWiG (2013) decided that for a survival endpoint, the upper limit of the 95% CI for RR must be below 0.85 to qualify for the "major" benefit category. To qualify for the "considerable" benefit category, the upper limit must be less than 0.95. If the upper confidence limit is greater than 0.95 but less than 1.0, the added benefit is classified as "minor."

**Table 2** Basis for classifying additional benefit (modified from G-BA, 2011a).

| Extent of the additional benefit (categorization) | Basis for the classification |
|---|---|
| Major | A sustained improvement of the therapy-relevant benefit that was previously unattained compared to the appropriate comparative therapy |
| Considerable | A significant improvement of the therapy-relevant benefit that was previously unattained compared to the appropriate comparative therapy |
| Minor | A moderate and not just small improvement of the therapy-relevant benefit that was previously unattained compared to the appropriate comparative therapy |
| Not quantifiable | Because the scientific data basis does not allow it |
| None | No additional benefit has been demonstrated |
| Smaller benefit | The benefit of the medicinal product to be assessed is smaller than the benefit of the appropriate comparative therapy |

**Table 3** IQWiG criteria for added benefit categories under different types of endpoints (IQWiG, 2013, Appendix A).

| | | Target figure category | | |
|---|---|---|---|---|
| | | Mortality | Severe symptoms (or consequential complications) and side effects and health-related quality of life [a)] | Nonsevere symptoms (or consequential complications) and side effects |
| Added benefit | Major | 0.85 | 0.75 Risk ≥ 5% [b)] | Not applicable |
| | Considerable | 0.95 | 0.90 | 0.80 |
| | Minor | 1.00 | 1.00 | 0.90 |

a) The endpoint should be a patient-relevant endpoint or a validated/established instrument. The comparative measure is the relative risk for an undesirable outcome between the new treatment and the comparator.

b) Percent of undesirable outcome is ≥ 5% in at least one of the two groups compared.

To qualify for a benefit category for an endpoint, the upper limit of a 95% confidence interval for the comparative measure needs to be lower than the value in the table.

Added benefit on severe symptoms and quality of life is perceived to have less value compared to that of survival. To qualify for the "major" category, the upper limit of the 95% CI must be below 0.75, which corresponds to a true RR of 0.17. In other words, a greater treatment effect is required of benefit on severe symptoms and quality of life to qualify for the "major" category. If benefit is measured by treatment effect on nonsevere symptoms, the benefit could never qualify for the "major" category. The upper limit of the CI for such an endpoint needs to be 0.80 and 0.90, respectively, to qualify for "considerable" and "minor" category. The requirements on the upper limit to qualify for different categories of added benefit under different types of endpoints are given in Table 3.

An advantage of Table 3 is its transparency in that a manufacturer knows what to expect. The approach has been criticized by some researchers (e.g. Röhmel, 2013). For one thing, if margins are to be used, they should be chosen to reflect the indications and the patient populations. It does not seem sensible to have one set of margins applying to all situations. Second, the margins are based on a relative risk measure. There are many situations where the impact of the benefit attributable to a new treatment is better measured on an absolute scale than a relative one. For example, a reduction in the mortality risk from 60 to 40% has a greater healthcare implication than a mortality reduction from 30 to 20%, but both reductions have the same RR. Third, the margins are based on the expectation of two large studies. Many development programs that investigate the effect of a new treatment on mortality and/or major morbidity (e.g. anti-coagulants or anti-platelets, oncology products) conduct only one large phase III trial. The expectation of two large outcome trials is especially unrealistic for a new indication on mortality or severe morbidity for an approved drug.

Skipka et al. (2015) proposed a methodological approach to determine major, considerable, and minor treatment effects in the early benefit assessment of new drugs. Additional discussions on the use of formal criteria to check clinical relevance in randomized clinical trials as part of a benefit assessment strategy is given by Vach and Gladstone (2015).

We believe this is one area where we especially need our academic colleagues to join in and conduct research to provide more insight into the choice of the comparative measure, thresholds to define benefit categories, one study against two studies, and how decisions on these issues will impact the design and analysis of registration trials. We understand that the magnitude of an observed treatment effect is also important to regulators when judging whether an observed benefit is clinically meaningful or not. However, the latter is typically done in the context of the disease population and the overall evidence instead of applying a prespecified set of thresholds to all situations.

### 2.5 Determination of net benefit

When a new treatment is associated with more adverse reactions than the comparator, the benefit category could be downgraded by 1 or 2 levels, based on the severity and seriousness of the adverse reactions. Lack of data could also result in downgrading. For example, the benefit of cabazitaxel on mortality in patients with advanced prostate cancer was initially judged to be considerable. However, severe side effects resulted in a net-benefit designation of "minor" for the product (G-BA, 2012a; IQWiG, 2012b). Ipilimumab's benefit on mortality was originally considered to be major. Because of considerable damage potential for serious and severe immune-mediated adverse events, the added benefit was downgraded to "considerable" by IQWiG (G-BA, 2012b; IQWiG, 2012c).

In both cases, the decisions did not seem to weigh adverse reaction data in a quantitative analysis. In recent years, the EMA has sponsored a benefit-risk methodology project that published several work package reports available at http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/document_listing/document_listing_000314.jsp&mid= WC0b01ac0580665b63. During the same period, the PROTECT project sponsored by Europe's Innovative Medicines Initiative and the European Federation of Pharmaceutical Industries and Associations has completed several case studies to illustrate benefit-risk assessment methodologies. These case studies are available at http://www.imi-protect.eu/benefitsRep.shtml. We feel that work by the above groups represent a good attempt at creating a structured framework to conduct benefit-risk assessment. In this special issue, Nixon et al. (2015) and Waddingham et al. (2015) discuss options to conduct structured benefit-risk assessment. We hope that IQWiG would consider taking advantage of the output from the above efforts in determining the net benefit.

### 2.6 Primary and secondary endpoints

The analysis and interpretation of clinical trial results depends on the trial design, which includes a deliberation of primary and secondary endpoints. When there are multiple primary endpoints, a multiple comparison procedure needs to be prespecified to control the Type-I error rate for the primary inference. If a treatment effect is found to be statistically significant at the prespecified significance level, the treatment is said to be efficacious. Analysis of secondary endpoints offers support to the primary endpoint(s) and provides helpful information to healthcare providers and patients (Stone and Chuang-Stein, 2013). The distinction between primary and secondary endpoints is important not only for interpreting trial results, but in the trial's ability to draw inference with adequate power.

According to Section 2.2.2 of ICH E9, the primary endpoint should generally be the one to determine the sample size of a study. In practice, primary endpoints for registration trials are selected in agreement with regulators and supported by medical experts. The same subsection in ICH E9 also states that the primary endpoint should be specified in the protocol. It further states that redefinition of the primary endpoint after trial results are known will almost always be unacceptable, since the biases this introduces are difficult to assess. As for secondary endpoints, ICH E9 states that their predefinition in the protocol is also important. There needs to be an explanation of the roles of secondary endpoints in interpreting trial results.

Because AMNOG focuses on patient-relevant endpoints, IQWiG does not necessarily accept prespecified primary endpoints or distinguish between primary and secondary endpoints. The case of removing asymptomatic embolism from the primary composite endpoint in the assessment of an anticoagulant mentioned in Section 2.2 is an example of disregarding a prespecified primary endpoint. Choosing a nonprimary endpoint as the basis for decision may not be an issue when decisions are made based on a meta-analysis of multiple studies, because the strength of the evidence comes from multiple studies. Unfortunately, the number of studies with IQWiG-required endpoints at the time of IQWiG's benefit assessment is typically limited. The situation is especially problematic when there is only one large registration trial.

We prefer that the original choices of endpoints in a trial be used in IQWiG's benefit assessment. We encourage pharmaceutical companies to seek early advice, which may lead to modifications to the development program. On the rare occasions when a component of a composite endpoint is removed, we propose in Section 2.2 a possible way to handle the situation. The same approach could be considered when another endpoint is used to make the benefit decision. The basic idea is to adjust the confidence level, if necessary, to compensate for the lack of precision in the estimated treatment effect because the study is not sized to have enough precision for estimating treatment effect on the endpoint(s) chosen by IQWiG.

### 2.7 Certainty of conclusion on additional benefit

The quality of the studies and data included in the Benefit Dossier needs to be described. IQWiG prefers data from head-to-head comparisons in randomized clinical trials. When a head-to-head comparison is not available, IQWiG will accept indirect comparison but states that this will lead to in more uncertainty about the results.

IQWiG's assessment of certainty of conclusions is aligned with the international standards of evidence-based medicine such as the GRADE guidelines (Atkins et al., 2004). When there is only one study, this will not automatically lead to a downgrading of the certainty level if the study has high quality, a large sample size, and a robust estimate for the treatment effect.

IQWiG gives the following criteria for Proof, Indication and Hint for conclusions on additional benefit (IQWiG, 2013, Table 2):

(i) Proof (highest certainty of conclusions): meta-analysis demonstrating a homogeneous and statistically significant effect with high qualitative certainty of the results; at least two independent studies demonstrating clear treatment effects in the same direction with high qualitative certainty of the results.

(ii) Indication (medium certainty of conclusions): meta-analysis demonstrating a homogenous and statistically significant effect or multiple studies demonstrating clear treatment effects in the same direction, both with moderate qualitative certainty of the results; individual studies demonstrating moderate treatment effects in the same direction with high qualitative certainty of the results; a single study demonstrating a statistically significant effect with high qualitative certainty of the results.

(iii) Hint (weakest certainty of conclusions): meta-analysis demonstrating a homogeneous and statistically significant effect or multiple studies demonstrating clear treatment effects in the same direction, both with low qualitative certainty of results; individual studies demonstrating moderate treatment effects in the same direction with moderate qualitative certainty of the results; a single study demonstrating a statistically significant effect with moderate qualitative certainty of the results.

Under the above criteria, there is almost no chance for an oncology product to receive a "Proof" level designation. When the G-BA-chosen comparator is not the control used in the registration trials, indirect comparisons (Bucher et al., 1997) and mixed treatment comparisons (Lu and Ades, 2004) that preserve randomizations within individual trials could be used to conduct benefit assessment. Unfortunately, indirect comparisons that preserve randomizations are not always possible due to differences in study design, differences in inclusion/exclusion criteria across studies, or simply the lack of a common treatment across the set of trials.

In our opinion, the 3-level categorization needs revision for at least two reasons. First, regulatory approvals for conditions with unmet medical needs may be based solely on single arm studies. This is especially the case for conditional approvals. Second, an indirect comparison that preserves study randomization may not be possible. In this case, an indirect comparison using methods such as matching-adjusted indirect comparison (Signorovitch et al., 2012) may still provide useful information.

For these two reasons, we consider a fourth category below "Hint" would be helpful to reflect the status of the available evidence in these two cases and could aid price negotiations. We hope IQWiG would be open to such an addition.

## 3 Summary

In this paper, we raised questions related to the implementation of AMNOG from an industry perspective. Many of our comments relate to the scientific basis for making pricing and reimbursement decisions. ICH E9 provides background for our comments concerning regulatory approval decisions. One of the tenets of statistical principles in ICH E9 is prespecification. This applies to primary and secondary endpoints, analysis methods, and the analysis sets. We contrast these principles with those adopted by IQWiG and the G-BA in conducting benefit assessments.

While we recognize that the objectives of the benefit assessment required under AMNOG are different from those underlining regulatory decisions, we feel that a stronger link is needed between the way IQWiG/G-BA conducts its benefit assessment and that adopted by the regulatory body (e.g. EMA). This could help reduce confusion and provide some level of consistency in benefit determinations from different decision-making authorities.

We have proposed some possible alternatives in Section 2. We encourage manufacturers to seek parallel consultations with regulatory and HTA decision-making bodies when designing registration trials. When conducting benefit assessments, we prefer following prespecified rules to protect the integrity of the statistical inference. On the rare occasion when deviations from prespecified choices are necessary, we propose a general concept of adjusting the confidence level to construct confidence intervals for treatment effect on endpoints that do not have enough precision under the existing study design. We invite our academic colleagues to conduct research to decide if this could be a viable alternative. Furthermore, we invite our academic colleagues to join in and conduct research that could help solve issues raised in this paper.

**Acknowledgments** The authors wish to thank Bertram Weber, Amelie Kohler, Julia Schiffner-Rohe, Peter Marx, Prof Joachim Röhmel, reviewers, and the Guest Editor for their helpful comments and suggestions that helped improve the quality of this paper substantially.

**Conflict of interest**

F.L. is an employee of Pfizer Deutschland GmbH. C.C.-S. is an employee of Pfizer Inc.

## References

Antman, E. M., Wiviott, S. D., Murphy, S. A., Voitk, J., Hasin, Y., Widimsky, P., Chandna, H., Macias, W., McCabe, C. H., Brauwald, E. (2008). Early and late benefits of prasugrel in patients with acute coronary syndromes undergoing percutaneous coronary intervention. *Journal of the American College of Cardiology* **52**, 2028–2033.

Atkins, D., Best, D., Briss, P. A., Eccles, M. P., Falck-Ytter, Y., Flottorp, S., Guyatt, G. H., Harbour, R. T., Haugh, M. C., Henry, D., Hill, S., Jaeschke, R., Leng, G., Liberati, A., Magrini, N., Mason, J., Middleton, P., Mrukowicz, J., O'Connell, D., Oxman, A. D., Phillips, B., Schünemann, H. J., Edejer, T., Varonen, H., Vist, G. E., Williams, J. W. Jr., Zaza, S.; GRADE Working Group (2004). Grading quality of evidence and strength of recommendations. *British Medical Journal* **328**, 1490.

Bender, R., Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology* **54**, 343–349.

Beresniak, A., Auray, J., Duru, G., Medina-Lara, A., Tarricone, R., Sambuc, R., Torbica, A., Lamure, M.; Echoutcome Study Group (2012). PRM14 European assessment of the validity of the QALY outcome measure: results from the experiment conducted by the Echoutcome project. *Value in Health* **15**, A462.

Bucher, H. C., Guyatt, G. H., Griffith, L. E., Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* **50**:683-91.

Buyse, M., Piedbois, P. (2008). Endpoints and surrogate endpoints in colorectal cancer: a review of recent developments. *Current Opinion in Oncology* **20**, 466–71.

Buyse, M., Piedbois, P. (2010). Meta-analysis based on individual patient data: example of advanced colorectal cancer. *Recherche en Soins Infirmiers* **101**, 25–28.

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., Van der Elst, W., Burzykowski, B. (2015). Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal* (this issue).

Djulbegovic, B., Kumar, A., Soares, H. P., Hozo, I., Bepler, G., Clarke, M., Bennett, C. L. (2008). 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. *Archives of Internal Medicine* **168**, 632–642.

Duru, G., Auray, J. P., Beresniak, A., Lemure, M., Paine, A., Nicoloyannis, N. (2002). Limitations of the methods used for calculating quality-adjusted life-year values. *Pharmacoeconomics* **20**, 463–473.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R. et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European Journal of Cancer* **45**, 228–247.

European public assessment report (EPAR) for Xalkori (2012). Available at http://www.ema.europa.eu/ema/index.jsp?curl=pages/human/medicines/002489/human_med_001592.jsp&mid=WC0b01ac058001d124 (accessed 3 September 2014).

European Medicines Agency (2013). Guideline on the evaluation of anticancer medicinal products in man. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/01/WC500137128.pdf (accessed 18 October 2013).

European Medicines Agency (2014). Guideline on the investigation of subgroups in confirmatory clinical trials (DRAFT). EMA/CHMP/539146/2013. Available at http://www.ema.europa.eu/ema/index.jsp?curl=pages/includes/document/document_detail.jsp?webContentId=WC500160523&mid=WC0b01ac05800 9a3dc (accessed 1 January 2015).

Federal Ministry of Health (2013). Das Gesetz zur Neuordnung des Arzneimittelmarktes (AMNOG). Available at http://www.bmg.bund.de/krankenversicherung/arzneimittelversorgung/arzneimittelmarktneuordnungs-gesetz-amnog/das-gesetz-zu-neuordnung-des-arzneimittelmarktes-amnog.html (accessed 14 October 2013).

G-BA (2011a). Chapter 5: assessment of the benefits of pharmaceuticals according to §35a SGB V. Available at http://www.english.g-ba.de/downloads/17-98-3042/Chapter5-Rules-of-Procedure-G-BA.pdf (accessed 24 October 2013).

G-BA (2011b). Anlage III - Vorlage zur Abgabe einer schriftlichen Stellungnahme zur Nutzenbewertung nach § 35a SGB V. Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ). 25.10.2011. Available at http://www.akdae.de/Stellungnahmen/AMNOG/A-Z/Ticagrelor/Ticagrelor.pdf (accessed 24 October 2013)

G-BA (2012a). Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V – Cabazitaxel. 29.03.2012. Available at http://www.g-ba.de/downloads/40-268-19 10/2012-03-29_AM-RL-XII_Cabazitaxel_TrG.pdf (accessed 24 Oct 2013).

G-BA (2012b). Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V - Ipilimumab. 02.08.2012. Available at http://www.g-ba.de/downloads/40-268-2010/2012-08-02_AM-RL-XII_Ipilimumab_TrG.pdf (accessed 24 October 2013).

G-BA Assessment (2013). Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel – Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V - Crizotinib. Available at https://www.g-ba.de/informationen/nutzenbewertung/44/#tab/beschluesse (accessed 3 September 2014).

G-BA (2014). Verfahrensordnung des Gemeinsamen Bundesausschusses; 5. Kapitel: Bewertung des Nutzens und der Kosten von Arzneimitteln nach §§ 35a und 35b SGB. 18. Dec 2008. Available at https://www.g-ba.de/downloads/62-492-873/VerfO_2014-03-20.pdf (accessed 8 September 2014).

ICH E9 Statistical Principles for Clinical Trials. 5 February 1998, Step 4. Available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step 4/E9_Guideline.pdf (accessed 6 October 2013).

IQWiG Reports – Commission No. A10-05 (2011a). Validity of surrogate endpoints in oncology (Version 1.1; Status 21.11.2011). Available at https://www.iqwig.de/download/A10-05_Executive_Summary_v1-1_Surrogate_endpoints_in_oncology.pdf (accessed 28 August 2014).

IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (2011b). Ticagrelor: Nutzenbewertung gemäß § 35a SGB V. IQWiG-Berichte – Jahr 2011 Nr. 96. Available at https://www.iqwig.de/download/A11-02_Ticagrelor_Nutzenbewertung_35a_SGB_V_.pdf (accessed 24 October 2013).

IQWiG (2012a). Auftrag A11-30; Dossierbewertung: Apixaban (neues Anwendungsgebiet) – Nutzenbewertung gemäß § 35a SGB V; 15.03.2012. Available at https://www.g-ba.de/informationen/nutzenbewertung/5/#tab/nutzenbewertung (accessed 3 September 2014).

IQWiG (2012b). Cabazitaxel – Nutzenbewertung gemäß § 35a SGB V. IQWiG-Berichte – Nr. 114. Available at https://www.g-ba.de/downloads/92-975-32/2011-04-15-D-003_Cabazitaxel_IQWiG-Nutzenbewertung.PDF (accessed 3 June 2014).

IQWiG (2012c). Ipilimumab – Nutzenbewertung gemäß § 35a SGB V. IQWiG-Berichte – Nr. 130. Available at https://www.g-ba.de/downloads/92-975-108/2011-08-01-D-014_Ipilimumab_IQWiG-Nutzenbewertung.pdf (accessed 3 June 2014).

IQWig Institute for Quality and Efficiency in Health Care (2013). General Methods Version 4.1 of 28 Nov 2013. Available at https://www.iqwig.de/download/IQWiG_General_Methods_Version_%204-1.pdf (accessed 28 August 2014).

Lassere, M. N. (2008). The biomarker-surrogacy evaluation schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Statistical Methods in Medical Research* **17**, 303–340.

Lu, G., Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23**:3105-3124.

National Institute for Health and Care Excellence (2013). Guide to methods of technology appraisal 2013. National Health Service, England. Available at http://www.nice.org.uk/article/pmg9/ (accessed 24 March 2015).

Nixon, R., Dierig, C., Mt-Isa, S., Stockert, I., Tong, T., Kuhls, S., Hodgson, G., Pears, J., Waddingham, E., Hockley, K., Thomson, A. (2015). A case study using the PrOACT-URL and BRAT frameworks for structured benefit risk assessment. *Biometrical Journal* (this issue).

Paget, M.-A., Chuang-Stein, C., Fletcher, C., Reid, C. (2011). Subgroup analyses of clinical effectiveness to support health technology assessments. *Pharmaceutical Statistics* **10**, 532–538.

Röhmel, J. (2013). Stellungnahme von Joachim Röhmel zu Teilschritt 1: Entwurf der Änderungen der Allgemeinen Methoden 4.0 Aktualisierung einiger Abschnitte. Available at https://www.iqwig.de/de/methoden/methodenpapiere/allgemeine-methoden.3020.html (accessed 3 September 2014).

Signorovitch, J., Erder, M. H., Xie, J., Sikirica, V., Lu, M., Hodgkins, P. S., Eric, Q., Wu, E. Q. (2012). Comparative effectiveness research using matching-adjusted indirect comparison: an application to treatment with guanfacine extended release or atomoxetine in children with attention-deficit/hyperactivity disorder and comorbid oppositional defiant disorder. *Pharmacoepidemiology and Drug Safety* **21**, 130–137.

Skipka, G., Wieseler, B., Kaiser, T., Thomas, S., Bender, R., Windeler, J., Lange, S. (2015). A methodological approach to determine minor, considerable and major treatment effets in the early benefit assessment of new drugs. *Biometrical Journal* (this issue).

Stone, A., Chuang-Stein, C. (2013). Strong control over multiple endpoints: are we adding value to the assessment of medicines? *Pharmaceutical Statistics* **12**, 189–191.

Vach, W., Gladstone, B. P. (2015). A framework to assess the value of application of formal criteria to check clinical relevance in RCTs as part of a benefit assessment strategy. *Biometrical Journal* (this issue).

Verband forschender Pharma-Unternehmen (vfa) (2012). Sample presentation for the "AMNOG—act for the restructuring of the pharmaceutical market in statutory health insurance". Available at http://www.vfa.de/de/inline/amnog/allgemeine-informationen.html#musterpraesentationen (accessed 31 October 2013).

Waddingham, E., Mt-Isa, S., Nixon, R., Ashby, D. (2015). A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment. *Biometrical Journal* (this issue).