BMJ Open

Exploratory analyses of clinical trial data used for health technology assessments: a retrospective evaluation

Björn J Oddens ,[1] Israel T Agaku ,[2] Ellen S Snyder ,[1] William Malbecq,[3] William WB Wang,[1] Karen M Kaplan,[1] Gary G Koch,[4] Frank W Rockhold [5]

To cite: Oddens BJ, Agaku IT, Snyder ES, *et al*. Exploratory analyses of clinical trial data used for health technology assessments: a retrospective evaluation. *BMJ Open* 2022;**12**:e058146. doi:10.1136/bmjopen-2021-058146

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-058146).

Received 07 October 2021
Accepted 22 June 2022

[1]MRL, Merck & Co, Rahway, New Jersey, USA
[2]Department of Oral Health Policy and Epidemiology, Harvard School of Dental Medicine, Boston, Massachusetts, USA
[3]MRL, MSD Europe, Brussels, Belgium
[4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[5]Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina, USA

Correspondence to
Dr Björn J Oddens;
bjorn.oddens@merck.com

ABSTRACT

Objectives To examine the validity and statistical limitations of exploratory analyses of clinical trial data commonly requested by agencies responsible for determining which medical products may be financed or reimbursed by a healthcare system.

Design This was a retrospective review of efficacy and safety analyses conducted for German Health Technology Assessment (HTA) evaluations with a decision date between 2015 and 2020, and an illustrative safety-related exploratory analysis of data from two phase III clinical trials of verubecestat (an anti-amyloid drug whose development was stopped for lack of efficacy) as would be mandated by the German HTA agency.

Results We identified 422 HTA evaluations of 404 randomised controlled clinical trials. For 140 trials (34.7%), the evaluation was based on subpopulations of participants in the originating confirmatory trial (175 subpopulations were assessed). In 57% (100 of 175), the subpopulation sample size was 50% or less of the original study population. Detailed analysis of five evaluations based on subpopulations of the original trial is presented. The safety-related exploratory analysis of verubecestat led to 206 statistical analyses for treatments and 812 treatment-by-subgroup interaction tests. Of 31 safety endpoints with an elevated HR (suggesting association with drug treatment), the HR for 81% of these (25 of 31) was not elevated in both trials. Of the 812 treatment-by-subgroup interactions evaluated, 26 had an elevated HR for a subgroup in one trial, but only 1 was elevated in both trials.

Conclusions Many HTA evaluations rely on subpopulation analyses and numerous post hoc statistical hypothesis tests. Subpopulation analysis may lead to loss of statistical power and uncontrolled influences of random imbalances. Multiple testing may introduce spurious findings. Decisions about benefits of medical products should therefore not rely on exploratory analyses of clinical trial data but rather on prospective clinical studies and careful synthesis of all available evidence based on prespecified criteria.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ All German Health Technology Assessment (HTA) evaluations with a decision date between 2015 and 2020 were reviewed.
⇒ Analysis of the full safety data sets from two pivotal phase III clinical trials of verubecestat, including multiplicity adjustment to control the false discovery rate; the availability of complete data sets from two similar trials allowed the consistency of the observed associations to be assessed.
⇒ Of 422 HTA evaluations reviewed, only 5 (chosen to illustrate the methodological challenges that can occur with exploratory analyses for HTAs) are presented in detail.
⇒ The verubecestat clinical development programme was stopped, therefore no post-marketing surveillance is available to support conclusions of the detailed safety analysis reported here.

INTRODUCTION

The strength of conclusions that can be reached by analysis of data from clinical trials of medicinal products differs markedly based on whether the analysis is confirmatory (ie, hypothesis confirming) or exploratory (ie, hypothesis generating). A typical phase III clinical trial is designed to confirm the efficacy and continue the safety assessment of a medicine that has already been evaluated in a series of preclinical and clinical studies. A number of fundamental principles are followed.[1] The study population is well defined to include patients for whom the results are relevant and who could benefit from the treatment. Treatment assignment is randomised to produce statistically comparable distributions of baseline variables among patients receiving the studied treatments. The number of recruited participants is prospectively determined to be large enough to detect differences between treatments for the outcomes of interest. Primary study endpoints are relatively few and specified a priori to limit spurious findings; that is, those emerging from coincidence (chance) or unknown factors, rather than causally related to treatment. To prevent a post hoc search for positive results (especially efficacy advantages), a statistical analysis plan is prepared before the clinical trial data are unblinded and analysed. For these

reasons, the conclusions derived from the predefined, statistical analysis of data from phase III confirmatory trials are considered robust, and inferences about cause (treatment) and effect (patient response) can be made from them.

Robust clinical trials also include supportive 'secondary' analyses, planned a priori, to demonstrate additional findings that are expected to be in harmony with the primary endpoints. These secondary analyses are part of the statistical analysis plan, and the potential for false positive results is rigorously controlled by standard statistical methodology (eg, type I error control or false discovery rate correction). In addition, cautious subgroup analyses are often used to examine the consistency of findings in subgroups of interest with those for the overall population[1]; however, if a subgroup analysis results in unexpected findings, the result requires support by other clinical or biological data, or confirmation in a new trial assessing the reproducibility of the unexpected observation, before the findings are used to change clinical practice.[2]

In contrast, post hoc or 'exploratory' analysis of data from phase III trials may not have internal validity (the ability of a study to correctly assess cause and effect for the population that was the target of the study) because of qualitative and quantitative deviations from the original trial protocol. The study population may no longer be well defined (eg, pooled analysis of several confirmatory trials with different populations); there may be an insufficient number of participants to detect differences; absence of randomisation can occur (eg, by excluding data for trial participants treated with a specific comparator treatment); and there is an increased risk of spurious findings—the problem of 'multiplicity' (examining numerous endpoints). These deviations introduce a risk of bias and 'noise' into the analysis, resulting in unreliable interpretations that must at best be considered exploratory hypotheses until tested in a designed-for-purpose confirmatory trial.[2]

Despite the risks inherent in post hoc exploratory analyses, such analyses are often requested by funding authorities charged with evaluating which medical products will be financed or reimbursed by a healthcare system. For example, the Federal Joint Committee (Gemeinsamer Bundesausschuss or G-BA), the German Health Technology Assessment (HTA) agency and highest decision-making body for health insurance funds in Germany, commissions HTAs by the Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). According to G-BA requirements, a product sponsor must identify and reanalyse data for a '[for Germany] relevant subpopulation from the clinical studies', as well as analyse subgroups within this subpopulation defined by gender, age, disease severity or stage, and country or geographical region of treatment—irrespective of whether the subpopulation or subgroups were considered in the study design.[3] The English National Institute of Clinical Excellence (NICE) calls for providing clinical (and cost-effectiveness) estimates separately for 'each relevant subgroup of patients'.[4] The French Haute Autorité de Santé (HAS) evaluates clinical trial subgroups for potential differences in expected health effects, and requires that the cost-effectiveness of the product is assessed for each subgroup analysed.[5] These requirements for analysis of subgroups—a form of exploratory analysis—can be driven by a variety of factors. There may be concern that the licensed indications of a medicinal product are too broad and not limited to those patients who, in the view of the HTA agency, have a need unmet by current treatments. For reimbursement purposes, HTA agencies may be interested in knowing which patient subgroups in a trial experienced the greatest clinical benefit from the new product. An agency may also have questions regarding the appropriateness of a comparator drug used in the trial, especially for large, multicountry studies. In such settings, HTA agencies may recommend removing from an analysis those patients treated with comparators not considered standard care in their country, or not in their list of recommended regimens.[6]

Paget *et al*[7] described statistical principles for subgroup analyses supporting HTA evaluations, emphasising the methodological limitations that must be addressed. These limitations apply both to subpopulation analyses (those which focus on a subset of the original trial population and discard the data from other trial participants) and subgroup analyses (those which cross-tabulate data from participants based on, for example, age, disease severity, prognostic factors or concomitant medication and which do not discard data from any trial participants defined as included by the original statistical analysis plan). Paget *et al* recommend focusing only on prespecified subpopulations or subgroups that have a clear biological rationale, adjusting for multiplicity in statistical analyses, conducting sensitivity analyses to understand the robustness of the data, quantifying the uncertainty of results and replicating results with independent data sources. They also outline how to present and report such results. In recent years, however, the reliance of HTA agencies on exploratory analyses has greatly expanded, as illustrated by the examples of the G-BA, NICE and HAS, increasing the risk of spurious findings. The number of countries requiring such analyses has increased, the purpose for which they are being required has expanded (initially for efficacy and more recently for safety endpoints) and the scale to which they are being required (in terms of number of variables being assessed) is unprecedented. For example, the G-BA and Agenzia Italiana de Farmaco (AIFA)[8] have added new requirements for analyses of adverse event (AE) outcomes within the context of the Medical Dictionary of Regulatory Activities. These analyses must classify AEs at the highly nuanced 'Preferred Term' level (specific symptom, sign or diagnosis; for example, 'arrhythmias', totalling 27 308 possible terms) and not just at the 'System Organ Class' (SOC) level (umbrella term for all conditions affecting a given system or organ, for example, 'cardiac disorders', totalling 27 terms), for the subpopulation that was deemed relevant

as well as by each subgroup within that subpopulation.[3] The G-BA explained the motivation of their new requirement as 'asking for more precise documentation of all endpoints, subgroups and data cuts from clinical studies. Moreover, all AEs and serious AEs should be reported for all data cuts to make the risk profile transparent'.[9] Consequently, hundreds of exploratory analyses including inferential statistics may be carried out on a data set from a trial specifically designed to assess only a few confirmatory endpoints. The G-BA posts the results of the exploratory analyses on their website for full transparency. For this agency, as well as AIFA, these might be viewed as supportive analyses assessing whether calculated levels of risk versus a comparator treatment (eg, relative risk or risk difference) are acceptable. But these calculations can yield spurious findings, both false positives, due to multiple statistical testing, and false negatives, due to the small sample sizes of the data cuts. These potential false positives and false negatives are thus also included in the publicly disclosed information about the drug, without consideration (or a warning) about the methodological shortcomings of the exploratory analyses that produced them.

In this paper, we illustrate the statistical issues associated with drawing conclusions from large numbers of exploratory analyses, first by reviewing several German HTA evaluations and then by analysing data sets from two confirmatory clinical trials according to current German HTA requirements.[3] We address the possibilities of spurious findings leading to erroneous conclusions from the mandated exploratory analyses. We conclude with recommendations for future directions.

METHODS

Review of actual evaluations by the German HTA agency and associated exploratory analyses

We used the example of Germany because HTA evaluation requirements in this country are defined by federal legislation, and because sponsor HTA dossiers and agency assessments are public documents. In addition, the German HTA requirements are among the most extensive worldwide, even when other agencies share elements of the G-BA approach (subgroups as in England and France and safety analyses as in Italy). We searched all evaluations completed by the G-BA, that is, all with a decision date in the period 2015–2020 (all are available at www.G-BA.de; date of retrieval 18 June 2021).

We identified the clinical trials considered by the G-BA during each evaluation and determined whether the entire study population from any relevant clinical trial was included in the evaluation, or instead, a subset of the entire study population or subpopulations of the individual study arms were used. We compared these patient numbers with the total size of the original trials (or size of the relevant study arms). Beyond this, it is impossible to compare the details of all the HTA evaluations carried out during the defined period in a uniform way—they include evaluations of new products and new indications for products that are already reimbursed, as well as a variety of data types (original clinical trials, mixed treatment-comparisons, literature data). Therefore, we chose five HTA evaluations from different sponsors and for different therapeutic areas that were based on subpopulations of the originating clinical trial population: two oncology products (MSD's pembrolizumab and another company's axitinib), two specialty care products (sarilumab, alirocumab) and one primary care product (tiotropium/olodaterol). We considered how the following characteristics were affected by the data reduction that was applied: (1) multiplicity; (2) uncontrolled influences of random imbalances; (3) precision and power from subgroup analyses; and (4) epidemiological and statistical interpretations.

Case study: exploratory analyses of verubecestat clinical trial safety data

To illustrate the implications of HTA authority requirements for a large set of post hoc analyses (such as the new German and Italian requirements related to drug safety), we conducted a series of analyses of data from two phase III clinical trials of verubecestat (NCT01739348 and NCT01953601) conducted by Merck & Co (Rahway, New Jersey, USA), which together formed the pivotal trial programme for the drug.[10 11] Verubecestat is an orally administered β-site amyloid precursor protein-cleaving enzyme 1 (BACE-1) inhibitor that blocks production of amyloid-β. The trials compared verubecestat, as a potentially disease-modifying treatment, with placebo. The drug was expected to prevent clinical progression in patients with mild-to-moderate dementia[10] and amnestic mild cognitive impairment due to Alzheimer's disease.[11] When results of these trials failed to support efficacy of treatment, clinical development was discontinued. We nevertheless analysed the two clinical trial data sets specifically according to the new German requirements as a case study, conducting the mandated analyses of AEs for the entire trial population and then for subgroups. We conducted the exploratory analyses of AEs by SOC and Preferred Term for events meeting the incidence criteria set by the G-BA (ie, incidence ≥10%, or ≥1% and in at least 10 patients in one or more treatment groups) using time-to-event Cox regression analyses. These analyses were performed for the entire trial population. AEs with elevated HRs (lower limit of the 95% CI exceeding 1) were identified. Although the G-BA formally reviews both elevated and reduced AEs versus the comparator drug to assess clinical benefit and value, the focus is usually on elevated AEs associated with the innovation drug, reducing its value.

We compared the number of results with an elevated HR (roughly corresponding to nominal one-sided $p<0.025$) expected due to chance versus the number observed. The expected number due to chance at an α-level (ie, type I error) of 2.5% (one sided because we looked only at elevated HRs) was derived by multiplying

0.025 by the number of tests performed. To resolve some of the uncertainty produced by these mandated safety analyses at SOC and Preferred Term level, we used multiplicity control of the false discovery rate according to the methods of Benjamini and Hochberg,[12] and Mehrotra and Adewale.[13] To assess the likelihood of a true association, all SOC and Preferred Term AEs exhibiting nominal association with treatment were medically evaluated based on consistency with the known safety profile of verubecestat, including knowledge of BACE physiology, and data from preclinical studies and previous clinical trials of the drug. The AEs were also classified based on HR thresholds of the G-BA for major, considerable and minor clinical harm (in accordance with G-BA guidance,[14] the inverse thresholds for clinical benefits were used). We evaluated the outcome of the statistical analyses, medical evaluation and G-BA classification using the criteria of Hill[15]: (1) strength of the association (effect size); (2) consistency of the association (reproducibility); (3) specificity (uniqueness of outcome to exposure); (4) temporality (exposure precedes outcome); (5) biological gradient (dose–response relationship); (6) plausibility (biological rationale); (7) coherence (eg, consistency of the observation with preclinical research); (8) additional experimental evidence; (9) analogy (with other known associations). We also performed treatment-by-subgroup interaction tests with Cox regression analysis. If both the main treatment effect and interaction terms for the endpoint had a nominal p<0.05 (two sided since the interaction can indicate an elevation or reduction), the treatment effect was estimated at each of the corresponding subgroup levels. The expected number of interactions due to chance was determined by multiplying 0.05 by the number of tests performed.

Patient and public involvement

There was no patient or public involvement with this research.

RESULTS

We present results of our review of German HTA dossiers, followed by examination of our case study findings. To provide context for the review, results are presented by methodological issue involved, after a brief overview of that issue.

Review of German HTA exploratory analyses

We retrieved and reviewed 422 German HTA evaluations. In 287 of these, at least one randomised controlled trial was included (404 trials). For 264 trials (65.3%), the total study population was considered for the value assessment. For 140 trials (34.7%), subpopulations were considered. For these 140 trials, 175 different subpopulations were analysed (eg, for partial application areas). In 51 cases (29.1%), the subpopulation sample size was 30% or less of the size of the original study population, and in 100 (57.1%), 50% or less of the original study population was considered (figure 1).

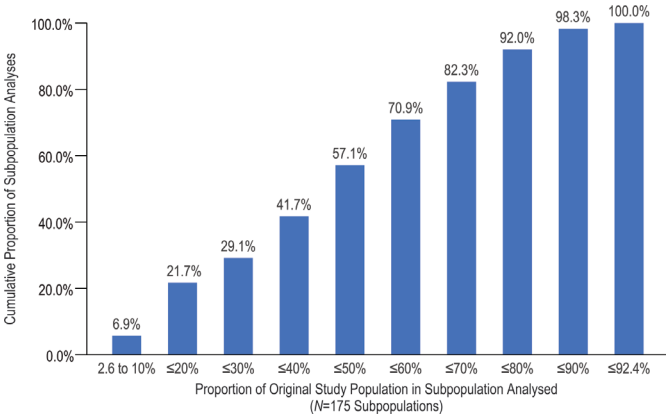

Figure 1 Cumulative frequency (Y-axis) of the proportion of the original study population (X-axis) considered for value assessment in Germany between 2015 and 2020. These proportions were calculated based on a total of 175 subpopulations taken from 140 trials.

The five HTA evaluations presented in table 1 illustrate the rationale used by the G-BA to reduce clinical trial data sets to subpopulations. The exploratory analyses included in these selected examples used approximately one-third to one-half of the original clinical trial sample size (range: 35%–58%). The ensuing methodological problems are discussed in greater detail below.

The problem of multiplicity

Overinterrogation of clinical trial data without the safeguard of controlling false positive findings has been extensively written about in the statistical literature. Numerous analyses of a data set introduce the risk of spurious findings and false conclusions since every statistical test has an error rate.[2 7] An α of 5% applied to a single test implies an error rate of 5%, meaning that 1 in 20 of such statistical analyses will inappropriately reject the null hypothesis when it is in fact true (concluding there is a treatment effect when there is none, a type I error or false positive). The test may also inappropriately fail to reject a null hypothesis when it is false (concluding there is no treatment effect when there actually is, a type II error or false negative).[16] Either type of error can lead to incorrect conclusions when evaluating either treatment efficacy or safety, and the likelihood of both is increased by multiplicity or multiple queries of a data set. Subgroup/subpopulation analyses are among the most common forms of multiple queries, and resulting treatment effect estimates and inferences are weakened by the multiplicity problem.[2 7 17]

Our review of past German HTA evaluations revealed multiple subgroup/subpopulation analyses by various clinical and demographic characteristics (table 1) from each primary data set, indicating that every subsequent conclusion may be increasingly compromised by multiplicity. It is, however, impossible to know the extent to which this querying of the data multiple times has

Table 1 German HTA evaluations: pivotal clinical trials considered, original trial population, HTA population (% of the original trial population) and clinical benefit assessment

Product (evaluation year)	Disease	Trial name/ number (phase)	Clinical trial population	Subpopulation evaluated	HTA population* (% of trial population)	G-BA clinical benefit assessment
Pembrolizumab (2020)[23]	Squamous cell head and neck carcinoma	Keynote-048 (phase 3)	882	Only patients with PD-L1 combined positive score ≥1%. Pembrolizumab monotherapy only	512 (58.0)	Considerable benefit
Axitinib (2017)[24]	Advanced renal cell carcinoma after failure of prior treatment with sunitinib or a cytokine	AXIS	723	Only patients with prior cytokine-based treatment	251 (34.7)	Minor benefit
Tiotropium/olodaterol (2015)[25]	Chronic obstructive pulmonary disease (COPD)	TONADO 1+2 (phase 3) — population of pooled analyses	2063	Two subpopulations were requested: ► Grade II and grade III–IV COPD with ≤2 exacerbation per year without inhaled corticosteroids use ► Grade III–IV COPD with ≥2 exacerbation per year with inhaled corticosteroids use Exclude olodaterol monotherapy comparator subjects and patients treated with 2.5 µg tiotropium	Subpopulation 1: 988 (47.9) Subpopulation 2: 144 (7.0)	1 subpopulation minor benefit, 1 subpopulation less benefit
Sarilumab (2017)[26]	Rheumatoid arthritis	MONARCH (phase 3)	369	Only patients for whom the doctor documented methotrexate intolerance were considered appropriate	169 (45.8)	1 subpopulation considerable benefit, 3 subpopulations no additional benefit
Alirocumab (2018)[27]	Atherosclerotic cardiovascular disease	ODYSSEY OUTCOMES (phase 3)	18 924	Only patients with a sufficiently high (maximum) dose of concomitant statin use were considered	8790 (46.4)	No additional benefit

*In evaluations including >1 subpopulation, the subpopulations were non-overlapping.
G-BA, Gemeinsamer Bundesausschuss; HTA, Health Technology Assessment.

introduced random errors which could lead to incorrect inferences.

Uncontrolled influences of random imbalances

Randomisation in clinical trials is performed in an effort to eliminate selection bias that could influence trial results for its overall population as that corresponds to the inclusion and exclusion criteria defined in the protocol.[18] It is achieved by assignment to a group using a random identification number rather than by location, time of enrolment or other criteria. Accordingly, randomisation enables such groups to have statistically comparable distributions of both observed and unobserved baseline variables.

However, for subgroups/subpopulations, the treatment groups have smaller, and possibly randomly different, sample sizes; and they are also vulnerable to random imbalances of prognostic baseline variables. An important consequence of these sources of randomness is much larger variability for the estimates of treatment comparisons for subgroups/subpopulations than for the overall population (ie, much wider CIs). Since adjustment for prognostic covariates can reduce such variability, assessments for subgroups/subpopulations need their use to the extent that is possible.

Within the German HTA evaluations reviewed, exploratory analyses typically involved segmentation of the trial

population (by exclusion) into subpopulations in systematic ways, using circumstantial factors likely to result in imbalances between the treatment arms, rather than common biological variables such as age and gender (table 1). Reasons for exclusion of patients varied, from consideration of only specific comparators (the case of tiotropium/olodatorol), to only specific prior treatments (axitinib), monotherapy (pembrolizumab) or disease severity (tiotropium/olodatorol), or to only patients intolerant to other treatments (sarilumab) or who were optimally pretreated (alirocumab). Such inclusions or exclusions could create prognostic differences between the analysed patient subpopulations and consequently bias the estimation of the treatment effect. Analysis mandates do not include comparisons of baseline characteristics between the subpopulations and the originating trial population to evaluate potential imbalances the subpopulation approach might have created, and if such comparisons were done, they were not reported. And covariate adjustment for imbalances in predictive variables was not applied. This lack of control for potential random imbalances in subpopulation characteristics increases the potential for invalid findings.

Reduced precision and power

The precision of a statistical analysis determines how wide (imprecise) or how narrow (precise) the CI of any given point estimate is. Power determines the ability of a statistical test to reject the null hypothesis that there is no difference in the values of a particular parameter between the treatment arms. When a CI of the treatment difference is available, the rejection of the null hypothesis of no difference is corroborated when the CI does not contain the 'no difference' value (ie, 0 for interval scale and 1 for ratio scale). It is important to note that statistical significance in clinical studies does not necessarily imply a clinically meaningful difference in the point estimates. If a difference is statistically robust but numerically small, clinicians might question the incremental value of the treatment intervention. Both precision and power are driven by sample size (in this case trial population size): a larger sample size provides a more precise estimate and a more powerful test of difference for a superiority assessment or exclusion of an unacceptable difference for a non-inferiority assessment. To calculate the minimum number of patients needed in each treatment arm to detect a treatment effect, statisticians consider the size of the expected treatment effect and the natural variation (independent of treatment) in the size of that effect.[1]

A subpopulation analysis by default uses smaller sample sizes than were enrolled in the original study and is thus frequently imprecise and underpowered. Yet, the wider CIs from subpopulation analyses can still be interpretable *if* they overlap with those for the overall trial population, indicating harmony or consistency with the overall trial. In our review of 422 German HTA evaluations, we noted that 34.7% analysed only subsets of the original clinical trial data. The large confirmatory trials were reduced to only half or even one-third of their original size. In accordance with G-BA requirements, these subpopulations were additionally analysed by subgroups of age, gender, disease severity, and country or region of treatment, further splitting the samples. In some cases, the sample sizes became very small, for example, in the sarilumab MONARCH trial and in subpopulation 2 of the tiotropium/olodaterol TONADO trial (table 1). In the TONADO trial, the sample size of subpopulation 2 was about 1/14 of that for each of the original arms. Given that the SE for an estimate for a treatment group has the square root of its sample size (N) in the denominator (ie, $SE=SD/\sqrt{N}$), reducing the sample to one-fourth doubles the width of the CI. In the example above, in which the sample size was reduced to 1/14, the width of the CI is almost quadrupled.

Problems with clinical benefit assessment

Our review of the five examples of German HTA evaluations summarised in table 1 revealed inconsistencies in observed clinical benefit of the new drugs. For these drugs, benefits were recognised only for a subpopulation of the clinical trial. In the cases of tiotropium/olodaterol and salirumab, benefit was noted for some subpopulations, but not for other (mutually exclusive) subpopulations. Given the wider CIs resulting from smaller sample size in these analyses, the conclusion that no benefit was discernible is not surprising but may not be true, considering that the overall study population results showed benefit.

Case study showing potential for erroneous conclusions from mandated safety analyses

Application of the new German HTA requirements for safety analyses to the phase III clinical trials of verubecestat led to 206 statistical analyses for AEs at the SOC or Preferred Term level for the *entire trial population:* for 16 of them the investigator had rated the event (at SOC or Preferred Term level) as serious and for 12 the company had rated the event as severe (ie, Common Terminology Criteria for Adverse Events [CTCAE] grade 3–5); the others were not additionally classified. One hundred and seventy SOCs and Preferred Term events across the two studies met the German HTA incidence criteria and were elevated. According to standard statistical assumptions, four of these comparisons are predicted to have a one-sided $p<0.025$ simply due to chance (ie, at one-sided α of 2.5%, 0.025×170 tests≈4). However, in our analyses, 20 Preferred Terms and 11 SOC terms were numerically higher in the active treatment group compared with placebo and had an elevated HR (lower limit of the 95% CI exceeding 1), yielding a total of 31 elevated HRs (table 2). 'Major harm', in the sense of the G-BA criteria, was not observed. However, 'considerable harm' was observed 10 times (four times in both trials; but in six cases appearing as 'considerable harm' in one trial and 'minor' (one AE) or 'no harm' (five AEs) in the other). The harm levels were for 20 events rated differently

Table 2 Incidence, HRs, G-BA classification, 95% CIs and nominal (two-sided) p values resulting from exploratory statistical analyses of the incidences of adverse events in phase III clinical trials of verubecestat for dementia and amnestic mild cognitive impairment due to Alzheimer's disease, as mandated for German Health Technology Assessment

Adverse event	Protocol 17[10] Incidence* HR G-BA classification Nominal p value (non-bold); adjusted p value (bold)	Protocol 19[11] Incidence* HR G-BA classification Nominal p value (non-bold); adjusted p value (bold)	Medical evaluation†
System organ class			
Nervous system disorders	31.7% vs 24.5% *1.40 (1.15 to 1.71)* Minor harm **p=0.0010**	34.5% vs 29.3% *1.28 (1.02 to 1.60)* No harm p=0.0302	Relation to the drug may be possible
Psychiatric disorders	37.4% vs 25.7% *1.64. (1.36 to 1.98)* Considerable harm **p=0.0010**	38.0% vs 27.7% *1.57 (1.26 to 1.96)* Considerable harm **p=0.0001**	Relation to the drug may be possible
Skin and subcutaneous tissue disorders	25.9% vs 16.7% *1.69 (1.34 to 2.14)* Considerable harm **p=0.0010**	38.6% vs 24.4% *1.89 (1.50 to 2.38)* Considerable harm **p=0.0001**	Relation to the drug may be possible
Musculoskeletal and connective tissue disorders	23.3% vs 16.3% *1.53 (1.20 to 1.94)* Minor harm **p=0.0010**	25.2% vs 30.2% 0.87 (0.68 to 1.11) No harm p>0.05	Uncertain
Injury, poisoning and procedural complications	26.6% vs 18.9% *1.54 (1.23 to 1.92)* Minor harm **p=0.0010**	26.0% vs 23.4% 1.21 (0.94 to 1.56) No harm p>0.05	Uncertain—higher rate of falls/ fractures may be drug related
Renal and urinary disorders	11.1% vs 7.0% *1.70 (1.19 to 2.43)* Minor harm **p=0.0035**	10.3% vs 9.5% 1.15 (0.77 to 1.72) No harm p>0.05	Unlikely to be related to drug
Metabolism and nutrition disorders	11.6% vs 7.1% *1.73 (1.21 to 2.46)* Minor harm **p=0.0024**	11.6% vs 9.7% 1.27 (0.86 to 1.87) No harm p>0.05	Primarily due to weight loss, relationship to drug possible
Gastrointestinal disorders	30.0% vs 23.7% *1.34 (1.10 to 1.65)* Minor harm **p=0.0044**	33.5% vs 29.8% 1.22 (0.97 to 1.52) No harm p>0.05	Unlikely to be related to drug
Infections and infestations	37.9% vs 33.2% *1.25 (1.05 to 1.49)* No harm **p=0.0130**	39.3% vs 40.3% 1.04 (0.85 to 1.27) No harm p>0.05	Unlikely to be related to drug
Blood and lymphatic system disorders	3.1% vs 1.4% *2.29 (1.08 to 4.83)* No harm p=0.0299	3.3% vs 3.3% 1.06 (0.53 to 2.11) No harm p>0.05	Unlikely to be related to drug
Investigations	14.7% vs 12.2% 1.25 (0.94 to 1.67) No harm p>0.05	13.8% vs 10.1% *1.49 (1.03 to 2.15)* No harm p=0.0345	Unlikely to be related to drug
Preferred terms			
Weight decreased	6.3% vs 3.0% *2.24 (1.33 to 3.76)* Considerable harm **p=0.0024**	6.6% vs 2.1% *3.50 (1.72 to 7.12)* Considerable harm **p=0.0005**	Relation to the drug may be possible
Anxiety	6.9% vs 3.8% *1.88 (1.17 to 3.01)* Minor harm p=0.0087	9.1% vs 4.3% *2.27 (1.35 to 3.83)* Considerable harm p=0.0019	Relation to the drug may be possible

Continued

Table 2 Continued

Adverse event	Protocol 17[10] Incidence* HR G-BA classification Nominal p value (non-bold); adjusted p value (bold)	Protocol 19[11] Incidence* HR G-BA classification Nominal p value (non-bold); adjusted p value (bold)	Medical evaluation†
Urticaria	1.9% vs 0.4% *4.46 (1.27 to 15.64)* Considerable harm p=0.0196	3.7% vs 0.8% *4.75 (1.61 to 14.04)* Considerable harm p=0.0048	Relation to the drug may be possible
Rash	3.7% vs 3.6% 1.06 (0.61 to 1.83) No harm p>0.05	6.6% vs 2.5% *2.84 (2.46 to 5.52)* Considerable harm p=0.0020	Relation to the drug may be possible
Depression	6.6% vs 5.5% 1.24 (0.81 to 1.90) No harm p>0.05	10.3% vs 5.2% *2.19 (1.35 to 3.54)* Considerable harm p=0.0014	Uncertain
Insomnia	4.7% vs 3.0% 1.65 (0.95 to 2.84) No harm p>0.05	6.2% vs 2.7% *2.49 (1.3 to 4.78)* Considerable harm p=0.0060	Relation to the drug may be possible
Gastro-oesophageal reflux disease	1.7% vs 1.6% 1.13 (0.5 to 2.57) No harm p>0.05	5.0% vs 1.9% *2.87 (1.33 to 6.18)* Considerable harm p=0.0070	Unlikely to be related to drug
Dry eye	Did not meet incidence criteria No harm	2.9% vs 0.8% *3.78 (1.24 to 11.48)* Minor harm p=0.0190	Unlikely to be related to drug
Suicidal ideation	6.4% vs 3.4% *1.99 (1.21 to 3.26)* Minor harm p=0.0065	9.3% vs 6.4% 1.55 (0.98 to 2.45) No harm p>0.05	Uncertain
Cough	3.7% vs 3.1% 1.22 (0.69 to 2.16) No harm p>0.05	6.0% vs 3.1% *2.09 (1.12 to 3.89)* Minor harm p=0.0207	Unlikely to be related to drug
Dizziness	8.3% vs 4.8% *1.80 (1.18 to 2.74)* Minor harm p=0.0067	9.1% vs 7.0% 1.37 (0.88 to 2.15) No harm p>0.05	Unlikely to be related to drug
Hallucination, visual	2.0% vs 0.9% 2.42 (0.93 to 6.31) No harm p>0.05	2.1% vs 0.2% *10.56 (1.35 to 82.47)* Considerable harm p=0.0246	Relation to the drug may be possible
Hypotension	1.4% vs 0.9% 1.74 (0.63 to 4.78) No harm p>0.05	2.1% vs 0.4% *5.34 (1.17 to 24.39)* Minor harm p=0.0305	Unlikely to be related to drug
Pain in extremity	3.0% vs 1.1% *2.74 (1.21 to 6.19)* Minor harm p=0.0152	3.3% vs 2.7% 1.32 (0.63 to 2.73) No harm p>0.05	Unlikely to be related to drug
Muscle spasms	2.4% vs 0.9% *2.96 (1.17 to 7.52)* Minor harm p=0.0222	2.9% vs 2.3% 1.33 (0.60 to 2.92) No harm p>0.05	Unlikely to be related to drug
Dyspepsia	1.6% vs 1.1% 1.42 (0.57 to 3.54) No harm p>0.05	2.5% vs 0.8% *3.15 (1.02 to 9.76)* No harm p=0.0470	Unlikely to be related to drug

Continued

Table 2 Continued

Adverse event	Protocol 17[10] Incidence* HR G-BA classification Nominal p value (non-bold); adjusted p value (bold)	Protocol 19[11] Incidence* HR G-BA classification Nominal p value (non-bold); adjusted p value (bold)	Medical evaluation†
Syncope	4.0% vs 2.0% *2.11 (1.11 to 4.01)* Minor harm p=0.0225	3.1% vs 2.1% 1.62 (0.73 to 3.60) No harm p>0.05	Unlikely to be related to drug
Decreased appetite	4.4% vs 2.4% *1.9 (1.05 to 3.44)* No harm p=0.0328	2.5% vs 2.5% 1.05 (0.47 to 2.33) No harm p>0.05	Unlikely to be related to drug
Fall	8.3% vs 5.7% *1.54 (1.03 to 2.30)* No harm p=0.0361	7.6% vs 7.0% 1.17 (0.73 to 1.86) No harm p>0.05	Relation to the drug may be possible
Back pain	6.6% vs 4.3% *1.61 (1.02 to 2.55)* No harm p=0.0425	6.4% vs 6.4% 1.07 (0.65 to 1.76) No harm p>0.05	Unlikely to be related to drug

Adverse events with HR (drug/placebo) for which the lower limit of the two-sided 95% CI exceeded 1 are *italicized*. P values that maintained p<0.05 after multiplicity adjustment using the methods of Benjamini and Hochberg,[12] and Mehrotra and Adewale[13] are in **bold**; p values not in **bold** had p>0.05 after multiplicity adjustment. All terms were classified by G-BA criteria[14] and medically evaluated.
*Drug versus placebo.
†Medical evaluation by Clinical Programme lead, M Egan, based on what is known about β-site amyloid precursor protein-cleaving enzyme 1 physiology, preclinical studies, pharmacodynamics and clinical trials.
G-BA, Gemeinsamer Bundesausschuss.

across the two trials. Comparing these results with the medical evaluation of the AEs, two events were unlikely to be related to the drug, whereas they were rated as 'considerable harm'; whereas for four 'minor' or 'no harm' AEs, an association with the drug might have been possible.

We also conducted the mandated *subgroup analyses* for AEs (not shown in table 2). For the first trial, we performed 352 statistical interaction tests (at p<0.05 since the interaction can go in both directions) for events at the SOC and Preferred Term level. By random error, we would expect to see 18 interactions between treatment and subgroups for an event at p<0.05, and 8 were observed. For the second trial, we performed 460 interaction tests. By random error, we would expect to see 23 interactions at p<0.05, and 18 were observed.

How should the results of these analyses be interpreted? Which of the 31 statistical safety observations at p<0.05 in the entire study population can be attributed to random error (type I errors)? Eleven of these events had previously been reported in publications of the trial results,[10 11] but many others listed in table 2 were not consistent with the totality of the evidence (physiology, preclinical studies, pharmacodynamics and previous clinical trials) about verubecestat, and 25 signals were found in only one trial.

To address many uncertainties produced by these mandated safety analyses, we also implemented control of the false discovery rate.[12 13] After adjustment, the number of SOC and Preferred Term events with p<0.05 was 10 (vs the original 31) (table 2). Again, some had been reported in the trial publications (eg, weight loss) but others had not (eg, renal and urinary disorders).

We used Hill's criteria,[15] including previous knowledge about verubecestat to assess which of the AEs with a statistical signal might be plausibly related to the drug. For certain events and classes of events, a relationship with the drug may have been possible (nervous system disorders, psychiatric disorders, skin and subcutaneous tissue disorders, weight decrease, anxiety, urticaria)—especially because they were reported in both trials. However, after adjustment, seven other statistical observations were found in only one of the two trials analysed here, failing the Hill criterion of 'consistency of the association', suggesting they were not likely to be caused by verubecestat. For example, lack of reproducibility seemed to refute gastro-oesophageal reflux disease (considerable harm), syncope, hypotension, dry eyes and cough (all minor harm) confirming medical evaluation based on biology, other research and the totality of the data. Without doing the analysis on two trials simultaneously, which may not be possible in all HTA evaluations (note that four of the five examples in table 1 were based on one trial), this could not have been evaluated.

Furthermore, only 1 of the 812 subgroup analyses (352 in the first trial and 460 in the second trial) of events at SOC and Preferred Term level showed a treatment-by-subgroup interaction with p<0.05 in both trials (the HR for dizziness differed between men and women in

both studies). This general lack of reproducibility might suggest that the appearance of a relationship between treatment and the AE occurrence in a specific subgroup in a single trial is due to chance.

DISCUSSION

Our review of past German HTA evaluations coupled with a case study illustrates how reliance on exploratory analyses can lead to uncertain, yet potentially influential conclusions about drug efficacy and safety. Although we reviewed German evaluations and used G-BA requirements for safety analyses, other HTAs also focus on clinical benefits (or presumed lack thereof) in trial subgroups (eg, England, France, Australia, Canada) or request exploratory analysis of AEs at Preferred Term level (Italy). If confirmatory clinical trials demonstrate efficacy for their overall populations, then efficacy would be expected to be homogeneous in subpopulations (or subgroups) of interest, particularly for the confirmed endpoints with multiplicity control. Sufficient overlap of CIs between a subpopulation or subgroup and the overall population is reasonable support for the efficacy for these patients. What can be of interest from subpopulation/subgroup analyses are findings suggestive of an absence of efficacy in a certain subgroup, but these analyses are vulnerable to type II errors, erroneously suggesting absence of effect because of sample sizes that are too small or because of an excessive number of analyses.[19]

We also demonstrated that extensive data-driven analyses of drug safety can result in an array of findings that may be due to chance. In these extensive safety analyses, the concern is about type I errors, or false signals, both for increased and reduced AE rates. Multiplicity adjustment can mitigate some of these observations but not completely resolve the issue. The aforementioned problems associated with exploratory subpopulation analysis are compounded by the growing number of specific *safety* questions being asked of the data, adding to the uncertainty of any conclusions that might be drawn. In our case study, we used two confirmatory trials, allowing assessment of the consistency of the associations, otherwise it would have been impossible to draw conclusions about the 'reasonable possibility of the AE being potentially related to drug'.[20]

Study limitations

We assessed 422 German HTA evaluations and only illustrated five in detail. Was there selection bias for these five? The German IQWiG, the institute that conducts the analyses for the G-BA, published on the same topic. They reviewed drugs entering the German market following regulatory approval between 2011 and 2017.[6] According to IQWiG, 89 of the 216 drugs had clinical benefit. For 37 of these (42%), IQWiG concluded 'no clinical benefit in the whole approved patient population' but recognised benefit only for a subpopulation. This number is consistent with our assessment (34.7%). The five illustrated examples are typical of how the German HTA process evaluates benefit based on fractions of original trials. Similarly, the approach used for the verubecestat analyses was not arbitrary but was done according to what the German processes prescribe.

Recommendations

HTA agencies evaluate the therapeutic value of a medicinal product. Based on consideration of the analyses we have performed, together with generally recognised criteria for internally valid data analysis, we have several suggestions for improving the quality of such evaluations. First, it should be recognised that a critical part of evidence-based patient care relies on high-quality data from well-conducted clinical trials and thorough regulatory agency review. If, based on earlier clinical studies, an HTA agency concludes that treatments, previously demonstrated to be sufficiently well tolerated and effective in a particular population, should be used by specific patients, it is not necessary to conduct additional analyses of data from a clinical trial of a novel treatment which happens to include that population. Rather, the agency can simply recommend which patient subpopulation should be reimbursed for the cost of the new treatment and which should be reimbursed only for previously reviewed regimens. In this way, an HTA agency would maintain the quality of the evidentiary standard (the data). If necessary, the agency should identify the need for new trials when prespecified analyses from available clinical trials are not sufficient to support inference for a population of interest.

Second, it should be recognised that the quality of evidence from exploratory associations is often insufficient for causal inferences. Any analysis must consider the original design and overall findings of a trial, put subpopulations/subgroups in perspective, and not assess them in isolation. Thus, any exploratory analysis that does not confirm the approved, prespecified analyses of clinical trial data is correspondingly hypothesis generating rather than conclusive.

Third, it should be recognised that the benefit/risk assessment of a new medicine or vaccine is the ongoing responsibility of regulatory agencies, and that it is unlikely that HTA agencies will unveil risks that have not already been considered and thoroughly evaluated by regulators. Their objective to compare treatments for the purpose of reimbursement is not served by these hundreds of exploratory analyses, prone to both false positive and false negative conclusions. For characterisation of a drug's safety profile descriptive statistical methods, supplemented by CI and graphical presentations of patterns of AEs (both within treatment groups and within subgroups) are provided in regulatory submissions for market authorisation.[21] However, if additional exploratory drug safety analyses are required for HTA purposes, we recommend the following modifications to the existing procedures:

1. Rigorously apply Hill's criteria to suggestive safety signals and consider effect size, reproducibility, specificity, temporality, dose–response relationship, biological plausibility, coherence of the observations, additional experimental evidence and analogy with other known associations.
2. Rely on multiplicity adjustment to reduce the false discovery rate, and also on careful medical evaluation

of the totality of evidence—as we illustrated for verubecestat. Multiplicity adjustment is standard practice to strictly control the overall type I error for primary and secondary *efficacy* endpoints in phase III (confirmatory) trials, reducing the probability of at least one false positive finding (ie, concluding there is efficacy when it does not exist) across these endpoints (eg, $\alpha \leq 5\%$). However, these methods are not suitable for *safety* endpoints since they severely limit the ability to detect a true signal: for Bonferroni's method the nominal p value for each AE would be compared with α divided by the total number of evaluated AEs, meaning that the p value required to identify a safety signal for a between-group difference becomes increasingly small, therefore increasingly difficult to meet, with each AE tested. However, Benjamini and Hochberg,[12] and Mehrotra and Adewale[13] developed methods to control the false discovery rate for clinical data. These methods reduce spurious discoveries while preserving power to detect true differences. It should be noted, however, that the associations these methods identify from post hoc analyses are still exploratory and should be viewed cautiously.

3. We recommend harmonising evaluation standards of regulatory and HTA agencies. HTA agencies and the European regulatory agencies should build a better understanding of each other's methods of assessment and work together to collectively determine which data they want collected and how these data should be analysed, *before* a trial is initiated. Parallel consultations between the European Medicines Agency and the new European Union (EU) approach to HTAs based on Regulation (EU) 2021/2282[22] demonstrate the potential of such an approach, as long as the new EU framework does not default to current HTA reliance on exploratory analyses of robust data. Rather than issuing a standard list of factors to analyse data against and mandating identification of 'relevant' subpopulations and subgroups *after* data read-out, as currently done, HTA agencies should outline questions requiring subpopulation/subgroup analyses *before* a trial is undertaken, so that treatment sponsors can provide suggestions for the approach to subpopulations/subgroups and the structure of such analyses. In this way, questions posed by HTA agencies, along with those from regulatory agencies, can be included in the design and statistical analysis plans of confirmatory trials.

Transparency

Both regulatory and HTA agencies release their findings to the public. HTA agencies in different countries disclose assessments to varying degrees, from manufacturer submissions and evaluations (eg, NICE in the UK), to entire statistical analyses packages (eg, the Australian PBAC and Canadian CADTH), or even a dossier of all original and exploratory analyses (eg, the German agency). Disclosure is laudable, but can, without clarification of methodological shortcomings and context (as we described here), lead to confusion and diminish transparency. Decisions to recommend financing, based on multiple exploratory analyses, can lead to the appearance that different conclusions have been reached by different agencies. Although this may reflect variation in healthcare decision choices, we suggest that lack of scientific robustness or statistical credibility of the analyses should also be considered as an explanation. As we noted, the overt reliance on exploratory analyses to draw firm conclusions about drug benefits often appears to refute regulatory assessments. When potentially spurious signals from data-driven safety analyses are reported by HTA agencies, and these do not appear to have been considered by regulators, a gap of trust may be created. It may be particularly difficult for healthcare professionals and patients to make sense of information that conflicts with safety information in product labels. HTA agencies can take steps towards resolving this concern by describing how findings were reached, which methods were used, how methodological limitations were considered, and how and why their conclusions might differ from those of regulatory agencies.

SUMMARY/CONCLUSION

In this paper, we illustrated that an over-reliance on subgroup analyses to discern added therapeutic value in a fraction of a clinical trial population is problematic: the resulting exploratory efficacy analyses of the data sets are prone to loss of precision and power, and they are more vulnerable to the uncontrolled influence of random imbalances. Therefore, they may miss important treatment benefits (false negatives). We further demonstrated that a large volume of exploratory analyses of safety data is likely to result in numerous spurious findings that inappropriately implicate treatment as a cause of AEs (false positives). Such findings can unduly and inappropriately influence patient perception and clinical practice, especially when associated methodological limitations on the analyses are not disclosed. When unsubstantiated findings from HTA-mandated analyses are posted in the public domain without qualification indicating the exploratory nature of the analysis and the subsequent unreliability of potential conclusions, healthcare professionals could potentially be misinformed. In summary, the numerous post hoc analyses of clinical trial data required by HTA agencies to assess added benefit of a product for pricing and reimbursement have numerous internal validity issues. This suggests a need for review of current approaches to HTA.

Acknowledgements Medical writing support was provided by Edward A O'Neill. The analysis of the G-BA evaluations between 2015 and 2020 was performed by Pharm-Analytics, Hamburg, Germany. Analyses of verubecestat AEs were performed by Hal Li with programming support by Yong Zhu, Xingji Han and Rinki Jajoo. We thank Michael F Egan for medical evaluation of the verubecestat AEs. We thank Lara Wolfson, David Strutton and Virginia Acha for their advice in developing this paper. Jennifer Rotonda and Michele McColgan provided administrative support. Edward A O'Neill, Hal Li, Yong Zhu, Xingji Han, Rinki Jajoo, Michael F Egan, Lara Wolfson, David Strutton, Virginia Acha, Jennifer Rotonda and Michele McColgan are employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Rahway, New Jersey, USA.

Contributors BJO, ITA, WM, WWBW and KMK contributed to the conception, design or planning of the study. BJO acquired the data. BJO, ESS, WWBW and KMK

analysed the data. BJO, ITA, ESS, WM, WWBW, KMK, GGK and FWR interpreted the results. BJO, ITA, ESS, WM, WWBW and KMK drafted the manuscript. BJO, ITA, ESS, WM, WWBW, KMK, GGK and FWR critically reviewed or revised the manuscript for important intellectual content. All authors provided final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. BJO acts as guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests BJO, ESS, WM, WWBW and KMK are current employees of Merck Sharp & Dohme LLC. a subsidiary of Merck & Co, Rahway, New Jersey, USA and may own stock/stock options in Merck & Co. GGK is the Principal Investigator of a collaborative biostatistics grant from Merck. He is also the Principal Investigator for biostatistics grants from other biopharmaceutical sponsors that have no relationship to the submitted work. FWR has grants/contracts from AstraZeneca and BMS, and consulting relationships with Merck KGaA, Frazier Life Sciences and Janssen; all have no relationship to the submitted work. ITA has no disclosures.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The data sharing policy, including restrictions, of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Rahway, New Jersey, USA is available at http://engagezone.msd.com/ds_documentation.php. Requests for access to the clinical study data can be submitted through the Engage Zone site or via email to dataaccess@merck.com.

ORCID iDs
Björn J Oddens http://orcid.org/0000-0001-5393-2555
Israel T Agaku http://orcid.org/0000-0002-5116-2961
Ellen S Snyder http://orcid.org/0000-0002-8097-2621
Frank W Rockhold http://orcid.org/0000-0003-3732-4765

REFERENCES

1 Friedman LM, DeMets DL, Furberg CD, *et al*. *Fundamentals of clinical trials*. 5th Edition. Springer International Publishing, 2015.
2 Fleming TR. Clinical trials: discerning hype from substance. *Ann Intern Med* 2010;153:400–6.
3 Gemeinsamer Bundesausschuss. Anlage II.6: Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen. Available: www.g-ba.de/downloads/17-98-4825/2019-02-21_Anl2_6_Modul4.pdf [Accessed Jun 2021].
4 NICE health technology evaluations: the manual. Available: https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741 [Accessed Mar 2022].
5 Choices in methods for economic evaluation – HAS. Available: https://www.has-sante.fr/jcms/p_3216041/en/methodological-guidance-2020-choices-in-methods-for-economic-evaluation [Accessed Mar 2022].
6 Wieseler B, McGauran N, Kaiser T. New drugs: where did we go wrong and what can we do better? *BMJ* 2019;366:l4340.
7 Paget M-A, Chuang-Stein C, Fletcher C, *et al*. Subgroup analyses of clinical effectiveness to support health technology assessments. *Pharm Stat* 2011;10:532–8.
8 Domanda di rimborsabilit e prezzo. Available: https://www.aifa.gov.it/domanda-rimborsabilita-e-prezzo [Accessed Mar 2022].
9 apmhealthEurope. Germany's G-BA to introduce new template for assessment files with 'clearer specifications' from 1 April 2020. Available: www.apmhealtheurope.com/story/17918/64797/germany-s-g-ba-tointroduce-new-template-for-assessment-files-with-clearer-specifications-from-1-april-2020 [Accessed Jun 2021].
10 Egan MF, Kost J, Tariot PN, *et al*. Randomized trial of Verubecestat for mild-to-moderate Alzheimer's disease. *N Engl J Med* 2018;378:1691–703. Clinicaltrials.gov: NCT01739348.
11 Egan MF, Kost J, Voss T, *et al*. Randomized trial of Verubecestat for prodromal Alzheimer's disease. *N Engl J Med* 2019;380:1408–20. Clinicaltrials.gov: NCT01953601.
12 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 1995;57:289–300.
13 Mehrotra DV, Adewale AJ. Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. *Stat Med* 2012;31:1918–30.
14 IQWiG, general methods, version 6.0 of 5 Novemebr 202. Available: https://www.iqwig.de/methoden/general-methods_version-6-0.pdf?rev=194070 [Accessed Sep 2021].
15 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295–300.
16 Dmitrienko A, D'Agostino RB. Multiplicity considerations in clinical trials. *N Engl J Med* 2018;378:2115–22.
17 Brookes ST, Whitely E, Egger M, *et al*. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004;57:229–36.
18 Rosenberger WF, Uschner D, Wang Y. Randomization: the forgotten component of the randomized clinical trial. *Stat Med* 2019;38:1–12.
19 Koch GG, Schwartz TA. An overview of statistical planning to address subgroups in confirmatory clinical trials. *J Biopharm Stat* 2014;24:72–93.
20 Efficacy Guidlines ICH. E2A - Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. Available: https://www.ich.org/page/efficacy-guidelines [Accessed Apr 2021].
21 ICH Efficacy Guidlines, E9 - Statistical Principlaes for Clinical Trials. Available: https://www.ich.org/page/efficacy-guidelines [Accessed Apr 2021].
22 Regulation (EU) 2021/2282 on health technology assessment and amending directive 2011/24/EU. Available: https://www.europeansources.info/record/proposal-for-a-regulation-on-health-technology-assessment-and-amending-directive-2011-24-eu/#:~:text=Regulation%20(EU)%202021%2F2282%20%2D%20adopted%20by%20the%20European,a%20text%20with%20EEA%20relevance [Accessed Mar 2022].
23 Gemeinsamer Bundesausschuss. Benefit assessment method for the active substance pembrolizumab (new field of application: squamous cell carcinoma head and neck area, PD-L1 expression ≥1%, first-line, monotherapy). Available: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/512/ [Accessed Jun 2021].
24 Nutzenbewertungsverfahren zum Wirkstoff Axitinib (Nierenzellkarzinom). Available: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/283
25 Gemeinsamer Bundesausschuss. Benefit assessment methods for the active substance tiotropium/olodaterole (COPD). Available: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/183/ [Accessed Jun 2021].
26 Gemeinsamer Bundesausschuss. Benefit assessment methods for the active substance sarilumab (rheumatoid arthritis). Available: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/305/ [Accessed Jun 2021].
27 Gemeinsamer Bundesausschuss. Benefit assessment procedure for the active substance alirocumab (renewed benefit assessment § 14: hypercholesterolemia or mixed dyslipidemia). Available: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/407/ [Accessed Jun 2021].