Online Supplemental Material 1

Definition of comparisons, trial designs and appraisal outcome

For this review, we have categorized all identified comparisons as follows (see table 2 for detailed description). First, we distinguish between quantitative and qualitative comparisons. Qualitative comparisons describe those instances where a "satisfactory method" (in the following: comparator) was described as an adjunct treatment to the investigational product, or alternatively, where it was shown that there was no complete overlap in indications between comparator and investigational product. Quantitative comparisons, on the other hand, were categorized into direct and indirect comparisons. All indirect comparisons were further sub-categorized into three types. The methodologically most simple type is the SBS comparison, also called naïve comparison, where treatment effect data on the same outcome variable across two or more independent trials are extracted for both the investigational product and the comparator. The difference in summary statistics between the treatment of interest and the comparator (e.g. difference between objective response rates from the respective trials) is then evaluated without any adjustment or quantifying the comparison's uncertainty (e.g. by displaying a confidence interval). In contrast, all other indirect comparison methods, that used a formal hypothesis test and quantified the uncertainty of the estimated effect, were termed "inferential indirect comparisons" in analogy to the formal statistical inference they facilitate. The outlined categorization was chosen to fit all identified comparisons, which is why qualitative comparisons were recorded, even though they were not the focus of this review.

The terms "main trial design" and "comparator trial design" used in this review describe the types of studies that were used as a basis for the comparisons, i.e., from which the data were extracted to perform the comparison between the investigational product and the approved product. The different trial designs were categorized as such for the purpose of this review:

- randomized controlled trial: all trials with multiple trial arms to which patients were randomly allocated;
- non-randomized trial: all trials with multiple trials arms, but non-randomized treatment allocation;
- single-arm trial (SAT): trials with a single (active) treatment arm;
- observational study: non-interventional studies that were not based solely on registry data;
- registry study: non-interventional studies specifically based on registry data;
- none: this label was used for all those qualitative comparisons which did not depend on trial data;
- multiple: this label was used for all aggregate data cited from multiple sources of literature;
- meta-analysis: the underlying design was categorized as such if the used data were pooled estimates extracted from meta analyses.

The COMP's appraisal was categorized as follows: a comparison could either be accepted, rejected, or not considered. The latter means that the comparison was presented to the COMP as part of the applicant's submitted documents, but no comment was made in the assessment report regarding the COMP evaluation of this comparison. Rejected comparisons were further categorized into the COMP's specific evaluation of the clinical significance and the methodological soundness, respectively, if this could be discerned from the assessment report. Accordingly, a comparison could be categorized as 'rejected' based on either lacking clinical significance or methodological soundness alone, or because of a lack of both. Further, if this was not specified clearly in the assessment report, the rejected comparison was categorized as 'rejected unclear', in other words based on a global assessment. Lastly, we recorded cases as 'unclear' where multiple comparisons were presented between the investigational product and the comparator, but it could not be discerned which of the comparisons were considered relevant for the positive COMP decision.