EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

20 September 2018
EMA/CHMP/579441/2018
Human Medicines Research and Development Support Division

Meeting Report:
Workshop on the draft reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development (EMA/CHMP/138502/2017)

3-4 May 2018, European Medicines Agency, London

30 Churchill Place ● Canary Wharf ● London E14 5EU ● United Kingdom
Telephone +44 (0)20 3660 6000 **Facsimile** +44 (0)20
Send a question via our website www.ema.europa.eu/contact

Table of contents

1. Introduction and background

The European Medicines Agency has published a draft Reflection Paper (RP) on statistical methodology for the comparative assessment of quality attributes in drug development (EMA/CHMP/138502/2017), for a 1-year public consultation until 31 March 2018.

Further to the external consultation, the Agency held a multi-disciplinary scientific workshop with representatives from interested stakeholders on 3-4 May 2018.

The main objective of the workshop was to discuss comments received during the public consultation phase and aspects related to methodological approaches and practical challenges for comparisons at quality level. This should lead to a better understanding of how statistical methods could aid to further progress in the area of data comparisons at quality level, but should also identify challenges related to the application of various suggested analysis techniques.

The workshop touched upon quality, manufacturing, statistics and methodology areas. Interested stakeholders were invited to nominate experts with experience and/or sufficient knowledge in previously mentioned fields, who were familiar with the content of the draft RP and had ideally contributed to comments during the public consultation phase.

An organising committee consisting of representatives from the Biostatistics Working Party (BSWP), the Biologics Working Party (BWP), the Biosimilar Medicinal Products Working Party (BMWP), Quality Working Party (QWP) and the Agency was formed and workshop participants were selected based on the above-mentioned criteria to ensure a fruitful and balanced scientific discussion.

The following topics of interest were identified based on the comments received during the external consultation:

- Can statistical inference support evidence of similar quality?
- Underlying assumptions presented in the EMA reflection paper
- Statistical equivalence of means and alternatives
- The importance of establishing similarity at quality level for the 'totality of evidence'
- Implications of the EMA reflection paper on small molecules
- Operating performance of frequently used similarity criteria
- New strategies and promising alternative methodological approaches for the comparative analysis of quality attributes data

2. Workshop format and conduct

Workshop participants were invited to submit presentation proposals based on the previously mentioned topic suggestions. The workshop was structured as follows:

Session A: Problem statements and challenges

Session B: Case studies with focus on pre- and post-manufacturing changes

Session C: Case studies with focus on Biosimilars

Session D: Closed session for regulators only

Session E: Operating characteristics of currently/frequently used similarity criteria

Session F: New strategies and alternative methodological approaches

Session G: Closed session for regulators only

The closed regulators sessions were used to reflect upon the discussions and to explore how the raised points could be taken into account in the finalisation of the reflection paper.

Each session was comprised of 4 presentations (20 minutes each) followed by a 30-minute discussion among participants. A detailed agenda[1] was published on the Agency's dedicated website[2].

The workshop was attended in person by 31 industry participants representing 7 industry associations and 32 regulators from national competent authorities and from the Agency (see list of participants[3]). Additionally, the workshop was followed remotely by 24 assessors from European national competent authorities.

This report summarises key points which were discussed during sessions A, B, C, E, and F of the workshop. It should not be understood as a consensus among participants or regulators, but rather as a description of topics raised and discussed. With the speakers' approval, the presentations have been published along with other documents on the EMA events' website[2], and individual links to the presentation slides are included in each session summary in section 3 of this report. A list of identified important issues to be considered for further work on the RP is given in section 4. Participants' comments and feedback on the workshop are summarised in section 5.

3. Session summaries

3.1. Session A: Problem statements and challenges (chaired by Ina Rondak, EMA)

Four presentations on One size doesn't fit all: Why it should be three different guidances (Bruno Boulanger, EFSPI, Arlenda), Medicines for Europe's view on the application of statistical methodology for comparability (Martin Schiestl, Medicines for Europe, Sandoz), Statistical tests, Bayesian analysis, or heuristic rules for demonstration of analytical biosimilarity? (Richard Burdick, AAPS, Elion Labs) and Challenging issues in biosimilar regulatory submission in the United States (Shein-Chung Chow, FDA (USA)) were followed by an audience discussion.

3.1.1. Overall summary of presentations

The presentations in the first session touched upon general concepts and reflected upon whether different rigor is desired for comparability (e.g. throughout the lifecycle of a product) and similarity (e.g. biosimilar vs reference medicinal product) assessments. Assay variability was discussed as an important source of variability along with factors that might need to be taken into account when comparing values from different assays (e.g. time, technology and prior knowledge). It was deemed important to define clear objectives of the comparison tasks and critique was raised on using the mean value as a sole summary characteristic to describe underlying distributions. Clear understanding of fundamental concepts and definitions is vital and better understanding of operating characteristics of methods applied is necessary.

Parallels/arising common views
Same fundamental statistical concepts (i.e. same "toolbox") are applicable in various contexts.

Divergent positions

[1] http://www.ema.europa.eu/docs/en_GB/document_library/Agenda/2018/07/WC500251470.pdf
[2] http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/events/2017/09/event_detail_001507.jsp&mid=WC0b01ac058004d5c3
[3] http://www.ema.europa.eu/docs/en_GB/document_library/Other/2018/07/WC500251471.pdf

It was debated whether the same rigor is necessary for all scenarios, i.e. whether criteria for biosimilar comparisons to reference medicinal products need to be more stringent than criteria for comparability throughout the lifecycle of products.

3.1.2. Summary of panel discussion

Main topics suggested for discussion

- Do same statistical concepts apply to all settings (possibly with tailored implementations for various settings)?
- Understanding of 'inferential' concept in light of consistent manufacturing (i.e. comparison of processes or products)?
- Role of specifications in choice of comparability ranges?
- Same priorities and operating characteristics necessary for various settings of data comparison?
- Relevance of conclusions from analytical similarity for whole biosimilar development?

Main topics discussed

- The concept of 'consistent manufacturing' and the control of shifts and drifts in means during the lifecycle of a product were discussed.
- Clarification was sought on the difference between 'comparison of processes' and 'comparison of products'.
- Discussion on known and unknown factors influencing assay variability and importance of assay validation.
- The use and role of specification limits in similarity assessment was debated.

Contentious issues

Mean shift and drift in reference medicinal product could hamper biosimilar development if different rigor is applied for comparability (i.e. pre- and post-manufacturing changes) and similarity (i.e. biosimilar compared to reference medicinal product) exercises.

Areas of (possible) consensus

A better understanding of operating characteristics is necessary and the mean might probably not be adequate as summary characteristic for underlying distribution in similarity exercises.

3.2. Session B: Case studies with focus on pre- and post-manufacturing changes (chaired by Martijn van der Plas, BWP, CBG-MEB (NL))

Four presentations on Considerations on biological APIs extracted from material of human and animal origin (René van Herpen, APIC, Aspen Oss BV), Comparability study performed to support a process change for a marketed biological product (Jochen Felix Kepert, EBE, Roche), Pre- and Post-Manufacturing Change for a Biological Product (Christophe Agut, EBE, Sanofi & Vivien Le Bras, EBE, Merck) and Comparability study to support commercial process change via stability study (Bianca Teodorescu, EBE, UCB) were followed by an audience discussion.

3.2.1. Overall summary of presentations

Stakeholder presentations stressed the inherent variability of biologicals themselves and the associated manufacturing process; although extracted ('biochemical') products may represent a worst case in this respect (see Van Herpen), presentations demonstrate that this applied to all products to some extent. It was flagged that 'consistent manufacturing' has no ICH definition and that this term could be understood differently by different experts (either as 'within certain limits' or as 'random variations around a fixed mean'; this has important methodological implications).

Furthermore, it was brought forward that range-based methods and a stepwise approach (see esp. Kepert, also Agut & Le Bras) are currently the standard approach. This seems a historically grown practice which is perceived as an efficient way to deal with the type of data, especially where criticality of parameters is often difficult to firmly establish (i.e. it is considered more efficient to assess criticality of a limited number of parameters where differences are observed, than assess criticality beforehand, see Kepert). Criteria based on means are often difficult to reconcile with the type of data available (see simulations of Agut & Le Bras). This is consistent with the methodological points brought forward by Stangler in session E.

3.2.2. Summary of panel discussion

The following topics were discussed, among others, while no firm consensus emerged on these issues:

- It was debated how to reconcile stepwise approaches that allow justifications afterwards with robust methodology which dictates that criteria are defined *a priori*.
- The importance and practicality of sampling in the pre- and post-manufacturing setting was discussed. “How to sample?”, “Should all batches be included, if not, how to select suitable batches (e.g. last 30 batches; batches from last 3-5 years)?”, “Is sampling especially important for extended characterisation?” were questions of interest in the discussion.
- It was pointed out that some critical quality attributes are more (inherently) variable than others and that physicochemical understanding of data is of importance for the choice of ‘meaningful’ characteristic (e.g. range or mean) utilised for comparative data analyses.

3.3. Session C: Case studies with focus on Biosimilars (chaired by Niklas Ekman, BMWP/BWP, Fimea (FI))

Four presentations on Case studies on statistical tools used for comparability assessment (HyungKi Park, Medicines for Europe, Samsung Bioepis), Practical considerations in the statistical evaluation of biosimilarity — a laboratory perspective (Henriette Kuehne, AAPS, Coherus BioSciences), Analytical similarity (José G. Ramírez, EBE, Amgen) and Comparison of Two Groups of Stability Data (Franz Innerbichler, EBE, Novartis) were followed by an audience discussion.

3.3.1. Overall summary of presentations

The presentations of session C gave an overview of issues encountered in the assessment of analytical similarity and also tried to discuss some solutions. The topics included considerations on the statistical tools available (both: ranging approach and equivalence testing) and the impact of reference medicinal product drifts and shifts on the applicability of the tools, issues related to sampling of the reference medicinal product, the sources of variability and uncertainty related to product knowledge (both for the reference medicinal product and the biosimilar candidate), the challenges related to conflicting results from interlinked quality attributes, as well as the possibility to use statistical tools for comparing stability data.

Arising common views in presentations

A common theme throughout the presentations was that equivalence testing of means is problematic, especially in the cases of quality attribute shifts or drifts. The example published by Kim et al (2017)[4] for trastuzumab was highlighted several times; in addition, the in-vitro potency of etanercept was shown as a case where a (clinically irrelevant) drift has taken place.

[4] Kim, S., Song, J., Park, S., Ham, S., Paek, K., Kang, M., Chae, Y., Seo, H., Kim, H.-C., Flores, M., 2017. *Drifts in ADCC-related quality attributes of Herceptin®: Impact on development of a trastuzumab biosimilar*. mAbs 9, 704–714. doi:10.1080/19420862.2017.1305530

Identified divergent positions in presentations

It was debated whether more emphasis should be put on the use of statistical methods for comparison of stability data. Although no firm conclusion could be reached, this was generally not found necessary. All reference medicinal product batches within the specified shelf life, could be considered as 'clinically qualified' and could thereby contribute to the quality target product profile (QTPP) of the biosimilar candidate.

3.3.2. Summary of panel discussion

Main topics suggested for discussion

- When and under which conditions could equivalence approaches provide added value for demonstrating analytical and functional similarity?
- When and under which conditions should a ranging approach be preferred?
- Is there a link between the overall manufacturing process control system and the chosen statistical approach to demonstrate similarity?
- Could the statistical approach used impact on the size of the required clinical programme?
- How to ensure that representative batches of the reference medicinal product are sourced and that the sourcing plan is transparent?

Main topics discussed

- It was argued that flexibility in the statistical approach should be maintained and that underlying test assumptions would need to be fulfilled, irrespective of the chosen statistical approach. A balanced discussion of the pros and cons of different approaches would be worthwhile.
- The term 'descriptive statistics' (used in the current *Similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues* guideline (CHMP/BWP/247713/2012)[5]) should be defined more clearly and the difference between inferential and purely descriptive methods should be clarified.
- It was argued that the statistical test should not be a pass/fail test for similarity. Scientific judgement of the data is always needed. It was suggested that a less stringent approach for demonstration of similarity could be applied if a certain quality attribute is demonstrated or known to be clinically not relevant.
- While the suggested equivalence test compares means of batches, the quality range approach compare values of individual batches. Thus the nature of comparisons is not the same.
- The difficulty of claiming similarity for bimodal distributions using equivalence testing of means was discussed. Equivalence test for intercept might be misleading for comparison of stability data. However, stability/degradation could be seen as a source of variability.
- The link between the overall manufacturing process control system and the demonstration of similarity was pointed out. It was suggested that specifications could be applied to ensure that future batches are aligned with the desired quality profile, at least with regard to the most critical quality attributes. All elements of control strategy should, however, be considered.
- It was argued that the extent of the clinical program should not be dependent on whether an inferential statistical method has or has not been used for determining analytical and functional similarity. Irrespectively of the approach taken to demonstrate similarity, all relevant quality attributes have to be identified and the similarity properly addressed.

[5] http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000886.jsp&mid=WC0b01ac058002956b

3.4. Session E: Operating characteristics of currently/frequently used similarity criteria (chaired by Andreas Brandt, BSWP, BfArM (DE))

After an Introduction (Andreas Brandt, BSWP, BfArM (DE)), three presentations on Operating characteristics of frequently used similarity rules (Florian Klinglmueller, BSWP, AGES (AT)), Performance characteristics of quality range methods and equivalence testing in the comparative assessment of quality attributes (Thomas Stangler, Medicines for Europe, Sandoz), Analytical Similarity Assessment (Shein-Chung Chow, FDA (USA)) were followed by an audience discussion.

3.4.1. Overall summary of presentations

The session focused on operating characteristics of frequently applied approaches to draw conclusions on similarity of test and reference. Frequently used approaches are equivalence testing and reference/quality range based criteria. Most important operating characteristics of a similarity criterion are the probability to correctly conclude on similarity for indeed similar products/product versions (power), as well as the probability to falsely conclude similarity for non-similar products/product versions (type 1 error or "false positive" finding). In absence of a generally agreed definition of 'similarity' in quantitative terms, the term 'probability for a similarity decision' is used.

It was highlighted that controlling a parameter (e.g. mean) or controlling a population are two different concepts. Based on the assumption that safety and efficacy within the reference product's variability can be assumed and mean ± 3 standard deviations (SDs) covers well the width of a population, a proposal for defining similarity in quantitative terms (for normally distributed quality attributes) was made: 3 SDs of test population in 3 SDs of the reference population. Frameworks for exploring, comparing and visualizing operating characteristics were introduced, which allow evaluation of the dependency of operating characteristics on the different parameters (or parameter combinations). Operating characteristics for frequently applied decision criteria to conclude similarity were presented (for normally distributed quality attributes), varying the number of test and reference batches, mean difference between test and reference, SD ratio of test and reference, and for a shift in the mean of the reference product. It was pointed out that tolerance intervals and k SD ranges often have undesirable properties: the power decreases with increasing sample size and probability for a conclusion of similarity increases for shifts of reference away from test distribution. Otherwise, equivalence testing of means appears to be the wrong tool to control a population in a population. Finally, comments on (recently withdrawn)[6] FDA draft guidance[7] were presented and discussed, with focus on scientific justification of 1.5 SD_{Ref} equivalence margin for tier 1, appropriateness of tier 2 quality range approach and requirements for number of batches to be included in the similarity exercise.

Arising common views in presentations
Equivalence testing and quality range criteria aim at different questions and are applied based on different underlying similarity definitions (equivalence of a parameter vs. population overlap).

Differentiation between 'descriptive' and 'inferential' statistical methods should not be made in dependency of the applied similarity criterion (equivalence test or quality range) but on the conclusions that are drawn. 'Descriptive' means that only the batches in the similarity exercise shall be described. 'Inferential' means that conclusions for the totality of test and reference product (or underlying production processes) are made (which goes beyond describing the batches in the similarity exercise), considering the batches as samples. In particular, decision criteria based on ranges/intervals can also be considered 'inferential'.

[6] https://www.fda.gov/Drugs/DrugSafety/ucm611398.htm, accessed on 13 August 2018
[7] Draft Guidance for Industry: Statistical Approaches to Evaluate Analytical Similarity

With regard to the FDA approaches outlined in the (recently withdrawn[8]) draft guidance[9], it was debated that not only difference in means but also variability may need to be taken into account. It was also noted that the tier 2 approach may be sensitive to violations of the assumption of same mean and variance for test and reference.

Identified divergent positions in presentations

There were divergent views on which similarity rules have undesirable operating characteristics; in particular with regard to the operating characteristics of k SD based methods.

3.4.2. Summary of panel discussion

Main topics suggested for discussion

Question what 'similarity' is in quantitative terms was proposed for discussion.

Main topics discussed

- The discussion focused on the definition of 'similarity', starting from the proposal given in one of the presentations to define 'similarity' as 3 SDs of test population in 3 SDs of the reference population.
- Other 'population-based' definitions of similarity could be considered, for example based on the quantiles of the distributions, e.g. % of population overlapping.
- If 'similarity' is defined, a k SD quality range could be used for making decisions, calibrating the k for a specific sample size to achieve the desired operating characteristics.
- Bayesian statistics could be a natural way of approaching the problem – the similarity exercise should not aim at making inference about a parameter of the distribution, but estimating the probability that a patient will receive a batch within an appropriate range.
- It was argued that biosimilars are not developed with the aim of having the same variance as the reference for a given quality attribute, and that under this assumption aiming to show similarity of variances seems to be an odd target.

3.5. Session F: New Strategies and alternative methodological approaches (chaired by Florian Klinglmueller, BSWP, AGES (AT))

Four presentations on EMA Case Study: Design of Experiments (Robert Shaw, VE, AstraZeneca), Statistical methodology for biosimilars, comparison of process changes and comparison of dissolution profiles. A perspective from EFSPI (Mike Denham, EFSPI, GlaxoSmithKline), The value of Bayesian statistics for assessing comparability (Timothy Mutsvari, EFSPI, Arlenda) and Establishing, Assessing, and Comparing Quality Attributes from a Small Sample of Development Batches through Full-scale Production (Kimberly Vukovinsky, ISPE, Pfizer) were followed by an audience discussion.

3.5.1. Overall summary of presentations

Several approaches for the comparison of quality attributes between test and reference products were presented. Both approaches that promise improvements on the experimental design level as well as approaches that take effect at the statistical analysis level were discussed.

On the design level strategies were presented that incorporate knowledge about sources of variation (e.g. cyclical drifts in the production process) and thereby reduce variation, improve sensitivity and avoid potential biases of subsequent statistical analyses. In addition, quality-by-design approaches were discussed. Here, especially the problem of deriving specifications from production processes that are still under active development and thereby subject to changes was considered.

[8] https://www.fda.gov/Drugs/DrugSafety/ucm611398.htm, accessed on 13 August 2018
[9] Draft Guidance for Industry: Statistical Approaches to Evaluate Analytical Similarity

On the analysis level, statistical decision rules that address the problem of population based biosimilarity were presented. Population based similarity was loosely defined as the biosimilar's quality attribute distribution being (at least partially) contained within an acceptance range derived from the originator product. The proposed approaches followed a common recipe: first an acceptance range is estimated from the reference samples, for example using a tolerance interval. Second, the scale of the test products quality attribute distribution is estimated, for example using a prediction interval. Similarity is concluded if the latter interval is included in the former. Different implementations of this approach were proposed based on various combinations of frequentist and Bayesian tolerance and prediction intervals.

The potential advantages of Bayesian inference for the conclusion of analytical similarity were highlighted. Frequentist hypothesis testing aims to control the long-term rate of erroneous similarity conclusions, under the assumption that the two products are in fact not biosimilar. Bayesian inference, in contrast, attempts to quantify the probability that the null- (dissimilarity) or alternative hypothesis (similarity) is true conditional on the observed data. This approach is by some considered more intuitive. However, Bayesian approaches require that additional assumptions about the model parameters' (prior) distributions are made.

3.5.2. Summary of panel discussion

Main topics suggested for discussion

- Estimating reference product specifications using tolerance interval methods overestimates the width of the targeted (inter-quantile) range. Consequently, derived range estimates may cover quality attribute values outside the true reference specifications. From a patient/safety perspective, however, a cautious approach would be to deliberately underestimate that range.
- The concept of population based similarity needs to be specified in concise mathematical terms, such that suitable decision criteria (e.g. non-inferiority of lower and non-superiority of upper quantiles of the test products quality attribute distribution) can be developed.
- The ability to transfer conclusions of similarity in terms of quality attributes to safety/efficacy highly depends on the applied concept of similarity and functional form of the quality attribute-safety/efficacy relationship. E.g. test products with quality attributes concentrated at upper/lower specification limits of the reference product could be considered similar on the population level. This could imply an inferior safety/efficacy profile of the test product.
- Bayesian methods offer opportunities in terms of modelling, interpretation and use of historical data. Framing analytical similarity as a Bayesian decision problem appears promising, as it could be tailored to reflect a variety of assumptions. Whether required assumptions can be justified in all/some cases is still an open debate.
- Bayesian approaches for confirmatory decision making do not directly control the type I error rate. Their results depend on the assumptions made about prior distributions. A justification for the plausibility of additional model assumptions and a comprehensive understanding of the frequentist properties, especially the false positive rate, will be pivotal for the acceptability of such approaches.

Main topics discussed

- Notwithstanding issues with extrapolation of safety/efficacy, it was argued that a conclusion of biosimilarity would impose different standards on biosimilars (when deciding analytical similarity based on equivalence testing of means) than on originators (when e.g. deciding pre- and post-manufacturing change comparability which is typically based on specification based criteria). The question whether the same standards should apply to pre- and post-manufacturing change comparability and analytical similarity for biosimilar products was discussed. Nevertheless, since population based similarity concepts also take into account the variability, in cases where the variability in quality attributes measurements is mainly due to assay variability, equivalence testing of means could be more appropriate.

- Critique that Bayesian procedures do not provide control over frequentist error rates was debated. Bayesian approaches that require only few assumptions as well as methods to implement plausibility checks were mentioned. It was recognised, however, that naïve approaches that make minimal assumptions often lead to the same conclusions as corresponding frequentist procedures.

4. Main workshop outcomes

The following items will impact upcoming decision making regarding the revision of the draft reflection paper:

- The majority of comments were supporting the view that different settings as mentioned in the scope of the RP (pre- and post-manufacturing change, biosimilars, special small molecules) need to be kept separated, there is no 'one-size fits all' solution.
- However, in the biosimilar setting, a special issue arises with drifts/shifts in the reference medicinal product (RMP), with an associated risk to impose more stringent similarity criteria for the biosimilar comparison as compared to manufacturing changes during the lifecycle of the RMP.
- Importance to improve the understanding that a differentiation between 'descriptive' and 'inferential' statistical methodology is required.
- Differentiation between 'descriptive' and 'inferential' statistical methods should not be made in dependency of the applied similarity criterion (equivalence test vs quality ranging method using specific intervals) but on the nature of conclusions that are drawn. E.g. methods were discussed where ranging/interval calculations were used to define a similarity criterion to be eventually used in an inferential manner.
- For quality attribute data comparison settings, where a similarity conclusion goes beyond the samples/batches actually analysed, any similarity criterion applied will have specific performance characteristics (or operating characteristics (OCs)), describing its suitability to control the risk for false positive/false negative conclusion on similarity.
- For each specific similarity criterion, these OCs may vary considerably, depending on the actual circumstances of data collection (i.e. number of batches, underlying data distributions, existence of shifts/drifts, differences in variability, etc.). In that sense, it might rarely be possible to generally categorise similarity criteria into 'conservative' or 'liberal', without context of the actual setting where the criterion is planned to be applied.
- It is evident that (best possible) knowledge of OCs for the application of a specific similarity criterion is essential to justify its use in the context at hand.
- Describing operating characteristics requires a quantitative definition of true/false similarity. In the workshop, several proposals for defining true similarity in quantitative terms were presented and discussed, particularly alternatives to defining similarity via equivalence of a parameter, such as population overlap.
- As expected, it was not possible during the workshop to generally restrict the set of adequate statistical methods for quality attribute data comparison. However, the presentations[10] contained information regarding performance deficiencies of frequently used similarity criteria (at least under some plausible underlying data distribution assumptions).
- In consequence, one possible option to progress could be to shift the focus of upcoming regulatory guidance away from the search for 'optimal' statistical decision criteria (proposal of specific statistical tests). Alternatively, focus could be put on improving the common understanding of the importance to have decision criteria in place for which OCs are understood for realistic assumptions regarding underlying data distributions. On the one hand, this would keep flexibility in the choice of methods; on the other hand this would call for frameworks in which different similarity criteria can be compared with respect to their OCs. In the workshop, some proposals for such a framework where presented.
- Against the background of the above-mentioned flexibility in choice of methods, the suitability of new/alternative similarity criteria, e.g. following Bayesian approaches, could be explored.
- The two areas, which are currently mentioned to be out of scope in the draft reflection paper, namely 'criticality assessment of quality attributes' and 'manufacturing process quality control' were repeatedly identified to stand in close relation to the issues discussed during the workshop. In upcoming revision activities, these inter-relations need to be taken into account as far as possible.

[10] http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/events/2017/09/event_detail_001507.jsp&mid=WC0b01ac058004d5c3

- Multiregional cooperation to be envisaged on regulators side, in particular with FDA and options for exchange with other regions shall be explored further.

Based on this report, the workshop outcome was debriefed and presented to the Biostatistics Working Party (BSWP), the Biologics Working Party (BWP), the Biosimilar Medicinal Products Working Party (BMWP), Quality Working Party (QWP) and the Committee for Medicinal Products for Human Use (CHMP) for information. CHMP supported the idea that a multi-disciplinary group representing these working parties is composed to revise and finalise the reflection paper, also taking into consideration the comments received during the public consultation phase. Given these comments, as well as input from the workshop, further initiatives beyond the revision of the reflection paper could be required in the future, with possible impact on other adopted or currently drafted/revised regulatory guidance documents.

5. Participants' comments received during the workshop

All participants were invited to provide further comments in writing, should they not have had the chance to share them during the discussions. Cards for "take home messages", "questions (?)" or "notes (!)" were available to participants and 61 such comments were handed in at the end of the workshop. An overview of these comments will be provided in the following. They were considered as capturing "general" or "methodological" aspects.

Most comments underlined that further considerations and guidance are needed in this field and that further understanding of performance or operating characteristics of applied criteria will be necessary.

General comments

Clarification was sought on the next steps of the finalisation of the reflection paper and its implications for other guidelines. Besides, a clear definition of terminology would be appreciated. It was suggested that further aspects, e.g. criticality assessment of quality attributes and their relation to clinical relevance as well as the evaluation of stability data, should be included in the reflection paper. The scope of the reflection paper and of the workshop was discussed on the note cards and it was pointed out that the workshop focused more on aspects related to biosimilar developments. Regarding the scope of the reflection paper, it was suggested to consider the different areas of application (i.e. biosimilar development, pre- and post-manufacturing changes and generics) and their methodological and regulatory implications separately from one another. A harmonisation of EU and FDA guidance was deemed desirable. On the organisation of the workshop, earlier interaction with stakeholders would have been appreciated and further interactions are hoped for.

Methodological comments

Participants commended the presentations and further reflected on the properties and applicability of various statistical methods. It was suggested that the patient risk, the quality of the data and the area of application should be taken into account for the choice of statistical method. Comments on the collection of data comprised reflections on the sources of variability and how to take them into account, on continuous data collection and on the potential of including pilot scale data in the comparison exercise. Regarding the statistical methods, a wide variety of reflections on the potential advantages and disadvantages of specific approaches were shared: e.g. graphical description only, equivalence test for means, methods to simultaneously assess means and variances, min-max criterion, Bayesian approaches and out-of-specifications ranges. Further reflections arose on the implications of violations of statistical assumptions and it was suggested to depict such assumptions in the reflection paper. It was furthermore suggested that the reflection paper could be clearer on the preferred approaches whilst at the same time not being too prescriptive on the choice of methods. Approaches on how to assess multiple critical quality attributes simultaneously (e.g. using funnel plots) and to frame the problem of similarity in terms of predictive modelling were proposed. Focusing on biosimilar development, reflections on methodological implications of shifts and drifts in the reference medicinal product (i.e. multimodal distributions) were shared along with reflections on how to deal with

different amounts of information from originator and biosimilar product and with increasingly better controlled processes.

Overall, valuable input was obtained from the feedback cards. The raised points were in great alignment with the comments received during the public consultation phase or with the discussions taking place during the workshop.