EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

26 July 2021
EMA/CHMP/138502/2017
Committee for Medicinal Products for Human Use (CHMP)

Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development

Draft agreed by Biostatistics Working Party	February 2017
Adopted by CHMP for release for consultation	23 March 2017
Start of public consultation	01 April 2017
End of consultation (deadline for comments)	31 March 2018
Agreed by Biostatistics Working Party	June 2021
Adopted by CHMP	22 July 2021

Keywords	***Statistical methodology, comparative assessment, quality attributes, drug development, manufacturing changes, biosimilars, generics, dissolution, inferential statistical methods, similarity assessment***

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

© European Medicines Agency, 2021. Reproduction is authorised provided the source is acknowledged.

Table of contents

Executive Summary

This reflection paper identifies specific areas where the quantitative comparative evaluation of drug product quality characteristics plays an important role from the regulatory perspective. This might involve decision making processes potentially leading to marketing authorisation as well as post-authorisation decisions during drug lifecycle. The document focusses on methodological aspects in relation to statistical data comparison approaches for pre- and post-manufacturing changes, biosimilar development, and generics' development. The reflection paper raises open issues from a statistical perspective, and addresses questions related to comparison objectives, sampling strategies, sources of variability and options (or limitations) for statistical inference.

This document is targeted at both experts from industry and regulatory assessors. This reflection paper complements other available regulatory guidance where comparative data assessment of quality attributes is discussed for certain contexts, but also provides more detailed guidance of how to actually carry out the comparison task based on empirical sample data.

From the methodological perspective, the reflection paper aims to establish a common language and to improve understanding among all experts concerned with quality characteristics' data comparison. It is also intended to trigger further discussion of realistic requirements to demonstrate 'similarity at the quality level' in the different contexts mentioned above. The reflection paper also discusses likely limitations hampering statistical inference, pointing towards meaningful, but expectedly less stringent, alternatives.

1. Introduction

The comparison of empirical data from quality characteristics of drug products (Quality Attributes, QA) is of importance in many areas of drug development. One common objective of such comparative data analyses is to demonstrate that two drug products (or drug substances) stemming from two different manufacturing processes are similar with regard to relevant QAs.

Many of these comparisons are carried out during early drug development, prior to the initiation of clinical trials. The outcome of such comparative investigations has an impact on decisions concerning subsequent development steps. In this early development setting, the adequacy of the (similarity) conclusions drawn from comparative analyses of QAs primarily relates to the producer's risk. Measures to control this risk are in the remit of the drug developer, and are hence not in the focus of regulation.

However, there are also many comparison settings involving comparative QA data analysis with the aim to demonstrate similarity, which have a direct (or sometimes even pivotal) impact on regulatory decision making. This reflection paper is intended to raise awareness that – in these cases – adequate control of the risk for a false positive similarity decision is crucial. From a regulatory decision making perspective, it is the regulator's responsibility to understand the comparison approaches that are applied, in order to adequately inform the decisions for approval, variation, and life cycle management of medicinal products.

Examples, where the comparative evaluation of quality characteristics plays a major role in regulatory decision making are:

- the comparison of a particular drug product in versions pre- and post-manufacturing change, e.g. when it becomes necessary to establish a bridge between clinical trials, where the two versions of the drug product were used exclusively,

- the comparison of a candidate biosimilar product to a reference medicinal product at the quality level, in particular in development programmes with (suggested) abbreviated clinical development parts,
- the comparison of a candidate generic product to the reference medicinal product, e.g. when a clinical (bioequivalence) trial is waived, and pivotal evidence to support similarity comes from comparative dissolution experiments.

In situations where pivotal evidence for similarity comes from comparative statistical analyses of clinical data, there is an established methodological framework in place, based on drawing inferences from samples (patients or healthy subjects). This involves agreed methods to justify equivalence limits for the variables of interest and statistical equivalence testing. Within the clinical data analysis framework, for each specific similarity criterion the (prospective) estimation of the risks to draw a false positive or a false negative decision is common practice. Once those risks are quantified, they serve as operating characteristics (OC) for the similarity criterion specified. OCs are usually used to assist in the choice of the most suitable statistical criterion (statistical test) to demonstrate similarity. However, in situations exemplified above, where decision making primarily relies on comparisons other than clinical data (e.g. analyses of QA data or in-vitro experimental data), regulatory decision making is currently often driven by heuristic approaches. In general, following such approaches precludes the estimation of risks to draw false decisions, i.e. the determination of OCs is precluded as well. This fact currently hampers exploitation of the potential of QAs' data comparisons to serve as a solid basis to draw similarity conclusions.

It has to be acknowledged that the inferential analysis approach applied to clinical data cannot be easily transferred to quality data comparison. Distinct differences exist between the two settings in the nature of the data-generating processes, as well as in the options to control for important factors influencing the data of interest.

Despite these differences, which will be addressed in more detail in subsequent sections, the goal of this paper is to reflect under which circumstances, and to what extent, the implementation of inferential statistical methodology can assist or even facilitate comparative evaluation of QA data. Following a problem description, a two-step approach is introduced: firstly, consideration needs to be given to what would constitute an agreeable 'similarity condition' based on assumed underlying data distributions; subsequently, a suitable similarity criterion needs to be identified and chosen to assess whether the similarity condition truly holds. This choice needs to be informed by exploration of OCs.

After the definition of the legal basis in Section 2, Section 3 touches upon two related topics of comparative QA data analysis, i.e. 'critically assessment' and 'manufacturing process control methodology'. Section 4 presents important fundamental methodological prerequisites which need to be considered when establishing a statistical framework for decision making based on QAs´ data comparisons. Section 5 discusses the options as well as the possible limitations related to the use of inferential statistical methods in the context of various analysis settings, which would typically result from comparison tasks as exemplified above. Section 6 provides guidance on how the various methodological aspects discussed can be reflected in prospective planning, and discusses the advantages of the preparation of a 'QA data comparison protocol'.

2. Legal basis and relevant guidelines

The legal basis and the procedures for making an application for a marketing authorisation are set out in Directive 2001/83/EC as amended and in Regulation (EC) No 726/2004. For generic applications the legal basis can be found in Article 6 of Regulation (EC) No 726/2004 and Article 10 of Directive 2001/83/EC as amended. The legal basis for similar biological medicinal products, also known as

biosimilars, can be found in Article 6 of Regulation (EC) No 726/2004 and Article 10(4) of Directive 2001/83/EC as amended. The Commission Regulation (EC) No 1234/2008, amended by Commission Regulation (EU) No 712/2012, outlines the legal basis of the examination of variations to the terms of marketing authorisations for medicinal products for human use and veterinary medicinal products.

Further information and relevant questions & answers on the eligibility and legal requirements of applications to the Centralised Procedure for generics and biosimilars are available on the pre-authorisation page of the Agency's website.

This reflection paper should be read in conjunction with all other relevant guidelines, especially with the current versions of the following:

- ICH guideline Q5E: Note for guidance on biotechnological/biological products subjected to changes in their manufacturing process (CPMP/ICH/5721/03)
- ICH guideline Q8(R2): Pharmaceutical Development (EMA/CHMP/ICH/167068/2004)
- ICH guideline Q9: Quality Risk Management (EMA/CHMP/ICH/24235/2006)
- ICH guideline Q10: Pharmaceutical quality system (EMA/CHMP/ICH/214732/2007)
- ICH guideline Q11: development and manufacture of drug substances (chemical entities and biotechnological/biological entities) (EMA/CHMP/ICH/425213/2011)
- Guideline on similar biological medicinal products (CHMP/437/04 Rev 1)
- Guideline on the investigation of bioequivalence (CPMP/EWP/QWP/1401/98 Rev. 1/ Corr)
- Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues (EMA/CHMP/BWP/247713/2012)
- Guideline on the pharmacokinetic and clinical evaluation of modified-release dosage forms (EMA/CHMP/EWP/280/96, Rev1)
- Note for guidance on the clinical requirements for locally applied, locally acting products containing known constituents (CPMP/EWP/239/95 final)
- 'Variations guidelines' - Guidelines on the details of the various categories of variations, on the operation of the procedures laid down in Chapters II, IIa, III and IV of Commission Regulation (EC) No 1234/2008 of 24 November 2008 (2013/C 223/01)

3. Definitions and working assumptions

Throughout the text the term ‘drug product’ is used to simplify reading, but it is evident that quality comparisons are also made on either the drug substance or on an intermediate level. The considerations made in this paper equally apply to all cases where the QA data analysis is of high relevance for regulatory decision making.

In this paper, the term ‘quality attribute’ (QA) is meant to describe any kind of physico-chemical characteristic, biological/activity characteristic, immuno-chemical property, purity/impurity characteristic, or any other in-vitro characteristic, which is identified a priori as a (sufficiently) important attribute to serve as candidate for the comparison task at hand. As regards the scale of measurement, the range is from numerically measured QAs (e.g. molecular weight) to qualitatively assessed QAs (e.g. colour). The scale of measurement will usually have an impact on the methodological options for the actual comparative data analysis to be planned.

Reflections made in this document are based on the assumption that a set of QAs has been identified *a priori* which is suitable and comprehensive for the intended similarity evaluation. This means that the selection process of important QAs – often associated to a 'criticality assessment' – is not within the scope of this reflection paper. However, it needs to be noted that this selection process in itself has an impact on the likelihood to eventually conclude on similarity based on QAs' data comparison. Hence, keeping criticality assessment out of scope of this paper should not lead to an interpretation that its importance is neglected. Criticality assessment is discussed in several other guiding documents (see also Section 2) from various perspectives and for different compound classes. In the recent past, plans for QAs' data comparison have been proposed which categorise QAs according to their criticality (e.g. into different 'tiers'), foreseeing different comparative analysis techniques (i.e. similarity criteria) with graded rigour for the categories defined. Of note, such proposals would implicitly suggest that it would indeed be possible to rank different inferential statistical comparison approaches according to their potential to correctly decide upon (dis)similarity. This directly relates to the exploration of OCs, and underlines the importance of the framework introduced in Section 4 of this document.

Performing a comparison at the quality level based on samples taken from two manufacturing processes can have the interpretation that there is interest in drawing conclusions on similarity between the entirety of the material produced by each of the two manufacturing processes. When addressing the 'entirety of material produced' by a certain manufacturing process, the question arises in how far the process is capable of producing material of 'consistent' quality over a certain time span. Whilst process-control methodology as well as release testing are usually foreseen to guarantee targeted quality standards within one specific manufacturing process, an assumption of 'consistent manufacturing' can be an oversimplification in some cases. For example, there might be shifts or drifts in some QAs within the control limits, which might not be irrelevant from the (process) comparison perspective. However, it is important to note that the interpretation of inferential statistical analysis would require some sort of assumption concerning the consistency within the entirety of the material produced, from which samples are drawn and taken for data analysis. This aspect is expected to complicate the implementation of inferential statistical methodology to support a similarity conclusion based on QAs' data comparison. On the other hand, process control and release testing might also be seen as sources of important information potentially contributing to the comparison task at the quality level.

In order to be able to reflect upon the potential of inferential statistical methodology in the context of QAs' data comparison, a simplistic view is followed in the remainder of the reflection paper: whenever it is mentioned that two products are compared, it is assumed that these products can be 'consistently' manufactured, guaranteed by adequate process-control measures. It needs to be kept in mind that the assumption of 'consistency' can be a very strong assumption, which will be hard to verify in many practical situations, in particular with regard to newly established manufacturing processes. It is also important to note that the 'consistency' assumption (i) is generally not in contradiction with a certain within-process variability of QA data, and (ii) should not be seen to conflict with the general goal to strive for 'Continual Improvement of Process Performance and Product Quality' as described in ICH Q10 (Pharmaceutical quality system, EMA/CHMP/ICH/214732/2007). However, changes introduced to improve product quality would be expected to alter some QAs (on purpose), and for the time periods where such changes are introduced, the 'consistency'-assumption might thus not be fulfilled, for reasons well understood.

4. Approaching the quality attributes comparison task from the inferential statistical perspective

Generally, inferential statistics aims at making conclusions about an 'underlying truth' that cannot be completely observed (e.g. the entirety of material produced by a manufacturing process) based on samples (e.g. batches). In contrast, descriptive statistics focuses on describing the observed samples only. Therefore, whether an approach is inferential or descriptive does not primarily depend on the specific method or criterion that is applied, but on the conclusions that are intended to be drawn – i.e., as soon as a conclusion is made that goes beyond the samples at hand, the approach is inferential by definition. In particular, as similarity exercises aim at making conclusions about similarity of products resulting from manufacturing processes in general, and not only about the batches at hand, a similarity exercise is generally considered to be an inferential approach, independent from the applied method.

From the methodological perspective, two main aspects need to be addressed (in sequential order) to enable statistical inference from QA sample data:

(a) Thinking about underlying data distributions, what constitutes an agreeable 'similarity condition'? (details in sections 4.1. and 4.2.)

(b) What is an adequate 'similarity criterion' to assess whether the similarity condition can be assumed to hold? (details in section 4.3.)

With the subsequent sections a framework for comparative statistical inference is introduced based on answers to questions posed under (a) and (b) above.

4.1 Understanding a manufacturing process as a data distribution

Performing a comparison at the quality level based on samples taken from two manufacturing processes can have the interpretation that there is interest in drawing conclusions on similarity between the entirety of the material (which will ever be) produced by each of the two manufacturing processes. A fundamental methodological aspect is that a manufacturing process can be viewed as a data generating process. QAs can be measured on an ongoing basis, e.g. from batch to batch. As a certain variability for the quality of the resulting manufacturing output is usually assumed, the collected data from individual units of observation (e.g. batches) cannot be expected to be all exactly the same for one specific QA all the time. The actually measured QA values over time vary to some extent, resulting in an empirical data distribution for each measured QA. In order to "formalise" the underlying data generation process, one usually searches for options to approximate the empirical (actually seen) distribution with a theoretical distribution, which corresponds to a mathematical function. Commonly, those functions contain one or more distinctive parameters describing the shape of the distribution that underlies the "best" approximation to the data. As an example, the normal distribution with its two parameters to determine the mean and the variance is often used as an approximate distribution. This has a strong mathematical justification, as variability that results from many small disturbances can be proven to behave closely to such a normal distribution. However, many other theoretical functions can serve to approximate empirical distributions.

A consequence of this 'model thinking' via a mathematical distribution function is that it is not the actual sample (e.g. the last three batches produced) itself that is of utmost relevance for decision making on similarity, but the underlying data distribution from which the actual material is 'sampled'. Hence, produced (and sampled) material needs to be understood as 'vehicle' to estimate some characteristics of the underlying distribution. In this context, it is important to understand that any similarity/equivalence claim eventually made after a comparative data analysis always refers to two underlying data distributions, and not to the two sets of samples taken from two processes (e.g. two

sets of batches). Such understanding corresponds to the general idea of statistical inference. In contrast, in order to describe the samples drawn, descriptive statistical methods are used.

The intention to approximate the empirical data distribution for a QA by making use of a specific statistical distribution function requires the simplifying assumption of a consistent manufacturing quality for that QA over time (see also Section 3). As manufacturing changes can alter a QA's empirical data distribution, the underlying assumptions of the data generating process (the model distribution) would also be required to change. If empirical QA data for one drug product (manufacturing) would stem from two (or even more) different distributions over a certain production period, this would usually hamper a straightforward inferential statistical data comparison. In the past, examples were discussed where shifts or drifts in some QAs occurred within control limits which were considered potentially relevant from the (process) comparison perspective. This difficulty with the consistency assumption can complicate the implementation of inferential statistical methodology to support a similarity conclusion based on QAs' data comparison. On the other hand, it should be clear that problems with potential inconsistency cannot be simply overcome by declaring a similarity exercise to be 'descriptive' only, as the intended conclusions are still inferential in nature.

4.2 Similarity condition

Although several criteria to support similarity claims with regard to a QA have been proposed in practice, the underlying question about what actually constitutes the 'similarity' that is intended to be concluded has rarely been addressed in an explicit and quantitative way. Following the understanding that any similarity/equivalence claim eventually made after a comparative data analysis always refers to the two data generating distributions underlying the manufacturing processes, such a claim requires in first place an agreement on a 'similarity condition', i.e. a concise description for when two data distributions allow a conclusion of 'similarity'. It is important to differentiate the 'similarity condition' from the 'similarity criterion': The former corresponds to a consensus between stakeholders involved in regulatory decision making (applicants as well as regulatory agencies' experts). The latter is a concrete instruction for how empirical data should be analysed to check whether the *a priori* agreed similarity condition can be assumed to hold or not.

To specify the similarity condition, considerations related to the maximum allowed difference between the two underlying data distributions for the specific QA are of interest. The question needs to be answered as to which differences would still be compliant with a statement that the material from the two processes can be considered sufficiently similar. Ideally, this should be guided by an understanding of what differences in a QA could have an impact on clinical efficacy or safety. In practice, however, such knowledge is usually very limited.

Another important aspect is the question of whether any deviations between two underlying distributions in a certain QA would be equally relevant in both directions, i.e. the question could arise of whether it would be sufficient to rule out marked differences in one direction only (e.g. rule out increase in impurity, or decrease in potency), or if the goal is to protect against differences in either direction. This question is not only related to the nature of the QA itself, but also to the regulatory context of the comparison task (i.e. it might be different in the pre/post-manufacturing change setting as compared to the biosimilarity setting, see Section 5).

For a particular QA, the type of underlying distribution will also depend on the scale of measurement. The spectrum of the scales of measurement ranges from nominal (e.g. colour of a liquid substance) to continuous (e.g. measured binding capacity).

In the following, the principle of the 'similarity condition' is exemplified assuming interest lies in a QA with an underlying continuous measurement scale. In the example below, the blue model function

approximates the QAs' data distribution for the reference product (R), and the green function depicts the model distribution for the test product (T).

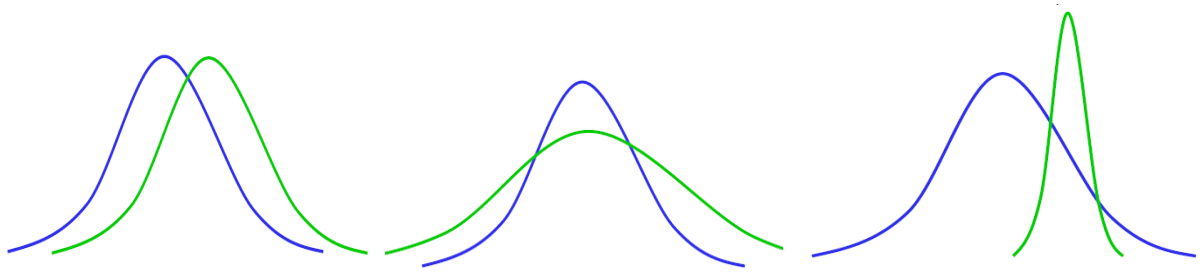

Figure 1 *Figure 2* *Figure 3*

Distributions can be different regarding location (Figure 1), spread (Figure 2) or combinations thereof (Figure 3). As 'similarity' is context-dependent, no universally applicable/agreeable similarity condition exists. However, for the comparability task for a specific QA, consensus regarding the 'similarity condition' is required. As distributions may be characterized by parameters, similarity of two distributions could be defined via similarity of parameter(s) – such as the mean – which determines the centre of a distribution. If similarity in means was the only requirement to conclude on similarity in general, an agreeable similarity condition in such a case could be described via Figure 2. Alternatively, following the idea that the entirety of the material produced by the reference manufacturing process (within given limits) represents acceptable quality, similarity could be defined based on the overlap of the test distribution with a reference range defined based on the underlying distribution for the reference product. Following this understanding of similarity, the distributions in Figure 3 would support a similarity conclusion, in contrast to the ones in Figure 1 and 2. These possibilities are only two intuitive examples to define the similarity condition. Many other possibilities exist. Furthermore, it is important to mention that one and the same similarity condition might not be equally meaningful for the whole range of QAs within one specific comparison task. Of note, the principle exemplified above can also be applied to other scales of measurement, although other statistical methods may be needed for evaluation.

In conclusion, it needs to be noted that for most comparisons of QAs there is no general agreement yet regarding what constitutes an agreeable similarity condition based on the underlying distributions. However, as long as this question remains open, any subsequent discussions regarding the adequacy of a certain similarity criterion (see Section 4.3) aiming to support a similarity claim based on samples falls short. In particular, operating characteristics of a similarity criterion such as the probability of correctly/falsely concluding similarity cannot be quantified when there is a lack of consensus and pre-specification of the similarity condition. Hence, the selection of the applied 'similarity criterion' (see Section 4.3) needs to be preceded by the definition of the 'similarity condition' at all times.

4.3 Using a similarity criterion to investigate if the pre-determined similarity condition holds for a single QA

Several different similarity criteria have been applied in the past to support similarity claims. In many instances however, when the suitability of those criteria was discussed in the context of regulatory decision making, a preceding discussion to reach a common understanding of what constitutes the condition of similarity was lacking. However, as mentioned above, a clear idea regarding the similarity condition for a specific QA in a given context is essential before a similarity criterion is applied based on observed (sample) data.

No similarity criterion can be generally considered 'right' or 'wrong', because no criterion will lead to a correct decision in all situations. In particular, as the underlying truth is not observable, it will never be

known whether a specific decision on the fulfilment of a similarity condition based on a similarity criterion is right or wrong. Nevertheless, the operating characteristics of similarity criteria regarding their capability to correctly decide on a pre-determined similarity condition can be characterized under certain assumptions. The next sections introduce and discuss those assumptions.

4.3.1 Sampling / Experimental Approach

The application of inferential statistical methods and the interpretation of their results require that samples of units taken for analysis are representative of the underlying data generating process(es). One selection strategy to realise the goal of representativeness would be random sampling. However, a random sampling approach might not be possible in practice, nor optimal given the nature of the manufacturing processes. A typical situation is the availability of a (limited) number of production batches, often produced consecutively. In such a scenario, the question about 'representativeness' of available batch material is clearly dependent on (i) the fulfilment of the assumption of a 'well-controlled consistency' in the manufacturing process(es) per se, and (ii) the available knowledge concerning sources of variability.

For an actual sampling plan, this knowledge needs to be taken into account to define the 'unit of observation', to also avoid repeated sampling of units carrying no further relevant information for the comparative analysis. One commonly used approach is to see the production batch as the unit of observation, which can be used for data analysis (one data point per batch). Nonetheless, it is important to strive for a thorough understanding of the sources which can cause variability in the actually measured values of the QAs of interest. One meaningful way to categorise sources of variability is to identify the level on which a certain factor will cause variation in the measured QAs. It is important to note that the actual spectrum of variability-generating factors can substantially differ between small/synthesised and biological molecules. With synthesised products, manufacturing in itself may be less prone to introduce much variability in comparison to production of biological products. Variability seen in QAs of small molecules is more likely to be attributable to variations in assays and measurement systems.

Overall, one/some of the following sources of variability may turn out to be relevant (non-exhaustive list):

- sources causing between-batch variability, e.g.
    - location of manufacturing (batch source),
    - scale of manufacture,
    - age of the batch (time since manufacturing),
    - source of starting materials;
- sources causing within-batch variability, e.g.
    - circadian effects,
    - time since batch-manufacturing start;
- sources causing within-sample variability, e.g.
    - use of different assays to measure one and the same QA,
    - ill-defined or variable sample preparation/storage;
- sources causing within-assay variability: e.g.
    - measurement error related to assay accuracy and assay precision.

It is generally not possible to identify all factors which cause variability in measurements. However, sufficient understanding of the potential primary sources of variability in the data is key to decide upon

what is actually taken as the unit of observation. The definition of the unit of observation is a necessary pre-requisite for sampling considerations.

There might be situations where the comparison task at the quality level can be approached following a prospective (experimental) strategy, allowing for a priori considerations regarding adequate sampling. This may include strategies for stratified sampling, representing the deliberate choice of certain sample units based on the assumption that these are representative for the underlying data generating process.

4.3.2 Revisiting frequently applied similarity criteria

Similarity criteria based on statistical intervals

Similarity criteria are often based on statistical intervals. Most frequently applied similarity criteria require that all test samples (or at least a given percentage) are included in an interval ('reference range') that is estimated based on the reference samples. A less frequent approach is a requirement that a statistical interval estimated based on test samples should be included in the reference range. The general idea behind this approach is to show that the manufacturing process for the test product will result in a product that lies within the 'range of variability' of the reference product, as the reference samples represent acceptable quality ('clinically qualified', see Figure 3 in Section 4.2).

The operating characteristics of such criteria in terms of false positive and false negative regulatory decision making will strongly depend on the number of test samples. Generally, the probability of fulfilling such a criterion increases with decreasing number of test batches. This means that a sufficiently large number of test batches will be required to limit the probability for a false positive conclusion. The probability for a false negative conclusion, on the other hand, will increase with an increasing number of test samples. This antagonism is usually explored in the investigation of operating characteristics of interval-based criteria. Operating characteristics will also depend on how well the reference range captures the true 'range of variability' of the underlying reference distribution, which will usually also depend on the number of reference samples.

Examples of interval-based similarity criteria, which feature the properties mentioned above, are the 'min-max'-approach, the '+/- x-sigma' approach as well as the 'Tolerance Interval' approach.

<u>Min-Max, X-sigma and Tolerance Intervals</u>

In many instances of QA data comparison, the interval determined by the range between the minimum and the maximum data point observed for the reference distribution is used as 'reference range'. Alternatively, x-sigma intervals (i.e. mean +/- x*SD, SD being the standard deviation) are also frequently proposed as reference range. This is often justified by mean +/- x*SD covering a given percentage of the reference distribution (under a normal distribution assumption), e.g. mean+/-3*SD covering 99% of the reference distribution. However, it is important to distinguish between the 'true' mean+/-x*SD range of the underlying distribution and the range estimated based on the sample mean and sample standard deviation. With the latter range it is generally not possible to accurately determine the true range of variability (e.g. the '99%-claim' as mentioned above) due to measurement error and other sources of variability.

As an alternative to x-sigma intervals, tolerance intervals (TI) that take sampling variability into account are sometimes proposed as reference range. A TI is usually derived to estimate a data range by which a specified proportion p (e.g. the central 99%) of the underlying distribution is assumed to be covered with a pre-specified degree of confidence (e.g. 95%). However, although TIs cover the central proportion of the distribution (e.g. the true mean +/- 3*SD range) with high certainty, this does not

imply that the TI is a good (narrow) estimator of this range. In contrary, as TIs take measurement error into account for quantification of uncertainty regarding the estimate of the central proportion of the reference distribution, this does imply that TIs are generally broader when the uncertainty is large (i.e. sample size is small).

Of note, these observations do not automatically disqualify the intervals mentioned for a QA data comparison task. From a pure conceptual perspective, defining a target range with such intervals appears intuitive. However, from the regulatory perspective, the concern for an unreasonably high risk for false positive conclusions on similarity would need to be addressed. Applicants should therefore discuss the operating characteristics of these approaches in regulatory submissions and justify that the risk of a false positive conclusion is acceptably low.

Prediction Intervals

Prediction intervals (PI) are estimated to describe a data range covering data outcome of future units with a pre-specified degree of certainty (% PI). PIs can be derived for one single future observation, for a set of k future observations, but also for a parameter characterising the underlying distribution of future observations, e.g. for the mean of future observations. In the context of comparative QA data analysis, it appears intuitive from a conceptual perspective to use PIs to characterise data distribution for the test-product (or post-manufacturing-change condition), as any similarity claims would be made in relation to future manufacturing. As for all other intervals mentioned so far, the context of use of a PI in the definition of a similarity criterion will strongly influence the operating characteristics, and hence the acceptability of prediction intervals.

Bayesian Credible Intervals

Computation of intervals following Bayesian principles have a fundamentally different interpretation compared to the intervals described above which follow a frequentist statistical approach. Bayesian intervals generally consider unknown distribution parameters (or other quantities of interest) as random. Credible intervals hence can be interpreted as region on a particular data scale, which contains the parameter of interest with a specified probability. The acceptability of Bayesian approaches to inform regulatory decision making remains controversial. In the context of QA data comparison, the application of Bayesian methodology, e.g. to characterise and (continuously) update distribution models of a certain manufacturing process, could be considered a reasonable strategy. However, acceptability of Bayesian approaches implemented in QA data similarity criteria would not only require persuasive operating characteristics, but also an acceptance of the conceptually different interpretation of data analysis in general.

Similarity assessment based on Equivalence testing

Equivalence testing aims to demonstrate that a difference/distance between parameters describing an underlying distribution (e.g. mean) larger than a given margin can be reasonably excluded. Hence, an important question to be addressed upfront is the choice of a suitable characteristic(s) to be compared. In addition, a metric to describe the difference/distance between the parameters for the two distributions needs to be defined. The definition of such a metric relates to the intention to derive one single measure to describe the difference of interest, and thereby to 'simplify' the analysis task. For the example of the comparative analysis of means, this metric could simply be the difference of means or the ratio of means. Equivalence is usually concluded if a confidence interval computed in terms of the given metric for the parameter of choice (e.g. the difference in means) is entirely contained within a pre-specified acceptance range. Thereby, the acceptance range constitutes the (agreed) similarity condition, i.e. two distributions are considered similar if the difference in the parameters of interest for the underlying distributions is smaller than the maximal difference allowed according to the specified acceptance range. The advantage of equivalence testing is that operating characteristics, in particular

the probability for a false positive similarity claim, are well understood under certain assumptions. However, equivalence testing is often criticized for the underlying 'similarity condition' not being compliant with the usual understanding of similarity, which does not require that distribution parameters are similar. In addition, the basis for justifying an acceptance range is often lacking, as the quantitative relationship between a QA and clinical outcomes is usually poorly understood.

The methodological approach of equivalence testing of means exemplifies a setting, where intervals get utilised for two clearly distinct purposes in the comparative data analysis: One purpose is the 'quantification of uncertainty of estimation'. This is done by computation of the confidence interval for the difference in means. The other purpose is the 'definition of an acceptance range'. However, in many comparative analyses other than equivalence testing of means, it turns out to be difficult to keep those two mentioned purposes separate. For example, in the case of computing a TI from reference data, the resulting interval is usually supposed to serve both goals: quantification of uncertainty and – at the same time – definition of an acceptance range. The fact that those two methodological aspects often mixed makes a general judgement of the suitability of certain statistical intervals for the QA data comparison task difficult.

4.3.3 Selecting a similarity criterion from a range of possible candidates

Any similarity criterion based on statistical analysis of QAs' data, which is supposed (or can be expected) to have a substantial impact on regulatory decision making, needs to be assessed for its capability to lead to the right decisions. From the methodological point of view, an *a priori* investigation is required to assess the chances for incorrect decisions associated with the application of a certain similarity criterion. Most of the time the actual outcome of a comparative QA data analysis will lead to a 'yes' (similar) or 'no' (not similar) outcome. Therefore, an understanding of the probability for a false positive decision on similarity (erroneously deciding for similarity of underlying distributions that are not compliant with the pre-specified similarity condition) is central. The same applies for the probability for a false negative decision (erroneously deciding against similarity of underlying distributions that are compliant with the similarity condition). These probabilities generally behave antagonistically, and their magnitudes depend on various factors such as the true underlying distributions, the pre-specified similarity condition and sample size. With a systematic theoretical investigation of how the probability for a similarity decision depends on those factors, the 'operating characteristics' of a suggested similarity criterion can be explored.

A framework could be established for such a systematic investigation of the operating characteristics of candidate similarity criteria. As the true underlying test and reference distributions are unknown, the probability of a similarity conclusion for a given similarity criterion should be investigated for a range of scenarios with different underlying distributions (for example two normal distributions varying the difference in means, and ratios of standard deviations). Influence of sample size (number of test and reference samples/batches) on operating characteristics should be investigated within the framework. The relevance of the chosen framework should be justified, particularly the plausibility of the underlying test and reference distributions and the feasibility of sample sizes.

Such exploration could generally include several candidate similarity criteria to eventually select the optimal criterion, given the expected/feasible setting for sampling and data collection. Similarity criteria may also be 'calibrated' to have the desired operating characteristics.

As a matter of principle, investments in generating more evidence should be rewarded. In other words, the quality of decision making should generally increase with increasing amount of information, e.g. with increasing number of batches. Vice versa, comparison approaches which would make it easier to pass a defined similarity criterion by decreasing the amount of information on which the decision is

based cannot be endorsed. Large uncertainties in the estimation of reference data (distributions) should never manifest in large acceptance ranges.

The computation of some intervals (introduced above) requires further parameterisation, e.g. central proportion coverage with TIs, etc. Such parameterisation should be done and justified at the planning stage, before actual QA sample-data is seen. Criteria with a low potential for misleading outcome interpretation associated with the arbitrary choice of actual values for such parameters would generally be preferred. The variety of comparison approaches may also comprise analyses requiring less (or no) specific *a priori* distributional assumptions, such as non-parametric techniques, bootstrapping or other re-sampling methods ('distribution free' intervals). Furthermore, it is important to note that different concepts for statistical intervals not only differ in their method of computation, but also (and importantly), in the interpretation of the resulting interval.

From the regulatory perspective, the main concern remains the false positive similarity conclusion. Therefore, characterisation of this risk is considered most important and the selection of a similarity criterion needs (also) be driven by minimisation of that risk. Any justification of an acceptable maximal magnitude of this risk for one single QA's data comparison needs to be put in context of the overall risk for a false conclusion on similarity based on the analyses of all QAs' comparisons. See next sub-section for further details.

4.4 Overall conclusion of similarity at the quality level

The considerations outlined in the previous sections describe an outline for inferential evaluation of similarity for a single QA. For many tasks of comparing data on the product-quality level, it is expected that the comparison will involve more than one QA. This would generally mean that all the methodological considerations explained above would need to be applied separately for each QA selected for the comparison task. The read-outs for different QAs are expected to be observed on different scales with varying quality of information, ranging from binary outcome to continuous measurements. Even if the assay read-outs for a set of QAs are all on a metric/continuous scale, the shapes of underlying data distributions can be rather different. In this context, it is generally unreasonable to assume that one and the same comparison approach will be suitable for comparative evaluation of all the QAs involved. In most instances, tailored approaches may be required to reflect the mentioned diversity in QAs.

For the case that adequate statistical frameworks can be identified and applied for the comparison of more than one QA of interest, an *a priori* specified concept ('overall success criterion') is recommended to describe the minimum requirements for a claim of similarity. Such a concept would ideally be put in an analysis plan which is prepared prior to sampling and conduct of the comparison analyses (refer to recommendations in Section 6).

Any post-hoc justifications that observed (unexpectedly large) differences in one or more of the analysed QAs would have no or only minor impact on clinical outcome might eventually be seen to contradict preceding criticality assessment of QAs and/or an adequate definition of the similarity condition. In addition, restricting such a justification to the observed samples/batches is generally flawed when the aim is to make inferential conclusions on an underlying distributions (i.e. it would need to be justified that it can be concluded from the samples at hand that differences in underlying distributions are unimportant).

The overall risk of a false positive conclusion on similarity following an inferential statistical comparison of several QAs will strongly depend on the assessments of operating characteristics that are possible to make for each separate QA data analysis. Whilst the quantification of this overall risk might be difficult, nonetheless it should be addressed already at the planning stage. Currently, only limited guidance can

be given regarding the adequate choice of acceptable probability for a false positive decision for the comparison of QAs' data. The 5%-significance level established in the context of clinical trials can serve as an obvious first threshold for orientation. Generally, considerations concerning the risk of a false positive conclusion of similarity at the quality level become more important, the more this comparison is expected to carry pivotal evidence to support regulatory decision making.

In this context, considerations on the probability for a false negative decision might eventually also become relevant from a planning perspective, as sample size constraints (e.g. low batch numbers) and associated risk for a false negative conclusion may lead to the need to use approaches other than inferential statistical comparison of QAs.

5. Implications for settings where the comparison at the quality level is of particular relevance in regulatory decision-making

This section provides reflections on a number of commonly occurring situations where a comparative evaluation at the quality level is highly relevant.

The various settings described below share the challenge to somehow predefine an appropriate similarity condition. The definition of such an appropriate condition and the assessment of (maximum allowed) truly existing differences in QAs would benefit from a good understanding of the impact such differences could have on clinical outcome. The extent of knowledge regarding this association between differences at the quality level and clinical outcome (efficacy and/or safety aspects) will already drive the criticality assessment of the QAs. However, frequently there is a high degree of uncertainty concerning the impact of certain QAs on the clinical outcome. Hence, a considerable degree of arbitrariness might be unavoidable in practice. Therefore, submissions which use criticality assessments to inform the similarity analysis approach should include a discussion of the impact of QAs on clinical outcome in supporting the choices made.

One frequently occurring limiting factor hampering the application of inferential statistical methodology is that sampling can be non-random. The samples used may therefore not be representative of an entire manufacturing process. In this case, any particular statistical model applied will fail to describe uncertainty in the desired manner, and the corresponding results have no inferential interpretation.

Overall, it is up to the manufacturer to describe which similarity condition is assumed and which similarity criterion is adequate to check this condition. These decisions need to be tailored to each specific QA and the specific comparison (manufacturing change, generic, biosimilar, etc.). Ideally, the chosen similarity criterion is robust and minimises the risk of a false positive decision (i.e. has appropriate operating characteristics). A justification for the chosen similarity criterion should be provided in regulatory submissions, based on the considerations given below and in Section 4 above.

5.1. Biologicals

5.1.1. General issues relating to Comparability exercises

Comparability exercises encompass both comparability after manufacturing changes and biosimilarity (see EMA Guideline on Similar Biological Medicinal Products[1] which confirms that both situations fall under the same scientific umbrella).

[1] CHMP/437/04 Rev 1 as revised

Comparability is not a stringently defined concept. According to ICH Q5E (CPMP/ICH/5721/03), '*comparable*' means that products have highly similar quality attributes before and after manufacturing process changes and that 'no adverse impact on the safety or efficacy, including immunogenicity, of the drug product occurred.' ICH Q5E (CPMP/ICH/5721/03)therefore does not require QAs to be identical or equivalent in a specifically stipulated inferential statistical sense. This flexible wording makes it possible to implement a similarity comparison approach which is most suitable to the specific situation, in line with the reflections offered in Section 4 of this Reflection Paper. This may include, but is not limited to, situations where differences may be acceptable if they represent improvements in safety (e.g. less impurities), or for example, formulation changes (which may induce intended or consequential changes in pH or osmolality), where requiring 'similarity' (including pre-specified similarity criteria) would defy the purpose of the change.

Sampling considerations

Sampling is one of the major differences between a manufacturing change and the biosimilarity situation. However, two common observations can be made, which relate to the impact of the control strategy on the data distribution:

Firstly, shapes of underlying data distribution are often not adequately approximated by normal distribution because of truncation. Empirical data distributions are often truncated by the control strategy, especially for those QAs which are tested at release. In other words, certain (extreme) values can occur but will not be able to reach the final product because of the selection/truncation, even if they would be predicted by formal intervals. It is often assumed that the observed min-max range will converge towards the specification limit interval if a large number of samples is taken. However, this assumption may be questionable in certain cases, especially where the specification limit interval is wide compared to process capability. In such cases an untruncated empirical data distribution may be observed.

Secondly, the number of 'test batches' (either post-manufacturing change or biosimilar batches) is usually relatively low. In this context, it has to be noted that there is no specific minimum number of required batches/units (e.g. 3 batches, as frequently suggested in practice for manufacturing changes) which could guarantee to capture the true underlying variability.

Based on assumed manufacturing consistency, the first 'test batches' could be interpreted as confirmatory that the new process yields 'comparable' batches; i.e. manufacturing consistency implies that the first produced batches (usually the process performance qualification [PPQ] batches) are representative for all future (not yet manufactured) batches. In addition, an adequate control strategy is a major contributor to manufacturing consistency. If all 'test batches' are not included in the comparability exercise, then the issue of sampling deserves special attention in any justification of a plan to utilise inferential statistical methodology for QA data comparison. In addition, it has to be noted that a very low number of available 'test batches' could *per se* represent the limiting factor to carry out a meaningful inferential assessment, e.g. because the desired precision for interval estimation cannot be achieved.

QA comparison framework

As noted above, the similarity condition based on the underlying distributions should drive the choice of statistical methodology for the eventually applied data comparison.

Different statistical approaches may be used for different QAs. The definition of the similarity condition is impacted by both the criticality of a specific QA and the physico-chemical properties described by it. Criticality in itself does not seem appropriate to drive the choice of statistical methodology; the

physico-chemical properties described by the QA (which impact the observed or assumed distribution) and the established or suspected impact on clinical performance should determine the criterion.

This has commonly been supported by using risk assessment tools to rank quality attributes in terms of their impact. It is noted that this approach is well developed in the assessment of biosimilarity; however, the scientific principles apply to all forms of comparability exercises, especially all types of major changes. Regardless of whether such ranking approaches are used, it is important to consider whether the chosen statistical similarity criterion is appropriate for the particular QA. Therefore the operating characteristics and risk of false positive conclusion should be considered and justified for the particular statistical method(s) for each QA. For most of the comparative analyses of QA data, the focus would usually be on two-sided comparative investigations. Exemptions could include potential improvements in specific QAs (e.g. reducing impurities) which might translate to safety advantages.

Statistical equivalence of means is rarely attainable, but may also not be necessary; this criterion seems most suitable for parameters which are not batch-dependent, like e.g. binding constants of non-glycosylated proteins to their receptors. For batch dependent parameters, the reflections with respect to various unequal but potentially acceptable distributions, as provided in Section 4.2, fully apply.

Risks of false conclusions

The question of adequate control of the risk for a false conclusion on similarity is of utmost importance. Hence, the exploration of operating characteristics of similarity criteria applied for QA data constitutes a key conceptual element to quantify the risk for a false conclusion on similarity. Therefore, in regulatory submissions, the operating characteristics of the chosen statistical approach should be outlined, and the risk of false positive and false negative conclusions clearly explained, given the description of the similarity condition, the choice of statistical similarity criterion and the number of batches sampled from the two manufacturing processes of interest.

5.1.2. Specific issues relating to manufacturing changes

The comparative evaluation of the quality of two product versions before and after a certain manufacturing change (including, but not limited to, transfers to other manufacturing sites, formulation changes, scale up, and changes during development) is a very common task during the lifecycle of a medicinal product.

In contrast to the biosimilar setting (see next sub-section), the typical starting point in the pre/post-manufacturing setting is usually based on easy access to knowledge regarding the 'reference' – here the pre-change manufacturing process. Such knowledge usually relates to the whole history of the product's manufacturing, the sensitivity to changes in the production setup in terms of excursions of important QAs, sources of variability when measuring QAs, sensitivity of assays used, etc. This means that available knowledge concerning different causes for within- and between-batch variability can inform the statistical comparison approach.

Sampling considerations

Concerning sampling, the manufacturer usually has access to the data from pre-change batches. All these data, or a well-defined complete subset (e.g. all batches since the last major process change) can be used as the sample set. However, it commonly occurs that only routine release test results are available for all these pre-change batches, and that it is necessary to 'draw' representative samples for extended characterisation. It should be justified in regulatory submissions that the limited number of ‘drawn’ batches is indeed representative.

As noted previously, in many instances only a low number of batches produced consecutively after the manufacturing change are available for the comparison task. In this context, there is no specific minimum number of required batches/units (e.g. 3 batches) which can be assumed to be acceptable for the comparison task without further justification.

5.1.3. Specific issues for Biosimilar setting

QA investigation of the reference medicinal product

In the biosimilar setting, a major challenge is the limited access to information regarding the manufacturing of the reference medicinal product. Hence, many sources of observed variability in the QAs of interest may remain obscure. Therefore, to demonstrate biosimilarity at the quality level, multiple reference product batches must be sampled and analysed to investigate the underlying data distributions of the reference product. Random sampling of reference product batches may be difficult to achieve; however, sampling should be such that obvious biases are avoided and the true range of variability can be estimated with some certainty.

Sources of variability in the reference product data require special attention, and cases have been described in the past where significant shifts or drifts for the reference medicinal product's data distribution have been observed for relevant QAs (e.g. in the extreme case leading to non-overlapping clusters of reference medicinal product batch series). In such cases, the target for biosimilarity assessment might not be easily identifiable without further considerations regarding the reasons for the within-reference medicinal product manufacturing differences. For those QAs which are impacted by clearly identifiable shifts/drifts, in addition to analysing the reference dataset as a whole, it is advisable to conduct a statistical analysis of the pre- and post- shift ranges separately, in order to understand whether the biosimilar QAs are more closely aligned with the pre- or the post-shift profile.

Similarity condition

For each selected QA the Applicant must demonstrate that the biosimilar candidate has "a highly similar" quality profile compared to the reference medicinal product. However, there is no strict definition of what constitutes “highly similar” in a statistical context. Therefore, developers should firstly strive to gain an understanding regarding an agreeable similarity condition (as outlined in Section 4.2) for each QA. This would need to reflect the maximum allowed difference between underlying data distributions which would still result in a safety and efficacy profile comparable to the reference product. It is recognised that identifying the similarity condition may be challenging, as the impact of differences at the quality level on clinical outcome (efficacy/safety/immunogenicity) is often hard to predict or quantify. It is recognised that statistically significant differences may be clinically irrelevant in some cases. From a methodological viewpoint, the optimal approach would be one where failing the similarity criterion in the QA analysis would by definition imply a clinically meaningful difference (i.e. where the allowable difference is established beforehand as part of the similarity condition). In practice, as clinically allowable differences are extremely difficult to establish based on theoretical grounds only, it seems currently unavoidable to use empirical batch data for informing which differences are acceptable. In line with what has been stated above regarding risk assessments and QA classification, these batch data should be combined with available (non-)clinical data about e.g. the mechanism of action, to avoid too stringent similarity criteria for QAs which would have no reasonable correspondence to relevant clinical differences.

In the biosimilar setting, any difference identified in any product characteristic could in principle be interpreted as a potential signal for non-similarity between the reference medicinal product and the biosimilar candidate. Nonetheless, it is recognised that the question in how far dissimilarity in QA data

can be seen compliant with a biosimilarity claim may not be based on the outcome of a single statistical test, but rather taking the entire biosimilar data package as a whole.

5.2. Small molecules

5.2.1 Specific issues for abridged/hybrid applications

Abridged or hybrid marketing authorisation applications for small molecules represent one further arena where data comparison at the quality level, including also on an ex-vivo/in-vitro data, could be of pivotal relevance for regulatory decision making. Locally-applied, locally-acting products represent one example where, under certain circumstances, similarity needs to be explored between a test- and a reference product. Examples are droplet-size comparison for aerosols/inhalation products or comparative assessment of data from permeability assays for transdermal products.

The Note for Guidance on the clinical requirements for locally applied, locally acting products containing known constituents (CPMP/EWP/239/95 final) mentions options to waive therapeutic equivalence trials if other models can be justified to generate sufficient evidence to support an equivalence claim. Similar to that, Appendix II of the CHMP Guideline on the Investigation of Bioequivalence (CPMP/EWP/QWP/1401/98 Rev.1/Corr) describes biowaiver conditions for the development of special pharmaceutical forms (e.g. eye drops, nasal sprays or cutaneous solutions). Here, waiver criteria are based on comparison analysis results involving data from QAs of the test and the reference product. In these documents no further detailed guidance regarding the methodological framework for the actual analysis of equivalence are provided. In lack of such guidance, equivalence criteria agreed to be suitable to compare PK data in the immediate release products' bioequivalence setting (estimating confidence intervals for the ratio of means and comparing to an acceptance range of 80%-125%) are occasionally suggested to support a similarity claim. However, in absence of additional (preceding) considerations what would constitute the similarity condition (see Section 4.2), the evaluation of operating characteristics of such criteria falls short.

Furthermore, the CHMP Guideline on the pharmacokinetic and clinical evaluation of modified release dosage forms (EMA/CHMP/EWP/280/96, Rev1) contains considerations regarding similarity of dissolution profiles regulating waivers and the need for bracketing approaches, but neither include stipulations regarding a broadly agreeable similarity condition nor further recommendations on reasonable approaches for using inferential statistical methodology.

The comparative analysis of dissolution profiles between two (versions of a) medicinal product(s) can be seen as a special case under the scope of this reflection paper. This special case is characterised by the fact that there is only one QA of interest, i.e. dissolution over time. As mentioned in Appendix I of the CHMP Guideline on the Investigation of Bioequivalence (CPMP/EWP/QWP/1401/98 Rev.1/Corr), comparative dissolution investigations are relevant for the justification to waive bioequivalence studies. In this context, the guidance introduces dissolution similarity assessment as 'Bioequivalence surrogate inference', which actually implies that inferential statistical methodology would ideally be applied to e.g. infer a similarity-in-dissolution claim from the tablet sample to the whole tablet population (all tablets ever produced by a given manufacturing process). The guideline recommendation towards a definition of the similarity condition is a 10% limit for the mean dissolution over time. The suggested criterion to investigate of whether this condition holds is the f2 metric, where differences in sample averages are suggested to be used for deriving a distance measure between reference and test compound. Alternative options for dissolution similarity assessment to handle situations where the f2 metric is not considered suitable comprise other model-independent distance metrics as well as model-based investigations of dissolution profile differences. It is interesting to note that, when following such alternative comparison strategies, the assessment of similarity in dissolution may go beyond the sole

evaluation of distribution means. This aspect confirms the observations described in Section 4.2 that common agreement around a broadly accepted similarity condition is an important pre-requisite for similarity investigations based on QAs' data.

As for all cases mentioned above, only limited guidance exists from the methodological point of view, the fundamental requirements as introduced in Section 4 would need to be considered, given an model/experiment identified to support a similarity claim based on empirical QAs' sample data. Some of the aspects described above for the biosimilarity setting to build a statistical framework might also be applicable to the broader field of abridged/hybrid applications. This might however also pertain to the challenges to attribute observed variation in the empirical sample data to potential sources of variability.

6. Quality Attributes data comparison protocol

Against the background of the topics discussed in this reflection paper, the concluding recommendation is given to pre-plan a certain QA data comparison task as far as possible. One reasonable option is to prepare a dedicated planning document for this purpose, e.g. a 'Quality Attributes Data Comparison Protocol'. With such a document, the data comparison task would ideally be contextualized with the whole product development, indicating and weighing the importance of demonstrating that two products (or manufacturing processes) are similar at a quality level. It is important to note that only adequate pre-planning will protect against the potential criticism related to data-driven analyses and biased post hoc decisions. Regardless of the actual (regulatory) context of a certain QA data comparability task, the following guiding principles for planning the content of a comparison protocol can be given:

- The objective as well as the context of the QA data comparison should be clearly stated. Descriptions would ideally include considerations regarding potential consequences for the two possible outcomes of the comparative investigation, namely either that similarity could be demonstrated, or not demonstrated. Examples for consequences based on demonstrated similarity are: continuation of manufacturing after an implemented manufacturing change, moving ahead within a biosimilar comparison task to the next stage in the recommended stepwise manner, or even to waive a clinical trial based on demonstrated similarity at the quality level (e.g. similarity in dissolution demonstrated).
- Considerations should be made regarding which aspects of the comparability task could be approached in a prospective manner. Also, unavoidable limitations and feasibility issues in relation to the conduct of comparative QA data analyses could be described. Even if the nature of the data comparison remains retrospective, several aspects of the comparison task could nonetheless be pre-planned before the actual data for inclusion in the analysis is collected. These aspects could cover in particular: the set of QAs subject to analysis, definition/elaboration of the similarity condition, the sampling strategy, the similarity criterion applied per QA.
- Elaborations concerning the similarity condition (see Section 4.2) need to be contained in the comparison protocol as a basis for any inferential statistical framework of QA data comparison. In this context, a discussion is required on whether it would be meaningful to assume the same similarity condition for all QAs involved in the similarity assessment.
- Representativeness of samples analysed is the key pre-requisite for a meaningful interpretation of results in inferential statistical methodology. Hence, considerations concerning the sampling strategy are of utmost importance, and are expected to include the decision regarding what the unit of observation will be: manufacturing batch, vial/pool of liquid formulation, tablet, etc.

Decisions in this regard would ideally also be driven by the knowledge of potential sources of variability in the QA data. Descriptions of sampling plans need to also include justifications regarding exclusion (non-selection) of batches/units, which were principally available for the comparison task. It is acknowledged that in some situations investigations will be limited to non-random samples or to samples for which information regarding the origin or specific manufacturing circumstances cannot be retrieved. In any case, regulatory assessors would ultimately need to verify that selection of batches/units was not data-driven.

- For each (Critical)QA which is selected for comparative data analysis, a similarity criterion – according to the descriptions in Section 4.3 – needs to be pre-specified. For QAs with similar scales of measurement and similar distributional assumptions, the same similarity criterion could be planned, in case the same similarity condition applies. Ideally, the choice of similarity criteria is based on investigations of operating characteristics with special focus on the risk for a false positive decision of similarity.
- The comparison protocol would ideally also contain a description of how the combined comparative evaluation of several QAs would eventually translate into an overall conclusion of similarity at the quality level.