EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

23 March 2017
EMA/CHMP/138502/2017
Committee for Human Medicinal Products (CHMP)

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

Draft

Disclaimer: This reflection paper has been written to provide current regulatory considerations regarding statistical aspects for the comparative assessment of quality attributes where these are used, or are proposed for use, in drug development and Marketing Authorisation Applications. It was also prepared to invite comments in relation to the opportunities and limitations related to inferential statistical methodology applied on quality attributes' data in the exploration of similarity of two drug products. Whilst in some parts the paper describes Frequentists statistical methods, the field is also open to explore alternative approaches, e.g. following Bayesian methodology.

The current document does not contain explicit guidance on which statistical approaches are most suitable. It rather tries to establish a framework and a common language to facilitate future discussions among stakeholders.

The content of this reflection paper and its implications shall be further discussed at a European Medicines Agency's public work shop at the end of the 12-month public consultation phase. A longer than usual consultation period will allow companies to come forward to EMA via interaction with the Scientific Advice Working Party with proposals that may include the principles and methods discussed in this document or alternative approaches that are not discussed in this document.

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

Comments should be provided using this template. The completed comments form should be sent to RP-stats-QA@ema.europa.eu

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

30 Churchill Place ● Canary Wharf ● London E14 5EU ● United Kingdom
Telephone +44 (0)20 3660 6000 **Facsimile** +44 (0)20 3660 5555
Send a question via our website www.ema.europa.eu/contact

An agency of the European Union

Table of contents

Executive Summary

This reflection paper identifies three areas of interest from the regulatory perspective, where the comparative evaluation of drug product’s quality characteristics plays an important role, either during drug development, during drug lifecycle, or during in decision making processes potentially leading to marketing authorisation. The document focusses on methodological aspects in relation to statistical data-comparison approaches for the settings of: pre- and post-manufacturing change, biosimilar developments as well as generics’ development. For all these settings defined, the reflection paper raises open issues from a statistical perspective addressing question related to comparison objectives, sampling strategies, sources of variability, options for statistical inference and acceptance ranges.

This document is targeted to both, experts from industry and regulatory assessors. The paper tries to connect to other available regulatory guidance where the issue of comparative data assessment concerning quality attributes is discussed for certain contexts, but where more detailed guidance of how to actually carry out the comparison task (based on empirical sample data) is lacking.

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

1. Introduction

Comparison of empirical data from quality characteristics of drug products (quality attributes) is of importance in many areas of drug development. There are at least three areas where the comparative evaluation of quality characteristics plays a major role in decision making on the manufacturer's as well as on the regulator's side:

- the comparison of a particular drug product in versions pre- and post-manufacturing change,
- the comparison of a candidate biosimilar product to a reference medicinal product,
- the comparison of a candidate generic product to the reference medicinal product;

In these areas, many different methodological approaches to set up a framework for the comparison of quality characteristics are followed and often require regulatory assessment. In many instances, the suggested comparison approach contains statistical elements in order to support the assertion that the quality profile of two (versions of a) drug products can be considered similar. This frequently involves the definition of 'similarity'-criteria, mostly based on information regarding known or expected variability of quality data associated with the underlying manufacturing processes. However, conclusions drawn from comparative data analyses (e.g. "a manufacturing change has not substantially altered the product quality") are often based on rather limited information available, e.g. a small number of manufacturing batches.

Making use of inferential statistical methods means quantifying uncertainties arising from the fact that claims (or decisions) are made based on limited data stemming from a sample. If comparative data analysis is limited to the sole description of the samples taken, it is evident that no clear inference can be drawn regarding drug material that was not sampled. Understanding the need and the options to quantify uncertainty related to decision making based on sample data is key to evaluate the capabilities statistical concepts may bring to the matter of comparing quality attributes. At first sight, it might seem straightforward to apply inferential statistical methods (like equivalence testing) for the purpose of comparing data from quality attributes, but often severe limitations exist regarding practical applicability, given the specific circumstances related to sampling and data collection. From the regulatory assessment perspective, it has become evident that the potential role of 'classical' inferential statistical methods (which are considered well established in the comparative analyses of clinical data) is currently not sufficiently clear in the context of comparison of quality data. Also, the lack of significant differences alone does not imply similarity. Hence, the question of whether the desired conclusion of similarity of products could indeed be inferred from often limited information from sample data remains difficult to answer in many occasions.

Therefore, the goal of this paper is to reflect under which circumstances, and to which extent the implementation of inferential statistical methods can assist or even facilitate comparative evaluation of quality attributes data. In many instances, fundamental limitations (e.g. in relation to the non-representative nature of retrievable sample data) would make the application of inferential statistical methodology not meaningful. In such cases, it will be important to identify and describe those obstacles.

Separate considerations are given to the regulatory areas introduced above, whenever possible also in context to other relevant regulatory guiding documents. After providing some working definitions and delineations in the next section, Section 4 will introduce these regulatory settings in more detail. Section 5 lists important fundamental methodological prerequisites which need to be considered when attempting to establish a statistical framework for decision making based on quality attributes' data comparisons. In Section 6, the settings as introduced in Section 4 are revisited and the options as well as the possible limitations related to the use of inferential statistical methods are discussed.

At the end of the document a summary of important issues is provided to support in planning, conduct and assessment of quality attributes' data comparisons. This reflection paper is hence targeted to both, industry and regulators to promote progress in the common understanding of meaningful application of statistical methodology in this specific area. There is neither the intention nor an option to strive for an automatism by introducing a purely 'technical' data comparison methodology which would remain un-reflected by the know-how of drug developers and regulatory assessors acting as experts in the field. It is, however, important to note that all decision criteria currently used to conclude on similarity on the quality level involve empirical considerations based on sample data. The use of sample data for reasonable decision making usually requires statistical considerations. Hence, understanding of some fundamental statistical concepts is key for development and assessment of such decision criteria to avoid mistakes in decision making. Furthermore, improved common understanding can be expected to facilitate consistent assessment on the regulatory side in the future.

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) 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) 726/2004 and Article 10(4) of Directive 2001/83/EC as amended.

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 (revision 1) (EMA/CHMP/BWP/247713/2012)
- Guideline on the pharmacokinetic and clinical evaluation of modified-release dosage forms (EMA/CHMP/EWP/280/96 Corr1)
- Note for guidance on the clinical requirements for locally applied, locally acting products containing known constituents (CPMP/EWP/239/95 final)

3. Definitions and delineations

Throughout the text the term 'drug product' is used to simplify reading, but it is evident that quality comparisons are also often made on either the drug substance or on an intermediate level. However, the considerations made in this paper need to be understood to equally apply to all of that terms.

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 be included in 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 impact on the methodological options for the actual data comparison.

This reflections paper does not touch upon the topic of criticality assessment of QAs. The reason is that criticality assessment is discussed in many other guiding documents (listed in Section 2) from various perspectives and for different compound classes. The issue of meaningful selection of QAs for comparative purposes is primarily driven by non-statistical reasoning, and is often judged on a case-by-case basis. Hence, the starting point in this document is actually the assumption that a set of QAs has been identified a priori which is found suitable for the purpose of a comprehensive comparison. Suggested concepts for the comparison on the quality level may differentiate/categorise QAs according to their criticality (different 'tiers'), foreseeing different comparative analysis techniques with graded rigor for the categories defined.

Another delineation which seems important in relation to the reflections given below is that this paper will also not touch upon process-control methodology. Adequately applied process-control will in general target at consistent manufacturing in a specific manufacturing process environment. Following a simplistic view, a process control system will flag marked deviations from expected product quality looking at several QAs, potentially triggering measures to counteract in order to be able to continue with 'consistent' production in the future. From a statistical perspective, this means that a specific quality control setup constantly compares empirical data from the production process against a predefined target range on an ongoing basis over time. Following this reasoning, it is evident that a specific manufacturing process is subject to (allowed) variation in itself, even if manufacturing is judged to be 'consistent' by help of process-control techniques over time. This reflection paper follows this understanding of a well-controlled manufacturing process. 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 is important to note that this assumption needs to be made for simplifying reasons, as discussed concepts for comparative data analysis will inevitably lead to misinterpretations if one or both processes to be compared are themselves subject to drifts in product quality over time. 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 set up manufacturing processes. It is also important to note that the 'consistency'-assumption 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). 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.

Performing a comparison on the quality level based on samples taken from two manufacturing processes usually means that there is interest in drawing conclusions on similarity for the entirety of the material produced by the two manufacturing processes. Hence sample material needs to be understood as 'vehicle' to estimate quality characteristics for the entirety of material produced in the past and the future, assuming consistency of the production processes as defined above. Therefore, the understanding that interest is not so much in the actual samples (e.g. batches) drawn, but in the underlying (actually unknown) data distributions of the entirety of the materials ever produced, is key to follow the considerations in this document.

4. Settings where the comparison on the quality level is of particular relevance in regulatory decision-making

This Section categorises the vast majority of occasions where a need is seen to have a comparative evaluation on the quality level. Described scenarios primarily focus on situations where two sets of available batches are subject to a comparison task. The simplest task of comparison of checking whether one specific manufactured batch fulfils certain release criteria ('within specification') is briefly addressed in Section 5.1.

It seems important to note that the described settings can be quite different, in particular with regard to the practical or scientific implications a conclusion of demonstrated similarity on the quality level could have. It is hence not straightforward to assume that the same rigor of evidence to support similarity would be required in these different situations. As a consequence, the range of potentially suitable approaches and methods to carry out comparative data analyses might differ in the different settings described in the following. All settings mentioned below would merit from further reflections concerning the options and limitations of inferential statistical methodology which might be considered suitable for application in the situations described.

4.1. Pre/post manufacturing change

The comparative evaluation of the quality of two product versions before and after a certain manufacturing change (and/or manufacturing transfer) is a very common task occurring during the lifecycle of a medicinal product. This might also include comparative investigations when moving from lab-scale to a larger manufacturing scale in the drug development, when changing the formulation, or when altering source or grade of starting materials.

In principle, a comparison task can arise for chemical (synthesized) as well as for biotech-derived/biological products. Whereas for handling the task for chemicals no dedicated methodological guidance exists, for biotechnological/biological products the ICH Note for Guidance (NfG) Topic Q5E describes the general goal of a comparability exercise for two product variants before and after a manufacturing change as "ensuring the quality, safety and efficacy of drug product produced by a changed manufacturing process." This NfG states that this does not necessarily mean that QAs of the pre-change and the post-change product are identical, but that the goal is to show that they are 'highly similar' in a sense that marked differences which would have adverse impact upon safety and efficacy of the drug product can be ruled out. Further interpretation of this wording might also indicate that manufacturing change-triggered differences in product quality, which are associated with positive impact on safety and/or efficacy, could eventually be accepted from the regulatory perspective. In this context, it is important to understand what type of differences on the quality level (in the selected QAs) would actually be associated to such positive impacts. As further explained in Section 5.1, such understanding would drive the choice of methodological statistical concepts used for the comparative analysis of QAs' data.

In contrast to the biosimilar setting (Section 4.2) the typical starting point in the pre-/post manufacturing setting is usually based on easy access to available 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. Most of the time, the limiting factor is the low amount of new batch material available after the manufacturing change. As a consequence, QAs' data from just a few 'post-manufacturing-change' batches are taken as single values and compared to 'data-ranges' describing the pre-change manufacturing condition.

A huge diversity of comparison approaches has been applied in the past, and some of them included statistical intervals, e.g. tolerance intervals, x-sigma, min-max range interval, etc. From a statistical perspective, the context of use of these intervals is however rarely clear in relation to the interpretation of conclusions drawn, i.e. whether the methods applied would really be suitable to support the claim that the post-change manufacturing process can generate material of sufficiently similar (or even "better") quality as compared to material produced by the pre-change process. Hence, some dedicated reflections will be made for this specific comparison setting (Section 6.1).

4.2. Biosimilar developments

The task to compare two biological medicinal products on the quality level is inherent to biosimilar developments. The CHMP Guideline on Similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues, rev.1 (EMA/CHMP/BWP/247713/2012) addresses the importance of this task within the whole biosimilar comparison and mentions physicochemical properties, biological activity and immunochemical properties as relevant sets of QAs for the comparison task. The guideline requests that "… analytical data submitted should be such that firm conclusions on the physicochemical and biological similarity between the reference medicinal product and the biosimilar can be made." In order to achieve that goal, an extensive (side-by-side) comparability exercise is deemed required to demonstrate that the biosimilar candidate has "a highly similar" quality profile as compared to the reference medicinal product. The guideline furthermore mentions the quality target product profile (QTPP) as a development tool for biosimilar manufacturing. The QTPP, corresponding to a set of quantitative ranges for key QA based on data collected on the chosen reference medicinal product, are also suggested to guide the comparability exercise.

From the general methodological point of view, the goal to demonstrate equivalence (in contrast to non-inferiority) is the focus in the biosimilar setting. As also mentioned in the Guideline, exemptions could be potential improvements in specific QAs (e.g. impurities) which might translate to safety advantages. However, for most of the comparative analyses of QA data between the biosimilar candidate and the reference medicinal product, the focus would usually be on some sort of equivalence investigations.

In most biosimilarity development programmes satisfactory similarity on the quality level is understood as the first important milestone to be achieved in the stepwise development approach. In this context it is important to note that the comparisons on the quality level is likely the most sensitive part of the whole comparison exercise to detect differences between the biosimilar candidate and the reference medicinal product. Many of the preclinical and clinical models used subsequently to continue the comparative development are often judged less sensitive to detect such differences. At the same time, however, the impact of differences at the quality level on clinical outcome (efficacy/safety/ immunogenicity) is often hard to predict or quantify. This usually aggravates the definition of meaningful equivalence-criteria in the QAs' data comparison and hampers biosimilar development approaches where a stronger emphasis is put on the evidence from the comparability exercise at the quality level.

Despite these difficulties, there is increasing interest in the question of whether the rigor of the comparative approach or the degree of similarity demonstrated on the quality level can determine the amount of additional evidence for similarity to be generated at later stages of development, in order to reduce remaining residual uncertainty. In particular, questions had been raised by developers whether comparisons in the clinical models can be abbreviated on basis of a robust comparison of selected QAs revealing compelling evidence of similarity. From related discussions between developers and regulators, it became evident that there is no common understanding what kind of statistical data analysis approaches would be considered suitable (if any) for comparison tasks involving data from QAs. It was found that the potential role, as well as the limitations of inferential statistical methods related to equivalence testing need further reflection in this setting. However, it can be seen likely that future methodological reflections would lead to data comparison methods which will eventually go beyond the descriptive statistical approach mentioned in Section 5.2 in Guideline EMA/CHMP/ BWP/247713/2012, if the basis for regulatory decision making would – to a large extent – be based on the demonstration of similarity on the quality level.

One further special aspect frequently arising in the biosimilarity setting is the need to bridge from non-EU sourced comparator products to the EU-sourced reference medicinal product. As such bridging usually involves data comparison on the QAs' level, Section 6.2 provides some related comments.

4.3. Other settings and generic developments

Abridged or hybrid marketing authorisation applications for small molecules represent one further arena where data comparison on the quality level, and possibly also on an ex-vivo/in-vitro level could be of pivotal relevance for regulatory decision making. Locally applied, locally acting products represent one example where under certain circumstances equivalence hypotheses based on data from certain QAs or data from ex-vivo/in-vitro experiments need 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, the 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 analyses' 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. In many instances however, these criteria turn out to be not sufficiently justified for the desired context of use.

The comparative analysis of dissolution profiles constitute another special case that fits into the framework of exploring equivalence hypotheses on the quality level. In the development of generic drug products circumstances exist, where the conclusion on equivalent dissolution profiles can serve as surrogate for in-vivo bioequivalence. Decisions for waivers, that alleviate the need to carry out comparative in-vivo (i.e. pharmacokinetic or even therapeutic equivalence) studies should then be based on results of analyses on "bioequivalence surrogate inference" according to the Appendix I of the CHMP Guideline on the Investigation of Bioequivalence (CPMP/EWP /QWP/1401/98 Rev.1/Corr). In this context, it is reiterated that the term 'inference' is used to reflect the actual expectation that data analysis on the quality level (here, dissolution) is in fact related to claims concerning the entirety of material produced by the manufacturing processes at hand, and not only to the samples tested in the dissolution experiment.

Furthermore, the CHMP Guideline on the pharmacokinetic and clinical evaluation of modified release dosage forms (EMA/CPMP/EWP/280/96, Corr1) contains considerations regarding similarity of dissolution profiles regulating waivers and the need for bracketing approaches, but does not include more detailed recommendations regarding reasonable approaches for using inferential statistical methodology. Hence, further reflections can be expected to be helpful for planning and analysis-purposes in this area as well (Section 6.3).

5. Approaching the comparison task from the statistical perspective and associated obstacles

5.1. The choice of characteristics to be compared and related comparison objectives

Following a statistical understanding, observed data for the QAs of interest coming from the selected batch-material need to be understood as actual realizations of underlying (unknown) data distributions. The interpretation is that, for each QA of interest, actually two unknown distributions corresponding to the two manufacturing processes are subject to comparison. Against this background, the question arises which characteristic(s) of these distributions should be taken for the comparison task. The choice of a suitable characteristic to be compared also depends on the scale of measurement of the QAs of interest (nominal to continuous scale). If underlying data distributions are parameterised, parameters of these distributions can be used for the comparison task. For QAs measured on a continuous scale, one option is to compare the means (as parameters) of the distributions. This would correspond to a comparison of the location of the distributions, leaving aside any comparative investigations concerning the spread/variance of the distributions. However, parameters describing the spread of the distributions can of course also be subject to comparative analyses.

In many instances in practice, no dedicated considerations are given regarding the choice of the distribution characteristic of interest. Instead, often single observed values representing the individual batches are directly taken for the comparative analysis. Whereas such an approach is not 'wrong' from a methodological perspective per se, careful interpretation is required based on the observed outcome of such comparisons (see examples in Section 5.5).

Hence, from a planning perspective, the issue of identifying the data distribution characteristic (or parameter) of interest to be compared needs to be addressed upfront. The other important question is related to the actual objective for each specific QA's comparison: e.g. if means are compared, is it sufficient to rule out marked differences in one direction only (e.g. rule out increase in impurity, or decrease in potency), or is it the goal to protect against differences in either direction? This question is closely related to the comparison scenario at hand given the regulatory context (see categories introduced in Section 4), but at the same time needs separate considerations for each QA foreseen for the comparison task. For example, in one and the same pre-/post-manufacturing change comparison, it may well be that that e.g. a reduction in mean post-change impurity could be acceptable (one-sided comparison), whereas for other QAs (e.g. potency) marked differences in pre-/post-change means in either direction need to be excluded (two-sided comparison), as such differences - depending on the direction - might relate to expected negative impact either on clinical efficacy or on safety.

Given the considerations above, it appears that a specific comparison task for one selected QA will fall under one of the following categories concerning the underlying objective:

5.1.1. Within-specification claim

This refers to the comparison of QAs of one given batch to a pre-defined specification range (e.g. for batch release purposes). A specification range could be one- or two-sided. In this case there is interest in the batch material at hand, and the question to be answered is whether the data observed for that particular batch is within the range of expectations for the underlying manufacturing process. Of interest from a methodological perspective is the question of how the specification intervals are derived. Several methods are applied in this context, and not all of them might be considered suitable to account for the uncertainty arising from the fact that specifications are calculated based on data from sampled batches.

It is important to note that methods applied in the context of comparisons against specifications do not automatically qualify for other comparison tasks involving quality data (i.e. pre/post manufacturing change evaluations, biosimilar setting), as a more general inferential interpretation related to the underlying manufacturing process is required for the latter.

5.1.2. One-sided comparison objective, non-inferiority claim

Such a claim could be based on the underlying understanding that actually two data distributions related to two manufacturing processes are subject to the comparability task, and produced material is understood as realisations of these processes. Often, two sets of batches coming from these manufacturing processes would serve as samples for statistical evaluation. The claim to be tested would be that one of the two processes (e.g. the manufacturing process after a manufacturing change) is able to produce batches with 'non-inferior quality' as compared to the other process (e.g. the pre-manufacturing change process), measured by the QA selected.

In statistical terminology, this corresponds to a 'one-sided' statistical test. One classical approach to carry out such non-inferiority investigations is the comparison of a one-sided confidence interval (e.g. for the difference in means) derived from actual sample data to an a priori defined acceptance region (non-inferiority margin). However, it has to be noted that such an approach already requires some assumptions to be fulfilled (see Section 5.4 and following sections).

5.1.3. Two-sided comparison objective, similarity/equivalence claim

The same conceptual understanding as described for the non-inferiority claim applies in principle also to the equivalence claim. The difference is that the claim to be tested would be that the two processes under consideration are able to produce material with equivalent quality (as measured by the QA at hand).

One classical way to carry out equivalence testing would be to derive a two-sided confidence interval (e.g. for the difference in means) and compare it to a pre-defined equivalence margin. But as mentioned in 5.1.2, pre-requisite conditions would need to be fulfilled, and such an analysis approach might not be feasible in many instances to compare QA data.

It is reiterated that any potential non-inferiority or equivalence conclusion drawn for one specific QA would not apply to the two actual sets of batches used for the analyses, but to the entirety of material produced by the manufacturing processes at hand. This differentiates inferential statistical testing from purely descriptive data comparison (which only refers to the samples drawn). This of course requires the assumption of consistent production processes to be fulfilled. However, the use of inferential methods requires further assumptions to be fulfilled. The following Sections of this chapter will discuss those.

5.2. Understanding sources of variability in quality data and 'the unit of observation'

In contrast to clinical research where usually the trial participant is the starting point for considerations regarding the unit of observation most suitable for statistical analysis, corresponding considerations for the comparisons of QAs characterising underlying manufacturing processes do not appear that straightforward. One commonly used approach is to see the production batch as the unit of observation which can be used for data analysis. Although this might be a meaningful strategy in many instances, 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 (non-exhaustive):

- 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

Sufficient understanding of the potential sources of variability in the data available is key to decide upon the unit of observation, and to explore the range of suitable statistical analysis methods. The definition of the unit of observation will also be important for sampling considerations. With thoroughly selected statistical methods it is possible to account for possibly existing dependence between observations.

Depending on the nature of the comparability task and the underlying objective, access to information describing the context of data collection for the QAs of interest may be limited. Such a limitation would hamper identification of potential sources of variability. In consequence, options for an inferential statistical analysis approach for the desired data comparison would be limited as well.

5.3. Random Sampling / Experimental Approach

Application of inferential statistical methods and the interpretation of their results require that samples of units taken for analysis are representative for the underlying data generating process(es). The ideal selection strategy would be random sampling. Implementing such a process would mean that generally each of the units available for selection would have an equal chance to be selected/sampled. In context of the comparison of QAs it is often realised in practice that a random sampling approach might not be achievable/feasible. One frequently encountered 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 e.g. to avoid repeated sampling of units carrying no further relevant information for the comparative analysis.

The non-random nature of samples used for the purpose to compare manufacturing processes, resulting in questionable 'representativeness', needs to be understood as one frequently occurring limiting factor hampering the desired application of inferential statistical methodology. If representativeness cannot be assumed, any particular statistical model applied will fail to describe uncertainty in the desired manner, and the corresponding results have no inferential interpretation.

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

5.4. Finding a metric to describe the difference between two manufacturing processes

Once the parameter of interest is selected for the comparison task (e.g. the mean, cf Section 5.1), the next step would be to find a method/metric to describe the difference/distance between the parameters for the two distributions. For the example of the comparative analysis of means, this metric could simply be the difference of means or the ratio of means. Defining this metric immediately leads to a corresponding optimal outcome of the comparison analysis. If e.g. an equivalence hypothesis is to be tested (i.e. a null hypothesis of non-equivalence would need to be rejected), the goal would be to generate sufficient evidence to demonstrate that the difference in means is sufficiently close to zero, or that the ratio of means is sufficiently close to 1 (evaluated by making use of confidence intervals, see Section 5.5).

The definition of such a metric to describe the difference/distance between the two unknown underlying distributions relates to the intention to derive one single measure to describe the difference of interest, and thereby to 'simplify' the analysis task. As already mentioned in Section 5.1, such reasoning can establish the bridge to statistical testing of equivalence or non-inferiority. In order to carry out such tests, it is not only necessary to derive a point estimate for the metric of difference defined. Two further elements are required: a method to quantify the uncertainty around the derived point estimate, and the definition of an acceptance range to describe the maximum allowed difference between the two distributions of interest, which would still be compliant with a statement that the material from two underlying manufacturing processes can be considered similar.

5.5. Statistical intervals to quantify uncertainty of claims based on sample data

With the computation of certain statistical intervals it is possible to quantify uncertainty in relation to drawing a conclusion from samples to the entirety of material ever produced by underlying manufacturing processes. It is important to note that this potential of quantification of uncertainty is the advantage of inferential statistical methods over simple descriptive data analysis. If data analysis is limited to description of the samples taken (e.g. solely reporting of sample means and ranges), it is evident that no clear inference can be drawn regarding drug material that was not sampled. In order to make full use of the inferential property of statistical intervals in the setting of comparative data analysis, it is essential that the objective of the comparison as well as the metric to characterise differences of underlying distributions is consciously chosen.

5.5.1. Comparison approaches based on intervals commonly seen

From a statistical point of view, a clear distinction has to be made between quantification of uncertainty by the use of statistical intervals on the one hand, and the goal of defining acceptance ranges for the framework of the comparability task to be accomplished (e.g. equivalence margin, non-inferiority margin) on the other hand. Frequently, measures to quantify uncertainty are either not applied at all or used in a wrong context (see next paragraph).

In practice, comparability ranges are frequently established based on a statistical interval, e.g. the min-max range or a tolerance interval calculated from characterisation data of the reference product. Although such intervals are considered useful for data-descriptive purposes, the methodological limitations related to these intervals when used as similarity decision criteria need to be understood.

Min-Max range

In its fundamental property, a min-max range describes the observed data range in a sample (e.g. for a selected set of batches), and has no direct interpretation per se for the quantification of uncertainty concerning the location of the unknown data distribution(s). In some comparison settings 'Min-Max ranges' have been suggested to compare selected QAs between two sets of batches (e.g. pre/post manufacturing change or reference/test in the biosimilar setting). Simple rules to claim similarity such as 'the min-max range of test is entirely contained in the min-max range of reference' seem flawed as the probability of fulfilling this criterion generally increases with decreasing number of test batches investigated. This actually means that chances are highest to claim similarity if only a few (or in the extreme just one) test batches are/is used for this kind of comparison. This is of concern, as such similarity criteria promote small-sample investigations to increase the likelihood to conclude similarity, and hence in parallel increases the chances of false positive conclusions on similarity. Of note, comparison of single batch data to a min-max range might be suitable in the context of batch-release (see Section 5.1.1).

Tolerance intervals and x-sigma approaches

A tolerance interval (TI) is usually computed to estimate a data range by which a specified proportion p (e.g. the central 90%) of the units from the underlying population is assumed to be covered with a pre-specified degree of confidence c (e.g. 95%); Similarity rules suggested in the past involving the TI concept were conditions like 'measured QA data from all test batches of the sample fall within the 90%/95% TI computed from the reference batches'. Whereas a TI is conceptually suitable to describe uncertainties related to a claim for an unknown data distribution, its application requires thorough consideration due to several reasons. First of all, standard methods to compute TIs assume normality for the underlying unknown distribution, and the validity of this assumption can actually not be checked in most practical instances. Further, the choice of the parameters p and c remains arbitrary, and – if applied in a decision criterion as mentioned above - high values for p and c (eg.99%/99%) wrongly suggest high precision and certainty for the decision making on similarity, whereas actually the opposite is the case due to associated widening of the TI if p and c approach 100%. Such a similarity assessment approach exemplifies the undesirable mix of 'quantification of uncertainty' and the 'definition of an acceptance range' by making use of one and the same statistical (TI) interval. Similar methodological concerns arise with the application of 'x-sigma rules' (where x is usually one of: 3, 4 or 6), in particular if applied to characterise the reference (or pre-manufacturing change) QA's data distribution to define a 'target range' for a similarity investigation. Hence, it is primarily the described methodological deficiency related to the actual application, rather than a too low number of samples which makes similarity decision rules based on TIs or 'x-sigma' approaches often unsuitable for a comparison task. As a consequence, there are usually no options to overcome such fundamental methodological deficiencies by increasing the sample size for the computation of TIs.

5.5.2. Guiding principles for the use/computation of statistical intervals for QA data comparison

Adequacy of the choice of a certain statistical interval to quantify uncertainty related to statistical sampling always depends on (i) the underlying comparison objective (Section 5.1), (ii) the choice of the characteristic/parameter describing the data distribution (Section 5.1), and (iii) the metric to describe the difference between the two data distributions (Section 5.4). Once the metric is decided upon (e.g. the difference of means), one further question relates to the assumed sampling distribution of that metric, e.g. whether normality can be assumed. Only if these aspects are clarified upfront, a proper choice can be made regarding the statistical interval method to be used to estimate the uncertainty related to the sampling approach.

Existing different concepts for statistical intervals not only differ in their method of computation, but also (and importantly) in the interpretation of the resulting numerical interval.

Prediction intervals

Prediction intervals (PI) are estimated to describe a data range covering data outcome of units drawn in the future with a pre-specified degree of certainty. PIs can be derived for a 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.

Confidence interval

Another important statistical interval concept is of course the confidence interval (CI). As mentioned earlier, CIs are frequently used in the context of equivalence/non-inferiority settings in clinical research settings. CIs usually describe a data range which is assumed to cover a parameter (e.g. the mean) of the unknown distribution with a given probability.

It is important to note that interval estimation techniques for CI, PI and also TI can be adapted to directly quantify uncertainty related to claims on differences (or ratios) in parameters of two underling distributions, e.g. a 95% CI for the difference between two means (e.g. between reference and test means) can be derived. In the technical computation of intervals (in particular confidence intervals) it might also be possible to account for the available knowledge regarding the sources of variability in the data material to be analysed. For instance, parametric statistical methods can be used to account for specific correlation structures as well as for factors associated to between/within batch variability. It is however beyond the scope of this reflection paper to provide a comprehensive overview of methodological approaches to adequately compute intervals to quantify uncertainty of claims based on sample data. It is neither considered possible nor necessary to categorically preclude any kind of statistical modelling approach for the data comparison task at hand. It is however seen required to justify the choice of applied methods against the background presented above. The variety of candidate methods may also comprise analysis approaches requiring less (or no) specific a priori assumptions such as non-parametric techniques, bootstrapping or other re-sampling methods ('distribution free' intervals).

5.6. Definition/justification of an equivalence/similarity criterion, acceptance range

Any inferential statistical comparison of QAs would require an a priori definition of an acceptance range or a correspondingly defined acceptance criterion. The definition of an acceptance range is usually not resulting from the analysis of actual sample data (cf. to the TI example in Section 5.5). It is rather the result of separate considerations related to maximum allowed difference between the two (unknown) underlying data distributions for a specific QA of interest, which would still be compliant with a statement that the material from the two processes can be considered similar/equivalent/non-inferior. For a specific comparison task involving QAs, acceptance limits/regions would need to be understood as an a priori fixed design element, and should hence conceptually be differentiated from statistical intervals derived from actual sample data.

As regards the scale of measurement (e.g. additive or ratio scale), the defined acceptance range should fit to the metric chosen to estimate the difference between the two underlying data distributions of interest.

As in many other settings of non-inferiority and equivalence testing (also in the area of clinical trials), the a priori definition of acceptance ranges can also be expected to be controversial in the comparison of QAs. The interpretation of (maximum allowed) truly existing differences in QAs will in most cases require a good understanding of the impact such differences could have on clinical outcome on the patient level. The extent of knowledge regarding this association between differences on the quality level and clinical outcome (efficacy and/or safety aspects) will already drive the criticality assessment of the QAs, and hence the selection of specific QAs for the comparison task. Moreover, there are also cases where pharmaceutical quality by itself would be a primary driver to define acceptance ranges, in particular if the range of 'good pharmaceutical quality' is narrow and associated potential related changes on the clinical level would be negligible. However, in many instances a certain degree of arbitrariness in the definition of acceptance ranges might be unavoidable in practice. Against this background, an interesting question is whether the development of agreeable standards (i.e. broadly agreed acceptance ranges as this is the case in e.g. the bioequivalence assessment) could be a meaningful goal for the future. For the moment, arbitrariness in the definition of acceptance ranges would need to be reflected in the eventual assessment of any comparative analysis carried out.

5.7. Defining an overall 'success criterion' to claim equivalence/similarity in presence of a large number of QAs

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 in Section 5 so far 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, underlying data distributions can be rather different. In this context it is unreasonable to assume that one and the same statistical concept will be suitable for comparative evaluation of all the QAs involved. In most instances, tailored approaches seem to 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 ('success criterion') seems necessary to describe the minimum requirement for a claim of similarity. Such a concept would need to be put in an analysis plan which is prepared prior to sampling and conduct of the comparison analyses. Any post-hoc justifications that observed (unexpectedly big) differences in one or more of the analysed QAs would have no or only minor impact on clinical outcome might be seen to contradict preceding criticality assessment of QAs and/or an adequate definition of corresponding acceptance ranges for single QAs.

The overall risk of a false positive conclusion on equivalence (or non-inferiority) following an inferential statistical evaluation will strongly depend on the type-1-error specifications (alpha, significance level) in each separate QA data analysis. Only little guidance can be given regarding the choice of nominal alpha for the comparison of QAs' data. Generally, a priori considerations concerning the risk of a false positive conclusion on equivalence (or non-inferiority) on the quality level would become more important, the more this comparison is expected to carry pivotal evidence in the whole comparison task within a specific drug development. Some case-specific comments are provided in Section 6.

In this context, power considerations might eventually also become relevant from a planning perspective, as sample size constraints (e.g. low batch numbers) and associated low power may lead to refrain from inferential statistical comparison.

6. Reflections of issues raised, implications for planning and assessment

General guiding principles can be inferred from the issues mentioned in Section 5 which are equally applicable to the different regulatory settings introduced in Section 4. They need to be considered in the order they are presented in.

- For any data comparison plan on the quality level involving several QAs the objective should be clearly stated. Describing the objective of the comparison task ideally includes considerations regarding potential consequences for the two potential outcomes, namely either that similarity could be demonstrated, or not. Examples for consequences based on demonstrated similarity are: continuation of manufacturing after an implemented manufacturing change, moving ahead within a biosimilar development programme to the next stage in the stepwise comparison, or to waive a clinical trial based on demonstrated similarity in dissolution behaviour. These considerations should already cover the question what characteristics of the underlying distributions shall be compared. One of the options could be the comparison of means. However, in some other situations the comparative evaluation of the variability (e.g. variance) might need to be targeted.

- The whole spectrum of options should be explored in how far the comparison setting has to exclusively rely on investigations of data collected retrospectively, or whether a prospective approach could be envisaged as well. Even if the nature of the data comparison remains retrospective, several aspects of the comparison task could nevertheless be pre-planned before the actual data for inclusion in the analysis is collected. Examples would be pre-specification of: the set of QAs subject to analysis, the sampling strategy, the data analysis (interval) method applied, the acceptance ranges, etc. Only adequate pre-planning will protect against the potential criticism related to data-driven planning and biased post-hoc decisions.

- Considerations concerning the sampling strategy are of utmost importance, and are expected to include the decision what the unit of observation will be: batches, lots, packages, tablets, vials/pools of liquid formulations, powders, etc. Decisions in this regard shall also be driven by the knowledge on potential sources of variability in the QA data. As representativeness of samples analysed is the key pre-requisite for a meaningful interpretation of results in inferential statistical methodology, efforts should be taken to adequately describe the chosen sampling strategy. Such a description should also include justifications regarding exclusion (non-selection) of batches/units which were available for the comparison task. In instances where random sampling is not possible, regulatory assessors need to verify that selection of batches/units was not data-driven. 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 such cases, very limited options may exist for a reliable interpretation of results from a statistical inferential procedure. As there is no use of inferential statistical approaches applied to data not being representative at all, this issue should be flagged as early in the development as possible, and options could be explored to base the comparison of interest on more representative samples, or other ways to support similarity will have to be used.

- The criterion defined to judge similarity is ideally based on a metric which allows to estimate the 'distance' between the two unknown distributions (or parameters). Examples could be the difference in means, the ratio of means, the difference in proportions, or even more complex measures of distance such as the f2-function suggested for dissolution comparisons (Guideline on the Investigation of Bioequivalence (CPMP/EWP/QWP/1401/98 Rev.1/Corr)). Similarity criteria solely based on plans to compare single observations (e.g. of test batches) to a pre-defined acceptance range (based on reference data) are usually unsuitable to allow for reliable inference to the underlying general manufacturing process. One guiding principle for setting up the comparison plan is the simple rule that with an increasing amount of information available for the comparison (e.g. number of batches), the quality of the resulting decision should improve. From the statistical point of view, this means that the amount of uncertainty should principally decrease with increasing information from the manufacturing processes to be compared. An increase of the amount of available data for analysis should necessarily lead to higher precision of estimates and consequently to less uncertainty in decision making. For example, a large extent of uncertainty in the estimation of reference data (distributions) shall not be misinterpreted as large acceptance ranges for test-batch data to 'fall in'.

- As mentioned earlier, different statistical methods to derive intervals will rely on a number of assumptions. Hence, the description of the choice of statistical methodology for the QAs' comparison task should address the question of whether underlying assumptions can indeed be considered fulfilled.

- Setting up acceptance ranges (e.g. equivalence margin, non-inferiority margin) shall be seen as a separate task in the plan for QA data comparison. According to standard statistical principles, acceptance ranges are usually not a result of the actual data analysis, but are specified a priori. Such a pre-specification will usually take into account the available knowledge concerning the variability in the QA data to be retrieved, but also the assumed association between differences on the quality level and clinical outcome (efficacy and/or safety aspects). Acceptance ranges should always be defined on the scale of the metric defined to compare the distribution characteristics of interest. For example, if the ratio of means was chosen to investigate equivalence, a corresponding acceptance range should set (usually symmetrical) limits above and below the value 1.

- If all pre-requisites (as listed in Section 5) for an inferential statistical approach are fulfilled and the analysis can be planned accordingly, the issue of controlling for a false positive decision of similarity would deserve dedicated consideration. From a regulatory perspective, it appears difficult to recommend a range for 'acceptable' type-1-error specifications, as the different settings described in Section 4 differ with regard to potential negative consequences of false positive conclusions on similarity.

6.1. Specific issues for the pre/post-manufacturing change setting

As mentioned in Section 4.1, any changes in the product quality during the life cycle of a drug product can be considered acceptable from a regulatory perspective, as long as it can be ruled out that changes have an adverse impact upon safety and/or efficacy. This clearly opens the door to measures improving the quality of a specific product by manufacturing changes (either deliberately or even unintentionally). Whereas this corresponds to a one-sided comparison approach concerning the clinical consequences, this does not necessarily imply one-sided testing on the QAs' level (cf. the example of two-sided approach for potency comparison in 5.1.). However, in many instances, the comparison objective would be to investigate whether QA data of the post-change product can be shown to be non-inferior to QA data from the pre-change product.

The necessity to obtain representative samples of manufacturing units foreseen for the comparative analysis can be one very limiting factor in this context. Whereas it might be possible to 'draw' a representative sample from a larger set of pre-manufacturing change units, options to draw such a sample from the post-manufacturing batches might be limited, depending on e.g. the time since the change, batch size and manufacturing speed. In many instances it is expected that only a low number of batches produced consecutively after the manufacturing change would be available for the comparison task. Whereas consecutiveness can somehow be helpful to investigate the question of consistency in the new production process, it might not be necessarily compliant with an adequate sampling concept. Against this background, 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) which could guarantee representativeness. The question of representativeness of the first available batches for the whole future manufacturing process depends on the manufacturer's ability to maintain consistency in the important QAs in the long run. This issue deserves special attention in any justification of a plan to utilise inferential statistical methodology. In addition, it has to be noted that a very low number of available post manufacturing units 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.

As regards the identification of potential sources of variability in QAs, manufacturers may have substantial experience based on the manufacturing history of the pre-change product. For a statistical comparison approach, this might be advantageous when it comes to set up a statistical model to analyse empirical QA data given the sources of variation identified in production and assay systems. This means that available knowledge concerning different causes for within- and between-batch variability can inform the statistical comparison approach.

Whenever the justification of acceptance ranges for a pre/post comparison refers to pre-manufacturing-change release specifications, a clear description of the methods to derive those specifications should be provided.

6.2. Specific issues for Biosimilar setting

In the biosimilar setting, the task to compare two drug products on the quality level can generally be understood as an equivalence problem from the statistical viewpoint (exemptions mentioned in Section 4.2). The objective of concluding on the physicochemical and biological similarity between the reference medicinal product and the biosimilar candidate is clearly set out by the applicable guidance as mentioned earlier. When it comes to the selection of distribution characteristics for the QAs' data comparison, it hence appears reasonable to investigate metrics describing the location (e.g. mean) as well as the spread (e.g. variance) of the underlying distributions. In the biosimilar setting, any difference identified in any characteristic would need to be interpreted as a potential signal for non-similarity between the reference medicinal product and the biosimilar candidate. For this particular comparison setting, statistical analysis strategies have been suggested in the past which could allow for the conclusion on similarity in cases where variability was estimated to be smaller in the biosimilar candidate's data as compared to the reference medicinal product. Unless justifiable in relation to the mentioned exemption (potential improvements in specific QAs might translate to safety advantages), a conclusion on similarity would not be considered reasonable under such circumstances. Conclusion on similarity should ideally result from equivalence analyses where information regarding data origin (e.g. what data set characterises the reference medicinal product batches and what data set the biosimilar candidate batches) does not need to be utilised.

However, it has to be acknowledged in this context that cases have been described in the past where significant shifts/changes 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. Referring to the biosimilars' QTPP, EMA guideline EMA/CHMP/BWP/247713/2012 also discusses this issue in Section 5.2, suggesting that "… ranges identified before and after the observed shift in quality profile could normally be used to support the biosimilar comparability exercise at the quality level, as either range is representative of the reference medicinal product." Furthermore, data-distributional differences within the reference medicinal product which are attributable to the sourcing origin are important to be reflected for the justification of analysis plans using non-EU sourced comparator product material.

From the biosimilar developer's perspective, one further challenge is the limited access to information regarding the manufacturing of the reference medicinal product. Hence, sources of observed variability in the QAs of interest may remain obscure. From a statistical perspective, a high proportion of unexplained variability generally lessens the likelihood for a reliable similarity conclusion due to the lack of desired precision of interval estimates.

It is usually unavoidable that the manufacturing process setup (e.g. scale of manufacturing) of the candidate biosimilar changes several times during pre-marketing development. EMA Guideline EMA/CHMP/BWP/247713/2012 mentions that "Process changes may occur during the development of the biosimilar product, however, it is strongly recommended to generate the required quality, safety and efficacy data for the demonstration of biosimilarity against the reference medicinal product using product manufactured with the commercial manufacturing process and therefore representing the quality profile of the batches to be commercialised." Against this background, bridging concepts are often utilised to bridge to results from experiments carried out with previous variants of the biosimilar product, also to avoid unnecessary repetition of (ex/in-vivo) investigations. Whilst such bridging can be supported in general, the question of whether the pre/post manufacturing process comparison for biosimilars requires the same methodological rigor as the comparison to the reference medicinal product deserves dedicated consideration. Without further justification, it cannot be assumed that the same statistical methodology would be equally suitable in these two different comparison settings. Similar issues are related to bridging plans based on quality data between EU-sourced and non-EU-sourced comparator drug material.

In the framework of regulatory decision making concerning drug licensure, the question of adequate control of the risk for a false positive conclusion is of utmost importance. As regards suitable methodology of type-1-error control for equivalence testing, there is reasonable common understanding in the context of clinical trials, also in the biosimilar clinical comparison setting. However, this is currently not the case when applying inferential statistical methods for comparison on the QAs level. This is important to note in particular in light of existing initiatives suggesting biosimilar development plans where substantial evidence for similarity is supposed to be inferred from quality-data comparisons. It can be expected that the acceptability of future 'abbreviated' biosimilar programmes with a scientific comparative focus on the quality data will not only be influenced by the degree of understanding of the association between quality characteristics and clinical outcome, but will also strongly depend on how the risk for a false positive conclusion on similarity can be controlled. It is hence strongly recommended that any biosimilar programme with a focus on quality data comparison is scrutinised to control the risk for a false positive conclusion.

6.3. Specific issues for generic/hybrid developments and dissolution comparisons

The area of equivalence investigations for special pharmaceutical forms (as introduced in section 4.3) is quite diverse as not only 'pure' QAs, but also a variety of different measurements from ex-vivo/in-vitro assays can be subject to the data comparison task. Against this background, the fundamental methodological requirements as introduced in Section 5 would need to be considered, given the model/experiment identified to support an equivalence claim based on empirical sample data. Some of the aspects described in 6.2 for the biosimilarity setting to build a statistical framework to enable equivalence testing can also be applicable to the broader field of abridged/hybrid applications. This pertains in particular to the choice of metrics describing the location as well as the spread of the underlying data distributions of the attributes selected for the comparison, but also to the challenges to attribute observed variability in the empirical sample data to potential sources of variability.

As mentioned in Section 4.3, demonstration of similar 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 the Appendix I of the CHMP Guideline on the Investigation of Bioequivalence, comparative dissolution investigations are not only relevant for quality control to ensure batch-to-batch consistency, but are also of importance for the justification to waive bioequivalence studies. For the latter purpose, 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). When it comes to checking the prerequisites needed to apply inferential statistical methodology, this specific comparison task can generally be handled following the issues raised in Section 5.

The objective to demonstrate 'similar dissolution' actually has a two-sided interpretation from a statistical perspective. As regards the identification of the units of observation, guideline recommendations for comparative dissolution testing provided for oral (immediate) release forms is quite clear, suggesting to consider dissolution profiles from single tablets/capsules/etc. as the basis for evaluation. However, it has to be mentioned that no specific requirements have been expressed so far concerning the sampling of the units foreseen for the dissolution experiments. Hence, all general points made in the first part of Section 5 regarding adequate sampling, also based on available/retrievable knowledge regarding potential sources of variability (in dissolution behaviour) shall be taken into consideration for planning and assessment purposes.

Concerning the choice of the distribution parameter of primary interest for the comparison, the guideline recommendation in CPMP/EWP/QWP/1401/98 Rev.1/Corr corresponds to a comparison of mean dissolution over time. This at least applies for the standard comparison carried out via the suggested f2 metric, where differences in sample averages are suggested to be used for deriving the distance measure (between reference and test). 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.

However, the f2-metric and the mentioned alternative analysis approaches do not only differ regarding the characteristics chosen to compare dissolution profiles, but also regarding the potential to draw inference from sample results to a broader population of units. The f2 metric - by itself insensitive to the shape of the dissolution profiles and the spacing between sampling time points - was shown to have unfavourable statistical properties which make standard inferential statistical approaches (e.g. estimation of confidence intervals around the estimated f2-value from the sample) de facto impossible. Whereas this difficulty could potentially be overcome by choosing another model-independent distance metric or an approach to statistically compare fitted model parameters in a model-based comparison setting, several additional methodological issues would need to be addressed in order to enable a meaningful interpretation of any potential statistical inference. E.g., when discussing alternative analysis approaches, the guideline mentions similarity acceptance limits as one important design element for the comparative analysis, saying that these limits should be pre-defined and justified and not be greater than a 10% difference. It remains unclear however whether implementation of this requirement (to exclude 10%+ differences in dissolution) would be straight forward in an alternative data analysis setting, and in how far expectations concerning the required rigor to conclude on similarity can be met.

Another aspect would be the suitable pre-specification of the type-1-error probability, which would in most alternative analyses approaches manifest in the specification of the coverage probabilities of confidence interval/region estimates. However, it has to be acknowledged from the regulatory point of view that currently, if standard comparative evaluation via f2 is carried out, no meaningful quantification of the risk to false positively conclude on 'similar dissolution' is possible.

7. Appendix

The summary below may assist during planning of tasks related to QAs' data comparison, but also help assessors to scrutinise suggested approaches in this context. It is suggested to follow the bullet points in a top-down manner to better identify which limitations could hamper to continue with inferential statistical analysis strategy. The symbol **#** indicates possible actions which might be meaningful to take in the situations described. Whenever a descriptive statistical comparison approach is mentioned as the only option for the analysis of available data, it should be clear to analysists and assessors that a sole samples' description does usually not allow for further more general similarity claims concerning the entirety of the material produced by the (two) underlying manufacturing processes.

Summary of items which merit reflection when planning data comparisons on the quality level

- ✓ **General description of comparison setting/comparison objectives**
- ✓ **Given the QAs of interest, categorisation of QAs regarding scale of measurement (binary to continuous)**
- ✓ **For each QA, decision upon the characteristic/parameter of interest by which underlying data distributions will be compared (e.g. mean, variance, etc.)**
  - o If no such characteristic/parameter can be identified, options for data comparison would be limited to methods using data from single observation as such: # plan for a descriptive comparison approach making use of tabular and graphical presentations of the data measured/observed;
- ✓ **Translation to statistical objectives, e.g. deciding upon one- or two-sided comparison approach per QA**
  - o If no statistical hypothesis can be formulated: # plan for a descriptive approach presenting the estimates derived for the chosen parameters (e.g. descriptive presentation of means);
- ✓ **Identification of the unit of observation; at the same time exploration of potential sources of variability in QAs' data to be retrieved**
  - o If sources of variability of the manufacturing process remain obscure, a straight forward definition of the adequate unit of observation for comparative data analysis will be hampered: # describe uncertainties related to sources of variability and based on that, justify any choice of the unit of observation in case further statistical comparison is planned;
- ✓ **Consideration for which potential sources of variability the data analyses can be controlled for**
- ✓ **Sampling strategy**
  - o Description of whether there are prospective considerations for the sampling of units for analysis, covering options for random sampling and deliberate selection approaches;
  - o Judgement concerning (expected) representativeness, if representativeness cannot be assumed: # plan for a descriptive approach presenting the estimates derived for the chosen parameters (e.g. descriptive presentation of means);
  - o After sampling: Description of actual sampling process (according to plan?); Justify any 'non-selection' of units which might have been available for investigation;
- ✓ **Definition of metric/method to describe difference/distance between the chosen parameters (e.g. difference in means, ratio of means, etc.)**
- ✓ **Evaluation whether the so chosen setup for QA data comparison would allow for inferential statistical approach**
  - o Estimation (of the defined metric) based on sample data, including methods to quantify uncertainty of estimation (e.g. by confidence intervals/regions, etc.);
  - o Choice and description of the selected statistical approach for comparison; Identification/Justification of (distributional) assumptions made with the methods applied;
- ✓ **Pre-specification of an acceptance range for the analysis of each QA separately (e.g. equivalence margin, non-inferiority margin)**

  - If knowledge about the association between quality characteristics and clinical outcome (efficacy/safety) is limited, the specification of acceptance ranges might remain arbitrary and controversial: **#** reconsider the whole inferential statistical approach, as interpretation of outcome might remain inconclusive;

✓ **Consideration regarding the risk for a false positive conclusion on similarity (equivalence / non-inferiority) based on the similarity decision criteria defined**

  - Reflection of the assumed rigor of similarity decision criteria seen required in context of the particular comparison setting;