Pragmatic rules for comparability of biological medicinal products

van der Plas, R. Martijn; Hoefnagel, Marcel H. N.; Hillege, Hans L.; Roes, Kit C. B.

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
van der Plas, R. M., Hoefnagel, M. H. N., Hillege, H. L., & Roes, K. C. B. (2020). Pragmatic rules for comparability of biological medicinal products. *Biologicals*, *63*, 97-100. https://doi.org/10.1016/j.biologicals.2019.11.002

Pragmatic rules for comparability of biological medicinal products

R. Martijn van der Plas[a,*], Marcel H.N. Hoefnagel[a], Hans L. Hillege[a,b], Kit C.B. Roes[a,c]

[a] *Medicines Evaluation Board of the Netherlands, Utrecht, the Netherlands*
[b] *Dep of Cardiology and Epidemiology, UMC Groningen, the Netherlands*
[c] *Department of Health Evidence, Section Biostatistics, Radboud University Medical Center, Nijmegen, the Netherlands*

A B S T R A C T

Comparability is a key concept in the evaluation of both manufacturing changes and biosimilars. It constitutes a pragmatic and flexible approach which recognises that biologicals are inherently variable and that minor variations in quality attributes are often clinically irrelevant. In this discussion paper, we argue that comparability exercises rely on a number of pragmatic criteria. These criteria have been remarkably robust for 20 years of comparability exercises; however, the increased scrutiny of biosimilar applications provides an impetus for both codification and improvement of criteria for establishing comparability. Such a more rigorous, methodologically sound, approach towards comparability seems both feasible and beneficial.

1. Introduction

The manufacturing process of a biological medicinal product usually undergoes many changes during the product's lifecycle, and regulators have an extensive experience with the assessment of such changes. The key concept is that of 'comparability', i.e. the assessment whether or not the changed manufacturing process produces different products post-change as compared to the pre-change products [1]. The same comparability concept is currently applied to assess whether or not a biosimilar product can be considered 'comparable' to the associated originator, albeit with a different perspective.

Surprisingly, notwithstanding its importance and long history of use, comparability remains a somewhat woolly concept. The first FDA guidance [2] stated that '*two products are comparable if the results of the comparability testing demonstrate that the manufacturing change does not affect safety, identity, purity, or potency*.' The current ICH guideline [3] echoes this point of view by defining '*comparable*' as '*a conclusion 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.*' The body text of the latter guideline further stipulates that '*The demonstration of comparability does not necessarily mean that the quality attributes of the pre-change and post-change product are identical, but that they are highly similar and that the existing knowledge is sufficiently predictive to ensure that any differences in quality attributes have no adverse impact upon safety*

or efficacy of the drug product.'

The remainder of the ICH guideline (and its predecessors) mainly emphasizes the importance of sensitive analytical technologies to determine whether or not physicochemical differences are present; however, the issue of how to assess the results of these analytical methods, receives only limited attention and is mostly left to the assessor to be done on a case-by-case basis [4].

This concept of comparability constituted a pragmatic and flexible approach for dealing with manufacturing changes (and more recently, biosimilarity). It recognised that biologicals are inherently variable and that minor variations in quality attributes are often clinically irrelevant. The concept of comparability allowed the implementation of necessary manufacturing changes without the undue requirement that biological products had to be physico-chemically identical and without an impractical requirement to conduct comparative clinical studies. The same concept also opened up the possibility to develop biosimilars, products from a second manufacturer deemed 'comparable' to an originator. It is a firm EU position that both forms of comparability are conceptually the same, and the Guideline on Similar Biological Medicinal products [5] explicitly confirms that 'the scientific principles of such a biosimilar comparability exercise are based on those applied for evaluation of the impact of changes in the manufacturing process of a biological medicinal product (as outlined in ICH Q5E).' This commentary explicitly includes comparability as applied to biosimilars unless otherwise stated.

The development of biosimilars and associated regulatory

---

* Corresponding author.
E-mail address: rm.vd.plas@cbg-meb.nl (R.M. van der Plas).

experience has triggered the need to become more precise on the concept of comparability and the regulatory requirements.

The complexity of the subject was highlighted by the extensive discussions on the draft EMA reflection paper [6], and by the fact that the FDA first published, then withdrew [7] and subsequently re-published draft guidance on the topic. It appeared that regulatory practice which has been in place for decades is criticized, and concepts that are better known from the *clinical* evaluation of medicinal products– such as equivalence testing – were suggested to be introduced in the setting of the evaluation of biosimilars. This commentary aims to reflect on the current status of assessment and proposes new directions, to make the point that before engaging in methodological discussions and discard current practices that seem to have served us well, the actual concept of comparability needs to be clarified and made more precise.

1.1. Reflections on current practice

Pharmaceutical Quality assessors employ a flexible case-by-case approach to comparability, and over time certain pragmatic rules have emerged and are being applied. Broadly, assessors do not formally assess whether attributes are 'statistically equivalent'; in fact, such an assessment is even perceived to be at odds with the statement in ICH Q5E that '*The demonstration of comparability does not necessarily mean that the quality attributes of the pre-change and post-change product are identical, but that they are highly similar and that the existing knowledge is sufficiently predictive to ensure that any differences in quality attributes have no adverse impact upon safety or efficacy of the drug product.*'

Historically, the assessment of comparability focuses heavily on the establishment of 'comparability ranges'. This is reflected in section 5.2 of the EMA quality biosimilarity guideline: '*Quantitative ranges should be established for the biosimilar comparability exercise, where possible.*' It is also reflected in a recent draft Q&A of the WHO on biosimilars, which states: '*Statistical methods are usually dealing with means whereas the analysis of quality data in the context of comparability is often based on acceptable ranges. The means may change within the acceptability range. Furthermore, working with probabilities, like confidence intervals is problematic as it is expected that each batch of the product will be in the pre-defined range*' [8].

Particularly the last point is worth emphasizing: It is expected that both in the manufacturing process of the originator, as well as in that of the similar, batches that are not "within specified range" are rejected and do not reach the patient. To quality assessors these ranges have a distinct meaning representing a clinically qualified range of this attribute. Any batch released within this range is expected to have a "similar" clinical effect and differences within these ranges are assumed not to have a clinically relevant impact on safety or efficacy. It is our point of view that identifying this core assumption is extremely important; it presumes that variability within a comparability range does not have a clinically relevant effect and that all batches which fall within this range are equally efficacious and safe.

Importantly, the "within specification" ranges, as applicable to the originator product, are usually based on the observed variability in quality attributes resulting from the originators manufacturing process. Since these specifications have been accepted, all product resulting from these processes are acceptable and those outside of these ranges (and thus outside "observed variability") are not.

Therefore, quality assessors often intuitively accept min-max ranges (and sometimes tolerance intervals or XSD ranges) because min-max ranges are expected to converge towards, but not exceed, the expected specification interval. Quality assessors also have (in a biosimilarity exercise setting) little difficulty to accept shifts and changes within ranges of an originator, because the whole range is deemed safe and efficacious. This assumption could in certain cases be questioned, but it is important to notice that this is a common (and often implicit) assumption for any comparability exercise.

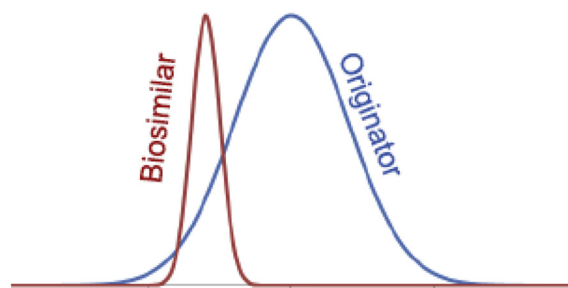Fig. 1 represents a theoretical distribution of a quality attribute (e.g.

Fig. 1. Most Quality assessors would accept that the 'red' product is indeed (for this attribute) comparable to the originator, because all the 'red' data points are within the 'blue' originator range. This despite the fact that the distributions (mean and variance) are different.

deamidation, aggregates, potency). Most Quality assessors would accept that the 'red' product is indeed (for this attribute) biosimilar to the originator, because the 'red' data points are within the 'blue' originator range, although the distributions (mean and variance) are clearly different. The key underlying assumption for this approach is that a patient is treated with a specific batch which represents a data point within the total range of the originator; that all these batches perform equally well in clinical terms; and that all these batches/data points therefore represent acceptable quality.

From this viewpoint and associated assumptions it is not surprising that both the current [9] and the withdrawn* [10] draft FDA guidance for biosimilarity accept the use of 'ranges'. These draft guidance documents suggest the use of a Quality Range (in the form of a XSD interval which should contain 90% of the data points, where the X depends on the estimated risks involved), thereby seemingly codifying an existing practice and suggesting that 'being within range' would be the most common criterion. However, the current version assumes that sponsors target the same mean as the originator, thereby indicating Fig. 1 as an *a priori* undesirable situation. Biosimilarity in the current draft guidance is understood as similar mean and standard deviation in the population. If then the same mean is targeted, the Quality Range can be used to verify that an attribute, as observed in the proposed biosimilar and the reference product, has a similar population mean and similar population standard deviation. This approach, although conceptually valid, is at odds with both the ICH Q5E endorsed regulatory practice, and with the original intent of the Quality Range.

Importantly, we believe that the proposed Quality Range criterion or similar range based criteria ignore the existence of other criteria, which are routinely applied in comparability exercises. Roughly speaking, these criteria are often framed in terms of 'comparability ranges' too, and may comprise the following:

- An attribute is not within range, but scientific knowledge supports that the attribute is not clinically relevant. For example, charge heterogeneity or C-terminal Lysines of MAbs. Currently, such a justification is often presented when a difference is found. However, from a methodological point of view, it would be preferable to predefine whether an attribute is clinically relevant -or not.
- An attribute is not within range, but the actual values are 'better' than those for the originator. For example, aggregate levels in MAbs are often lower for biosimilars, due to technical improvements in manufacturing over the last decade. The EMA biosimilar quality guideline [11] (section 5.2) states with respect to process-related impurities: '*Differences that may confer a safety advantage (e.g. lower levels of impurities) should be explained but are unlikely to preclude biosimilarity*', and this scientific principle is also sometimes applied

---

* Although this guidance has been withdrawn, it provides a relevant example of evolving regulatory thinking in this area.

to product related attributes. Again, from a methodological point of view, it would be preferable to predefine if a relative increase or decrease in levels can be anticipated and whether it is considered beneficial or not.

- An attribute is present at such a low level, that this level is considered clinically irrelevant. The EMA biosimilar quality guideline (section 5.2) states in this respect: '*In contrast, process-related impurities may differ [qualitatively] between the originator and biosimilar products, although these [impurities] should be minimised.*' Although this scientific principle was originally developed for process related impurities, it is sometimes also applied to product related low level attributes, e.g. certain minor glycosylation species. Again, from a methodological point of view, it would be preferable to explicitly define 'low level'.

Finally, some type of data are more amenable to an 'expectation' approach, where results should be equal to an expected value taking into account analytical variability (atomic mass by MS is an excellent example); or to visual examination (spectrograms and chromatograms).

1.2. Moving forward

For almost two decades, this approach towards comparability assessment constituted a workable situation for both regulators and industry. However, the advent of biosimilars has led to renewed interest in the establishment of more rigorous, statistically robust criteria [12]. There is no controversy that indeed a more rigorous approach is welcome, and that statistical methods and sampling strategy considerations would be an essential part. Unfortunately, two factors complicate the development of such a formalised, rigorous approach:

Firstly, statistical evaluation (by confidence intervals, hypothesis testing or otherwise) requires a clear, operational definition on what constitutes similarity, *independent of the samples at hand*. If statistical inference methods like statistical equivalence testing are considered, which attribute of the distribution they are applied to (difference in mean, variance, or combinations) should be driven by such an operational definition. It is not *a priori* clear that statistical equivalence testing of means provides the most adequate approach to evaluating similarity. The original FDA approach as described above [13,14] uses a statistical derivation (1.5 times the SD of the originator's values for a certain parameter), which reflects manufacturing consistency of the originator but is therefore not independent of the samples at hand. In our view, these calculations bypass the assumption that differences observed should be clinically irrelevant. For example, for protein content (strength) a specification of 90–110% of label claim is commonly accepted because variations of 10% around target are not expected to have an impact on clinical outcome. However, protein content can be precisely controlled during manufacture and the actual SD may be much lower. Which limit should the biosimilar comply with; the one statistically derived from manufacturing consistency, 10% around label aim, or some compromise between these two? And if biosimilars are supposed to comply with the more stringent manufacturing consistency derived limit, then this begs the question if the originator should also comply with this more stringent one, effectively making the 90–110% specification null and void.

Secondly, rationally, the sampling strategy would have to follow the concept and definition of similarity, the matched statistical approach and the required confidence (acceptable uncertainty). Sources of variation would be a key ingredient (batch-to-batch, within batch, analytical), hence simple random sampling is unlikely to be optimal. However, practical limitations to sampling and sample size constitute a major road block towards statistical assessment. Sample sizes are often small; three post-change batches are the norm in a manufacturing change comparability exercise and 'a dozen' biosimilar batches may already constitute extensive manufacturing experience. In addition, a biological product has multiple quality attributes, leading to multiplicity issues in a formal statistical analysis.

Although the number of pre-change or reference batches may be much higher, several issues remain. Whilst US FDA suggests that 10 batches may suffice, > 30 batches is often feasible, and European experience suggests that such a number may be necessary for a reliable estimate of the true range, especially if an attribute is highly variable. However, these batches may not be truly independent; e.g. two drug product batches may be manufactured from one drug substance batch. This is especially a point of attention if a biosimilar company investigates reference medicinal products taken from the market.

Finally, it is often unclear whether the relevant components of variation can adequately be represented (between batch, within batch, assay), and thus actual variation is difficult to estimate. This is specially a problem for 'extended characterisation' as performed for manufacturing changes; the number of pre- and post-change batches (data points) may then be as low as three plus three.

Restrictions in available knowledge, sampling structure and sample sizes as indicated above make a formal statistical evaluation problematic. For these reasons, the Min-Max range and calculated ranges like 3SD or Tolerance Intervals, became the *de facto* decision tools of choice for quality assessors dealing with comparability. Formalised statistical testing, especially equivalence testing, has never seriously been considered as a tool in comparability exercises. Its recent popularity is therefore remarkable, and seems more based on the use of equivalence testing in the setting of clinical trials for efficacy. It has not yet been adequately investigated and engaged to comparability exercises and Quality data, although first attempts have been made [15–17].

Furthermore, codified guidance should also address two related issues:

Some form of risk-ranking/filtering of Quality Attributes resulting in a tiered approach, is necessary to create order in the chaos of enormous data sets. Although fingerprint approaches may deliver detailed data, such detailed data can only be interpreted with some tool to discriminate between critical and non- or less-critical quality attributes. However, it is debatable whether such a tiered approach should contain two (critical or non-critical), or three discrete tiers, or should be more continuous.

Finally, the measurement of quality attributes is based on sampling from batches. Thus, sampling strategy, within batch variability and analytical variability do potentially impact the variation of manufactured product that will reach the patient as well. The issue of analytical variability is well-recognised and especially relevant when results of biological assays are to be interpreted; the variability of physicochemical methods is usually less problematic. Other sources of variability are less well understood.

2. Concluding remarks

In summary, comparability exercises rely on a number of pragmatic criteria. These criteria have been remarkably robust for 20 years of comparability exercises; however, the increased scrutiny of biosimilar applications provides an impetus for both codification and improvement of criteria for establishing comparability. This codified guidance should also comprise some form of risk-ranking/filtering of Quality Attributes. Such a more rigorous, methodologically sound, approach towards comparability seems both feasible and beneficial.

Disclaimer

The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of, or reflecting the position of, the MEB, the EMA or one of its committees or working parties. The article is intended to initiate regulatory discussion and should not pre-empt any future regulatory decision.

Funding

References

[1] Chirino AM, Mire Sluis A. Characterizing biological products and assessing comparability following manufacturing changes. Nat Biotechnol 2004;22:1383–91.

[2] US Food and Drug Administration. Guidance concerning demonstration of comparability of human biological products, availability. Fed Regist 1996;61:18612–3 https://www.govinfo.gov/content/pkg/FR-1996-04-26/pdf/96-10426.pdf, Accessed date: 2 October 2019.

[3] International Council on Harmonisation. Q5E Note for Guidance on Biotechnological/Biological products subject to changes in their manufacturing process. https://www.ema.europa.eu/documents/scientific-guideline/ich-q-5-e-comparability-biotechnological/biological-products-step-5_en.pdf; 2005, Accessed date: 2 October 2019.

[4] European Medicines Agency. EMA/CPMP/BWP/3207/00 rev1 Guideline on comparability of medicinal products containing biotechnology-derived proteins as active substances: quality issues. http://www.triskel.com/1%20Guideline%20biotech%20derived%20proteins.pdf; 2003, Accessed date: 2 October 2019.

[5] European Medicines Agency. CHMP/437/04 rev1 Guideline on similar biological medicinal products. https://www.ema.europa.eu/documents/scientific-guideline/guideline-similar-biological-medicinal-products-rev1_en.pdf; 2010, Accessed date: 2 October 2019.

[6] European Medicines Agency. Workshop on the reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development. https://www.ema.europa.eu/en/events/workshop-reflection-paper-statistical-methodology-comparative-assessment-quality-attributes-drug; 2018, Accessed date: 2 October 2019.

[7] US Food and Drug Administration. FDA withdraws draft guidance for industry: statistical approaches to evaluate analytical similarity. https://www.fda.gov/drugs/drug-safety-and-availability/fda-withdraws-draft-guidance-industry-statistical-approaches-evaluate-analytical-similarity; 2018, Accessed date: 2 October 2019.

[8] World Health Organisation. WHO Questions and Answers: similar biotherapeutic products. https://www.who.int/biologicals/expert_committee/QA_for_SBPs_ECBS_2018.pdf; 2018, Accessed date: 2 October 2019.

[9] US Food and Drug Administration. Development of therapeutic protein biosimilars: comparative analytical assessment and other quality-related considerations DRAFT GUIDANCE. https://www.fda.gov/media/125484/download; 2019, Accessed date: 2 October 2019.

[10] US Food and Drug Administration. FDA: approaches to evaluate analytical similarity guidance for industry DRAFT GUIDANCE sept 2017. https://www.pbwt.com/content/uploads/2018/06/UCM576786.pdf; 2017, Accessed date: 2 October 2019.

[11] European Medicines Agency. EMA/CHMP/BWP/247713/2012 Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues (revision 1). https://www.ema.europa.eu/documents/scientific-guideline/guideline-similar-biological-medicinal-products-containing-biotechnology-derived-proteins-active_en-0.pdf; 2014, Accessed date: 2 October 2019.

[12] European Medicines Agency. EMA/CHMP/138502/2017 Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development (draft). https://www.ema.europa.eu/documents/scientific-guideline/draft-reflection-paper-statistical-methodology-comparative-assessment-quality-attributes-drug_en.pdf; 2014, Accessed date: 2 October 2019.

[13] Yi Tsong, Dong Xiaoyu, Shen Meiyu. Development of statistical methods for analytical similarity assessment. J Biopharm Stat 2015. https://doi.org/10.1080/10543406.2016.1272606, Accessed date: 2 October 2019.

[14] Chow SC. On assessment of analytical similarity in biosimilar studies. Drug Des 2014;3:e124. https://doi.org/10.4172/2169-0138.1000e124, Accessed date: 2 October 2019.

[15] Burdick Richard K, Neal Thomas, Cheng Aili. Statistical considerations in demonstrating CMC analytical similarity for a biosimilar product. Stat Biopharm Res 2017;9(3):249–57. https://doi.org/10.1080/19466315.2017.1280412.

[16] Mielke J, Innerbichler F, Schiestl M, et al. The assessment of quality attributes for biosimilars: a statistical perspective on current practice and a proposal AAPS. https://doi.org/10.1208/s12248-018-0275-9; 2019, Accessed date: 2 October 2019 J 21: 7.

[17] Stangler and Schiestl. Similarity assessment of quality attributes of biological medicines: the calculation of operating characteristics to compare different statistical approaches AAPS Open 5:4. https://doi.org/10.1186/s41120-019-0033-9; 2019, Accessed date: 2 October 2019.