November 8, 2017

Submission of comments on 'Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development' (EMA/CHMP/138502/2017)

Comments from: AAPS Biosimilar Focus Group

Name of organisation or individual

Please note that these comments and the identity of the sender will be published unless a specific justified objection is received.

When completed, this form should be sent to the European Medicines Agency electronically, in Word format (not PDF).

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

1. General comments

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	We very much appreciate EMA opening up discussion on this important topic. We also acknowledge that the document discussing statistical methodology for comparative assessment of quality attributes is presented as a reflection paper, and is therefore not intended to offer practical guidance to drug developers. We hope that the desired outcome of these activities will result in such a guidance, and offer some suggestions for your consideration. The current version of the paper emphasizes the limitations and difficulties associated with the use of inferential statistical methodology in this field, which may be received as dissuading stakeholders from using such methodology. We believe that appropriate inferential statistical assessment of comparative data will result in better decision-making, and should be promoted as a better option than some of the descriptive or qualitative approaches that are conventionally used today (some of which are discussed in the Reflection Paper). This goal may not be achieved unless the Agency offers practical guidance to industry on selection or development of fit-for-purpose statistical methodology.	
	The document should be solely concerned with the assessments described in Section 4.2 for Biosimilar products. The uniqueness of such assessments requires particular focus on this issues. There are several characteristics that distinguish demonstration of similarity between an originator reference listed drug product (RLDP) and a biosimilar test product (TP) from other assessments of comparability including the following: 1.Lack of RLDP knowledge in a similarity assessment translates into increased risk in any statistical analyses based on the RLDP data, arising from such issues as:	

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	a. RLDP data may include intentional process changes and/or unintentional (but within specification and expected for the process (due to campaign effects, etc.)) process shifts which may make pooling of data inappropriate for statistical analysis, b. RLDP process deviations resulting in quality within permitted specifications but outside expected variability, inclusion of which inappropriately inflates the RLDP variability in statistical analyses. For example, a sampled RLDP lot may have a measured value that is out of trend with respect to other RLDP values, even if the release of the lot was justified based on impact to quality, safety, and efficacy. c. Linkage between drug substance (DS) and drug product (DP) lots is not identifiable from sampled RLDP lots in similarity assessments. If sampled DP lots were manufactured with the same DS, they are correlated, and the assumption of independence required in many statistical calculations is not appropriate. 2. RLDP target specifications and in-process control (IPC) limits are not known for the majority of the analytical methods in a similarity assessment. This lack of knowledge makes the selection of meaningful acceptance criteria more difficult, and comparisons based on originator process capability are impossible. 3. The sampling process used to collect the RLDP lots has an inherent bias that leads to RLDP lots being generally older than newly manufactured TP lots. This bias is especially problematic for stability indicating methods. 4. In addition, for stability indicating methods, the time since	

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	manufacturing of the RLDP lots can only be roughly estimated, making it difficult (at best) to account for changes in the quality attribute during shelf life. This will cause potential bias in mean estimates and/or inflation of variability estimates for the RLDP data. These issues generally will not exist for the TP, because the biosimilar manufacturer has complete knowledge of the TP lots used for statistical analysis. Thus, any comparisons between the RLDP and TP are subject to increased statistical risk in assessments of similarity, due to the imbalance in information between the two sets of data.	
	There have been several statistical methods proposed for demonstrating analytical similarity in the literature. These strategies include: 1.Comparison of means and/or standard deviations individually based on equivalence or non-inferiority testing. 2.Comparison of properties of the entire process distribution. a. Percentage overlap of the TP and RLDP distributions. b. Bayesian predictive probability distributions as proposed by the European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) Working Group. 3. Comparisons of individual values and/or ranges of TP and RLDP based on heuristic rules of similarity. It would be a useful exercise for EMA to review these methods based on the following set of criteria for a statistical test for analytical similarity: 1. Protect patients from consequences of concluding similarity when products are not similar. 2. Protect sponsors from consequences of concluding lack of similarity when products are in fact similar (the consequences also include a lack of patient access to lower cost treatments) 3. Incentivize sponsors to acquire process knowledge sufficient to	

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	understand similarity of the TP to RLDP. 4. Enable decision making with small samples, while rewarding larger samples with lower patient and/or producer risk. 5. Examine entirety of the process distribution of product. 6. Demonstrate robustness to violations of assumptions (e.g. Normality, independence). 7. Be easy to understand and interpret. 8. Define acceptance criteria that are scientifically relevant and consistent across sponsors of the same RLDP. It is realized that no single approach will necessarily satisfy all the criteria listed above, but approaches meeting more of them should be preferred over those meeting fewer.	

2. Specific comments on text

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
190-203		Comment: As alluded to in the current document, a "well controlled" process does not necessarily remain constant over time, and thus does not necessarily infer a single distribution. This has important implications for the use and interpretation of inferential statistical methodology. Because of non-random use of sources of variability (e.g. starting/raw materials), a "consistent" process can be the combination of multiple sub-populations, and results will not be independent and identically distributed. This does not preclude the use of statistical comparisons, but requires care in applying and interpreting the statistical methodology. Proposed change: add description of this characteristic of typical manufacturing and cautionary advice regarding implications, such as the text above.	
252-255		Comment: It may be that frequently for pre- to post-change comparisons the data from a few post-change batches are "taken as single values and compared to 'data-ranges' …" from the pre-change process, but this is not a statistically valid or recommended approach. Proposed change (if any): Either delete or include a caution that this is not valid and note the many options for statistical inference in the next paragraph.	
258		Comment: Readers may misunderstand this line to mean	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		statistical intervals are not useful. Proposed change: Reword to suggest they are often misused, and encourage sponsors to ensure that use of these intervals meets the criteria noted above (see 2nd general comment).	
288-293		Comment: It is true that the impact of differences in QAs on clinical outcomes is typically not well-understood. At the same time, as specifications often are set based on process performance rather than clinical relevance, it is inconsistent reasoning to have approved the reference product specifications based on manufacturing variability (process performance) but use a lack of clinical relevance as a reason to not use reference product variability as the basis for an equivalence margin for a biosimilar. Proposed change (if any): Note the lack of understanding of the impact of QA differences on clinical performance, but (1) acknowledge the underlying lack of understanding of the relevance of the reference product specifications to clinical performance, and (2) note that this is a reason that well-thought out statistical methodology for comparability is critical to ensure risk-based decision making by regulators. Also note that where clinical impact is understood, this is preferred as the basis for setting comparability acceptance criteria.	
327-331		Comment: Bioequivalence statistical criteria (two one-sided test (TOST) of the ratio of means within 80%-125%) are	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		insufficient for demonstrating biosimilarity or equivalence/non-inferiority of a process post-change, etc. This methodology focuses on comparison of the average performance only. In order to determine whether two populations are equivalent (non-inferiority being a subset of equivalence, as noted in the document), it is insufficient to simply test differences in the mean. Further, separate tests of the means and standard deviations or variances do not address this shortcoming well, because (1) they focus on comparability of statistical parameter estimates rather than comparability of *results*, and (2) they can lead to confusion and misinterpretation (thus increasing patient risk) when one test meets a criterion and the other does not. Proposed change (if any): Note the shortcomings of focusing on tests of differences in means with or without a separate test for increase in variability. Discuss the fact that numerous statistical methods exist which can be used to quantify the distance between or overlap of two distributions – tolerance intervals, Bayesian inference, the overlap statistic or proportion of similar response (OVL or PSR), and non-parametric methods such as Kolmogorov-Smirov or Mann-Whitney U tests. The development of one or more of these for comparability of QAs should be encouraged and welcomed by regulators. Proposed criteria for comparing these methods is offered in our general comments.	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
367-370 (related to 535-537)		Comment: Disagree that single observed values being used for a comparative analysis is "not wrong." Perhaps this is meant just for the case of comparing release tests to specifications (?), lines 535-537. It is confusing to have this statement in a paper that is primarily concerned with comparisons of processes/products as a whole. Proposed change (if any): Delete this discussion.	
379		Comment: The word "that" was repeated twice.	
391-393		Comment: Specifications are not always derived from sample data. Proposed Change: Change "from the fact" to "when"	
399-401; 420-426; 461-482 (also see next row); 505-513		Comment: This language is repeated in part or in whole numerous times throughout the document. Proposed change (if any): Make a section toward the front of the document that *briefly* notes the underlying assumptions/requirements for statistical inference	
461-472		Comment: With regard to sampling, true random sampling is not only often infeasible, it is quite often not *optimal/ideal* for biopharmaceutical processes because true random sampling for such processes often cannot guarantee a *representative* sample – the true requirement – as well as stratified random sampling or even systematic sampling. Proposed change (if any): "pseudo-random" sampling is	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		mentioned in 480 as a possibility, inferring limited application. It is actually common to be the most appropriate sampling strategy. Note the frequent need for stratified or systematic sampling and emphasize representativeness at least as much as randomness, with references (Bergum or Wheeler).	
520-521		Comment: A min-max range is not truly a statistical interval, at least not comparable to prediction or tolerance intervals, as it does not incorporate any quantification of statistical uncertainty Proposed change (if any): If min-max ranges are included, it should be noted that they do not incorporate any statistical uncertainty and cannot be used for inference. They are merely descriptive of the sample.	
539-559		Comment: the limitations described for TIs apply to all statistical intervals derived to assess uncertainty. Proposed change (if any): in reference to so-called deficiencies, include other intervals such as confidence intervals, or merely describe more generally as "statistical intervals." Also note that these are not really deficiencies of the intervals, but characteristics of them that must be fully understood in order to properly use and interpret them.	
555-560		Comment: it is not a "methodological deficiency" and understanding the effect of sample size can result in suitable TIs	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		Proposed change (if any): remove this description	
585-587		Comment: This is actually rather difficult or impossible for TIs using frequentist methodology – with more than one level of variability, must use Bayesian techniques to calculate TIs Proposed change (if any): Note need for/opportunities for Bayesian inference throughout the document, highlighting applicability in some places such as this. Also, as the complexity increases, no matter whether the methodology is Bayesian or frequentist, must note the need for a statistician to be involved.	
599-604		Comment: As noted above, this is ideally true, but often not possible due to lack of clinical relevance of the original (pre-change or reference product) specifications. Thus, while any statistical textbook will say that criteria should not be based on the data, it is almost always impossible to do anything else. Arbitrary criteria such as mean difference 80-125% or biosimilar/post-change data within mean +/- 3SD are no better at making valid decisions and protecting patients (and are often worse) just because they are not data-based. Proposed change: Delete or note that what is described is an ideal approach to setting criteria, but that this is rarely possible in the context of pharmaceutical QAs. Also note that heuristic (non data-based) criteria are not any better at relating a conclusion of comparability to clinical outcomes, and may be worse because they do not even take into account	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		what is known about the already approved product.	
599-607 and 611-625		While it is ideal to select an acceptance range based on the "maximum allowed distance", for most QAs there is rarely an objective criterion upon which to base this distance because the clinical effect of changes in QAs is rarely known. There are currently "agreeable standards" for many QAs that have no scientific basis. The authors seem to suggest that this "arbitrariness" from "agreeable standards" is preferred to using sample data. Certainly using the sample data provides a foundation that is not arbitrary and allows a comparison against historical data that is presumably from an effective drug. It seems that using non-arbitrary sample data would allow a stronger conclusion of equivalence/similarity. We believe it would be a worthy effort to develop universal criteria that is based on the realities of manufacturing practice. Effect size is one useful metric for this purpose. Although such criteria may lack scientific meaning, they provide a universal standard that can be used across many attributes. As such, they would serve a purpose, much as the 80-125 ratio used in bioequivalence testing has served the community for many years.	
746		Comment: Is the use of "speed" synonymous with "volume"?	
778-781		Mathematically we can say being "smaller" is not equivalence. However, practically speaking, we actually want biosimilars to be either equivalent to or better (if better can be defined)	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		than the reference product. Also, it is hard to conclude that biosimilars have smaller variance than the reference product due to limited sample size and limited knowledge about the representativeness and independence of the reference product sample, even if the sample estimate of variability for the biosimilar turns out smaller than for the reference. Proposed change: An equivalence test on the ratio of variances should be one-sided and not two-sided, especially as technologies improve and provide potentially greater control over the process and assay. If the biosimilar sponsor can develop a less variable process or less variable assay, that should be rewarded and not penalized. The only worry for the biosimilar would be a much higher variance which would be an argument for non-similarity.	
946		Comment: what is meant exactly by "reconsider the whole inferential statistical approach"? Proposed change: delete	

Please add more rows if needed.