tests is not about the sample size (as large as required to pass the test) but about the width of the equivalence interval (the ELs). When dealing with bioequivalence studies, the ELs often chosen are ±20% of the reference, which is acceptable within the context of the clinical testing variability. However, setting a universal limit is not possible for quality attributes because they are based on a great variety of analytical methods and parameters.

The most rigorous statistical approach to set ELs, as well as the other interrelated parameters used in inferential statistical tests such as Type I and Type II errors (α and β), is to rely on decision theory in combination with inference. Decision theory serves to determine these parameters in a way that corresponds to the problem at hand. The derivation of these parameters is based on an objective function (minimize a loss function or maximize its negative, the reward function) and in the hazard rate, which is the expectation of the objective function. The decision theory relies on the quantitative concept of utility, which, in many applications and particularly in the field of health and medical sciences, is hard to quantify. In the absence of universal limits and decision theory-derived parameters, a third approach consists of setting limits on the basis of experimental data variability and employs the reference interval (RI). RI depends only on the standard deviation of the reference samples (= square root of the reference variability). Common applications of tolerance intervals for comparisons also use RI as a criterion.

Therefore, it makes sense to use RI as EL, with the justification that RI fixes the EL on the basis of the variability, sometimes expressed as a coefficient of variation (CV), which is an intrinsic summary characteristic of the concrete system.

Since power is not a consumer's risk, when assessing a formal test of equivalence, regulators are *a priori* less interested in the test's power. As for the biosimilar manufacturer, its best interest is to increase the power of the TOST since it will increase the probability of demonstrating the equivalence of the products. The most important and feasible way to increase power is to increase sample size. There is no consensus about what should be the sample size unit: several batches or repetitions (more than one sample of the same batch), respectively, reflecting the "between-batch variability" (σ^2_b) and the "within-batch variability" (σ^2_w). As the total variability σ^2 shall be the sum of σ^2_b and σ^2_w, both several batches and/or several repetitions per batch could be used with a repetition as a sample size unit. However, considering the significance of the "between-batch variability" (σ^2_b) for assessing similarity, it can be stated with some certainty that several batches of each reference and test product should be tested. Based on the suggestion from the stability guidelines (ICH Q1A(R2), Stability Testing of New Drug Substances and Products), it is assumed that three batches per product would be the minimum requirement to assess the between-batch variability. Power is also highly dependent on the variability of the methods and drops sharply when CV is close to the EL. Thus, for the biosimilar manufacturer, there are two not mutually exclusive options for increasing the probability of demonstrating equivalence: either by increasing the sample size and preferably the number of different batches or by reducing the variability of the methods used to assess the QA. The equivalence tests (e.g., TOST) thus appear to be the appropriate approach for demonstrating similarity between the biosimilar and its reference and does not suffer the deficiencies that were identified in the TI methodology. Indeed,