Cleaning Memo for January 2021
Statistics for Cleaning Validation?

There is an increasing call for the use of statistics for use in cleaning validation programs. Statistics can be appropriately used to evaluate data to determine variability. Statistics conceivably could be used in the following situations related to cleaning processes: Statistics for evaluating data in the design of a cleaning process.

A. Statistics for evaluating data in the *design* of a cleaning process
B. Statistics for evaluating data in a *qualification* protocol *in relation to meeting acceptance criteria*.
C. Statistics in the evaluating data in a *qualification protocol relating to the robustness* of the cleaning process.
D. Statistics in the evaluation of data in the *routine monitoring* of a validated cleaning process.

In this Cleaning Memo I will *not* cover the issue of statistics for the design of the cleaning process (item "A" above).

First, let me clarify that I am not a statistician.

Second, my favorite pharmaceutical statistician is Lynn Torbeck, and my favorite quote from him is the following from his 2011 *PharmTech* article "The Role of Statistical Significance Tests":

> "Practical significance comes from comparing a difference (i.e., a signal) to an absolute reference. Statistical significance comes from comparing a difference to a relative reference that contains noise or random variability. Practical significance always takes precedence over statistical significance. In fact, statistical significance should not be checked until practical significance is found."

As I understand this, it says that if I look at data and the variation is not practically significant, there is little value in determining statistical significance. This idea will be explored later in several examples.

Third, the value of statistics comes from having a significant number of data points *from the same population*. Why I state this is that I question the value of statistical measures like a mean and a standard deviation for data that is not the same population. The most common example of this is evaluating multiple swab locations in given equipment, averaging them, and determining the standard deviation (or another measure of variability). Unless the swab locations are the same population, that doesn't make sense. And in many (if not most) cases, we have selected our swab sample not on a statistical basis (such as one swab sample for every XXX square centimeters of surface area), but rather because those sampling locations represent different *worst case locations* (that is, locations most difficult to clean or most likely to have higher levels of residues if cleaning is marginal).

Now I have heard it said that the FDA in its Process Validation Guidance (2011) calls for use of statistics multiple times. And that is true (I counted 15 times in that guidance). But

in many cases those statements on the use of statistics are modified by phrases like "where appropriate", "where feasible and meaningful", and "whenever appropriate and feasible". I clearly agree that the *principles* of that FDA guidance for process validation (PV) should be applied to cleaning validation (CV), because CV is another type process. However, what is done for sampling in a manufacturing process is significantly different from what is done for sampling in a cleaning process. In a manufacturing process, I set specifications (usually a range) for the concentration of the API in the drug product. That specification is *not a limit*, but is a *target* that I *must* achieve, and must achieve *consistently*. When I sample drug product in PV, I expect to get the *same* values for API throughout the manufacture of one batch, as well as in other batches. Therefore the drug product has a consistent amount or concentration of API in each unit. So clearly in that situation I am sampling the same population and should evaluate the data statistically.

Let's take a look at what happens in qualification protocols for CV? We'll cover *both* determining whether we pass (or fail) *and* determining the robustness of our validated cleaning process. For a CV protocol, I also set a specification for the API in each swab sample (assuming I am doing swab sampling). However, that specification is *not a target* (or if it is a target, it is a target I am trying to miss!). No, unlike PV the specification is an *upper limit* (an upper value) that I want to be *below*. So the additional question about statistics in this situation is *how far below that limit do I want to be* (how robust do I want the process to be). And the answer is that while I want to be *consistently below that limit*, how far below depends on how much *robustness* I want in my cleaning process. I might have a limit of X per swab sample, in one situation my swab values are ranging from 0.2X to 0.5X. If these are truly different swab sampling locations, I should be able to accept that variation, because I have demonstrated the cleaning process gives swab results consistently below my limit. Now I would prefer to have a process where all my swab samples gave values in the range of 0.1X to 0.2X, with lower values and with less variation. But if it were me, I would validate that first process and then in a continuous improvement program, see if I can make changes to get lower and/or more consistent data by tweaking my cleaning process (an example of such tweaking might be changing my manual cleaning SOP such that more time was spent on scrubbing certain locations where higher sampling results were originally obtained). However, in both those cases with the practical significance of my results being *significantly below my limit*, what is the value of demonstrating *statistical* significance?

With any of this swabbing data, I can certain run statistics, but how would I *appropriately* use such statistical results since the sample locations are not the same population. Note that if I had taken seven *identical* locations on the equipment, then over three qualification runs it may be more applicable to use statistics. Suppose I used rinse sampling instead of swab sampling. The amount of sampling data I collect might be much less (one sample per equipment), so in a validation protocol I would have only three data points to treat statistically (unless I decided that rinse samples on *different* equipment items in a train could be treated together as the same population).

Now let's consider using statistics *for routine monitoring after a successful qualification protocol.* The main purpose of doing routine monitoring to confirm that the cleaning process is performing correctly. From an analytical point of view, this might include a final rinse sample for API by a specific method, a final rinse sample for TOC, a final rinse sample for conductivity, and/or one (or more) swab sample for the API. Let's address the *rinse sample issue first*. Obviously if the rinse sample involves testing the API by the same analytical method used in the qualification protocol, then if I exceeded the protocol limit (which I might retreat as my *action level* for routine monitoring) I have an unacceptable situation (assuming my OOS did not find the result invalid). What I am really looking for is (a) any value that might exceed an alert level or (b) any value which shows a clear trend that the process might be changing such that it might eventually result in exceeding my action level. This is a case where I could use statistics providing I had a *sufficient number* of runs to get anything meaningful in an evaluation to establish a mean value and a standard deviation. Based on that data, I might use the value of "the mean plus two standard deviations" as my *alert* level. The *action* level might be set at "the mean plus three standard deviations" or at my acceptance rinse limit. The key question is "what is a sufficient number of data points?" I might try to include data for any meaningful engineering runs, but I probably would like to have sufficient routine monitoring cleaning events to have a number such as twenty data points. On the other hand, I might set tentative action/alert levels until I achieve that number of runs that would give me more statistical confidence.

What if I am trying to do the same for swab sampling? As already mentioned, care has to be used to make sure the swab sampling locations are the same population. I can't (or shouldn't) just combine together ten different swab locations from a given equipment on three different qualification runs, and then conclude that I have thirty data points for a good statistical evaluation. Some may argue that if the cleaning process is robust enough, and on that basis they expect all swab samples to be <LOD (or <LOQ); therefore they would have sufficient data points from that number of sample location from three runs (assuming a total of thirty swab locations for the three runs) to be statistically meaningful. On the other hand, if my cleaning process was that robust and that was the data obtained, what is the point of doing a statistical analysis? I can just look at the data and conclude *practically* that my cleaning process is consistent and in a state of control. Furthermore, would I get upset or significantly concerned if on routine monitoring I got one sample location with a value slightly above the LOQ. Unless I consistently found the same location giving me a slightly higher value, I probably wouldn't spend too much time in an investigation; as the FDA states in it Process Validation Guidance, one "should *guard against overreaction to individual events* as well as against failure to detect unintended process variability" [emphasis added].

Some may object to my statement that the purpose of routine monitoring is to demonstrate that the cleaning process is in a state of control (or to discover that it is *not* in a state of control or to discover that it may be *trending* out of control). They would point to the 2018 EMA Q&A that an analytical evaluation (presumably for the API) should be done for routine monitoring to confirm that the HBEL is being met (unless justified by a

rigorous QRM evaluation). I have written on this elsewhere (July 2018 Cleaning Memo). However, this objective of assuring that the measured API values are *below* the limit set by the HBEL of the API probably does not require statistical evaluation. For a given routine monitoring of a cleaning process, the measure API is either above the HBEL-based limit or it is not. The EMA Q&A (at least as it is now written) does not consider statistical evaluation of measured values over multiple cleaning events. In that were the concern, then what is written above about routine monitoring as an indication of the state of control should apply. Furthermore, it probably should be the case that if the API is analytically measured every cleaning event as part of routine monitoring, then it is probably the case that it is not just the PDE/ADE limit that must be achieved, but that the cleaning validation limit should be achieved. The difference between a "HBEL limit" and a “cleaning validation limit” is something that is referred to in the EMA Q&A in Question #6, but still does not seem to be readily accepted (or understood) by many in the industry. But that is another story!!

Hopefully this Cleaning Memo will provide some insight into the appropriate (and inappropriate) use of statistics for evaluation of cleaning validation data.