Minutes

Data Integrity Work Group (DIWG)

Thursday, June 13, 2024

11:00AM-12:00PM

Meeting Materials Link

This meeting was recorded for internal use only to assure the accuracy of meeting notes.

Actions/Next Steps:

- ✓ Once they complete their paired sample analysis, ODU and DCLS will share this data with Elgin, and based on that Elgin Perry and Durga Ghosh will come up with a framework for other labs to use for moving forward.
- ✓ Elgin Perry will talk to Mike Lane about the data analysis.
- ✓ August will share a poll with the group to find a time for a DIWG meeting later in the summer.

Minutes:

11:00 AM Introductions, Announcements
Monitoring and Laboratory Analysis updates

Participants: August Goldfischer (CRC), Carol Cain (MD DNR), Cindy Johnson (VA DEQ), Durga Ghosh (USGS), Elgin Perry, Emily Young (ICPRB), Heather Wright (ODU), Jake Kilczewski (MDH), Jay Armstrong (DCLS), Jimmy Webber (USGS), Kristen Heyer (MD DNR), Lara Phillips (MDH), Laura Lockard (DE DNREC), Meg Maddox (UMCES), Meighan Wisswell (VA DEQ), Mike Mallonee (ICPRB), Nianhong Chen (UMCES Horn Point Lab), Renee Karrh (MD DNR), Samira Azemati (MDE), Sidney Anderson (UMCES), Suzanne Doughten (ODU), Tracee Cain (DE DNREC).

11:20 AM Coordinated Split Sample Program – *Mike Mallonee (ICPRB)*

Jay Armstrong (DCLS) said he sees divergence in the nitrite split sample results in December and in March. He looked into it for Division of Consolidated Laboratory Services (DCLS) and hasn't seen anything in their results to indicate that anything was off. All results are below the reporting limit. In March there is a split between all the labs. That may have to do with the high Total Suspended Solids (TSS) from that time and Jay will keep any eye on that. Mike said the replicates look good despite the extremely high TSS time period.

11:30 AM Lab comparison for instrumentation switch out – *Durga Ghosh (USGS) and Elgin Perry*

Durga shared that most of the labs have provided the requested information on their minimum requirements for protocol when making an instrumentation switch. She is waiting to hear back from a few of them. It appears that most will be making the instrumentation switch soon or already have. Some of the lab groups have started running the paired samples and have

collected adequate data as well. It appears there is not a lot of variation based on results so far. The number of required samples is not yet determined.

Elgin said that he did not receive information from the labs from Durga yet, but he can share his thoughts without having seen that. He said the split sample data replication is remarkably tight which is a good sign, and the lines are very parallel with little crisscrossing among lines. That suggests that differences among labs is due to random factors around calibration. That is going to be an issue as well with a comparison between old methodology and new methodology. It would be nice to design an experiment so that when comparing the old and new methods, part of the error term observed between the two methods captures this calibration component. One way to do that would be to have a lab recalibrate their device and run the same set of samples over again with both devices. Since there are multiple laboratories making the same change, it may be reasonable to assume that interlaboratory variability is the same as the calibration variability and that would help create an error term for comparing the old and new methods.

Jay Armstrong commented that with several of the analytes, the concentrations are at or below reporting limits. At that point, running a straight line, the equation is y = mx + b and b is the intercept and b is getting added if it's a positive or subtracted if it's a negative from the result. Getting down to concentrations 10ppb or lower, that will make a difference in what is being read. As the concentration ranges goes up it becomes less of a factor, and the factor that is the intercept from the curve may not even be noticed. Those intercepts will run pretty consistently around zero, though one run may be negative and one may be positive. It will depend on a lot of factors such as reagents. Calibration standards will be different. Jay said he's not sure how to capture that error and have it be consistent.

Elgin clarified that goal is not for the error to be consistent, just to be realistic. If 20 samples are run with one calibration curve with one device, and 20 samples are run with another device using a different calibration curve, and one gives a positive intercept and one gives a negative intercept, one might falsely conclude these two devices are giving different results. It may just be the intercept causing the difference between the two, rather than a true difference between devices.

Jay responded that he thinks that people in their studies will be doing more than one run. A Method Detection Limit (MDL) study is required, and part of an MDL study is a limit of quantitation and per EPA guidelines that requires a minimum of 3 runs. If samples are being run while doing the study and validation, there will be a minimum of 3 runs with different calibration curves for each run. If people are doing their parallels while doing their MDL studies that should get that variance. Jay said labs do what's required by the NELAC Institute (TNI) first, MDL studies and Limit of Quantitation (LOQ) verifications. Labs make sure it's in a clean matrix performing the way they expect. After they determine that then they'll go into parallel studies which are 3 or more runs; they don't put all the samples onto one run.

Suzanne Doughten (ODU) commented that if labs are doing something similar with SEAL and Lachat, when doing their comparisons they're using the same calibration curve on each instrument. They're running 30 samples on each instrument on the same day.

Jay said DCLS doesn't do it that way because they want to see the variability because they don't always have the same person running the test. He added he doesn't think either way is incorrect, however.

Elgin said from a statistics point of view that way of capturing the variability would be preferable, and asked for clarification if there are 30 samples, is that within each instrument?

Suzanne explained her lab is making up the standards. They're doing runs on separate dates, they're not running 100 paired samples on one day. Like Jay said they do MDL studies first, drift studies, reporting studies, etc. They do a lot before starting the comparison to make sure everything is working. They do analyze samples on the same day. Occasionally they have to re-run samples but that will be noted. Mike Lane is doing the analysis.

Elgin said it sounds like they do multiple calibrations with the full set of samples in both labs. Elgin said he would just ask they keep track of that so the calibration curve can be put in a statistical model as a random factor along with the sample-to-sample variability. That would enable a fair comparison of the two instruments.

Jay reminded the group of the main question which was whether 100 paired samples were needed or would 50 samples work? What is the size of these studies need to be going forward?

Elgin said that's a good question and has to do with what random factor is creating the biggest source of variability. Elgin said his guess is once an instrument is calibrated and runs a sample, if a sample were to be re-run the variability would be small.

Jay responded that would probably depend on the scale and absolute magnitude.

Elgin said if the process was repeated on a different calibration curve a bigger difference would be expected than if the sample was re-run on the same calibration curve.

Jay said he agreed although keeping in mind that was an assumption that would need to be tested.

Elgin said if that's true than the calibration variability is probably larger than the measurement tolerance within a calibration. From a statistical standpoint it's important to get more samples from different calibrations than getting more samples on the same calibration. Are the 100 samples run on one calibration? Jay and Suzanne both responded no.

Elgin said that measurement tolerance is how accurately you can reproduce the same results within one calibration. He thinks that the calibration variance is bigger than measurement tolerance and the level of replication is different for those two things. It would be necessary to have some numbers to estimate how many samples are needed of each type, for how much

replication is needed at each one of these random factors. The other approach is it sounds like some labs already have lot of data and Elgin could take a look at that to see any differences between the data from the two instruments.

Durga said that part of this discussion was that although ODU and DCLS have done some of these paired samples already, there are other labs earlier in the process of making this switch and want to know the best path forward. The question is what a unified approach will look like. All the labs are staggered when they are making the switch. Are there any minimalistic approaches that need to be considered when they make the switch even before they get to comparing the data? At this point there's no data that can be used to build models on. What they have is an estimate of the number of samples that would be impacted in the case of each lab, but they don't have any data. Knowing this and knowing that only two of the labs have done some analyses what would Elgin recommend as a path moving forward for labs trying to assimilate this into their work?

Elgin said he would recommend identifying what the random factors are to try and identify, and designing an experiment that gives some replicability on these random factors. Elgin said he sees three random factors: the measurement tolerance, which is how much variability when the same sample is run with the same calibration; calibration variability, which is variability when the same sample is repeatedly run on different calibrations; and inter-laboratory variability. The split sample data shows the inter-laboratory variability is pretty small. Ideally, an experiment would allow for estimating the variance associated with each of these random factors. How much replication is needed for each of those factors is dependent on how much variability each of those factors create, and that is hard to figure that out without data. Intuitively measurement tolerance would be expected to be a small component of error, calibration variance would be expected to be a larger component of error, and among-lab variability would be expected to be a larger yet component of error, but without data there is no way to verify that.

As far as a path forward, the sources of variability are known, so the data sets should be created so in the analysis process each of those sources of variability can be measured. Elgin recommended keeping track of when recalibrations happen and when samples are run on the same calibrations, which will help with quantifying each of those sources of variability. Then once data is available from labs who have already run the paired sample comparisons, Elgin can take the data and do some analysis to come up with estimates of how much replication is needed at each level. That could serve as guidance for labs coming along later.

Durga said that sounds reasonable. Having this in mind when the studies are conducted on both instruments before the switch is made would give some sort of an answer for going forward in terms of how many actual samples would have to be run in parallel. Durga said she thinks it can be done on a lab to lab basis but it would be nice if once data is obtained from ODU and DCLS they could share it with Elgin, and then based on that Elgin and Durga will come up with a framework for other labs to use.

Suzanne said that Mike Lane is working on the reports now.

Elgin said he would like to talk with Mike on what he's finding and share ideas on treating this as a nested random design and see if he thinks that's working.

11:50 AM Poll for scheduling next meeting.

12:00 PM Adjourn