

POTW Feedback on ILS2 Data Packet & Draft Study Report

Paul Bedore

Joint Stakeholder Advisory Committee & Expert Science Panel Meeting

September 6, 2023

POTW Research Interests & Need for Study

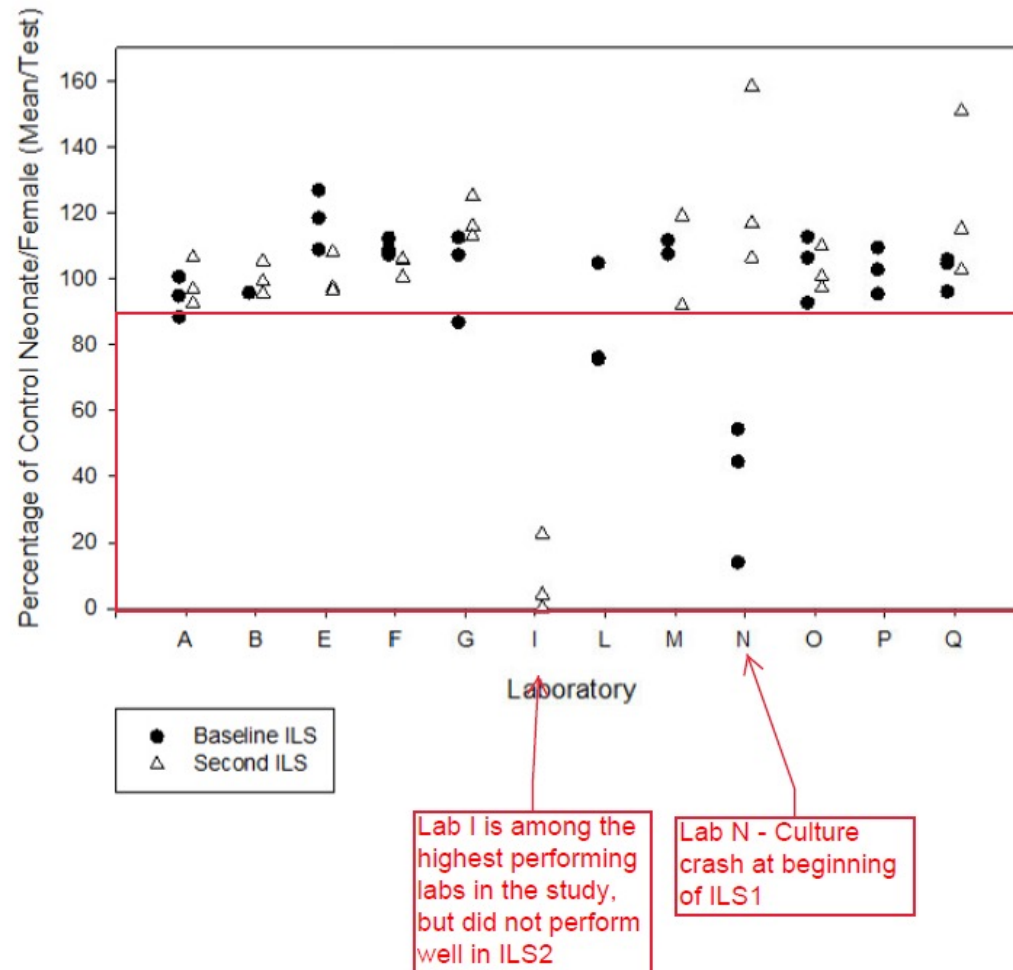
- POTW experience with low-level intermittent toxicity
- Led to split testing and observations that labs disagree whether samples are toxic
- POTWs were major proponent of a research study
- Study Goals:
 - Improve consistency of test method execution
 - Improve the consistency and comparability among labs
- POTWs – concerns with consistency and comparability
 - Factors to enhance labs correctly identifying non-toxic samples as not toxic
 - Identify factors that will enhance agreement among labs on split samples (not-toxic and toxic) –
 - Same answer from multiple labs
 - Study incorporating synthetic blank samples & spiked samples
- Summary of some key issues on following slides
- Additional detailed feedback in forthcoming report comments

Key Issues – Samples 1 and 2A

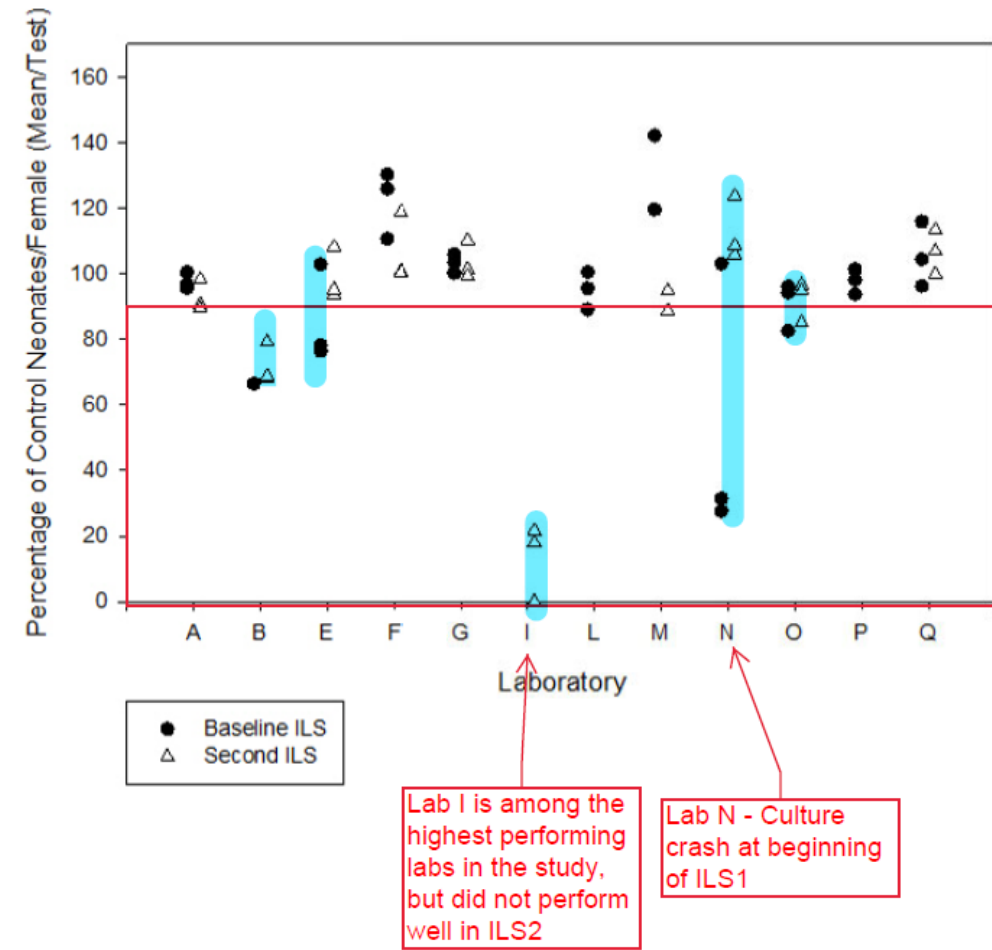
- Section 5.5, page 33 – Discussion of Samples 1 and 2A normalized by control reproduction
 - 10% effect is referred to as “toxicity” by this section – Analysis focuses only on “magnitude”
 - *“The only laboratory reporting toxicity consistently was Lab N...”*
 - **Request that the report statistically compare Samples 1 and 2A to their controls following EPA Test Method**
 - Accounts for magnitude and variability — conclusions can then be drawn about whether the effect magnitude (% control) or variability (CV) drive statistically significant differences
 - QAPP - references this statistical comparison, but not evaluating the data with the 10% effects threshold
 - Previous stakeholder comments requested this statistical analysis – POTW feedback on ILS2 data packet 8/7/23
- **Comment on/identify if there are effects from water quality/sample recipe**
 - Further evaluation and “digging” into the data is requested – are there effects from water quality?
 - Include figures of the control-normalized Sample 1 and 2A reproduction – current writeup cannot be followed without seeing the data in figures
 - Report – Lab N “only lab reporting toxicity consistently”
 - Figures provided to P. Bedore by SCCWRP (see following slide)
 - > 10% effect in 1/3 of 2A Samples tested at Lab B, E, I, N, and O
 - Better “agreement” on the Sample 1-to-control comparison
 - Less “agreement” on the Sample 2A-to-control comparison
 - Is it appropriate to expect 10% or less difference between sample 2A and its control?

Key Issues – Samples 1 and 2A

Sample 1



Sample 2A

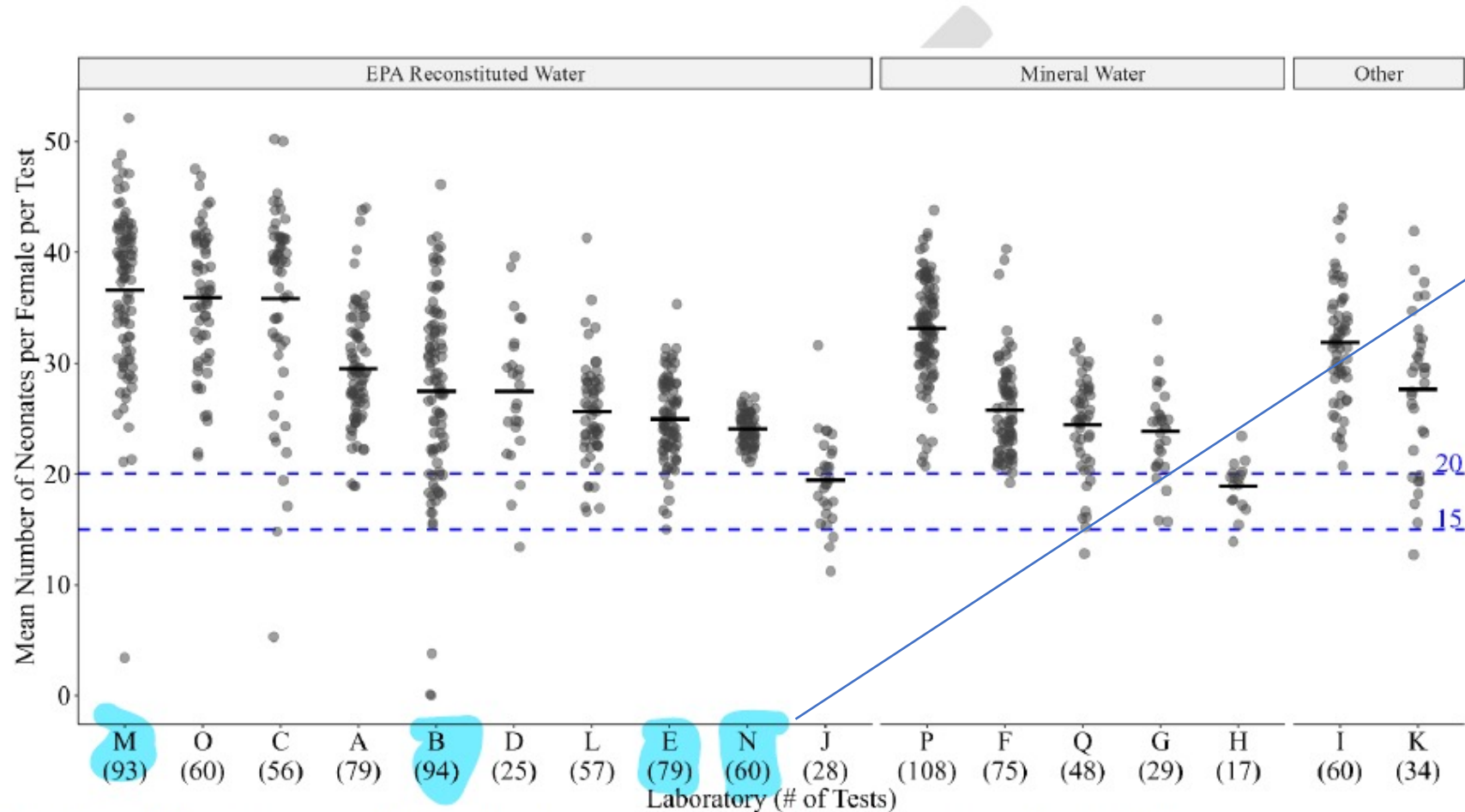


Key Issues – ILS1-to-ILS2 Comparisons

- Report presents a graphical analysis & assesses “magnitude” of change between ILS1 to ILS2 data
 - What statistical approach could be used and why was the graphical approach preferred?
 - i.e., comparison CVs, mean reproduction, etc. between ILS1 and ILS2
 - Intra-lab comparison
 - Inter-lab comparison
- Changes between ILS1 & ILS2 need to be better contextualized with historical data
 - Culture performance will naturally have changed between ILS1 and ILS2
 - Lab issues that occurred
 - Sensitivity to reference toxicant will fluctuate
- Report Examples
 - Marked improvement ... three laboratories did not meet one or both TAC - report
 - Labs B, E, M, N – Not meet TAC in ILS1, did better in ILS2
 - How were they performing on their own controls just prior to ILS2?
 - Higher frequency of labs achieving $CV \leq 20\%$. Lab B, E, O showed some improvement - report
 - Lab B did not improve
 - Lab E, M, and O improved, but the %CV was squarely in each lab’s historic CV range
 - CVs at some labs increased (Labs N, G, Q)

Key Issues – ILS1-to-ILS2 Comparisons

Figure 5-1. Mean number of neonates per female for each submitted test from the historical dataset. Data is organized by the type of dilution water used by the laboratory for their test controls. Laboratories are in order of high to low mean values within each water type.



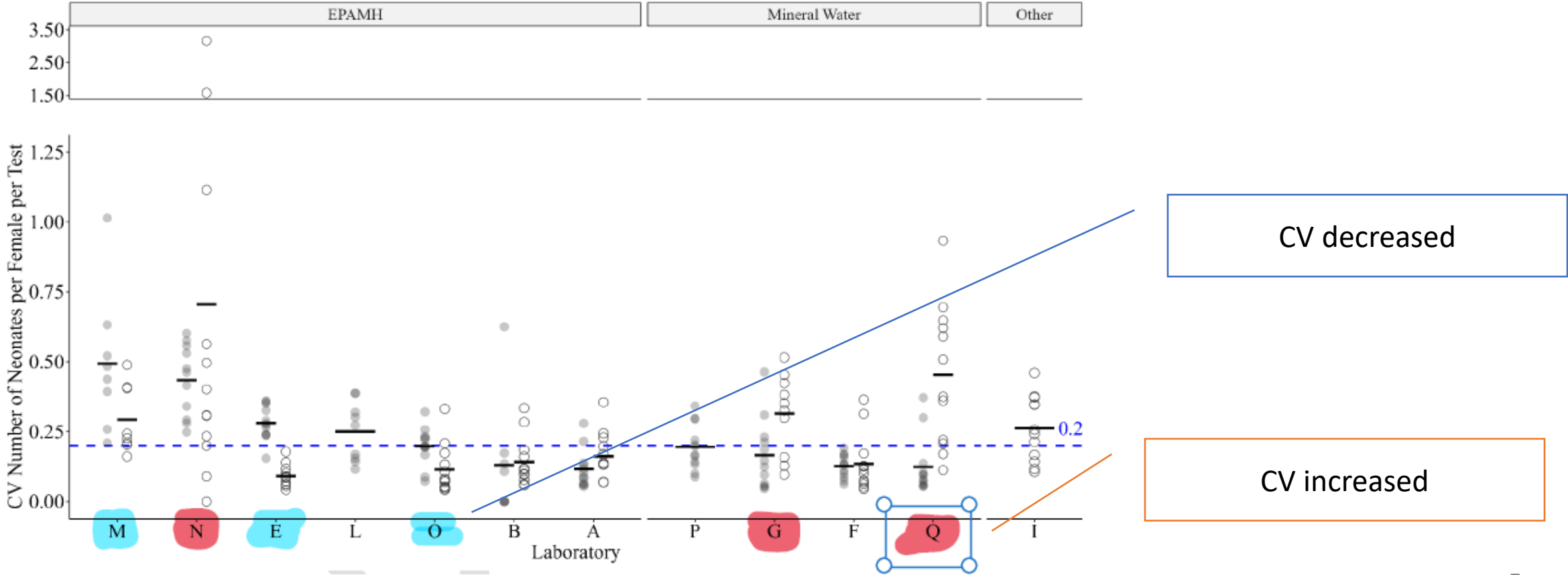
Failure to meet TAC in ILS1 for these labs is atypical

Meeting TAC in ILS 2 is typical

Note that most laboratories using the EPA reconstituted water add selenium and/or vitamins (see Appendix A, Table A2).

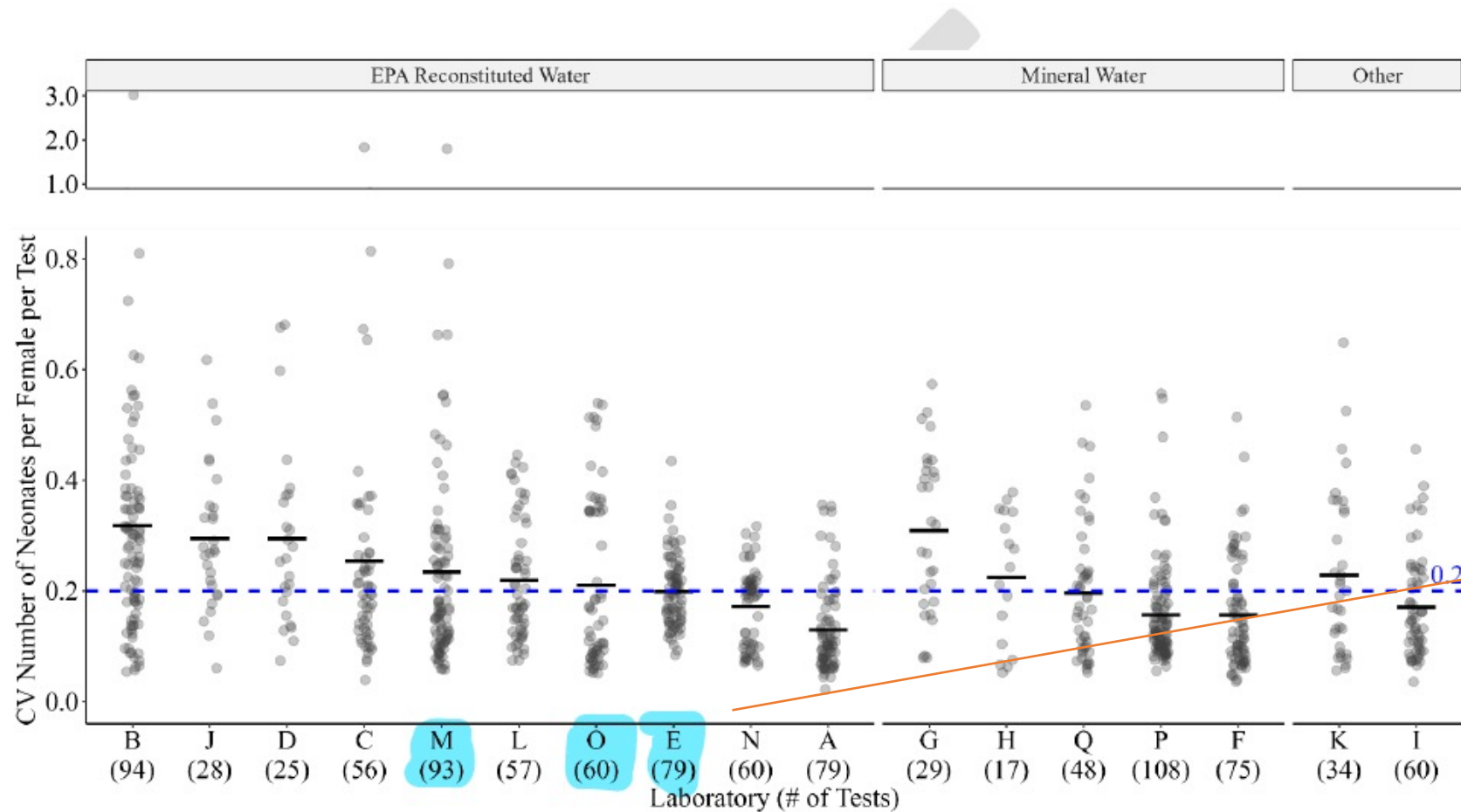
Key Issues – ILS1-to-ILS2 Comparisons

Figure 5-6. Coefficient of variation for the mean number of neonates per female in laboratory controls for the baseline and second split-sample exercises. Data is organized by the type of dilution water the laboratory uses in their controls. Labs are in the same order as mean neonate plot. Closed symbols are for the Baseline and Open are for the second ILS. The larger colored symbols are the mean values for each ILS.



Key Issues – ILS1-to-ILS2 Comparisons

Figure 5-2. Coefficient of variation for the mean number of neonates per female for each submitted test from the historical dataset. Data is organized by the type of dilution water the laboratory uses in their controls. Laboratories are ordered from high to low mean values within each water type.



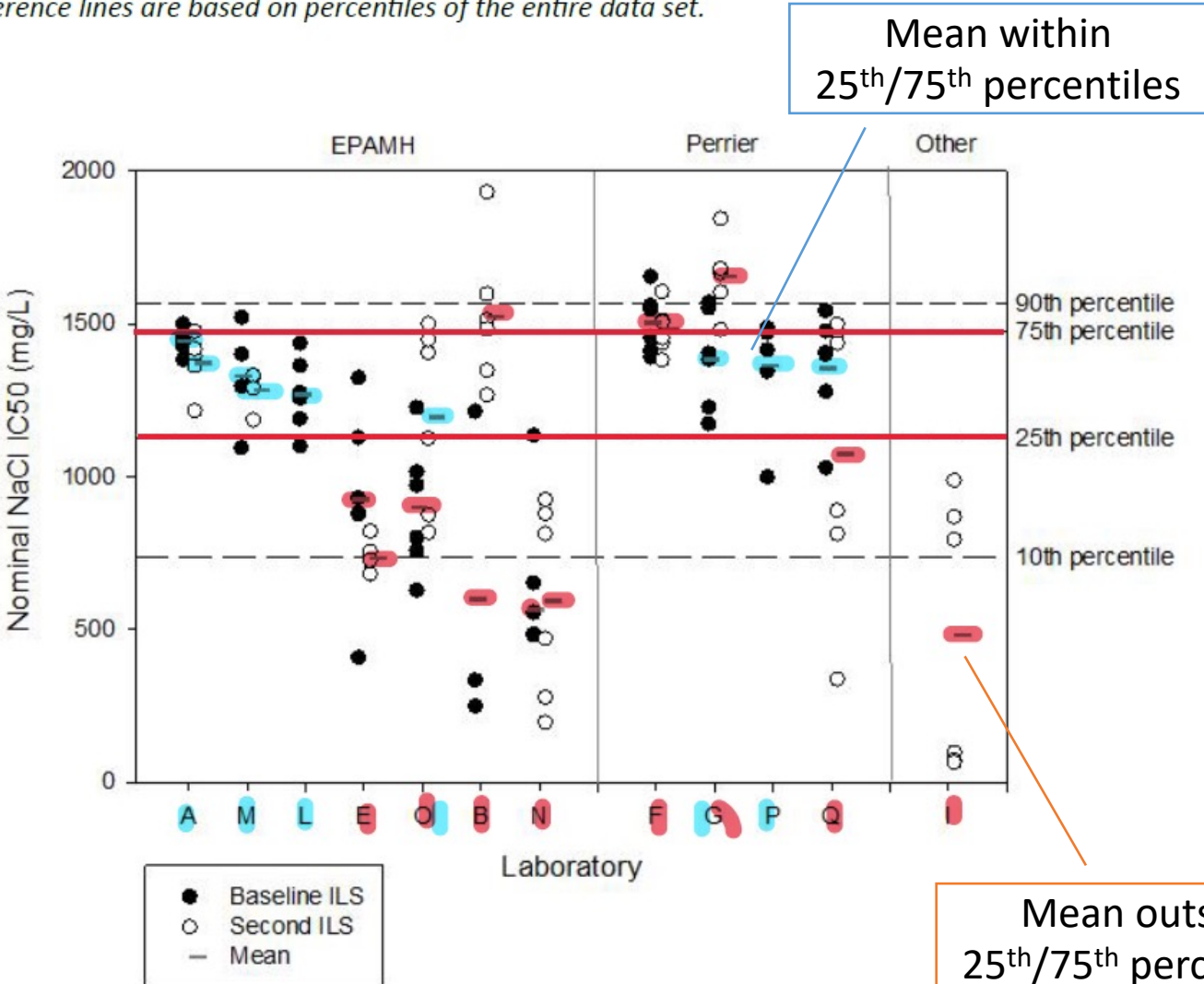
ILS2 CVs for these labs are within historic CV range for each lab

Key Issues – Spiked Samples

- IC25 vs. IC50
 - “While the Panel focused on IC50s that tended to be less variable, data quality programs may consider applying the percentile ranking approach to ... IC25s or LC50s” - report
 - 25% effect is more relevant & tangible to the public & stakeholder concerns – variability in ambient/compliance tests
 - Like using 25% PMSD threshold
 - IC50 – less variable, which conclusions may not be relevant/comport to 25% effect level
 - Request revise section and include conclusions/findings based on IC25
- Compare ILS1 to ILS2 –
 - Elaborate on findings of whether labs are going in the right direction after implement standardization in ILS2
 - Is there a statistical difference in IC25/50 between ILS1 and ILS2?
 - By individual lab
 - Among all labs
- Demonstrate report conclusion that there is no effect on IC50/25s from water quality
 - Figure showing IC25 and IC50 data by sample type (e.g., 2b-f & 3) to support conclusion (“no difference based on water quality”)
 - Provide statistical comparison to support conclusion
 - Discuss data nuances in light of a lab’s performance on Samples 1 & 2A

Key Issues – Spiked Samples

Figure 5-8. Nominal IC50 values from both Sample 2 and Sample 3 concentration series from both ILS. The reference lines are based on percentiles of the entire data set.



- Report concludes that Lab B showed improvement in ILS2. What does improvement mean?
- How much agreement should there be among labs and within labs?
- How does this data inform regulated community concerns?
 - Will a sample sent to multiple labs get the same answer?
 - What additional analysis could be used to elaborate on this?
 - i.e., for a given concentration of NaCl, what is the probability it will be determined to be toxic by 1 lab, 2 labs, 3 labs, etc.?

Key Issues – Conclusions and Limitations

- Conclusions
 - Include summary table
 - Clearly describe each performance metrics (these are never summarized in the report)
 - List each lab and evaluate relative to each metric
 - Lab B – put forward as primary example that standardization was effective
 - Improvements are subtle at Lab B
 - Compare to historic data
 - Describe how this study addresses stakeholder concerns
- Limitations –
 - If you had more time, what analyses would the you do to "*...refine laboratory metrics, including developing additional guidance to define consistently performing laboratories*"?
 - How long will it take to fully implement the recommendations?
 - How long will it take to observe changes in lab performance and iterate so that enhance lab performance is verified?

Key Issues – Executive Summary, Conclusions, Recommendations

- 4 of 12 labs met the expert's performance standards (Exec. Sum.)
 - Describe proportion of CA testing covered by these 4 labs
 - Clarify how many labs are in CA
- Include a timeline and steps for implementing the recommendations and verifying change to lab performance
- Investigating background water quality – Refinements to limitation/ recommendation per last ESP meeting
 - “High hardness” – focus of report limitation language
 - Main issue is variation in sample quality from lab's culture water
 - Could not investigate given current performance of 12-lab pool
 - Could investigate with the 4 labs that are meeting expectations