Article:
Joe M El-Khoury, Tony Badrick, Elvar Theodorsson.
Time to Reevaluate the 95% Inclusion Criteria for Defining Reference Intervals?
Clin Chem 2024; 70(5): 700-702. https://doi.org/10.1093/clinchem/hvae026

Guest: Dr. Joe El-Khoury from the Yale School of Medicine in New Haven, Connecticut.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett. In the clinical laboratory, "normal" for most analytes has traditionally been defined as the central 95% of a healthy population. When results outside of this central 95% reflect an ongoing disease state, this approach helps establish a diagnosis and allows the initiation of appropriate treatment. But what about values just outside the reference interval in a healthy patient undergoing routine testing for an annual physical?

Characterizing these values as abnormal can cause real harm by increasing patient stress and prompting unnecessary further evaluation. How did the central 95% approach become standard practice and is it still appropriate for the current model of healthcare delivery? If we choose to make a change, what criteria can laboratorians use to identify tests that would benefit from a different approach? An opinion article appearing in the May 2024 issue of *Clinical Chemistry* reviews the evolution of what we consider to be normal limits, proposes alternatives to the central 95% rule, and describes scenarios in which alternate approaches could be considered.

In this podcast, we're excited to welcome the opinion article's lead author. Joe El-Khoury is an Associate Professor of Laboratory Medicine at the Yale School of Medicine and director of the Clinical Chemistry Laboratory and Fellowship Program at Yale New Haven Health. He serves on the editorial board of this journal and is the recipient of the 2023 Young Investigator Award from the International Federation for Clinical Chemistry and Laboratory Medicine. Dr. El-Khoury, to start off, I want to get a little personal. What specific challenges have you encountered with current reference interval practices and how have these challenges influenced your perspective on the need for re-evaluation?

Joe El-Khoury: Thank you, Bob. I think one of the biggest issues I have with the current way we practice, in terms of implementing these reference intervals, is that it's a cookie-cutter approach. We basically said we're going to take this central 95th percentile of whatever population we use to derive these reference intervals after excluding outliers and then apply that to every test, with no specific consideration that each test may require based on the clinical utility. And so, as a practicing lab director, I've seen examples and I'll use examples when I ran into issues with reference intervals.

One of them, for example, I'll say is ALT, which I've talked about. It's a liver function test that's commonly used to assess how well your liver functioned. And what we've seen is that in fact, those intervals are often too wide as set by reference labs because the population used to derive those intervals are essentially including people who are overweight, who are obese, or including people who consume regular amounts of alcohol. And then -- So in this case, you have an example of a test where you're taking the central 95th percentile but you artificially widened, and you've changed what is truly the right definition of "healthy" because you're including these individuals as part of your population who may not be appropriate.

On the other extreme of things, which is kind of more important to this discussion when we're talking about reference intervals that are too narrow. And the example I used for that and I've talked about that separately in this journal is TSH. So, TSH which is commonly used to assess thyroid function as part of a thyroid function panel, is often used as a first step to check if you need to do follow-up testing and have a thyroid issue. So in this case, laboratories or manufacturers who derive reference intervals have done what we were told to do, which is get usually over 120 individuals.

And usually they do this at one point in time and then take the central 95% but the problem here is that they didn't recruit these individuals over different times of the year and TSH is known to vary by season. So you end up having these changes that are not accounted in those reference intervals. And on top of that, you are artificially nailing them even further by cutting out 2.5% on the top, 2.5% on the bottom, which is what the central 95th percentile is all about. And now you end up with a narrow reference interval that's calling people falsely high or falsely low, inappropriately, and causing all this unnecessary anxiety, unnecessary follow-up testing for people who are otherwise normal.

So, aside from those two examples, which I talked about separately excessively, focusing now on things we do often for our patients, like I'll use potassium as an example, which we often measure as part of a basic metabolic panel or comprehensive metabolic panel. We're basically getting people to derive reference intervals of that who are otherwise normal. In the outpatient setting, we've used typically a study or a form of survey to tell us they're normal, they're not on any meds, and then we're excluding outliers, and then still excluding 5%--2.5% on the top, 2.5% on the bottom--

then saying that's our reference interval. The problem here is, again, if we're testing potassium on everybody, we are purposefully flagging 5% of the unhealthy population.

Otherwise, no other disease and just saying that, you know, we don't know what that means but based on our reference interval which is based on that central 95th percentile, which is flagging those 2.5% on either end, that's again causing unnecessary anxiety and people who may need follow-up testing or in some cases, they don't even know what to do with that high level. So all of that, you know -- And I went with three different test examples using ALT which is too wide, TSH which is too narrow, and then this case focusing on the common tests like potassium, which are now being tested on everybody as part of these big panels.

We're basically over-diagnosing people as having too high or too low without any concept of what it means clinically. So looking at what we're talking about in this article is we're basically asking that every test deserves to be looked at and asked, "What does it mean if we're flagging at this level? And if there's no action to be taken or it's not really clinically significant, should we be flagging it?" And so, allowing us to now even expand that further and using this 99th percentile approach would help eliminate all of these unnecessary flags.

And so, this is the real issue I have and I think laboratories need to stop making these decisions in silos. We basically need to talk to our clinical teams, ask them when would they like us to flag potassium is high or potassium is low. And these are conversations we're actively having now as a result of this paper and basically, we're having a new concept in terms of defining reference intervals that moves away from this cookie-cutter 95th percentile, where the labs decide alone what is high and what is low and I just don't think that's the right way to do it.

Bob Barrett: Okay. Well after all that, could you elaborate on the origins of the 2.5% exclusion rule and how this historical context informs our approach to re-evaluating reference intervals today?

Joe El-Khoury: And that is another pet peeve. So, the problem with the current approach is that it really goes back, in terms of deciding on the central 95th percentile, to astronomical observations. It's crazy to think that your physicians today are making decisions based on astronomical observations in the 19th century. And so, the origin of this central 95th percentile basically is what was used in astronomy to decide the location of a planet or a star that's observed within a certain amount of accuracy.

So, the central 95% is basically saying it's the location of that star is X with this much confidence around that position, right? So, then we've applied this in medicine, and in many ways to the common idea that there is a normal change in man, and this again was introduced also in like the 19th century, where there should be a normal distribution of values related to man. And that concept has been blown out also, you know, in lab medicine even as far back as the 60s where we basically said there is no normal. That's why we call them reference intervals, not normal ranges. Of course, this hasn't caught up in the medical field.

People still call them normal ranges, but I would argue it's very important as laboratories to emphasize that point that normal, relative to what? Because you know, if you have somebody who's an outpatient, their distribution is going to be different than somebody who's an inpatient lying down in a bed because their volume fluids shifts and all of a sudden, you have 6% to 12% changes that are otherwise normal for that population, but you're using this concept of reference intervals that are derived based on that 95% that's going to cause issues. So basically, long story-short -- And we can dive into more detail about this in the paper, you know, these *P*-value derivations which originated in astronomy and later were adopted in many disciplines of science, people still use *P*-value today which is less than 0.05 to say a finding is significant.

And even now, that's being challenged in papers in *Nature*, saying that took it out of context. Sometimes *P* less than 0.01 is appropriate, which is exactly what we argue here in using the central 99th percentile. So, this is a problem that affected a lot of disciplines, not just us in lab medicine, but it's interesting to relate, the problems we're seeing with our reference intervals today goes back to decisions made by astronomers. And so, this is why it's really important to question this and figure out a way to adapt to the changing world.

Bob Barrett: The article mentioned the proposal to move from a central 95% to a central 99%. How can laboratorians determine which tests warrant such modification and how would this approach address the issue of false positive while minimizing false negatives?

Joe El-Khoury: You know, the devil is in the details. So I love this question because really, it forces us to say, "Well, how does this theoretical approach we're proposing really can be applied in a meaningful way for labs?" So, we didn't specify tests because these are conversations we're having now. So, there will hopefully be a follow-up on this that says, "These are the tests that need it," but we broadly outlined the thinking that this should follow. So, tests that are commonly ordered on patients that don't need it as part of, for example, the wellness movement.

So, if you have a comprehensive metabolic panel, or basic metabolic panel, those are tests that you potentially need to look at and ask, "Is it worth still being too tight or do we need to use 99th percentile at least instead of 95?" So, the number one rule we say in the various tests that are commonly ordered. The second thing we added is basically saying tests that are known to have a high false positive rate or we've already been using these things for so long and we know that there's no value in flagging this test at such a level, so we need to revisit and expand those ranges.

So, that's key to say that it's not -- You know, it's tests that are widely used today and then tests that already have shown that there is no need to be flagging at such a low level and there's evidence in the literature that suggests that those need to be revisited. So, this is broadly speaking. To be specific, again, we're looking at every test, at least we're starting with the comprehensive metabolic panel, and re-examining these and deciding which ones would be appropriate to widen because there is no action that needs to be taken at that level, and which ones should be kept because it is clinically relevant and so we shouldn't change that.

Bob Barrett: Well again going back to the article, it discusses alternate approaches to population-based reference intervals such as clinical decision limits, personalized reference intervals and reporting z-scores. Can you discuss the potential advantages and drawbacks of each of these approaches and how they compare to using the central 99%?

Joe El-Khoury: Absolutely. So, starting with the drawbacks of the central 95 or central 99%, the problem with both of these approaches, which is using 95 or 99, is you're assuming that there's a point where you're in and a point where you're out of the range, like as if you're healthy and all of a sudden unhealthy. And that's not the real world. We know that biology is more -- there's more of a spectrum of results and you don't necessarily fall out or in. It all really depends on context.

So that's where, for example, z-scores have an advantage. Z-scores are simply a way of interpreting the results in terms of how far you are from the median without really imparting an out or in knowledge. It's just saying you're plus-one standard deviation or plus-two standard deviations. The other advantage of a z-score is also, you don't need to know what the reference interval even is for every task, so you can just look at sodium like you do at potassium. All you'll see is +1, +0.3. For all of the tests, you're essentially normalizing how you interpret those results because you're always converting it and dividing it basically by the mean of results.

So in a way, z-scores offer the advantage of not needing total reference intervals and you can basically scan the report and see if a patient is kind of leaning farther away from the distribution or closer to the distribution of results and so it offers that advantage. The downside of a z-score is you really cannot then apply it to tests that are not standardized and harmonized across labs because then if you have changes in reference intervals or different antibodies that are used by tests, it may reflect it within your lab, but then your result may not apply as well if you basically are interpreting it to another lab because a z-score in your lab may be actually a completely different value in another lab.

But in a way, that does help you in making decisions in your lab and in a silo, but it would be ideal for tests that are harmonized and standardized because basically there is no now differences among labs and you can scan in the report and know whether your patient's far away from the distribution or close enough. So that covers z-scores. Looking at clinical decision limits, which is personally my favorite when possible, that basically says instead of making these decisions just by looking at a healthy population and saying what is normal and what is not, we are now looking at what when disease starts.

So, clinical decision limits like we use for glucose for example, we say it's 100, or vitamin D or for cholesterol, we say it's 200, that's basically recognizing at what point should we start making lifestyle changes? At what point should we start giving statins or taking action or giving insulin? Those are clinical decision limits that have been proven to be useful instead of choosing whatever set of 120 people that we call "healthy" and then deciding on that, we're basically deciding based on these clinical metrics to say, "This is the time that we need to flag patients because we need to follow up on them."

And I think that is the most powerful one to use among all of them except, sadly, it's not an option for all tests because not all tests have a single clinical decision use or we basically understand the spectrum of how to use. The last one I'll mention is basically the personalized reference intervals, which it looks at an individual's own result and what is the significant change around that using reference change values, which is a concept that you can apply for each test to account for what is normal biological variation and what is analytical variation, and then seeing a change in their own result over time. You can then say, "Well, they're abnormal relative to their own previous results."

So what you're seeing is changes in the individual that are biologically relevant instead of comparing them to somebody else in a population as we do today, which is not as relevant.

That also has its own challenges, of course. You're still going to have that in-or-out, which isn't great. The world is more gray than black and white. But it's been proposed as a way to eliminate reliance on these populations and gives you more of a personalized reference interval. So hopefully, that gives the audience an understanding of like these are the four major categories, essentially. There is no perfect one. It's going to be hopefully a combination of all that help us kind of tailor the approach we use for each test.

Bob Barrett: Well finally, Dr. El-Khoury, what steps should be taken to implement these proposed changes effectively?

Joe El-Khoury: So for us, like I mentioned, what we're doing is we've basically engaged our clinical providers in making decisions on reference intervals. We are right now having conversations about potassium, for example. We're part of a hyperkalemia signature care pathway, which many hospitals are starting to engage in, where you're basically trying to outline a standardized approach to treating and diagnosing a disease. And in this case, you know, we're having a discussion that is like, "Why aren't we flagging on 4.8, 4.9?" which is what the traditional 95th percentile would say should be something you flag on for potassium or a general reference interval depending on if you're using plasma or serum is around I'll say roughly 3.3 to around 5.

But what's the point of flagging at 5.1, 5.2, 5.3, if there's no clinical action, right? So for at least inpatients, we're really looking at this and starting to say, "Okay, if your first action in the hospital starts at 5.5, then let's flag the 5.5. Let's not cause unnecessary anxiety, trigger these things. There's no action to be taken." So, I recommend that laboratories really engage their clinical team and providers and physicians in trying to understand how they use these and ask them when would you like them to flag.

And try to see what you could learn and of course, find a way that would work for the whole team. I recognize there are some regulations that we have to worry about on the laboratory side that we need to consider. And also like, essentially, I really do think by having these conversations, we will have better outcomes and be able to determine and provide reference intervals that are more meaningful when they flag.

Bob Barrett: That was Dr. Joe El-Khoury from the Yale School of Medicine in New Haven, Connecticut. He served as lead author of an opinion article in the May 2024 issue of *Clinical Chemistry* describing a new approach to defining reference intervals and was our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.