

With TSH testing, no lack of discord

William Check, PhD

June 2013—In 219 BCE, after he had unified the seven warring states to establish the nucleus of the Chinese empire, the First Emperor of China promulgated uniform administrative practices throughout the land. One section of his decree read:

All men under the sky toil with a single purpose. Tools and measures are made uniform. The written script is standardized. Wherever the sun and moon shine.

Perhaps the field of thyroid function testing could use a return visit from the First Emperor to make its tools and measures uniform. Disagreement swirls around even the most basic aspect of thyroid function testing—reference intervals for measurement of thyroid stimulating hormone.

“Reference intervals for TSH are all over the place,” Carole Spencer, MT, PhD, professor of medicine at the University of Southern California and technical director, USC Endocrine Laboratories, says, adding that there are several reasons. One is that the reference interval is very sensitive to individuals in the cohort who skew the upper limit, which makes the distribution non-Gaussian and requires log transformation of values. Older persons, in whom higher levels are seen, are an example. “The upper limit of the TSH interval may be 7.5 mIU/L for a group of healthy 80 year olds, whereas it might be down around 3 for healthy 20 year olds,” Dr. Spencer says.

Obesity is another example. “If you include obese individuals, that will increase TSH values independent of thyroid function,” Dr. Spencer continues. “You only have to have three or four really obese individuals in your cohort to skew the TSH upper limit. Many people don’t appreciate the effects of obesity on TSH. When a person loses weight, TSH comes way down.” And if those with thyroid autoimmunity are not screened out prospectively, she says, by testing for antibodies to TPO [thyroid peroxidase], that will also skew the upper limit.

“Everyone gets exercised about the accuracy of the TSH reference interval, but they fail to understand who this interval was calculated on. What it comes down to,” Dr. Spencer says, “is that clinical judgment is more important [for diagnosing hypothyroidism] than the TSH reference interval.”



Dr. Killeen

Another source of uncertainty in thyroid function testing is the great difference between the population interval and individual variation, says Anthony Killeen, MD, PhD. “We don’t know each individual’s setpoint—their relation between TSH and free T4 [FT4],” Dr. Killeen, director of clinical laboratories and professor of laboratory medicine and pathology at the University of Minnesota Medical Center, Fairview, explains. “So we need to use the much wider population reference interval.” That requires clinical judgment. “At the upper end of normal, values around 4 mIU/L, we need to look at other aspects of the patient,” Dr. Killeen says. For instance, women older than age 55 will be at the upper end of the interval, where hypothyroidism is not uncommon. In Dr. Killeen’s view, the non-Gaussian skew at the upper end of the reference interval suggests that it also includes people with subclinical hypothyroidism. “Persons with higher values are more likely to have anti-TPO antibodies and to develop overt hypothyroidism,” he says.

A second controversy centers on the tools, the possibility of radically altering the method of measuring thyroid

hormones. Immunoassays should be replaced by ultrafiltration (or equilibrium dialysis) followed by tandem mass spectrometry (UF/MS/MS), says Steven J. Soldin, PhD, senior scientist in the Department of Laboratory Medicine in the Clinical Center at the National Institutes of Health and adjunct professor of endocrinology and metabolism at Georgetown University Medical Center. Many years ago Dr. Soldin noticed that endocrinologists asked him about samples in which FT4 did not match TSH. He sent those to a reference laboratory for equilibrium dialysis followed by immunoassay.

“We saw that free T4 by [the reference laboratory method] correlated with log TSH, whereas the initial direct analogue free T4 by immunoassay did not,” Dr. Soldin says. “Everyone knows that direct analogue free T4 methods are suboptimal and correlate poorly with log TSH in patients with hypo- and hyperthyroidism. We have had two to three decades of frequently reporting the wrong results for free T3 and T4 by immunoassay, which clearly impacts the clinical diagnosis.”

Dr. Soldin has been working for more than 10 years on improving the UF/MS/MS method, which he has demonstrated to be analytically superior to immunoassay and which he has put into place at the NIH, Children’s National Medical Center, NMS Laboratories, and Georgetown University for clinical application.

Controversy No. 3 in thyroid function testing is whether to screen all pregnant women. The American Thyroid Association does not recommend universal screening, while other groups do, Dr. Killeen notes. Another important issue is that TSH levels in pregnancy are lower than those in nonpregnant women and that FT4 declines with gestation. Dr. Killeen and others recommend trimester-specific intervals for TSH.

Of the uncertainty regarding the reference interval for TSH, Dr. Spencer says that, in addition to the cohort used to define the population interval, “methodological factors also come into play.” Immunoassays employ monoclonal antibodies, which have limited and varying specificity to detect the epitopes of TSH. “Circulating TSH is heterogeneous, especially with respect to glycosylation,” Dr. Spencer says. “So which TSH molecules an assay detects will depend on the monoclonal antibody you select.” Antibody variability raises a whole different issue for clinical interpretation, she says. “Because of heterogeneity in glycosylation, not all molecular forms of TSH are bioactive.” In particular, although the upper limit of TSH is higher in older individuals, not all TSH in older persons may be bioactive.

Dr. Spencer’s summary: “There are a lot of unknowns here. We could spend the whole day arguing where the TSH upper limit should be set.”

To resolve this problem, Dr. Spencer invokes the point Dr. Killeen made. “In reality,” she says, “TSH population reference intervals are not a sensitive parameter for detecting thyroid dysfunction because all thyroid tests have a low index of individuality—the relationship between the between-person variation and the within-person variation.” One study that measured TSH in a group of subjects every month for a year found that, for a normal person, TSH levels varied by 0.5 to 0.75 mIU/L across the study. “So the reproducibility of TSH measurements within an individual is much narrower than the interval you see when you combine data among individuals to get a population reference interval,” Dr. Spencer explains. “If an individual starts to develop hypothyroidism, their TSH could rise to 2.7, which might be highly abnormal for that person but still well within the reference interval of the population.

“Do you want to treat that individual?” Dr. Spencer asks. “Whether you do would not be determined by the TSH reference interval, but by the person’s lipids, the presence or absence of anti-TPO antibodies, family history, and a whole number of other issues as to why that patient came in to see the doctor.” Say the patient was a pregnant young woman. Considering that the upper limit of normal TSH in the first trimester is 2.5, “If her TSH is 2.8 or 3, you might well treat her,” Dr. Spencer says. On the other hand, a 70-year-old woman with no antibodies, in an age group with an upper limit of normal of 5 or 6, probably wouldn’t be treated.

Dr. Spencer calls a population reference interval “a very insensitive way” to assess thyroid dysfunction in individuals. “It is necessary to look at each patient in a specific way. I totally think, unless you are dealing with a very elderly patient, a general reference interval of 0.3 to 3 is a good starting point. Then factor in patient-specific

factors.”

Whether to treat subclinical hypothyroidism is also contentious. Subclinical hypothyroidism is usually due to autoimmune thyroid disease, such as Hashimoto’s thyroiditis. “If TSH is on the high side, 3 to 10 mIU/L, your next test should be anti-TPO antibodies. If you detect antibodies, likely there is some degree of clinical hypothyroidism.” These persons are at increased risk of progressing to overt hypothyroidism. Again, treatment would depend on the severity of the patient’s symptoms.

However, if the antibody test result is negative, Dr. Spencer says, “You must remember there are a number of reasons why TSH is high.” To the causes already mentioned, she adds polymorphism of TSH receptors in thyroid cells. “In these people it takes a higher level of TSH to do the job,” she explains. “So TSH may be high, TPO antibody negative, and nothing is wrong with them. They are euthyroid.”

Like Dr. Spencer, Dr. Killeen advocates weighing symptoms when interpreting TSH values. “Tiredness is a very common and vague complaint,” he notes. “Usually it does not indicate thyroid disease.” He suggests thyroid function testing only if tiredness is prolonged or debilitating. Dr. Killeen also points out that “textbook” symptoms of hypothyroidism—dry skin, diminished sweating, weight increase, periorbital puffiness—were described with more advanced hypothyroidism detected with older, less sensitive assays, not with the early forms detected with today’s more sensitive assays.

As for treating subclinical hypothyroidism, he notes that, while no association with mortality has been demonstrated, “Coronary heart disease events begin to rise above 7 mIU/L, becoming significant above 10 mIU/L.

“There has been a lot of disagreement on key questions” regarding TSH reference intervals, Dr. Killeen says. He favors the existing reference interval—0.3 to 4.0 mIU/L.

Between-assay variability is also a problem. Dr. Killeen cites a study sponsored by the International Federation for Clinical Chemistry that concluded, “Harmonization of TSH measurements would be particularly beneficial for 3 of the 16 examined assays” (Thienpont LM, et al. *Clin Chem*. 2010;56:902–911). Reference materials are available to standardize total T4 and T3 assays. However, “harmonizing” TSH assays from commercial manufacturers so they give the same result on the same sample is the optimal goal.

A CAP study of thyroid function testing performance using fresh frozen serum evaluated bias in methods for thyroid hormones among 3,900 clinical laboratories. The authors concluded, “A majority of the methods used in thyroid function testing have biases that limit their clinical utility. Traditional proficiency testing materials do not adequately reflect these biases” (Steele BW, et al. *Arch Pathol Lab Med*. 2005;129:310–317). Dr. Killeen says this accuracy-based Survey will be repeated in the next 12 months. “Accuracy-based Surveys for thyroid hormones enable us to compare results from different manufacturers in a more relevant way than conventional proficiency testing,” he explains.

Dr. Soldin’s experience on a CAP resource committee during the 1990s made it obvious to him that there was a problem with immunoassays in thyroid function testing. On proficiency tests, there was about a twofold divergence between the mean of the highest method and the lowest method for T4 and T3. “That’s one criterion that says maybe this method needs help,” he says.

Dr. Soldin’s work, as well as that of others, has showed very low correlation coefficients between FT4 by immunoassay and log TSH for many widely used commercial platforms (Soldin SJ, et al. *Clin Chim Acta*. 2010;411:250–252; Deventer HE, et al. *Clin Chem*. 2011;57:122–127; Gu J, et al. *Clin Biochem*. 2007;40:1386–1391; Serdar MA, et al. *Clin Chem Lab Med*. 2012;50:1849–1852). One study found that FT4 by immunoassay correlated with albumin and thyroid binding globulin, “suggest[ing] that this FT(4) method depends on binding protein concentrations and does not accurately reflect FT(4).”

“It’s fair to say that all analyzers we checked had poor correlations between free T4 and log TSH,” Dr. Soldin says.

In contrast, FT4 determined by UF/MS/MS correlated well with log TSH in several populations—pediatric, post-

thyroidectomy, and pregnancy (Kahric-Janjic N, et al. *Thyroid*. 2007;17:303-311; Soldin OP, et al. *Thyroid*. 2009;19:699-702). Free T4 determined by immunoassay following ultrafiltration also correlates with log TSH. "So I'm not saying that one has to do mass spectrometry," Dr. Soldin says. "I am saying that one has to separate the binding proteins and then do either immunoassay or mass spectrometry." Dr. Soldin does recommend determining FT4 and FT3 by UF/MS/MS for all specimens in which the TSH is greater than the 90th percentile or less than the 10th percentile.

Since 2006, Dr. Soldin has introduced UF/MS/MS into clinical diagnosis in the thyroid function testing laboratories he directs. "To my knowledge," he says, "Children's National Medical Center is the only hospital that has stopped doing free T4 by immunoassay and shifted totally to mass spectrometry. Since then the endocrine people have not complained about discrepancies between free T4 and TSH." Capital cost of the instrument was about \$500,000. "Within eight months that instrument was paid off just by generating free T3 and free T4 tests at that institution," Dr. Soldin says. "If immunoassay is giving you the wrong answer in most patients, it doesn't matter if it's cheaper. It's the wrong answer. That's not going to help your patients.

"Why is FDA approving methods that do not correlate well with log TSH and don't accurately measure what they're supposed to measure?" he wants to know.

Dr. Soldin says he submitted a study in which his group took samples from patients with "subclinical hypothyroidism" (FT4 measured by immunoassay) and retested the FT4 values by UF/MS/MS. "Three-fourths of them gave results that were low and agreed with TSH," he says. "It is clear that what was wrong was predominantly the free T4 method. Immunoassay is precise, but can give precisely the wrong values."

Dr. Soldin raises another major failing of immunoassays: measurement of T3 and free T3. "I'm getting phone calls and samples from people who want tests, and we're identifying a group of people who don't convert T4 to T3 very well," he says. When these people are put on T4 replacement therapy, TSH normalizes, as do free T4, T4, T3, and FT3 by immunoassay. However, the patients still do not feel well.

Measurement of T3 and free T3 by UF/MS/MS in these patients shows that T3 and free T3, the highly active thyroid hormone levels, are low. "When you treat with T3 replacement, many women feel much better," Dr. Soldin says. He is working with endocrinologist Jacqueline Jonklaas, MD, PhD, of Georgetown, to study a large cohort of women with what has been labeled subclinical hypothyroidism.

ATA and Endocrine Society Pregnancy Guidelines

TSH is the primary thyroid test to use for evaluating pregnant patients except when TSH is undetectable in hyperthyroid patients taking antithyroid drug Rx in whom an assessment of thyroxine status is needed.

- ◆ FT4I measurements, or TT4 measurements (using the nonpregnant reference range multiplied by a factor of 1.5 to compensate for elevated TBG), may be superior to FT4 immunoassay measurements.
- ◆ When using FT4 immunoassays to monitor treatment for Graves' hyperthyroidism, the goal is to maintain maternal FT4 at, or just above, the nonpregnant reference range for the method.
- ◆ FT4 immunoassay results should be interpreted with caution. It is recommended that each laboratory establish method-specific, trimester-specific reference ranges. (*Dr. Spencer considers this impractical.*)

Stagnaro-Green, et al. *Thyroid*. 2011;21:1081. De Groot, et al. *JCEM*. 2012;97:2543.

Dr. Killeen acknowledges that mass spectrometry will be the reference method against which other assays will be compared. He acknowledges, too, that Dr. Soldin has provided analytical evidence that free levels of hormone may be measured more accurately by mass spec. However, "There also needs to be clinical outcomes data with mass spectrometry," he says. "I'm not convinced it is necessary to switch at this time." He notes that UF/MS/MS is a more complicated technique than immunoassay and requires more skill on the part of laboratory staff.

Dr. Spencer says it's not practical for every hospital lab to be running free T4 with mass spec. "It is wonderful to have it there as a reference method," she says, "but the technique is very demanding and the equipment very expensive." Dr. Spencer calls physically separating free from bound T4 by ultrafiltration a "technically demanding and tricky method," and notes that Dr. Soldin has been successful at it. "But it is unrealistic to assume that most free T4 in the U.S. will be done by mass spec," she says, adding, "I don't see any way to do that for \$12 to \$15, which is the Medicare reimbursement rate.

"My big concern," she continues, "has always been why the FDA does not insist that the manufacturers of these free hormone immunoassays call them free T4 *estimate* assays. FDA did not stand up to the kit manufacturers and allows them to sell the assays as free T4 assays. Whereas Dr Soldin's work has clearly shown that they have a very poor inverse correlation with TSH and a positive correlation with thyroid binding globulin and albumin, which they should not do." She describes free T3 immunoassays as "useless" and "not even worth performing."

In practice, problems with free T4 measurement don't usually hinder diagnosis, she says. "Most of the time TSH is very solid. It is only a minority of times that you need a good free T4. We still run the old free thyroxine index, which is very robust because total T4 is far more robust than free T4. You can always measure TBG directly and calculate a total T4-to-TBG ratio to overcome binding issues."

"In fact," Dr. Spencer continues, "in pregnancy, where you have high TBG, there is a very predictable rise in total T4 to 1.5 times pre-pregnancy values. So you can actually use total T4 to estimate thyroxine status in pregnancy by merely adjusting the T4 reference interval by 1.5." This method is mentioned in new pregnancy guidelines, she says, as one way to overcome free T4 immunoassay problems when a good T4 estimate is needed in pregnancy.

Pregnancy is a special condition with regard to thyroid function. As Dr. Killeen notes, thyroxine-binding globulin rises, albumin falls, total TSH rises and, in the first trimester, TSH falls due to the thyrotropic effect of human chorionic gonadotropin. Free T4 declines with gestation. "Trimester-specific reference intervals for TSH should be applied," Dr. Killeen says. In the ATA's recent guidelines, recommended ranges for each trimester are: 0.1-2.5; 0.2-3.0; and 0.3-3.0 (Stagnaro-Green A, et al. *Thyroid*. 2011;21:1081). In these guidelines, the ATA does not recommend universal screening of pregnant women. At the University of Minneapolis, Fairview, there is no formal policy now. "It depends on the individual clinician," Dr. Killeen says.

Dr. Spencer emphasizes that maintaining an adequate supply of thyroid hormone is particularly important in the first trimester, when the only source of thyroxine for the developing fetal brain is the mother's supply. Measuring thyroid hormone accurately is crucial at this time. "In the last couple of years there have been increasing recommendations to screen younger women who may become pregnant," she says. "Pregnancy is a thyroid stress, and pregnancy will often unmask occult thyroid insufficiency due to autoimmune thyroid disease." When the first trimester TSH is above 2.5 mIU/L, current guidelines recommend that levothyroxine (L-T4) treatment be considered. "Fortunately, there is less downside to low-dose L-T4 treatment than not treating," Dr. Spencer says. "Studies suggest that iatrogenic subclinical hyperthyroidism will have no negative effect on the mother or fetus."

As for using trimester-specific reference intervals, Dr. Spencer notes that, from a laboratory point of view, this policy presents practical challenges. "Few laboratories would be able to develop a HIPAA-compliant protocol, have it approved by the hospital IRB, and recruit 120 pregnant patients in each of the three trimesters to study," she says.

Whether to screen women looking to get pregnant for antibodies is controversial. "Most experts recommend case finding using a long list of risk factors, including being over age 30 and fatigue," Dr. Spencer says. "Once you factor that in, you will pretty much be screening every woman—what pregnant woman isn't tired? Even though the American College of Obstetricians and Gynecologists does not recommend universal screening, more obstetricians screen women for thyroid function and treat a TSH above 2.5."

For this application, ambiguity about thyroid function testing may be clarified soon. Dr. Killeen says the NIH is currently conducting a randomized placebo study in which thyroid hormone is given to hypothyroid pregnant women. "Results are expected in 2015," he says. □

William Check is a writer in Ft. Lauderdale, Fla.