

Worldwide Quality Control Set to Tame Biomarker Variation

18 November 2009. Call it the end of creative individualism. (Just kidding!) But seriously, to hear leading scientists tell it, the field of Alzheimer disease research has reached a point where, on the translational goal of cerebrospinal fluid diagnosis, it's time for individual centers to set aside their local modus operandi and agree to march in lockstep with the collective. Teasing aside, this is a story about how differences among individual centers' CSF A β and tau measurements are posing a serious problem just as the field prepares to deploy those measures for multicenter drug trials in the earliest pre-dementia stages of AD that critically depend on CSF testing. This is as much a story about a potential solution. A large, external quality-control initiative led by **Kaj Blennow** at Sahlgrenska University Hospital in Molndal near Göteborg, Sweden, has started operating last month. It offers a standard protocol, and it welcomes centers around the world to join. (It's free, too.)

Funded by the Alzheimer's Association, the initiative aims to reduce site-to-site and batch-to-batch CSF test variation. Broadly speaking, this is intended to help CSF diagnostics mature single-center testing (aka "It works fine here!") into robust procedures that yield the same results everywhere. This is necessary, research leaders in academia and industry agree, to enable direct comparison in large research databases and therapeutic trials.

The problem and proposed solution are not unique to AD research. Standardization coupled with external quality control (QC) have brought about consistency and uniform cutoff values in other areas of medicine such as blood glucose or hormone tests, and the same can be done for AD, too, researchers say. Moreover, second-generation biomarker candidates that are beginning to emerge from basic proteomics research in CSF and plasma are likely going to face the same challenge once they have more research evidence behind them. They could at that point be added to the ongoing external QC program, said Blennow's colleague **Henrik Zetterberg** of Sahlgrenska University Hospital.

Fine points in protein measurement, technical details of sample collection, analytical procedure, and test manufacturing—oh, a writer can fairly hear those readers eager for the big new gene or clever new concept quietly sigh, "booring." But wait—the topic is important because the devil in the details of CSF testing could make a hash of the field's collective push to implement biomarker-enhanced diagnostic criteria for a prodromal diagnosis (e.g., [Dubois et al., 2007](#)). Likewise, it could derail expensive multicenter drug trials that use such criteria.

"As we start to use CSF biomarkers in clinical trials to either identify patient populations at risk of disease or to measure progression of disease, it is critically important that we have reproducibility and consistency among sites and investigators around the world. Without standardization of these assays, misinterpretation of data is possible if not likely, and interpretation of data among studies and sites will be challenging if not impossible," **Menelas Pangalos** wrote to Alzforum. After Pfizer's acquisition last month of Wyeth, Pangalos became Chief Scientific Officer of Pfizer's Neuroscience Research Unit.

The issue came to a head at the International Conference on Alzheimer's Disease (ICAD) held last July in Vienna, Austria. In his lecture, **Howard Feldman** of Bristol-

Myers Squibb, in describing the first such [Phase 2 trial](#), mentioned reliability and quality control of the CSF component at the trial's 75 study locations as his biggest concern for the trial. Also at ICAD, **Niek Verwey** of VU University Medical Center in Amsterdam, the Netherlands, presented results from his group's recently published comparison of some 18 centers in Europe and the U.S. Established labs in the field had measured the A β , tau, and phospho-tau content of identical test samples with mean variations of up to 37 percent ([Verwey et al., 2009](#); [ARF related news story](#)). And in a plenary lecture, Zetterberg reported that a massive clinical study, in which 12 international centers in Europe and the U.S. followed 1,583 people for two years, had achieved less accuracy in predicting incipient dementia within the MCI cohort based on CSF testing than do single-center studies ([Mattsson et al., 2009](#)).

“As CSF measurement of protein levels in the brain becomes more validated as a biomarker of early identification of AD, it is disturbing that when different labs take those measures you may not get robust and comparable results,” **Maria Carrillo** of the Alzheimer's Association told this reporter. The problem is acute because variation of this magnitude—in the 20, even 30 percent range—may be as large as the hoped-for effect of the treatment under trial, effectively swallowing up a potential efficacy signal. Scientists generally agree they have to keep test variation reliably below 10 percent.

At ICAD in Vienna, Zetterberg called for worldwide standardization of procedures in observational and therapeutic studies. And things have moved quickly. Carrillo invited some 20 experts in AD fluid biomarkers to meet there, hoping to gauge their interest in creating, and then validating, a standard protocol of CSF collection and analysis. Sixty-six people came. The crowd included many pharma researchers whose companies are increasingly collecting CSF in their trials. “Everyone was for the idea of doing this together. The energy in the room was phenomenal,” Carrillo recalled. Because this many heads don't make a nimble planning group, a handful of representatives from leading institutions formed an e-mail working group that created a [consensus protocol](#) for sample collection, handling, and analysis over the next month. By the time the Society of Neuroscience meeting rolled around last month, a grant proposal to validate this protocol in an external QC program was funded, and some 31 centers in academia, pharma, and biotech had signed up to participate, Carrillo said in Chicago. And already in the last week of October, Blennow's group began sending out QC samples of pooled CSF to the participating centers in 12 countries and three reference laboratories. The program is open for additional participants; for inquiries, contact the QC Programme Coordinator at neurokemi.su@vgregion.se or Henrik Zetterberg at henrik.zetterberg@clinchem.gu.se.

Blennow and others interviewed for this article said they hope the QC program will help participating labs synchronize their procedures and enable them to see how their local performance of a given assay compares relative to an independently established reference range for that assay. If run continuously, the program may also spur test manufacturers to minimize batch-to-batch variation of the assay kits they sell. In addition, the program might offer a new quality claim for studies worldwide, whereby a study would demonstrate in its methods section that its tests were performed within the range established by this external quality control. There are hopes and early discussion about using this QC program to help the Food and Drug Administration validate testing labs faster, Carrillo added.

This worldwide QC program is independent of the Alzheimer's Disease Neuroimaging Initiative (ADNI). The 800-pound gorilla of AD biomarker studies is itself expanding around the globe. The two initiatives are linked indirectly, though. The QC program incorporates some steps toward standardization that ADNI1 has already worked out, and vice versa; the results of the QC program may inform which assay is eventually chosen for use in ADNI2, Carrillo said. The grant for ADNI2 was submitted in October; if funded, enrollment will begin a year from now. Several CSF tests are in wide use and some scientific discussion surrounds their relative strengths and weaknesses. The QC program includes them all, and over the course of the next year is expected to show which ones perform most reliably across sites. "Robustness of CSF measurement is a necessity as we move forward into global trials, including ADNI2 and any other biomarker efforts that take place across the world," Carrillo said.

Importantly, the QC program offers one additional service, Zetterberg told ARF. For QC purposes, the Göteborg group has established large pools of carefully calibrated CSF reference samples. They span the range of A β , tau, and phospho-tau concentrations scientists can expect to encounter in their studies. Besides supplying the QC protocol validation program, this reference CSF is also available upon request to investigators who want to run it alongside study samples in their drug trials. The Göteborg group is known for using such pooled samples as internal controls for their own research and for service measurement of research and clinical samples that are sent to them from across Sweden. Building on that, the group has made enough reference CSF to cover entire therapeutic trials by outside drug sponsors, Blennow said. If each study plate included a QC sample, investigators could monitor the performance of their sites and fix any problems the QC sample might flag. Sponsors could also normalize the data to the QC standard; this would make it easier to compare separate trials to each other. "This is the newest part of the QC program. It offers a practical way of dealing with the variation even now in ongoing studies, before the field has achieved worldwide standardization," Zetterberg added.

This QC program comes free of charge to participating labs. In contrast, ongoing QC programs for other assays, such as glucose, hormones, liver enzymes, cardiac and tumor markers, etc., can cost participating labs a pretty penny. "This is an important service from the Alzheimer's Association," Blennow wrote. What's free, exactly? Receipt, every three months, of QC samples that participating sites can analyze as part of their routine assays or planned studies, as well as inclusion in the ongoing QC reference analysis.

"The QC program is the right way forward," concluded Amsterdam's Verwey, who was among the first researchers to put his finger on the problem when he compared site performance and published the results.

CSF Testing for AD: Single-center Bliss, Multicenter Woe?

19 November 2009. The first generation of CSF biomarkers for Alzheimer disease comprises those that measure the component proteins of the disease's signature pathologic lesions. They have been explored for almost 15 years in some 40 published studies, and the broad consensus is that they basically work. "There are more and more data showing that A β 42 and tau are good diagnostic and predictive markers to

identify AD very early,” said **Kaj Blennow** of Sahlgrenska University Hospital near Göteborg, Sweden. The catch is that most of these studies are done at single centers on patients from the same region, and all analyses for a given paper are typically done on one batch of assay kits. When scientists compare different centers doing things their way, they see a large variation. Even at a single site, variation occurs, such that internal controls run alongside the study samples can show a marked drift over time, Blennow said.

For a single site’s day-to-day clinical testing, the situation is quite workable, scientists interviewed for this article agreed. “If a person comes in and gets a lumbar puncture, we know nearly for certain based on this CSF measure whether this person has AD or not,” said **Niek Verwey** of VU University Medical Center in Amsterdam, the Netherlands. “Within our hospital the test is very good, and within some other hospitals it is also very good, but when you compare one hospital to another, it is not good anymore. And that is a serious problem.”

That sites vary in CSF measurements already came up in an early meta-analysis ([Sunderland et al., 2003](#)). Since then several groups have directly compared test performance in more depth, and their work has revealed that variation comes from many sources. For example, a three-country survey by **Jens Wiltfang**, then at the University of Erlangen-Nürnberg, Germany, noted that the consistency within a given assay (i.e., the same assay run today and again tomorrow on the same samples) was low as specified by the manufacturer, ranging from 7.5 to 3 percent for A β 42, total tau, and p-tau assays. But between labs, those percentages were 29, 26, and 27 percent, respectively ([Lewczuk et al., 2006](#)). The Amsterdam group led by **Rien Blankenstein** ran a larger comparison of 13 centers in 2004 and 18 centers in 2008. The Dutch scientists sent the same test samples out to participating labs across Europe and the U.S., and each lab analyzed the samples with the assays it uses on site for their own clinical and research purposes. The results were concerning. As the centers gained more experience with CSF testing between 2004 and 2008, results on tau did improve somewhat from a 21 to 16 percent variation; but for A β 42 variation widened from 31 to 37 percent between-center difference. That was partly because centers use different assays. When the scientists restricted their analysis to the most widely used test among the 18 centers, the Innotech ELISA from the Belgian company InnoGenetics, mean variation improved from 30 percent in 2004 to 22 percent in 2008, thanks to some standardization. However, that is still too high for large-scale multicenter and drug studies, said Verwey. In addition, this study showed that variation is about as large within a given center as among different centers ([Verwey et al., 2009](#)).

Leslie Shaw of University of Pennsylvania Medical Center in Philadelphia has led an international comparison among seven sites as part of quality control for ADNI1. This round robin found significantly lower variation of the test used in ADNI1, near the needed 10 percent range ([Shaw et al., 2009](#); [ARF related ADNI story](#)), and this is now generally viewed to be the CSF measurement variation in ADNI1. In this round robin, assay kits were shipped to the participating labs along with CSF samples, meaning the seven labs were using not only the same test, but also kits from the same production batch, said Blennow. This captures analytical variation arising from how different labs actually perform the test, but it misses differences between one batch of a given test and the next; nor does it capture differences among the different types of assay that are routinely used in different cities. Shaw’s study, as well as a report by a

German group on a European pilot study of the ADNI protocol, has created some confidence that variation in multicenter studies may be controllable if participating centers ship their study samples to a central reference lab that is highly versed in standardized procedures ([Burger et al., 2009](#)). However, pharma researchers have cautioned that for a CSF test to gain regulatory approval, generally speaking it must perform robustly in many routine settings.

The Three Sources of Variation

“If I am an AD patient and I go to Amsterdam and have a lumbar puncture, my A β is 500. If instead I go to London, it is 400. It will be different again if I go to Boston and again in Chicago. Why is that?” asked Verwey.

The reasons fall into three categories: the samples, the analysis, and the assays, scientists said. The first concerns how samples are collected and handled prior to actual protein measurement. This can vary in numerous ways: from whether the sample is frozen and thawed multiple times before being analyzed or analyzed first after a single freeze-thaw cycle, to when and how it is centrifuged, how long it is stored, down to what kind of containers hold the CSF, and more fine points like this. Scientists showed that A β sticks to polystyrene tubes, necessitating the use of polypropylene tubes. These seemingly persnickety details can jinx the protein measurement ([Schoonenboom et al., 2005](#)), and ideally they should be performed in exactly the same way at every participating site. Another source of variation comes from what time of day people undergo the spinal tap and whether they have eaten. For example, research has shown recently that CSF A β levels fluctuate over the course of the day in healthy people, though that pattern seems to wane in AD ([Bateman et al., 2007](#)); early morning spinal taps control for this source of variation.

Sample collection and handling have been an active topic of discussion for some time (e.g., see reports of [Antecedent Biomarker Working Group](#)). The issue has caused its share of hiccups, such as prompting a re-run of baseline in ADNI1 (see [ARF related news story](#)). However, by now many of the kinks appear to have been largely ironed out, and the [protocol](#) to be used for the QC initiative reflects best practices, said **Maria Carrillo** of the Alzheimer’s Association in Chicago. A separate kind of variation even upstream of sample handling stems from clinical differences among centers, for example, how they classify mild cognitive impairment (MCI) and what ages of patients they include ([Mattsson et al., 2009](#); [Petersen and Trojanowski, 2009](#)).

The second category of error arises from how people actually perform the testing itself, Blennow said. There are many small ways in which one analyst’s procedure differs from another’s, though overall, sites tend to get more skilled at running these tests over time. In the October 15 issue of *Clinical Chemistry*, **Cees Mulder**, Verwey, and colleagues from the Amsterdam group reported that as they monitored their own site’s performance over the course of six years, their results became more stable in the second half as they gained more experience ([Mulder et al., 2009](#)). Like the Göteborg group, the Amsterdam group routinely provides CSF testing for external healthcare providers in the country.

To get a close-up view of this second source of differences, the Amsterdam group invited technicians—the folks who actually run the tests their lab chiefs then present at conferences—to a workshop of side-by-side, elbow-rubbing CSF analysis. On October 19 and 20,, 26 analysts from 17 different European centers did exactly that at

the VU's Alzheimer's center. U.S. labs had received invitations but were unable to send a representative, either for lack of funding or because they use a different assay from the one used at the workshop, Verwey said.

Two analysts who did not know each other paired up into 13 groups; one analyst conducted a widely used A β 42, tau, and p-tau ELISA on three separate CSF pool samples, while the other watched, took exact minutes of each procedural step, and discussed the differences. The technicians received a protocol based on the manufacturer's publications but otherwise followed the procedure they use at their home institution. Everyone analyzed the same samples in the same lab at the same moment and the same temperature, using the same assay batch and the same reagents. By holding all these variables constant, the workshop isolated for detection intra-assay, interpersonal differences inherent to analytical procedure. "It was fantastic fun to do this with an enthusiastic group of people who don't get to travel to meetings very often," Verwey said.

VU's statisticians are still working out the results, but already during the workshop, it became clear that individual technicians do things differently in myriad little ways. Here's a partial list: Some people use a second, transfer ELISA plate that comes with the test kit while others do not; some use all given dilution steps in the ELISA standard line, others skip a dilution they consider superfluous. "People don't necessarily follow all the steps of the manufacturer's instruction," Verwey said. There's more: People used different amounts of sulfuric acid to stop the ELISA reaction, some use reverse pipetting while others don't; some shake the plate while others don't. Some people leave samples on the table between steps, others put them in the fridge; some incubate the ELISA plate at 25 degrees, others at room temperature. "We spotted some 20 points of difference, and we'll work on synchronizing these procedures," Verwey added. The Amsterdam group is a reference site in the Alzheimer's Association-funded QC program, and is working collaboratively with its leaders in Sweden.

The third category of measurement error appears to arise from the assays themselves, several researchers pointed out. On the perhaps most frequently used test, the Innotech ELISAs, several academic groups have reported that the intra-test variability at their site is low within a given production batch. But they have also noticed that the performance of a given test changes from batch to batch. ELISAs contain monoclonal catching and detecting antibodies that are typically generated in hybridoma cell culture. In producing a new batch, slight differences in the medium and other culture conditions, concentration, or purification of the antibodies—even in the reagents and plastics the manufacturer purchases for ELISA production—could all lead to batch-to-batch variability. This is published in a longitudinal study of CSF tau measurement ([Verwey et al., 2008](#)). It introduces uncertainty: Did Joe Smith's tau readout go up a notch because there is more tau in his CSF this year than last, or because the lab used a different assay batch? In an interview, Verwey told ARF that Blankenstein, who heads Clinical Chemistry at the VU Medical Center, purchased enough ELISA all at once for the year 2009. "The results were very stable at the hospital throughout 2009. I think it's probably because we used a single batch number," he said, adding that his group will continue to study this issue in the future.

Confirming this observation, Blennow noted that the external QC initiative, by demonstrating the long-term performance of a given ELISA across production lots,

will encourage companies to ensure that one batch of a given test corresponds completely with the previous one. “Not only sample collection and analysis will become more standardized. In time, assay production will, too, and both site and batch differences will become smaller,” Blennow said.

Batch-to-batch changes may explain, in part, why the cutoff values that Alzheimer’s centers calculate to decide whether a person has AD or not have shifted in recent years. In Amsterdam, the cutoff for A β 42, for example, over the past six years has crept up from 450 ng/L to 500, and is now approaching 550, said Verwey (see also [Mulder et al., 2009](#)). Creeping cutoffs could pose a problem for longitudinal studies. More broadly, the current situation where each center at present has to set its own cutoff due to the center-to-center variation is a challenge for multicenter studies, as well.

The CSF assays puzzle researchers in other ways, too. For example, they don’t show dilution linearity, meaning that if a given CSF sample measures in at 500 ng/L A β , diluting the sample by two will not generate a reading of 250 ng/L. On the contrary, the reading goes up, and then drops with further sample dilution, Verwey said. This may reflect how finicky and dynamic a peptide A β 42 is.

In the past three years, a growing number of laboratories have switched to using a newer, multiplex test that captures readings for A β 42, tau, and p-tau simultaneously in one run. Called INNO-BIA AlzBio3, this test generates different absolute values on a given protein from the corresponding Innostest. For example, the same sample that generates a 470 reading in the Innostest may generate a 160 reading in the AlzBio3. That in itself is not unusual, or troublesome. However, besides large differences in absolute CSF levels between these two methods, the scientists reported a lack of linearity between the assays ([Verwey et al., 2009](#)). This indicated to them that the differences cannot be attributed solely to standardization, and that comparison of these two methods is not useful. “The tests do not correspond tightly. There is no parallelism,” Verwey said.

This can be a problem for multicenter studies if different participating sites use different tests and all results are to be analyzed in one large database. Such studies would be well advised to choose a test in the beginning and stick to it, Blennow said. In addition, study sponsors could consider running a Swedish QC sample alongside whichever company kit they use.

When used clinically to distinguish AD from control, the AlzBio3 works well, scientists interviewed for this article agreed. In other areas of clinical laboratory practice, for example, troponin T measurement following a heart attack, both high-sensitivity and lower-sensitivity tests are helpful so long as each is used with its own reference range, Zetterberg added.

Worldwide Quality Control of CSF Biomarkers—How Does it Work?

20 November 2009. [Part 1](#) in this series laid out the rationale for a new initiative called the Alzheimer’s Association QC Program for CSF Biomarkers, and [Part 2](#) detailed the possible reasons why centers come up with such different results when testing how much A β 42, tau, or phospho-tau float in a given person’s cerebrospinal

fluid. Here, now, is a summary of the nuts and bolts of the QC program itself. The program combines a standardized protocol for sample collection that participants are encouraged to adhere to with periodic measurement of Swedish surplus CSF pool samples.

The [protocol](#) is freely available for download. The pooled CSF is available from **Kaj Blennow** and **Henrik Zetterberg**, at Sahlgrenska University Hospital in Molndal near Göteborg, Sweden.

This group has established pools yielding several thousand samples of human CSF with defined concentrations of these proteins. This is possible because, unlike in the U.S. where fear of lumbar puncture is quite common, in Sweden this procedure has long been part of routine neurologic practice. It is frequently performed in psychiatric, geriatric, and even some general medical practice, as well. Spinal taps have proven to be safe if performed correctly. The one reported side effect is headache; its frequency ranged from 0.9 to 4.1 percent in several published studies of consecutive taps ([Blennow et al., 1993](#); [Andreassen et al., 2001](#); [Peskind et al., 2005](#)). In Sweden, infection is routinely ruled out with spinal taps, and every medical student learns how to perform them. Swedish medical practice, therefore, generates a large amount of surplus CSF that used to be discarded, Blennow said, but now is captured for use in reference pools.

Three times a year, the Sahlgrenska group will ship three quality control samples to each participating site. Two samples are different, one is the same. The identical one will be stored on site and analyzed later to track longitudinal deviation over time. Each site analyzes the pool samples with the same assay they use for their research projects and clinical practice, and then sends the pool samples' results back to the Sahlgrenska group. At each timepoint, three designated reference labs—**Les Shaw**'s at the University of Pennsylvania, Philadelphia; **Rien Blankenstein**'s at VU Medical University in Amsterdam, the Netherlands; and **Piotr Lewczuk**'s at University of Erlangen, Germany, receive and analyze an additional six copies of each sample. The Sahlgrenska site serves as a reference site, too. The results of these combined runs will serve to establish a reference range for each sample.

Blennow's group will summarize the participating sites' results in a report, where each lab can compare how it fares. "Then they know if they are within or outside acceptable limits, and can track down the problem if indeed there is one," Zetterberg said. These problems are all fixable, he added. Sometimes it is as easy as replacing a dying lamp in the ELISA plate reader. The key is knowing when the readout begins to go astray, Zetterberg said.

The Alzheimer's Association so far has funded this initiative for 3 years with the possibility for extension. Blennow hopes that the QC program will continue indefinitely. External quality control accompanies most established testing in other medical disciplines; in heart disease, for example, QC programs have been running for 30 years. "We will need external QC as long as we use CSF biomarkers clinically and in drug trials," Blennow said.

Medical areas such as liver and kidney disease have had similar early problems with site variation and have brought them under control by analyzing the same aliquots of control sample several times a year and comparing results between labs in a broad-based effort. Likewise, tests for prolactin and thyroid-stimulating hormone have

overcome variability problems and now test with around 10 percent variance among hospitals.

Last but not least, this kind of shared effort can lead to universal cutoffs, scientists agreed. At present, site and assay variations require that each Alzheimer disease center calculates its own cutoff. For blood glucose, for example, a worldwide cutoff exists to delineate normal from abnormal. “That was possible only because a QC program has been in place for many years, and it is why we intend to keep this QC program going,” Blennow said. Readers interested in participating in the program can contact the QC Programme Coordinator at neurokemi.su@vgregion.se or Henrik Zetterberg at henrik.zetterberg@clinchem.gu.se.—Gabrielle Strobel.