test: comparison of gas chromatography-mass spectrometry and five automated immunoassays. *Clinical Endocrinology*, **78**, 673–680.

3 Clark, P.M., Neylon, I., Raggatt, P.R. *et al.* (1998) Defining the normal cortisol response to the short synacthen test: implications for the investigation of hypothalamic-pituitary disorders. *Clinical Endocrinology*, **49**, 287–292.

4 Klose, M., Lange, M., Rasmussen, A.K. *et al.* (2007) Factors influencing the adrenocorticotropin test: role of contemporary cortisol assays, body composition and oral contraceptive agents. *Journal of Clinical Endocrinology and Metabolism*, **92**, 1326–1333.

5 Chatha, K.K., Middle, J.G. & Kilpatrick, E.S. (2010) National UK audit of the short Synacthen(r) test. *Annals of Clinical Biochemistry*, **47**, 158–164.

Response to: 'Determining the utility of the 60 min cortisol measurement in the short synacthen test'

Dear Sir,

With respect to the recent article by Chitale *et al.* – Determining the utility of the 60 min cortisol measurement in the short synacthen test,[1] I agree that despite many years of clinical use, the interpretation of the short synacthen test (SST) is still debated. The adrenal response to synacthen is a continuous variable which is dichotomized into adequate/inadequate by the decision limit used in the SST. As the authors rightly point out the only time point that has been validated to a 'gold standard' is the 30 min cortisol sample.

In this study, the authors fail to verify the status of the patients included in the study, they have arbitrarily classified those patients who recorded levels of cortisol <550 nmol/l at 30 min and >550 nmol/l at 60 min as false positives, failing the synacthen test but having 'normal' adrenal reserve; however, this classification is based solely on the index test without mention of a reference test.

The SST is a screening test for adrenal insufficiency. The decision limit chosen is a balance between sensitivity and specificity. This balance depends on the relative importance of missing a diagnosis, false negatives, versus the burden of false positives. Screening tests are generally biased towards sensitivity, as it is deemed more important not to miss a diagnosis, at the cost of accepting lower specificity. Raising the decision limit typically increases the specificity of a test at the expense of lower sensitivity, some patients with the condition will be missed.[2] In those cases where the clinical findings and laboratory testing do not lever enough evidence to make a diagnosis, a second tier test with higher specificity should be performed – in the case of suspected adrenal insufficiency either a metyrapone test or insulin tolerance test (ITT).

I would caution against calculating the sensitivity and specificity of the cortisol thresholds using the other time point as the 'gold standard' as these values are not independent of each other.[3] As the authors point out no patient who passed the test at 30 min, failed at 60 min and this is reflected in the statistics presented in Tables 3 and 4.

Guidelines are available for the reporting of diagnostic accuracy studies, and the STARD initiative provides a checklist for authors.[4]

Richard I. King
Department of Clinical Biochemistry, Mater Pathology, South Brisbane, Qld, Australia
E-mail: richard.kingnz@gmail.com

doi: 10.1111/cen.12289

References

1 Chitale, A., Musonda, P., McGregor, A.M. *et al.* (2013) Determining the utility of the 60 min cortisol measurement in the short synacthen test. *Clinical Endocrinology*, **79**, 14–19.

2 Florkowski, C.M. (2008) Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clinical Biochemist Reviews*, **29** (Suppl. 1), S83–S87.

3 Greenhalgh, T. (1997) How to read a paper: papers that report diagnostic or screening tests. *British Medical Journal*, **315**, 540–543.

4 Bossuyt, P.M., Reitsma, J.B., Bruns, D.E. *et al.* (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *British Medical Journal*, **326**, 41–44.

The Authors' Reply

Dear Prof. Bevan,

We are very grateful to be given the opportunity to reply to the letters by Evans et al. and King about this study.

Evans *et al.* mention that there are several potential sources for error and misinterpretation when using cortisol readings, unadjusted for either gender or assay type. However, in clinical practice, cortisol measurements are often reported as only 'time zero', 'time 30 min' and 'time 60 min', with no additional data on how these should be adjusted according to gender and assay, and thus, how these values should be interpreted without these sources of bias being known. Most decisions would be based only on the clinical interpretation of the three provided values. Therefore, whilst we accept that there are methodological influences on the actual values provided by the test; to the jobbing clinician, these may play a minor role in influencing individual patient management. We would welcome a debate as to whether the additional information on gender and assay differences should be provided by different laboratories for every short synacthen test report to see if this changes treatment decisions. This, however, would necessitate a prospective study to address the validity of the proposed upper limits of normal for the different assays.

We acknowledge that our study was also limited by the fact it was a retrospective analysis. As Evans *et al.* describe, there was a reformulation of the Siemens assay during the time, the short synacthen tests were carried out, and this could have influenced the results. We accept that the results may change because of this, however, only a minority of the results would have been affected and it is difficult to know by how much.

Finally, Evans *et al.* suggest we made a transcription error in our manuscript. We quoted data from the UK NEQAS 2011 Steroid Annual Review but unfortunately, a transposition of the method biases occurred which we then used in good faith. Finlay Mackenzie inadvertently provided us with these incorrect data and he is keen to have the error corrected. He has thus supplied the correct information to Dr Evans and colleagues for inclusion in their letter to Clinical Endocrinology. We are happy to clear up any confusion this may have caused.

We agree with King that the interpretation of a test rests with the balance of specificity and sensitivity of the test itself, and that when there is doubt about the validity of a test that data from an additional, more specific test (in this case, an insulin tolerance test or metyrapone test) are beneficial. However, as mentioned in the study, this was a retrospective case notes analysis of data, and thus, this analysis was not possible. We agree that a prospective study comparing the values obtained by the various tests in subjects suspected of having adrenal insufficiency would be the way to answer the question.

Ketan Dhatariya
Elsie Bertram Diabetes Centre, Norfolk and Norwich University
Hospitals NHS Foundation Trust, Norwich, UK
E-mail: ketan.dhatariya@nnuh.nhs.uk

doi: 10.1111/cen.12304

Evidence of systematic and proportional error in a widely used glucose oxidase analyser: Impact for clinical research?

Real-time glycaemia is a cornerstone for metabolic research, particularly when performing oral glucose tolerance tests (OGTT) or glucose clamps. From 1965 to 2009, the gold standard device for real-time plasma glucose assessment was the Beckman glucose analyser 2 (Beckman Instruments, Fullerton, CA, USA), which technology couples glucose oxidase enzymatic assay with oxygen sensors. Since its discontinuation in 2009, today's researchers are left with few choices that utilize glucose oxidase technology. The first one is the YSI 2300 (Yellow Springs Instruments Corp., Yellow Springs, OH, USA), known to be as accurate as the Beckman.[1] The YSI has been used extensively for clinical research studies and is used to validate other glucose monitoring devices.[2] The major drawback of the YSI is that it is relatively slow and requires high maintenance. The Analox GM9 (Analox Instruments, London, UK), more recent and faster, is increasingly used in clinical research[3] as well as in basic sciences[4] (e.g. 23 papers in *Diabetes* or 21 in *Diabetologia*).

Although a report from the Analox manufacturer shows good linearity in a wide range of glucose concentrations; data assessing its reliability and agreement in clamp and OGTT conditions are scarce. The aim of this study was to assess whether or not the Analox is accurate to serve as a replacement for the YSI during clamp and OGTT studies. Our goal was to analyse the association, reliability and agreement between the two devices, in order to confirm their interchangeability for clinical research.

Two hundred ninety-three plasma specimens from 13 OGTT and hyperinsulinemic euglycaemic clamps from subjects recruited in our ongoing research study were used for this comparison. All subjects signed the IRB-approved consent.

Immediately after drawing, 0·4 ml of blood was placed in microtubes containing 30 I.U. of lithium-heparin and 1 mg sodium fluoride per ml of blood as glucose preservative. Both of these chemicals are known not to interfere with glucose oxidase measurements. Microtubes were spun in a microcentrifuge, and plasma was loaded simultaneously on both the YSI and the Analox. These were previously calibrated as specified by the manufacturers. Calibrations were repeated throughout the OGTT or clamps. Manufacturer's standards of various known concentrations were used to assess quality of calibration throughout the tests. All solutions were kept at 4 °C as suggested by the manufacturers.

To analyse absolute differences, paired-sample *t*-tests were performed between YSI and Analox results. Simple linear regression was used to confirm linear relationship, and its dispersion was assessed by standard error of estimation (SEE). To assess repeatability, the regression line was compared with the identity line. Concordance correlation coefficient (CCC), which contains both measurements of precision (p, Pearson's correlation coefficient) and accuracy (Cb, bias correction factor), was also computed. To confirm agreement, a Bland–Altman plot was carried out. Reliability was assessed using intraclass coefficient correlation (ICC), technical error of measurement (TEM) and coefficient of reliability (R). The percentage of TEM (%TEM) was considered as a measure of interdevice coefficient of variation. All analyses were performed using PASW for Windows version 20.0 (SPSS Inc., an IBM Company, Chicago, IL, USA) and MEDCALC Statistical Software (12.4.0.0; MedCalc Software, Ostend, Belgium). For all tests, statistical significance was set at $P < 0{\cdot}05$.

A mean significant difference of 1·05 mM was found between Analox and YSI ($P < 0{\cdot}001$), indicating a systematic error. Pearson's correlation ($r = 0{\cdot}777$, $P < 0{\cdot}001$) and linear regression ($R^2 = 0{\cdot}604$, $P < 0{\cdot}001$) are presented in Fig. 1(a). The SEE was 0·83 mM. A broad dispersion was observed in the euglycaemic range ($r = 0{\cdot}341$, $R^2 = 0{\cdot}116$, both $P < 0{\cdot}001$) with SEE = 0·86 mM. In values higher than 6·1 mM, dispersion was lesser ($r = 0{\cdot}994$, $R^2 = 0{\cdot}987$, both $P < 0{\cdot}001$) with SEE = 0·19 mM.

Repeatability results indicated weak precision ($P = 0{\cdot}777$) and accuracy (Cb = 0·712), with a low CCC (0·554). The Bland–Altman plot (Fig. 1b) illustrated that the higher the glycaemia, the higher the difference. There was a significant proportional bias (Kendall's Tau = −0·403, $P < 0{\cdot}001$).

Reliability was weak with an ICC of 0·865 ($P < 0{\cdot}001$) and high TEM and %TEM (1·89 mM and 31·9%, respectively). The coefficient of reliability was very low ($R = 0{\cdot}234$).