Alberto Falorni*, Vittorio Bini, Corrado Betterle, Annalisa Brozzetti, Luis Castaño, Marta Fichna, Olle Kämpe, Gunnar Mellgren, Pärt Peterson, Shu Chen, Johan Rönnelid, Jochen Seissler, Claudio Tiberti, Raivo Uibo, Liping Yu, Åke Lernmark and Eystein Husebye

Determination of 21-hydroxylase autoantibodies: inter-laboratory concordance in the Euradrenal International Serum Exchange Program

DOI 10.1515/cclm-2014-1106
Received November 11, 2014; accepted February 11, 2015; previously published online March 26, 2015

***Corresponding author: Alberto Falorni,** MD, PhD, Department of Medicine, University of Perugia, Via E. Dal Pozzo, 06126 Perugia, Italy, Phone: +39-75-5783588, Fax:+39-75-5783940, E-mail: alberto.falorni@unipg.it
Vittorio Bini and Annalisa Brozzetti: Department of Medicine, University of Perugia, Perugia, Italy
Corrado Betterle: Endocrine Unit, Department of Medicine, University of Padua, Padua, Italy
Luis Castaño: Cruces University Hospital, UPV/EHU, Ciberdem, BioCruces, Bilbao, Spain
Marta Fichna: Department of Endocrinology and Metabolism, Poznań University of Medical Science, Poznań, Poland; and Institute of Human Genetics, Polish Academy of Sciences, Poznań, Poland
Olle Kämpe: Department of Medical Sciences, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; and Centre of Molecular Medicine, Department of Medicine (Solna), Karolinska Institutet, Stockholm, Sweden
Gunnar Mellgren: Hormone Laboratory, Haukeland University Hospital and Department of Clinical Science, University of Bergen, Bergen, Norway
Pärt Peterson: Molecular Pathology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia
Shu Chen: FIRS Laboratories, RSR Ltd, Parc Ty Glas, Llanishen, Cardiff, UK
Johan Rönnelid: Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
Jochen Seissler: Medizinische Klinik und Poliklinik IV, Diabetes Zentrum, Klinikum der Universität München, Munich, Germany
Claudio Tiberti: Department of Experimental Medicine, University of Rome "Sapienza", Rome, Italy
Raivo Uibo: Department of Immunology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia
Liping Yu: Barbara Davis Center for Diabetes, University of Colorado Denver, Aurora, CO, USA
Åke Lernmark: Department of Clinical Sciences, Lund University, Skåne University Hospital, Malmö, Sweden
Eystein Husebye: Department of Clinical Science, University of Bergen and Department of Medicine, Haukeland University Hospital, Bergen, Norway

Abstract

Background: 21-Hydroxylase autoantibodies (21OHAb) are markers of an adrenal autoimmune process that identifies individuals with autoimmune Addison's disease (AAD). Quality and inter-laboratory agreement of various 21OHAb tests are incompletely known. The objective of the study was to determine inter-laboratory concordance for 21OHAb determinations.
Methods: Sixty-nine sera from 51 patients with AAD and 51 sera from 51 healthy subjects were blindly coded by a randomization center and distributed to 14 laboratories that determined 21OHAb, either by an "in-house" assay (n=9) using in vitro-translated ^{35}S-21OH or luciferase-labeled 21OH or a commercial kit with ^{125}I-21OH (n=5). Main outcome measures were diagnostic accuracy of each participating laboratory and inter-laboratory agreement of 21OHAb assays.
Results: Intra-assay coefficient of variation ranged from 2.6% to 5.3% for laboratories using the commercial kit and from 5.1% to 23% for laboratories using "in-house" assays. Diagnostic accuracy, expressed as area under ROC curve (AUC), varied from 0.625 to 0.947 with the commercial kit and from 0.562 to 0.978 with "in-house" methods. Cohen's κ of inter-rater agreement was 0.603 among all 14 laboratories, 0.691 among "in-house" laboratories, and 0.502 among commercial kit users. Optimized cutoff levels, calculated on the basis of AUCs, increased the diagnostic accuracy of every laboratory (AUC >0.9 for 11/14 laboratories) and increased the Cohen's κ of inter-rater agreement. Discrepancies in quantitation of 21OHAb levels among different laboratories increased with increasing autoantibody levels.
Conclusions: The quality of 21OHAb analytical procedures is mainly influenced by selection of cutoff value and correct handling of assay materials. A standardization program is needed to identify common standard sera and common measuring units.

Keywords: Addison's disease; adrenal antibodies; autoimmunity; diagnosis; RIA; standardization.

Introduction

The appearance of circulating autoantibodies against the steroidogenic enzyme 21-hydroxylase (21OHAb) is a marker of an ongoing adrenal autoimmune process that may ultimately lead to clinical primary adrenal insufficiency (PAI), also known as autoimmune Addison's disease (AAD) [1]. Assay of 21OHAb is currently used to sub-classify PAI into AAD [1–4], but is also considered the best single immune marker of autoimmune oophoritis in women with primary ovarian insufficiency (POI) [5–8]. Accordingly, analysis of this marker should be offered to all patients with either PAI or POI of unknown origin [1–8]. Furthermore, detection of 21OHAb in healthy subjects or in patients with other autoimmune diseases, in the absence of clinical signs of adrenal insufficiency, defines the so-called preclinical AAD, a condition with increased risk for the development of hypocortisolism [1].

Because of the several clinical applications of the 21OHAb assay, standardization among different laboratories is needed. Currently, the most widely used 21OHAb assays are based on modifications of two original protocols of fluid-phase immunoprecipitation of either ^{125}I-21OH [9] or in vitro-translated ^{35}S-21OH [10, 11].

The First International Serum Exchange for the determination of 21OHAb evaluated inter-laboratory concordance among four independent laboratories in Europe and the USA, using immunoradiometric assays [12]. Although a high rate of positive/negative agreement was observed between laboratories [12], concordance in quantitation of autoantibody concentrations was not satisfactory. It has been proposed that 21OHAb levels in positively scored samples would correlate with the degree of adrenal dysfunction in preclinical AAD [13, 14] and would influence the degree of accuracy of a correct diagnosis of AAD in patients with PAI [4]. Indeed, occasionally, low-level 21OHAb have been detected in patients with unequivocal post-tuberculosis adrenal insufficiency [15, 16], which strengthens the need for an international standardization of 21OHAb measurement.

With the aims of expanding the results of the First International Serum Exchange and paving the way to the identification of standard serum samples to be used in programs of autoantibody standardization and harmonization, a Second International Serum Exchange to evaluate 21OHAb inter-laboratory concordance among a larger group of European and US laboratories was performed.

Materials and methods

Study design

Fourteen laboratories participated in the study with their chosen 21OHAb assay method and were identified by anonymous two-digit codes (AA, ED, EH, EV, GH, ID, IL, IN, NI, ON, GR, SS, TA, and WI) (Table 1). Sample size was calculated on the basis of the accuracy of the estimate of area under curve (AUC) of the receiver-operating characteristic curve (ROC curve) for diagnostic sensitivity and specificity [17]. A sample size of 42 positive subjects and 42 negative subjects was adequate to reach a standard error of 5% at an estimated AUC of 0.85.

Table 1: 21OHAb assays participating in the interlaboratory agreement program.

Laboratory code	21OH antigen	Assay type	Immunoprecipitation	Labeling	Upper level of normal	Antibody titer
AA	Full length	In-house	Protein A	^{35}S	48	Relative index
EH	Full length	In-house	Protein G	Luciferase	47	Relative index
EV	Full length	In-house	Protein A	^{35}S	0.147	Relative index
GR	Full length	In-house	Protein A	^{35}S	8.5	Relative index
ID	Full length	In-house	Protein A	^{35}S	0.07	Relative index
IL	Last 230 C-terminal aa	In-house	Protein G	^{35}S	5	Relative index
SS	Full length	In-house	Protein A	^{35}S	0.06	Relative index
TA	Full length	In-house	Protein A	^{35}S	45	Relative index
WI	Full length	In-house	Protein A	^{35}S	0.150	Relative index
ED	14–495 aa	Commercial kit	Protein A	^{125}I	1	Arbitrary U/mL
GH	14–495 aa	Commercial kit	Protein A	^{125}I	1	Arbitrary U/mL
IN	14–495 aa	Commercial kit	Protein A	^{125}I	1	Arbitrary U/mL
NI	14–495 aa	Commercial kit	Protein A	^{125}I	1	Arbitrary U/mL
ON	14–495 aa	Commercial kit	Protein A	^{125}I	1	Arbitrary U/mL

Arbitrary units in the commercial kit used by laboratories ED, GH, IN, NI, and ON were calculated on dilutions of a high-titer 21OHAb-positive serum. The relative index in the remaining laboratories was: (sample–negative control)/(positive control–negative control). Relative index was ×1000 for laboratories AA, EH, and TA and ×100 for laboratory GR and IL.

Accordingly, sera from 51 Caucasian patients with known AAD (30 women and 21 men; median age and range, 45 years and 19–64 years; median disease duration and range, 6 years and 0–18 years) and from 51 Caucasian healthy control subjects (28 women and 23 men; median age and range, 42 years and 21–64 years) were consecutively collected at the out-patient clinics in Padua and Bergen for the study. Patients with non-autoimmune causes of AAD were excluded from the selection. Sera from 9 of the 51 AAD patients were coded in triplicate samples to evaluate reproducibility and intra-assay coefficient of variation (CV) of each laboratory. Hence, the total number of AAD sera redistributed to each center was 69. A randomization laboratory, which did not participate in the inter-laboratory concordance study, prepared 14 identical sets formed by one hundred twenty 200-μL serum aliquots. The samples were sent deep-frozen in dry ice to each laboratory. The study was approved by the local ethics committees at the participating centers and was conducted in compliance with the World Medical Association Declaration of Helsinki regarding ethical conduct of research involving human subjects.

21OHAb assays

21OHAb determination was performed using either a commercial immunoradiometric kit (RSR, Cardiff, UK) (laboratories ED, GH, IN, NI, and ON) or an "in-house" assay (laboratories AA, EH, EV, ID, IL, GR, SS, TA, and WI) [9–12, 18–24]. Eight "in-house" assays were based on immunoprecipitation of ^{35}S in vitro-translated 21OH (either full length or truncated) and one assay was based on immunoprecipitation of luciferase-labeled antigen (Table 1). The laboratories scored the samples as either positive or negative. In addition, laboratories calculated 21OHAb values for each serum, expressed as either arbitrary units (calculated on dilutions of a high-level 21OHAb-positive serum) or as a relative index: (sample−negative control)/(positive control−negative control) (Table 1). In the laboratories using the commercial kit, 21OHAb levels were derived from the provided standards using the suggested spline log/linear curve. In the case of very high antibody levels, arbitrary units were extrapolated above the highest standard, using the log/linear curve to avoid the "plateau" effect that would have been generated by simply assigning the value of the highest standard. Upper level of normal was calculated as the 99th percentile of a set of healthy control sera by laboratories EV and IL, as the 100th percentile of a set of healthy control sera by laboratory WI, and as mean+3 standard deviations of a set of healthy control sera by all other laboratories. For additional statistical analyses, each laboratory provided also the duplicate cpm values of each sample.

Statistical analyses

The CV of each laboratory was calculated by taking into consideration the total variability of the duplicates for each sample analyzed. To obtain the total CV, both the average of the CV of each duplicate and the square root of the average of their quadrates $\sqrt{(CV^2)}$ (https://www-users.york.ac.uk/~mb55/meas/cv.htm) were calculated because the distribution of CVs was not uniform and was proportional to the size of the measure. Only the estimation of the intra-assay CV, but not inter-assay CV, was possible, as samples were blindly coded and laboratories did not provide information on whether individual samples had been analyzed in the same analytical run or not. Reproducibility was estimated by calculating the free-marginal multirater κ [25] for those serum samples that had been repeated (triplicate samples in nine AAD subjects).

Diagnostic sensitivity was calculated as the percentage of AAD sera that scored positive and diagnostic specificity as the percentage of healthy control sera that scored negative. Diagnostic accuracy was calculated as AUC for a binary diagnostic test (positive/negative) [26]. Subsequently, the cutoff value that offered the best diagnostic accuracy (maximal accuracy cutoff) was recalculated for each laboratory, according to the ROC curves generated from index values, and diagnostic accuracy on quantitative data was recalculated according to that optimized cutoff value. Differences in AUC were tested with modified Z-test [27]. Partial AUC (pAUC), an alternative measure of the diagnostic accuracy that considers regions of the ROC curve with clinically relevant values of sensitivity or specificity, was calculated [28]. More specifically, pAUC was assessed for the high range of specificity between 95% and 100% in the R v.3.0.1 environment (The R Foundation for Statistical Computing, Vienna, Austria, 2014).

Analysis of concordance of qualitative results among different laboratories was performed using the Cohen's κ of inter-rater agreement [29], with Fleiss-Cuzick extension [30] when appropriate, according to the classification of sera as positive or negative provided by each laboratory. The Cohen's κ-test of inter-rater agreement provides a measure of the overlapping of classifications by different methods and/or operators and/or laboratories, and the gradation of the Cohen's κ was <0.2, poor; >0.2–0.4, fair; >0.4–0.6, moderate; >0.6–0.8, good, >0.8–1, very good [31].

For the analysis of concordance of quantitative results among different laboratories, ranked samples from each laboratory were plotted against the ranked samples listed according to increasing 21OHAb levels on the basis of the total results obtained in all the laboratories (samples from 1 to 51 are healthy control sera and samples from 52 to 120 are AAD sera), as previously published for other organ-specific autoantibody assays [32]. Intra-class correlation coefficient (ICC value ranging from 0 to 1), defined as the proportion of variance of an observation due to between-subject variability in the true scores, assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. The type of ICC was 2,1 [33]. A high ICC indicates that there is little variation between the scores given to each item by the raters. Gradation of ICC was 0–0.2, poor agreement; 0.3–0.4, fair agreement; 0.5–0.6, moderate agreement; 0.7–0.8, strong agreement; >0.8, almost perfect agreement [31]. Kendall's τ correlation coefficient between intra-subject standard deviation and intra-subject mean, used to assess the interdependence of these two parameters, was also calculated. A p-value <0.05 was considered statistically significant.

Results

The CVs, calculated by taking into consideration the total variability of the duplicates for each sample analyzed, ranged from 2.6% to 5.3% for the commercial kit and from 5.1% to 23% for "in-house" methods (Table 2). Since the distribution of CVs was not uniform and appeared to be proportional to the size of the measure, $\sqrt{(CV^2)}$ was also calculated (Table 2). $\sqrt{(CV^2)}$ ranged from 3.3% to 8.6% for

Table 2: Variability and reproducibility of 21OHAb assays.

Laboratory	Variability		Reproducibility[a]
	CV	√(CV²)	κ[b]
"In-house" assays			
AA	0.087	0.118	1
EH	0.187	0.229	1
EV	0.054	0.073	1
GR	0.079	0.113	0.703
ID	0.103	0.139	0.259
IL	0.106	0.144	1
SS	0.132	0.226	1
TA	0.230	0.309	0.703
WI	0.051	0.069	1
Commercial kit			
ED	0.028	0.036	1
GH	0.053	0.086	0.703
IN	0.032	0.041	0.703
NI	0.026	0.033	0.851
ON	0.035	0.048	0.703

[a]Calculated on single laboratory replicate sera. [b]Free marginal multirater κ.

the commercial kit and from 6.9% to 30.9% for "in-house" methods. Overall, the commercial kit showed CVs that were lower than most "in-house" assays, with the exception of EV and WI, which showed CVs similar to those obtained with the commercial kit (Table 2). In summary, the total CVs were very good (<5%) for the laboratories using the commercial kit (ED, GH, IN, NI, and ON) and for two laboratories that used "in-house" assays (EV and WI) and good (between 8% and 12%) for most of the other laboratories using "in-house" assays (AA, GR, ID, IL, SS). Two laboratories (ID, TA) showed total CVs approximately 20%. These values doubled when we calculated √(CV²) (Table 2).

Free-marginal multirater κ for the agreement of positive/negative results of nine AAD replicate sera (a measure of the reproducibility of the result for each laboratory) ranged from 0.703 to 1 for the commercial kit and from 0.259 to 1 for the "in-house" methods (Table 2). Median ranks and range of each serum that was distributed in triplicates are shown in Supplementary Material Table 1 that accompanies the article at http://www.degruyter.com/view/j/cclm.2015.53.issue-11/cclm-2014-1106/cclm-2014-1106.xml?format=INT.

Diagnostic sensitivity, specificity, and diagnostic accuracy, likelihood ratio for a positive or negative result (LR+, LR−), positive predictive value (PPV), and negative predictive value (NPV), according to the results provided by each laboratory using internally defined cutoff values, are shown in Table 3 (top). Diagnostic sensitivity ranged from 91.3% to 95.7% for the commercial kit and from 59.4% to 95.7% for the "in-house" methods. Diagnostic specificity varied from 29.4% to 98% for the commercial kit and from 52.9% to 100% for the "in-house" methods. Diagnostic accuracy, expressed as AUC, varied from 0.625 to 0.947 for the commercial kit and from 0.562 to 0.978 for the "in-house" methods.

Subsequently, the cutoff value that offered the best diagnostic accuracy (maximal accuracy cutoff) was recalculated for each laboratory, according to the ROC curves generated from index values (Table 3, bottom). After adjusting for the maximal accuracy cutoff value, AUCs improved for all laboratories, and increased to over 0.9 in 11 (79%) of 14 laboratories, as compared to only 6 (43%) of 14 when using internal cutoff values (Table 3). More specifically, a significant improvement of AUC was observed for laboratories GH ($p<0.001$), NI ($p=0.026$), ON ($p<0.001$), and SS ($p=0.015$) when using the optimized cutoff value.

Figure 1 is graphically representing AUCs of each assay subdivided for cutoff level and for 21OHAb assays used. Diagnostic accuracy of laboratory ID was very low also when using the maximal accuracy cutoff value, resulting in very close to random assignation. Hence, the results of laboratory ID were excluded from the subsequent analysis of inter-laboratory comparison of autoantibody levels.

One AAD serum was found negative in all assays. Another AAD serum was found negative in 13 of 14 assays. A total of 6 of 51 healthy control sera scored negative in all 14 assays. Meanwhile, 29 of 69 AAD samples scored positive in all 14 assays. These samples may be useful for future more refined standardization and harmonization procedures.

The analysis of concordance of qualitative results (classification of subjects as positive or negative by laboratories) showed a Cohen's κ (with Fleiss-Cuzick extension) of 0.603 for the general agreement with 14 laboratories per subject ($p<0.0001$). The Cohen's κ was slightly higher when taking into consideration the nine laboratories that used "in-house" assays (0.691, $p<0.0001$) and lower when considering only the five laboratories that used the commercial 21OHAb assay (0.502, $p<0.0001$). According to the gradation of Cohen's κ [30], the overall concordance among the 14 laboratories was good, but not excellent. Similarly, the concordance among the nine laboratories using "in-house" assays was good, whereas the concordance among the five laboratories using the commercial kit was moderate. The analysis of concordance for couples of laboratories showed an extremely high variability as κ ranged from 0.07 when comparing laboratory ID vs. laboratory NI to 0.966 when comparing laboratory IL vs. laboratory WI.

Table 3: Diagnostic accuracy, likelihood ratio, and predictive value of 21OHAb assays.

Laboratory ID	AUC	95% CI	pAUC between 95% and 100% specificity	Criterion	Sensitivity, %	95% CI	Specificity, %	95% CI	LR+	LR−	PPV	95% CI	NPV	95% CI
Actual cutoff														
"In-house" assays														
AA	0.951	0.896–0.982	0.766	>48	94.2	85.8–98.4	96.1	86.5–99.4	24.0	0.06	97.0	92.9–100	92.5	85.3–99.6
EH	0.879	0.807–0.931	0.736	>47	79.7	68.3–88.4	96.1	86.5–99.4	20.3	0.21	96.5	91.7–100	77.8	67.5–88.0
EV	0.920	0.857–0.962	0.920	>0.142	84.1	73.3–91.8	100	93.0–100	48.0	0.16	100.0	100–100	82.3	72.7–91.8
GR	0.879	0.806–0.931	0.549	≥8.5	85.5	75.0–92.8	90.2	78.6–96.7	8.7	0.16	92.2	85.6–98.8	82.1	72.1–92.2
ID	0.562	0.468–0.652	0.503	>0.07	59.4	46.9–71.1	52.9	38.5–67.1	1.3	0.77	63.1	51.3–74.8	49.1	35.9–62.3
IL	0.964	0.913–0.989	0.971	>5	92.8	83.9–97.6	100	93.0–100	48.8	0.07	100	100–100	91.1	83.6–98.5
SS	0.899	0.870–0.946	0.899	≥0.06	79.7	68.3–88.4	100	93.0–100	9.6	0.2	100	100–100	78.5	68.5–88.5
TA	0.903	0.836–0.950	0.829	>45	82.6	71.6–90.7	98.1	89.5–99.7	42.1	0.18	98.3	94.9–100	80.6	70.8–90.5
WI	0.978	0.933–0.996	0.971	≥0.15	95.7	87.8–99.0	100	93.0–100	48.8	0.04	100	100–100	94.4	88.3–100
Commercial kit														
ED	0.947	0.890–0.979	0.864	>1	91.3	82.0–96.7	98.0	89.5–99.7	46.6	0.09	98.4	95.4–100	89.3	81.2–97.4
GH	0.625	0.532–0.712	0.505	>1	95.7	87.8–99.0	29.4	17.5–43.8	1.4	0.15	64.7	55.4–74.0	83.3	66.1–100
IN	0.886	0.815–0.937	0.599	>1	87.0	76.7–93.8	90.2	78.6–96.7	9.2	0.14	92.3	85.8–98.8	83.6	73.9–93.4
NI	0.795	0.711–0.863	0.521	>1	94.2	85.8–98.4	64.7	50.1–77.6	2.7	0.09	78.3	69.4–87.2	89.2	79.2–99.2
ON	0.741	0.653–0.816	0.514	>1	91.3	82.0–96.7	56.9	42.2–70.6	2.1	0.15	74.1	64.8–83.4	82.9	70.4–95.3
Maximal accuracy cutoff														
"In-house" assays														
AA	0.961	0.909–0.988	0.870	≥59.0	94.2	85.8–98.4	98.0	89.5–99.7	48.0	0.06	98.5	95.5–100	92.6	85.6–99.6
EH	0.920	0.856–0.961	0.767	≥15.4	89.9	80.2–95.8	90.2	78.6–96.7	9.2	0.11	92.5	86.2–98.8	86.8	77.7–95.9
EV	0.971	0.923–0.993	0.971	≥0.064	94.2	85.8–98.4	100	93.0–100	48.0	0.06	100	100–100	92.7	85.9–99.6
GR	0.918	0.853–0.960	0.840	≥12.8	85.5	75.0–92.8	98.0	89.5–99.7	43.6	0.15	98.3	95.1–100	83.3	73.9–92.8
ID	0.578	0.485–0.668	0.504	≥0.008	66.7	54.3–77.6	51.0	36.6–65.2	1.4	0.65	64.8	53.7–75.9	53.1	39.1–67.0
IL	0.978	0.933–0.996	0.978	≥0.56	95.7	87.8–99.0	100	93.0–100	48.8	0.04	100	100–100	94.4	88.3–100
SS	0.971	0.923–0.993	0.957	≥0.025	94.2	85.8–98.4	100	93.0–100	9.6	0.06	100	100–100	92.7	85.9–99.6
TA	0.913	0.847–0.956	0.852	≥28.8	88.4	78.4–94.8	94.1	83.7–98.7	15.0	0.12	95.3	90.1–100	85.7	76.5–94.9
WI	0.978	0.933–0.996	0.971	≥0.043	95.7	87.8–99.0	100	93.0–100	48.8	0.04	100	100–100	94.4	88.3–100
Commercial kit														
ED	0.957	0.903–0.985	0.957	≥1.1	91.3	82.0–96.7	100	93.0–100	23.3	0.09	100	100–100	89.5	81.5–97.4
GH	0.886	0.815–0.937	0.601	≥6.4	87.0	76.7–93.8	90.2	78.6–96.7	8.9	0.14	92.3	85.8–98.8	83.6	73.9–93.4
IN	0.901	0.832–0.947	0.603	≥0.96	89.9	80.2–95.8	90.2	78.6–96.7	9.2	0.11	92.5	86.2–98.8	86.8	77.7–95.9
NI	0.886	0.815–0.937	0.745	≥2.6	81.2	69.9–89.6	96.1	86.5–99.4	20.7	0.20	96.6	91.9–100	79.0	68.9–89.2
ON	0.905	0.838–0.951	0.677	≥1.77	87.0	76.7–93.8	94.1	83.7–98.7	14.8	0.14	95.2	90.0–100	84.2	74.7–93.7

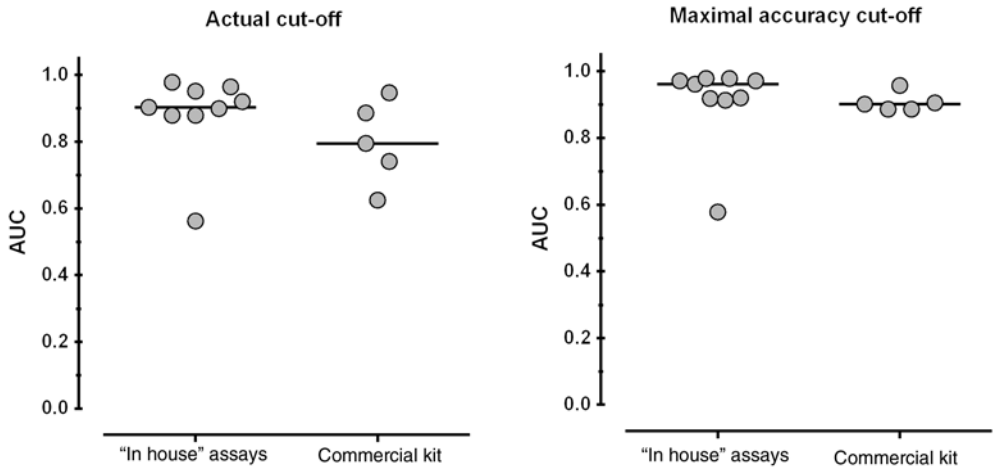

Figure 1: Dot-plot of AUC of the in-house assays and the commercial kit for 21OHAb determination according to cutoff level.

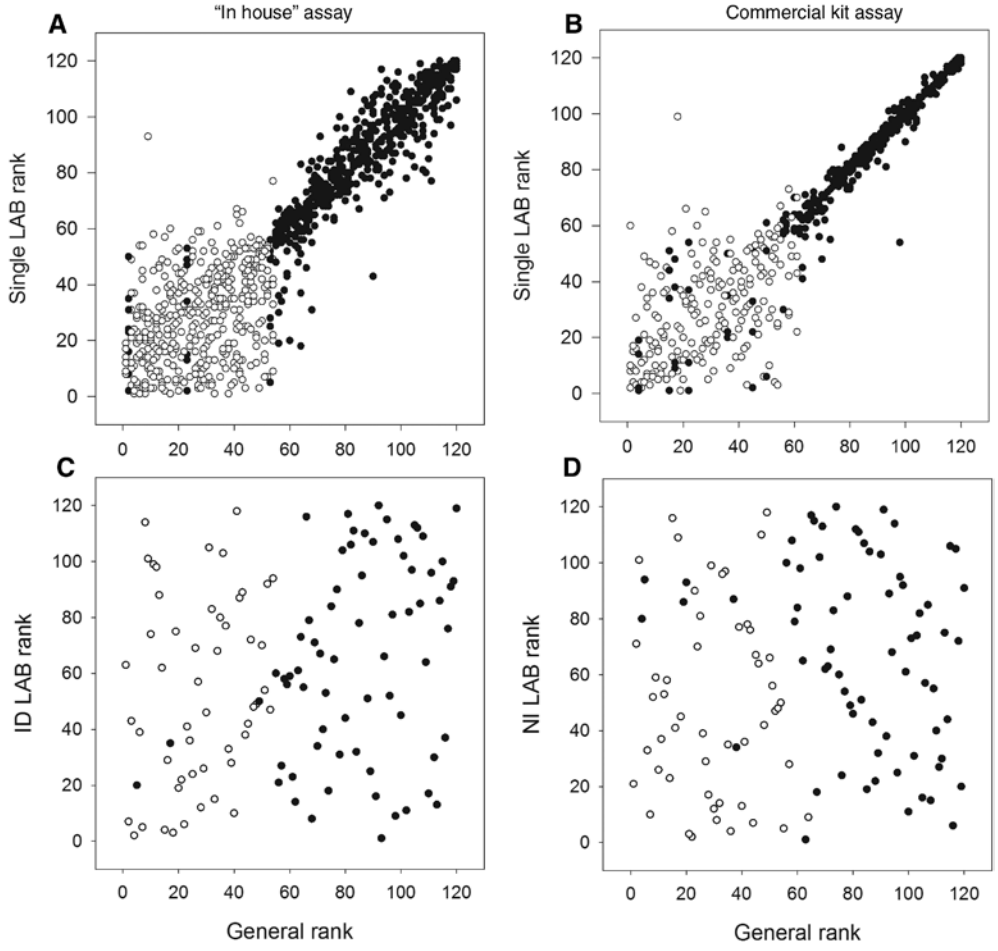

Figure 2: Rank of 21OHAb level in each laboratory plotted against general rank calculated according to the results provided by all the laboratories.

(A) Eight laboratories using "in-house" assays (ID not included); (B) four laboratories using the commercial kit (NI not included); (C) laboratory ID; (D) laboratory NI. In the *x* axis, samples 1–51 are healthy control subjects and samples 52–120 are patients with AAD. In the graph, white circles are healthy control sera and black circles are AAD sera.

When the analysis of concordance was repeated using the optimized cutoff values derived from the ROC curves of each laboratory, the Cohen's κ (with Fleiss-Cuzick extension) increased to 0.760 ($p<0.0001$) for the overall analysis of 14 laboratories, to 0.836 ($p<0.0001$) for the five laboratories that used the commercial 21OHAb assay, and to 0.738 ($p<0.0001$) for the nine laboratories that used "in-house" assays.

To graphically represent the distribution of 21OHAb levels among the different laboratories, ranked samples from each laboratory were plotted against the ranked samples listed according to increasing 21OHAb levels on the basis of the total results obtained in all the laboratories (Figure 2: samples from 1 to 51 are healthy control sera and samples from 52 to 120 are AAD sera). Given the very low diagnostic accuracy of laboratory ID and the high dispersion of 21OHAb levels of laboratory NI, the results of these two laboratories were plotted separately (Figure 2C and D, respectively) from those of the other laboratories (Figure 2A: "in-house" assays, n=8; Figure 2B: commercial kit, n=4). Although the statistical strength of this representation is limited, Figure 2A and B shows a high dispersion of results for the 51 samples from healthy individuals (expected to be negative for 21OHAb) and a lower dispersion for the AAD sera, with values that tended to concentrate along the ideal regression line, which would indicate that every laboratory is generating the same rank. Nevertheless, several results obtained with the AAD sera appeared distant from the ideal line, with a higher dispersion observed for samples with medium-high autoantibody levels in eight "in-house" assays (Figure 2A, samples 80–120) as compared to the same samples analyzed with the commercial kit and as compared to samples with low autoantibody levels. On the contrary, in four assays using the commercial kit, a higher dispersion was observed for PAI samples at low autoantibody level (Figure 2B, samples 52–75) as compared to those with high level.

To better analyze the inter-laboratory concordance for 21OHAb levels, ICC (ranging from 0 to 1) were calculated for all laboratory couples (Table 4). In this type of analysis, a high ICC indicates that there is little variation among 21OHAb levels given to each serum by the laboratories. This analysis showed a strong to almost perfect agreement in 21OHAb levels for the five laboratories using the commercial kit (ICC ranging from 0.658 to 0.957) (Table 4). Meanwhile, the ICCs obtained with the internal comparisons among the nine laboratories using "in-house" assays were extremely variable, as they were extremely low for ID and SS and highest for WI (with a minimum ICC of 0.038 for the comparison between ID and GR and a maximum ICC of 0.906 for the comparison between WI and IL).

All Kendall's τ correlation coefficients between intra-subject standard deviation and intra-subject mean were statistically significant ($p<0.001$) (with values ranging from 0.248 for the comparison SS-IL to 0.923 for the comparison ED-GH), which indicates that the differences in absolute 21OHAb levels between laboratories were increasing with increasing autoantibody level (data not shown).

Table 4: ICCs of the comparisons of 21OHAb levels.

Laboratory	"In-house" assays									Commercial kit				
	AA	EH	EV	GR	ID	IL	SS	TA	WI	ED	GH	IN	NI	ON
"In-house" assays														
AA														
EH	0.735													
EV	0.709	0.800												
GR	0.617	0.721	0.785											
ID	0.059	0.152	0.125	0.038										
IL	0.801	0.814	0.699	0.523	0.211									
SS	0.167	0.229	0.241	0.471	0.055	0.118								
TA	0.437	0.593	0.561	0.781	0.059	0.430	0.426							
WI	0.879	0.835	0.85	0.651	0.154	0.906	0.161	0.523						
Commercial kit														
ED											0.698	0.795	0.907	0.658
GH												0.934	0.876	0.928
IN													0.957	0.853
NI														0.802
ON														

Discussion

21OHAb is the main autoantibody marker to classify PAI as autoimmune [1]. Accordingly, the result of 21OHAb determination is of clinical relevance. First, if negative, further evaluation by adrenal imaging and determination of very-long-chain fatty acids in men is required [3, 4]. Second, patients with AAD almost invariably require replacement therapy with both hydrocortisone and fludrocortisone, whereas patients with other forms of PAI may not require fludrocortisone [2, 34]. Third, AAD patients require screening for other autoimmune diseases, as about two-thirds of them have another autoimmune disease [35]. Finally, the presence of 21OHAb in a healthy person predicts increased risk for development of adrenal and ovarian (in women) insufficiencies in the future [13, 14, 36]. Thus, it is essential that 21OHAb determination be standardized to ensure an accurate classification. In addition, the autoantibody level could be of importance for the accuracy of the diagnosis of AAD [4] and to estimate the risk for future development of PAI [13, 14].

To date, no program for the standardization of adrenal autoantibody determination has been implemented. In a First International Serum Exchange, four laboratories showed a high concordance in the positive/negative score and an overall adequate diagnostic accuracy using immunoradiometric assays [12]. However, relevant discrepancies were detected in the quantitation of autoantibody levels among the four laboratories [12]. Furthermore, the differences among the laboratories in calculation of 21OHAb level were higher for samples with high-autoantibody levels [12]. Based on that first experience [12], the current larger international serum exchange program has been performed by the European Addison network (FP7 Euradrenal). Fourteen laboratories were invited to participate in the program and left free to use their preferred 21OHAb assay to analyze the serum samples. The laboratories used either a commercial kit based on immunoprecipitation of ^{125}I-21OH (n=5) or in-house assays based on immunoprecipitation of in vitro-translated ^{35}S-methionine 21OH (n=8) or luciferase-labeled 21OH (n=1) [9–12, 18–24]. Immunocomplexes were separated using either protein A-Sepharose (n=12) or protein G-Sepharose (n=2) (Table 1). Differences existed between the commercial assay and in-house assays in 21OH construct, antigen labeling procedure, primary incubation time, serum volume requirement, volume of the antigen-antibody reaction, and immunoprecipitation separation (Table 1). In addition, each in-house assay was independently developed by each laboratory while the same commercial kit was used in parallel by five laboratories.

We here report that the commercial kit for 21OHAb determination used by five independent laboratories has a good intra-assay CV, clearly better than that of nine "in-house" assays. Meanwhile, using internally calculated cutoff values, diagnostic specificity was higher for "in-house" methods. Furthermore, in-house assays (with the exception of one laboratory) tended to provide better overall diagnostic accuracies. Accordingly, inter-laboratory concordance was higher among "in-house" assays than among laboratories using the commercial kit.

Diagnostic accuracy of all assays improved when using optimized cutoff values derived by ROC curves, and this phenomenon was more evident for the laboratories using the commercial kit, leading to an overall higher agreement among these laboratories. In addition, the best laboratory using the commercial kit (ED) had a diagnostic accuracy similar to those of the best laboratories using "in-house" assays. Taken together, a major proportion of the discrepancies observed among different laboratories was related to the choice of the cutoff value for positivity and all assays were similarly intrinsically valid. Differences among laboratories related mainly to the way the assay and results were handled. After optimization of the cutoff values, most laboratories showed good or very good diagnostic accuracies and AUC >0.9. Interestingly, when using the optimized cutoff value, four "in-house assays" provided a diagnostic sensitivity higher than 94% with a specificity of 100%; the best commercial kit showed a diagnostic sensitivity higher than 91% with a specificity of 100%. Hence, it is evident that the currently available assays, either developed "in-house" or commercial, may potentially guarantee a diagnostic accuracy that approaches 100% accuracy when a proper cutoff level is established. Accordingly, our study paves the way to the potential large-scale dissemination of 21OHAb determination to routine clinical laboratories with experience in immunoradiometric analysis. In addition, the results of our inter-laboratory serum exchange have important practical implications for 21OHAb analytical procedures in routine clinical laboratories, in which the quality of the results will likely be related to the selection of the cutoff value and correct handling of assay materials. Hence, the need for every laboratory to recalibrate the threshold of positivity locally. Meanwhile, as endocrinologists are expected to use the results of 21OHAb analysis in clinical practice for diagnosis of AAD and for identification of at-risk individuals, it is essential that they are informed of the potential discrepancies among results provided by different laboratories as well as of the diagnostic accuracy of this analysis in routine laboratories.

The reasons why a single laboratory, using an "in-house" method similar to those used by other laboratories (ID), performed so badly in terms of both diagnostic sensitivity and specificity are unclear. However, one cannot exclude the possibility that pre-analytical mistakes in handling of samples or exchange of sample codes may have generated those striking results. Therefore, the data generated by laboratory ID were removed from part of the statistical analysis. Laboratory GH was in a different situation, as it provided the highest diagnostic sensitivity and the lowest diagnostic specificity. In this case, it was clear that the selection of a low cutoff value was the major reason for the low diagnostic specificity. Indeed, when the cutoff value was increased from 1 to 6.4 units, diagnostic specificity of laboratory GH increased from 29.4% to 90.2% and AUC increased from 0.625 to 0.886, a value similar to those observed in other laboratories.

A major additional issue raised by our current study is the different results in quantitation of 21OHAb levels in autoantibody-positive sera. Interestingly, the five laboratories using the commercial kit provided a strong to almost perfect agreement in ranking the samples according to 21OHAb level. However, as already previously observed in the First International Serum Exchange [12], inter-laboratory discrepancies in absolute autoantibody quantitation increased significantly with increasing antibody level, as demonstrated by the statistically significant Kendall's τ coefficients in our present study, which confirms that 21OHAb quantitation of a given laboratory cannot yet be interchanged with that of another and that future standardization programs will be needed to identify common standard sera and common measuring units.

Acknowledgments: A.F. is chairing a Sub-Committee on Organ-Specific Autoantibodies of the IUIS Committee on Quality Assessment and Standardization (www.iuisonline.org) and received an unconditioned support from IUIS. We wish to thank Jose Ramón Bilbao, Åsa Hallgren, Hege Hoff Skavøy, Kai Kisand, Belinda Lind, Dongmei Miao, Maire Pihlap, Bernard Rees-Smith, Koit Reimand, and Ingrid Wigheden for valuable discussion and technical help.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Financial support: The study was supported by EU FP7 (grant number 201167), Euradrenal, and in part by NIH grant DK32083.

Disclosure statement: RSR Ltd is a manufacturer of medical diagnostics including kits for 21OH autoantibodies.

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Winqvist O, Karlsson FA, Kämpe O. 21-Hydroxylase, a major autoantigen in idiopathic Addison's disease. Lancet 1992;339:1559–62.
2. Husebye ES, Allolio B, Arlt W, Badenhoop K, Bensing S, Betterle C, et al. Consensus statement on the diagnosis, treatment and follow-up of patients with primary adrenal insufficiency. J Internal Medicine 2014;275:104–15.
3. Arlt W. The approach to the adult with newly diagnosed adrenal insufficiency. J Clin Endocrinol Metab 2009;94:1059–67.
4. Falorni A, Laureti S, De Bellis A, Zanchetta R, Tiberti C, Arnaldi G, et al. Italian Addison Network Study: update of diagnostic criteria for the etiological classification of primary adrenal insufficiency. J Clin Endocrinol Metab 2004;89:1598–604.
5. Falorni A, Laureti S, Candeloro P, Perrino S, Coronella C, Bizzarro A, et al. Steroid-cell autoantibodies are preferentially expressed in women with premature ovarian failure who have adrenal autoimmunity. Fertil Steril 2002;78:270–9.
6. Dal Pra C, Chen S, Furmaniak J, Smith BR, Pedini B, Moscon A, et al. Autoantibodies to steroidogenic enzymes in patients with premature ovarian failure with and without Addison's disease. Eur J Endocrinol 2003;148:565–70.
7. Bakalov VK, Anasti JN, Calis KA, Vanderhoof VH, Premkumar A, Chen S, et al. Autoimmune oophoritis as a mechanism of follicular dysfunction in women with 46,XX spontaneous premature ovarian failure. Fertil Steril 2005;84:958–65.
8. La Marca A, Brozzetti A, Sighinolfi G, Marzotti S, Volpe A, Falorni A. Primary ovarian insufficiency: autoimmune causes. Curr Opin Obstet Gynecol 2010;22:277–82.
9. Tanaka H, Perez MS, Powell M, Sanders JF, Sawicka J, Chen S, et al. Steroid 21-hydroxylase autoantibodies: measurements with a new immunoprecipitation assay. J Clin Endocrinol Metab 1997;82:1440–6.
10. Falorni A, Nikoshkov A, Laureti S, Grenbäck E, Hulting AL, Casucci G, et al. High diagnostic accuracy for idiopathic Addison's disease with a sensitive radiobinding assay for autoantibodies against recombinant human 21-hydroxylase. J Clin Endocrinol Metab 1995;80:2752–5.
11. Colls J, Betterle C, Volpato M, Prentice L, Smith BR, Furmaniak J. Immunoprecipitation assay for autoantibodies to steroid 21-hydroxylase in autoimmune adrenal diseases. Clin Chem 1995;41:375–80.
12. Falorni A, Chen S, Zanchetta R, Yu L, Tiberti C, Bacosi ML, et al. Measuring adrenal autoantibody response: interlaboratory concordance in the first international serum exchange for the determination of 21-hydroxylase autoantibodies. Clin Immunol 2011;140:291–9.
13. Laureti S, De Bellis A, Muccitelli VI, Calcinaro F, Bizzarro A, Rossi R, et al. Levels of adrenocortical autoantibodies correlate with the degree of adrenal dysfunction in subjects with preclinical Addison's disease. J Clin Endocrinol Metab 1998;83:3507–11.

14. Coco G, Dal Pra C, Presotto F, Albergoni MP, Canova C, Pedini B, et al. Estimated risk for developing autoimmune Addison's disease in patients with adrenal cortex autoantibodies. J Clin Endocrinol Metab 2006;91:1637–45.
15. Nomura K, Depura H, Saruta T. Addison's disease in Japan: characteristics and changes revealed in a nationwide survey. Intern Med 1994;33:602–6.
16. do Carmo Silva R, Kater CE, Dib SA, Laureti S, Forini F, Cosentino A, et al. Autoantibodies against recombinant human steroidogenic enzyme 21-hydroxylase, side-chain cleavage and 17α-hydroxylase in Addison's disease and autoimmune polyendocrine syndrome type III. Eur J Endocrinol 2000;142:187–94.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
18. Fichna M, Fichna P, Gryczyńska M, Walkowiak J, Zurawek M, Sowiński J. Screening for associated autoimmune disorders in Polish patients with Addison's disease. Endocrine 2010;37:349–60.
19. Seissler J, Schott M, Steinbrenner H, Peterson P, Scherbaum WA. Autoantibodies to adrenal cytochrome P450 antigens in isolated Addison's disease and autoimmune polyendocrine syndrome type II. Exp Clin Endocrinol Diabetes 1999;107:208–213.
20. Peterson P, Uibo R, Peränen J, Krohn K. Immunoprecipitation of steroidogenic enzyme autoantigens with autoimmune polyglandular syndrome type I (APS I) sera; further evidence for independent humoral immunity to P450c17 and P450c21. Clin Exp Immunol 1997;107:335–40.
21. Yu L, Brewer KW, Gates S, Wu A, Wang T, Babu SR, et al. DRB1*04 and DQ alleles: expression of 21-hydroxylase autoantibodies and risk of progression to Addison's disease. J Clin Endocrinol Metab 1999;84:328–35.
22. Myhre AG, Undlien DE, Løvås K, Uhlving S, Nedrebø BG, Fougner KJ, et al. Autoimmune adrenocortical failure in Norway autoantibodies and human leukocyte antigen class II associations related to clinical features. J Clin Endocrinol Metab 2002;87:618–23.
23. Söderbergh A, Myhre AG, Ekwall O, Gebre-Medhin G, Hedstrand H, Landgren E, et al. Prevalence and clinical associations of 10 defined autoantibodies in autoimmune polyendocrine syndrome type I. J Clin Endocrinol Metab 2004;89:557–62.
24. Burbelo PD, Goldman R, Mattson TL. A simplified immunoprecipitation method for quantitatively measuring antibody responses in clinical sera samples by using mammalian-produced Renilla luciferase-antigen fusion proteins. BMC Biotechnol 2005;5:22.
25. Warrens JM. Inequalities between multi-rater kappas. Adv Data Anal Classif 2010;4:271–86.
26. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Making 2000;20:468–70.
27. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
28. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
29. Altman DG. Practical statistics for medical research. Chapman & Hall: London, 1995:403–9.
30. Fleiss JL, Cuzick J. The reliability of dichotomous judgements: unequal numbers of judges per subject. Appl Psychol Meas 1979;3:537–42.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
32. Verge CF, Stenger D, Bonifacio E, Colman PG, Pilcher C, Bingley PJ, et al. Combined use of autoantibodies (IA-2 autoantibody, GAD autoantibody, insulin autoantibody, cytoplasmic islet cell antibodies) in type 1 diabetes: Combinatorial Islet Autoantibody Workshop. Diabetes 1998;47:1857–66.
33. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–7.
34. Falorni A, Minarelli V, Morelli S. Therapy of adrenal insufficiency: an update. Endocrine 2013;43:514–28.
35. Betterle C, Scarpa R, Garelli S, Morlin L, Lazzarotto F, Presotto F, et al. Addison's disease: a survey of 633 patients in Padova. Eur J Endocrinol 2013;169:773–84.
36. Falorni A, Brozzetti A, Calcinaro F, Marzotti S, Santeusanio F. Recent advances in adrenal autoimmunity. Exp Rev Clin Endocrinol Metab 2009;4:333–48.

Supplemental Material: The online version of this article (DOI: 10.1515/cclm-2014-1106) offers supplementary material, available to authorized users.