Side by side comparison of three fully automated SARS-CoV-2 antibody assays with a focus on specificity

Running head: Specificity of automated Anti-SARS-CoV-2 assays

Thomas Perkmann[1], Nicole Perkmann-Nagele[1], Marie-Kathrin Breyer[2], Robab Breyer-Kohansal[2], Otto C Burghuber[3], Sylvia Hartl[2,3], Daniel Aletaha[4], Daniela Sieghart[4], Peter Quehenberger[1], Rodrig Marculescu[1], Patrick Mucher[1], Robert Strassl[1], Oswald F Wagner[1], Christoph J Binder[1], and Helmuth Haslacher[1*]

[1]Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria
[2]Department of Respiratory and Critical Care Medicine and Ludwig Boltzmann Institute for COPD and Respiratory Epidemiology, Otto Wagner Hospital, Vienna, Austria
[3]Sigmund Freud University, Medical School and Ludwig Boltzmann Institute for COPD and Respiratory Epidemiology, Vienna, Austria
[4]Divison of Rheumatology, Department of Medicine III, Medical University of Vienna, Vienna, Austria

*Please address all correspondence to:
Helmuth Haslacher, MD PhD MSc BSc BA
Medical University of Vienna, Department of Laboratory Medicine
Waehringer Guertel 18-20, A-1090 Vienna
Phone: +43 1 40400 53190
Fax: +43 1 40495 15547
E-Mail: helmuth.haslacher@meduniwien.ac.at

Key words:
SARS-CoV-2; serology; specificity; laboratory automation; positive predictive value; seroprevalence;

List of abbrevations:

COVID-19	Coronavirus disease 2019
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
RT-PCR	Reverse transcriptase-polymerase chain reaction
ECLIA	Electrochemiluminescence assay
95% CI	95% confidence interval
CMIA	Chemiluminescent microparticle immunoassay
NPA	Negative percentage agreement
CLIA	Chemiluminescence immunoassay
ROC	Receiver-operating-characteristics
AUC, AUROC	Area under the (ROC-)curve
LOD	Limit of detection
PPV	Positive predictive value
NPV	Negative predictive value
nAbs	Neutralizing antibodies
EUA	Emergency Use Authorization

ABSTRACT

Background: In the context of the COVID-19 pandemic, numerous new serological test systems for the detection of anti-SARS-CoV-2 antibodies have become available quickly. However, the clinical performance of many of them is still insufficiently described. Therefore we compared three commercial, CE-marked, SARS-CoV-2 antibody assays side by side.

Methods: We included a total of 1,154 specimens from pre-COVID-19 times and 65 samples from COVID-19 patients (≥14 days after symptom onset) to evaluate the test performance of SARS-CoV-2 serological assays by Abbott, Roche, and DiaSorin.

Results: All three assays presented with high specificities: 99.2% (98.6-99.7) for Abbott, 99.7% (99.2-100.0) for Roche, and 98.3% (97.3-98.9) for DiaSorin. In contrast to the manufacturers' specifications, sensitivities only ranged from 83.1% to 89.2%. Although the three methods were in good agreement (Cohen's Kappa 0.71-0.87), McNemar's test revealed significant differences between results obtained from Roche and DiaSorin. However, at low seroprevalences, the minor differences in specificity resulted in profound discrepancies of positive predictability at 1% seroprevalence: 52.3% (36.2-67.9), 77.6% (52.8-91.5), and 32.6% (23.6-43.1) for Abbott, Roche, and DiaSorin, respectively.

Conclusion: We find diagnostically relevant differences in specificities for the anti-SARS-CoV-2 antibody assays by Abbott, Roche, and DiaSorin that have a significant impact on the positive predictability of these tests.

Introduction

COVID-19 is a new disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), which was first described by Chinese scientists in early January 2020 (1). On March 11, the WHO officially declared the novel SARS-CoV-2 infections a pandemic, which has now spread rapidly across the entire globe, with almost 6.5 million confirmed cases and over 375,000 confirmed deaths (2). COVID-19 is characterized by a broad spectrum of individual disease courses, ranging from asymptomatic infections to the most severe cases requiring intensive medical care (3).

The reliable detection of infected persons and, subsequently, their isolation is essential for the effort to prevent the spread of the SARS-CoV-2 virus quickly and efficiently. Therefore, reverse transcriptase-polymerase chain reaction (RT-PCR) testing is required for direct detection of the pathogen. Unfortunately, RT-PCR testing does not always give a clear answer to whether the SARS-CoV-2 infection is currently present or not (4,5).

On the other hand, serological testing for SARS-CoV-2 specific antibodies can be used as an additional diagnostic tool in case of suspected false-negative RT-PCR results (6) or for individual determination of antibody levels. Moreover, cross-sectional serological studies provide essential epidemiological information to allow a correct estimation of the spread of the disease within a population (7,8). The first commercially available serological SARS-CoV-2 tests, mostly standard ELISA tests or lateral flow rapid tests, have not always proved to be sufficiently specific and sensitive (9,10). Recently, the first tests for fully automated large-scale laboratory analyzers have been launched. The present evaluation aims to compare three of these test systems manufactured by Abbott (11), DiaSorin (12), and Roche (13), with particular emphasis on specificity, which is crucial for an adequate positive predictive value given the current low seroprevalence worldwide.

Materials and methods

Study design and patient cohorts

The present study aims at a detailed comparison of three automated SARS-CoV-2 detection methods with a particular focus on specificity and positive predictability. A total of 1,154 samples from three cohorts of patients/participants with sampling dates before 01.01.2020 were used to test specificity. The samples derived from three different collections: a cross-section of the Viennese population, LEAD study (14), preselected for samples collected between November and April to enrich seasonal infections (n=494); a collection of healthy voluntary donors (n=302; 269 individuals, 11 donors with a 4-fold repetition of the donation within a median period of 4.5 years [3.6-5.5]); a disease-specific collection of samples from patients with rheumatic diseases (n=358).

For estimation of test sensitivity, samples of 65 COVID-19 donors/patients with a symptom onset to analysis time of ≥14 days (median time interval of 41 [28-49] days) were evaluated in parallel on all three analysis platforms. For asymptomatic donors (n=6), SARS-CoV-2 RT-PCR confirmation to analysis time was used instead. We subjected only a single serum sample per patient to sensitivity analysis to avoid data bias due to uncontrolled multiple measurement points of individual patients.

Supplementary Table 1 gives a comprehensive overview of characteristics and cohort-specific inclusion and exclusion criteria; Supplementary Table 2, 3, and Supplementary Fig. 1 provide additional descriptive statistics on donors/patients included in the cohorts.

All included participants gave written informed consent for donating their samples for scientific purposes. From patients, only left-over material from diagnostic procedures was used. The overall evaluation plan conformed with the Declaration of Helsinki as well as with relevant regulatory requirements. It was reviewed and approved by the ethics committee of the Medical University of Vienna (1424/2020).

Biomaterials

Used serum samples were either left-over materials from diagnostic procedures (Department of Laboratory Medicine, Medical University of Vienna) or part of a sample cohort processed and stored by the MedUni Wien Biobank. All pre-analytical processes were carried out according to standard operating procedures in an ISO 9001:2008/2015-certified (MedUni Wien Biobank, Department of Laboratory Medicine) and ISO 15189:2012-accredited (Department of Laboratory Medicine) environment. Standard sample protocols were described previously (15).

Antibody testing

SARS-CoV-2 specific antibodies were measured according to the manufacturers' instructions on three different automated platforms at the Department of Laboratory Medicine of the Medical University of Vienna.

1. The Elecsys® Anti-SARS-CoV-2 assay (Roche Diagnostics, Rotkreuz, Switzerland) was applied on a Cobas e 801 modular analyzer. It detects total antibodies against the SARS-CoV-2 nucleocapsid (N) antigen in a sandwich electrochemiluminescence assay (ECLIA). For the suggested cut-off of ≥1 COI, the manufacturer reports specificities of 99.80% (95% CI: 99.58 – 99.92) for samples derived from diagnostic routine, 99.83% (99.51 – 99.97) for blood donors, and 100% (91.19 – 100) for both a common cold panel and a Coronavirus panel. Sensitivity was estimated as 65.5% (56.1 – 74.1) during days 0 – 6 post RT-PCR confirmation, 88.1% (77.1 – 95.1) from day 7 to day 13, and 100% (88.1 – 100%) from day 14 on (13). According to the manufacturer, the system delivers qualitative results, either being reactive or non-reactive for anti-SARS-CoV-2 antibodies.

2. IgG antibodies against SARS-CoV-2 nucleocapsid (SARS-CoV-2 IgG) were quantified employing a chemiluminescent microparticle immunoassay (CMIA) on the Abbott ARCHITECT® i2000sr platform (Abbott Laboratories, Chicago, USA). For the cut-off of ≥1.4 Index (S/C), the manufacturer gives a negative percentage agreement (NPA, corresponding to specificity) of 99.60% (95% CI: 98.98 – 99.89) calculated from samples collected before the COVID-19 outbreak and of 100.00% (95.07 – 100.00) in patients with other respiratory illnesses. Regarding diagnostic sensitivity, 0.00% (0.00 – 60.24) are reported <3 days after symptom onset, 25.00% (3.19 – 65.09) on days 3 – 7, 86,36% (65.09 – 97.09) on days 8 – 13, and 100.00% (95.89 – 100.00) from day 14 on (11). According to the manufacturer, the assay is designed for the qualitative detection of IgG antibodies to SARS-CoV-2.

3. The LIAISON® SARS-CoV-2 S1/S2 IgG test detects IgG-antibodies against the S1/S2 domains of the virus' spike protein in a chemiluminescence immunoassay (CLIA). The test was applied to a LIAISON® XL Analyzer (DiaSorin S.p.A., Saluggia, Italy). The manufacturer reports a diagnostic specificity of 98.5% (95% CI: 97.5 – 99.2) in blood donors and 98.9% in presumably SARS-CoV-2 negative diagnostic routine samples (94.0 – 99.8). Applying a cut-off >15.0 AU/mL (borderline results 12.0 – 15.0, require a re-test algorithm), the test's sensitivity is reported time-dependently with 25.0% (14.6 – 39.4) ≤5 days after RT-PCR-confirmed diagnosis, 90.4% (79.4 – 95.8) from day 5 to day 15, and 97.4% (86.8 – 99.5) after >15 days (12). Samples that repeatedly tested borderline were classified as positive. The manufacturer indicates to provide quantitative measurement results on the system.

Statistical analysis

Unless stated otherwise, continuous data are given as median (quartile 1 – quartile 3). Categorical data are given as counts and percentages. Diagnostic sensitivity and specificity, as well as positive and negative predictive values, were calculated using MedCalc software 19.2.1 (MedCalc Ltd., Ostend, Belgium). 95% confidence intervals (CI) for sensitivity and specificity were calculated according to Clopper and Pearson ("exact" method) with Standard logit

confidence intervals for the predictive values (16). Receiver-Operating-Characteristic (ROC)-curve analysis was used to evaluate test accuracy and compare the diagnostic performance of the three test systems, according to DeLong et al. (17). Between-test agreements were assessed by interpretation of Cohen's Kappa-statistics, and further evaluated with McNemar's tests. Statistical significance was assumed at $p<0.05$. Figures were produced with MedCalc software 19.2.1 and GraphPad Prism 8 (GraphPad Software, San Diego, USA).

Results

Specificity

To describe assay specificity, we used a total of 1,154 serum samples collected before SARS-CoV-2 circulated in the population and which are, by definition, negative for SARS-CoV-2 specific antibodies. The three different specificity cohorts A-C (described in detail in Supplementary Table 1-3 and Supplementary Figure 1) presented with different rates of false-positives (Table 1) - cohort C (cohort of rheumatic diseases) showing the highest reactivities. We found in total 3, 9, and 20 false-positive samples for Roche, Abbott, and DiaSorin, leading to an assay specificity of 99.7% (95%CI: 99.2-100.0), 99.2 (95%CI: 98.6-99.7), and 98.3% (95%CI: 97.3-98.9) respectively (Figure 1A-C). Median and 90th percentile values of negative samples were 0.025 and 0.115 Index for Abbott, 0.0815 and 0.0927 COI for Roche, and below LOD and 5.52 AU/ml for DiaSorin. False-positive samples yielded median values of 1.65 COI (1.47-1.72) for Roche Elecsys® Anti-SARS-CoV-2 (cut-off: ≥1 COI), 2.21 Index (2.14-2.67) for Abbott SARS-CoV-2 IgG (cut-off: ≥1.4 Index), and 22.4 AU/ml (17.38-57.35) for DiaSorin LIAISON® SARS-CoV-2 S1/S2 IgG (negative <12.0 AU/ml, equivocal 12.0 AU/ml <= x <15.0 AU/ml, positive ≥15.0 AU/ml).

Sensitivity

To estimate assay sensitivity, we used serum samples from 65 donors/patients at later time points following SARS-CoV-2 infection, at least ≥14 days after symptom onset (median interval of 41 [28-49] days). In this late phase, we assumed the majority of donors/patients having reached prominent and constant levels of SARS-CoV-2 specific antibodies. Surprisingly, we could find a relatively high percentage of samples that were testing negative for SARS-CoV-2 antibodies: DiaSorin 11, Abbott 10, and Roche 7 false-negatives leading to calculated sensitivities of 83.1 (71.3-91.2)%, 84.6% (73.6-92.4), and 89.2% (79.1-95.6), respectively. Five serum samples were consistently tested negative in all three assays despite being derived from individuals tested positive for SARS-CoV-2 by RT-PCR. All seven false-negatives in the Roche test

overlapped with false-negatives in the Abbott test (both nucleocapsid-antigen based assays), whereas DiaSorin was negative for an additional six serum samples exclusively (S1/S1-domain antigen-based assay).

PPV and NPV

Although specificity and sensitivity are essential criteria for assessing the quality of a test procedure, they have little informative value about the probability of a positive/negative test result, indicating the presence/absence of SARS-CoV-2 specific antibodies without taking prevalence into account. Therefore, a comparative overview for specificity, sensitivity, as well as positive and negative predictive values at 1%, 5%, and 10% SARS-CoV-2 antibody seroprevalence is shown in Figures 2A and 2B and summarized in Table 2. While the differences between the test systems for different seroprevalences do not have a significant impact on NPV (range 98.1%-99.9%), the consequences for PPV are pronounced. At seroprevalence rates of 10%, all three systems show acceptable PPVs of 97.4%, 92.3.%, and 84.2% for Roche, Abbott, and DiaSorin, but at 1% seroprevalence these drop to unsatisfactorily or even unacceptably low values of 77.6% (52.8-91.5), 52.3% (36.2-67.9), and 32.6% (23.6-43.1) for Roche, Abbott, and DiaSorin.

ROC Curve Analysis

As shown in Figure 3A-C, all three ROC curves presented with areas under the curves (AUC) above 0.97 (Abbott: 0.994 [95% CI: 0.987-0.997], Roche: 0.989 [0.981-0.994], DiaSorin: 0.977 [0.967-0.985]). Comparison of ROC-AUCs, according to DeLong et al., did not reveal significant differences (Differences: Abbott/Roche p=0.487, Abbott/DiaSorin p=0.112, Roche/DiaSorin p=0.395). In the next step, we aimed to assess whether modifying the cut-off values could improve the explanatory power of the ROC-curves.

Cut-offs associated with the Youden's index (maximum sum of sensitivity and specificity) of >0.42, >0.355, and >8.76 for Abbott, Roche, and DiaSorin lowered the PPV considerably, being as low as 26.4 (20.3-33.6), 62.1 (43.9-77.4), and 24.8 (18.9-31.9) at 1% seroprevalence for Abbott, Roche, and DiaSorin (Supplementary Table 4).

Between-test agreement/disagreement

Correlation analysis of measurement values between the different platforms showed only moderate to weak concordance. The Pearson correlation coefficient was r=0.66 (p<0.001), for both Abbott/DiaSorin (both IgG-assays) and Abbott/Roche (both nucleocapsid-antigen based assays). In contrast, Roche/DiaSorin, with a coefficient of r=0.25 and a hardly reached significance of 0.044, could only show a very weak correlation.

Therefore, the test systems' agreements were studied in a pairwise fashion applying inter-rater agreement statistics (Cohen's Kappa). The agreement between Abbott and Roche was very good (0.87 [0.81-0.94]). Agreement between Abbott and DiaSorin, and DiaSorin and Roche was good: 0.71 (0.62-0.80), and 0.76 (0.67-0.84), respectively (Table 3). Despite a good overall inter-rater agreement, significant differences could be shown using McNemar's test for DiaSorin and Roche (Supplementary Table 5).

Discussion

To the best of our knowledge, this is the first side-by-side comparison of three fully automated SARS-CoV-2 antibody tests applying more than 1,200 distinct donor/patient samples. We identified significant differences between two of the three systems, especially regarding positive predictability at the expectable low prevalence rates.

SARS-CoV-2 is a new virus closely related to the betacoronaviruses SARS-CoV and MERS. Like SARS-CoV-2, those highly virulent pathogens cause severe respiratory syndromes, often with lethal outcome (18). In contrast, infections with other members of the coronavirus family usually present with mild colds, including 229E, OC43, NL63, and HKU1 (19). Compared to SARS-CoV (which is no longer circulating), cross-reactivity between SARS-CoV-2 and endemic seasonal coronaviruses is low. To date, with few exceptions (24), no accumulation of cross-reactivities between anti-SARS-CoV-2 antibodies and seasonal coronavirus antibodies has been found. We have therefore refrained from screening a coronavirus panel for possible cross-reactivity.

Specificity

To best describe the specificity of a serological test, it is essential to have a reliable reference, i.e., to ensure that the samples used are negative for the target analyte. For SARS-CoV-2, this means using serum/plasma samples obtained before the first appearance of the new virus. Therefore, we have compiled large pre-COVID-19 cohorts, which have the following characteristics: A) samples of an age and sex-controlled population-based cohort of more than 11,000 participants (LEAD-Study) (14), randomly chosen from Vienna and surrounding areas (n=494). B) samples of healthy voluntary donors (n=302), which are typically used at our Department for the evaluation of new assays, and C) samples of a disease-specific collection of patients with rheumatic diseases including rheumatoid arthritis and systemic lupus erythematodes (n=358), known to have a high prevalence of autoantibodies and other atypical immune activities, enhancing the potential of interference with serological testing. We found several false-positives in the rheumatological cohort (n=13), and to a lesser extent in the other two cohorts (n=9 in the healthy donor cohort and n=10 in the LEAD study). Notably, an overlap of samples tested false-positive in the different systems did not typically occur, and only one of

32 samples tested positive in more than one assay (Abbott and DiaSorin, one sample from the LEAD study). Since these two test systems use different antigens (nucleocapsid vs. S1/S2 proteins) but the same detection method (IgG), this false-positive reaction is likely associated with interference of the IgG measurement. Calculated specificities are strongly dependent on the spectrum and the size of a selected specificity cohort. If we calculated the specificities of each cohort separately, we would be able to report variable specificities: cohort A (Roche 100%, Abbott 99.2%, DiaSorin 98.8%), cohort B (Roche 99.7%, Abbott 99%, DiaSorin 98.3%), and cohort C (Roche 99.4%, Abbott 99.4%, DiaSorin 97.5%). Roche would range from ideal 100% down to 99.4%, the same level as the best result for Abbott, and DiaSorin would be nearly as good as the worst Abbott specificity or show a 2.5% difference to the best Roche value. This would have an enormous impact on prevalence dependent parameters like PPV. A recent evaluation of the DiaSorin LIAISON® SARS-CoV-2 S1/S2 IgG assay with 1,140 pre-COVID-19 samples reported a specificity of 98.5% (20), nearly perfectly matching the specificity of 98.3% we found when calculating the average of all three cohorts. In contrast, another recent study reported a specificity of 100% for DiaSorin. However, the authors used only n=81 samples for specificity testing (21). Similarly, a further evaluation comparing all three SARS-CoV-2 tests by Abbott, Roche, and DiaSorin found quite different specificities, namely 100%, 98%, and 96.9% for Abbott, Roche, and DiaSorin, respectively. Again, the specificity cohort was very small (n=100, and n=98 for DiaSorin) (22). This underlines the importance of selecting adequately sized testing cohorts to obtain reliable and comparable results. In summary, the specificities of 99.7%, 99.2%, and 98.3% found in the present study are very close to the values given by the manufacturers of 99.8%, 99.6%, and 98.5% for Roche, Abbott, and Diasorin, which were also established on large collectives.

Sensitivity

The COVID-19 positive cohort used in this study for the estimation of sensitivities is relatively small (n=65). However, it has three distinctive features:

1. each patient/donor is represented in the collective with only one serum sample, avoiding bias of the data by multiple measurements of the same individuals,
2. the median time of blood sampling was 41 days after onset of symptoms and thus in the plateau phase of antibody formation, and
3. 80% of the cohort were non-hospitalized COVID-19 patients (two-thirds of them with mild symptoms), and only 20% were intensive care patients.

As sensitivity within the first 14 days after symptom onset is highly variable for most SARS-CoV-2 antibody assays but becomes better >14 days (23,24), we expected high sensitivities for all tested assays in the plateau phase of antibody formation. Surprisingly we found multiple RT-PCR confirmed COVID-19 patients displaying very low antibody titers that did not surpass the respective assay-specific cut-offs and therefore were considered negative. Five samples were negative in all three assays: all were RT-PCR confirmed cases, 4/5 non hospitalized (42-51 days after symptom onset), two with mild and the other two with moderate symptoms, and symptom duration of <1 week for all. None of these patients had a known immune dysfunction or other severe diseases. One patient was an ICU patient with an underlying hematological disease, and the sample was taken at day 15 after symptom onset. This patient mounted a partial antibody response starting from day 21 after symptom onset becoming positive in the Abbott assay (2.21 Index) and reaching positivity on day 30 also in the DiaSorin assay (29.1 AU/ml). At this late point, Abbott measured 4.19 Index, whereas Roche remained (since day 21) at a level around 0.2 COI. This example illustrates that although the vast majority of patients show compatible and plausible results in different test systems, single cases can display a complex picture of time-courses and reactivities in specific assay systems that are still poorly understood. Another interesting observation was that in six patients with positive detection of SARS-CoV-2 specific antibodies in the Roche and Abbott test, DiaSorin failed to detect antibodies. This observation, combined with the claim that the detection of S1/S2 protein-specific antibodies is equivalent to the detection of neutralizing antibodies (nAbs) (20), raises the fundamental question of whether nAbs are detectable in all patients with confirmed COVID-

19 infection. There is evidence that neither the assumption of equivalence of nAbs and S1/S2 protein-specific antibodies (25) nor the assumption that all COVID-19 patients produce measurable titers of nAbs is universally valid (26). A clear answer to the question of whether antibody measurements against nucleocapsid- or spike protein-associated antigens are more sensitive and specific, and how these behave in relation to nAbs assays (the postulated gold standard in terms of sensitivity and specificity) is not possible based on the data currently available.

PPV, NPV, ROC-Analysis, test agreement

Specificity and sensitivity alone are not sufficient to judge the performance of a diagnostic test; prevalence-dependent accuracy measures like PPV and NPV are necessary, and especially PPV, in times of low prevalence (27). For most regions affected by the pandemic, the prevalence of SARS-CoV-2 antibody-positive individuals is unknown but can be estimated to be below 5%. Therefore, for all SARS-CoV-2 EUA approved antibody tests, the FDA compares the performance of the assays based on a 5% seroprevalence (28). At this rate, the results presented here show PPV values of 94.8% (85-98), 85.1% (74.7-91.7), and 71.6% (61.7-79.8) for Roche, Abbott, and DiaSorin, respectively. The PPV values between Roche and DiaSorin differ so clearly that not even the 95% CI intervals overlap. Therefore, we must assume that these two assays differ significantly from each other in terms of positive predictability. Using these two tests at lower seroprevalences, such as 1%, leads to an even more pronounced difference between Roche and DiaSorin (77.6% vs. 32.6%) and an unacceptable low PPV of 32.6% (23.6-43.1) for DiaSorin. Although the area under the curves (0.994, 0.989, and 0.977 for Abbott, Roche, and DiaSorin) did not differ significantly from each other (sensitivity cohort size was too small), modeling of the cut-offs according to Youden's index revealed interesting insights: only Roche could increase the sensitivity without losing specificity dramatically (Sensitivity: 89.2% → 98.5%; Specificity: 99.7% → 99.4%; cut-off: >0.355 COI). In contrast, DiaSorin at the suggested cut-off of >8.76 AU/ml (similar to (20)) increased the sensitivity from 83.1% to 90.8% but worsened the specificity from 98.3 to 97.2%. In line with this, despite a good overall agreement between Roche and DiaSorin results (Cohen's Kappa 0.76 [0.67-0.84]), the McNemar's test still showed

significant differences, indicating disagreement (in particular in false-positives) more often than expected by chance.

The strength of this study is the side by side evaluation of three assays with a large number of negative samples to give reliable and comparable specificity data (no missing data). Limitations are the moderate numbers of positive samples. Moreover, obtained sensitivities cannot easily be compared to other studies because of the unique feature of our COVID-19 cohort, including 80% non-hospitalized patients with mainly mild symptoms. The latter is highly relevant for a potential use of antibody tests to assess seroprevalence in large populations.

Conclusion

We find diagnostically relevant differences in specificities for the anti-SARS-CoV-2 antibody assays by Abbott, Roche, and DiaSorin that have a significant impact on the positive predictability of these tests. We conclude that low seroprevalences require an unusually high specificity for SARS-CoV-2 antibody tests, which pushes some test systems to their limits earlier than others. Therefore, the choice of the test must depend on the respective seroprevalence, and strategies such as confirmation of possible false-positive test results with additional testing must be considered.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. Massachusetts Medical Society; 2020;382:727–33.

2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. Elsevier; 2020;20:533–4.

3. Chen Q, Zheng Z, Zhang C, Zhang X, Wu H, Wang J, et al. Clinical characteristics of 145 patients with corona virus disease 2019 (COVID-19) in Taizhou, Zhejiang, China. Infection. Springer Berlin Heidelberg; 2020;92:401–9.

4. Hase R, Kurita T, Muranaka E, Sasazawa H, Mito H, Yano Y. A case of imported COVID-19 diagnosed by PCR-positive lower respiratory specimen but with PCR-negative throat swabs. Infect Dis (Lond). Taylor & Francis; 2020;52:423–6.

5. Liu R, Han H, Liu F, Lv Z, Wu K, Liu Y, et al. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. Clin Chim Acta. 2020;505:172–5.

6. Yong G, Yi Y, Tuantuan L, Xiaowu W, Xiuyong L, Ang L, et al. Evaluation of the auxiliary diagnostic value of antibody assays for the detection of novel coronavirus (SARS-CoV-2). Journal of Medical Virology. 2020;:jmv.25919.

7. Farnsworth CW, Anderson NW. SARS-CoV-2 Serology: Much Hype, Little Data. Clin Chem. 2020.

8. Theel ES, Slev P, Wheeler S, Couturier MR, Wong SJ, Kadkhoda K. The Role of Antibody Testing for SARS-CoV-2: Is There One? J Clin Microbiol. 2020.

9. Infantino M, Grossi V, Lari B, Bambi R, Perri A, Manneschi M, et al. Diagnostic accuracy of an automated chemiluminescent immunoassay for anti-SARS-CoV-2 IgM and IgG antibodies: an Italian experience. Journal of Medical Virology. 2020;:jmv.25932.

10. Whitman JD, Hiatt J, Mowery CT, Shy BR, Yu R, Yamamoto TN, et al. Test performance evaluation of SARS-CoV-2 serological assays. medRxiv. Cold Spring Harbor Laboratory Press; 2020;:2020.04.25.20074856.

11. Abbott Diagnostics. SARS-CoV-2 IgG For Use With ARCHITECT, Revised April 2020. 2020.

12. DiaSorin S p A. LIAISION®SARS-CoV-2 S1/S2 IgG package insert 2020-04. 2020.

13. Roche Diagnostics. Elecsys Anti-SARS-CoV-2 package insert 2020-04, V1.0. 2020.

14. Breyer-Kohansal R, Hartl S, Burghuber OC, Urban M, Schrott A, Agusti A, et al. The LEAD (Lung, Heart, Social, Body) Study: Objectives, Methodology, and External Validity of the Population-Based Cohort Study. J Epidemiol. Japan Epidemiological Association; 2019;29:315–24.

15. HaslacherHelmuth, GernerMarlene, HoferPhilipp, JurkowitschAndreas, HainfellnerJohannes, KainRenate, et al. Usage Data and Scientific Impact of the Prospectively Established Fluid Bioresources at the Hospital-Based MedUni Wien Biobank. Biopreservation and Biobanking. Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2018;16:477–82.

16. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case–control studies. Statistics in Medicine. John Wiley & Sons, Ltd; 2007;26:2170–83.

17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44:837.

18. Rockx B, Kuiken T, Herfst S, Bestebroer T, Lamers MM, Oude Munnink BB, et al. Comparative pathogenesis of COVID-19, MERS, and SARS in a nonhuman primate model. Science. American Association for the Advancement of Science; 2020;:eabb7314.

19. Weiss SR. Forty years with coronaviruses. J Exp Med. 2020;217.

20. Bonelli F, Sarasini A, Zierold C, Calleri M, Bonetti A, Vismara C, et al. Clinical And Analytical Performance Of An Automated Serological Test That Identifies S1/S2 Neutralizing IgG In Covid-19 Patients Semiquantitatively. bioRxiv. Cold Spring Harbor Laboratory; 2020;25:2000082–35.

21. Tré-Hardy M, Wilmet A, Beukinga I, Dogné J-M, Douxfils J, Blairon L. Validation of a chemiluminescent assay for specific SARS-CoV-2 antibody. Clinical Chemical Laboratory Medicine. 2020.

22. Ekelund O, Ekblom K, Somajo S, Pattison-Granberg J, Olsson K, Petersson A. High-throughput immunoassays for SARS-CoV-2, considerable differences in performance when comparing three methods. medRxiv. Cold Spring Harbor Laboratory Press; 2020;:2020.05.22.20106294.

23. Tang MS, Hock KG, Logsdon NM, Hayes JE, Gronowski AM, Anderson NW, et al. Clinical Performance of Two SARS-CoV-2 Serologic Assays. Clin Chem. 2020.

24. Bryan A, Pepper G, Wener MH, Fink SL, Morishima C, Chaudhary A, et al. Performance Characteristics of the Abbott Architect SARS-CoV-2 IgG Assay and Seroprevalence in Boise, Idaho. J Clin Microbiol. 2020;:1–19.

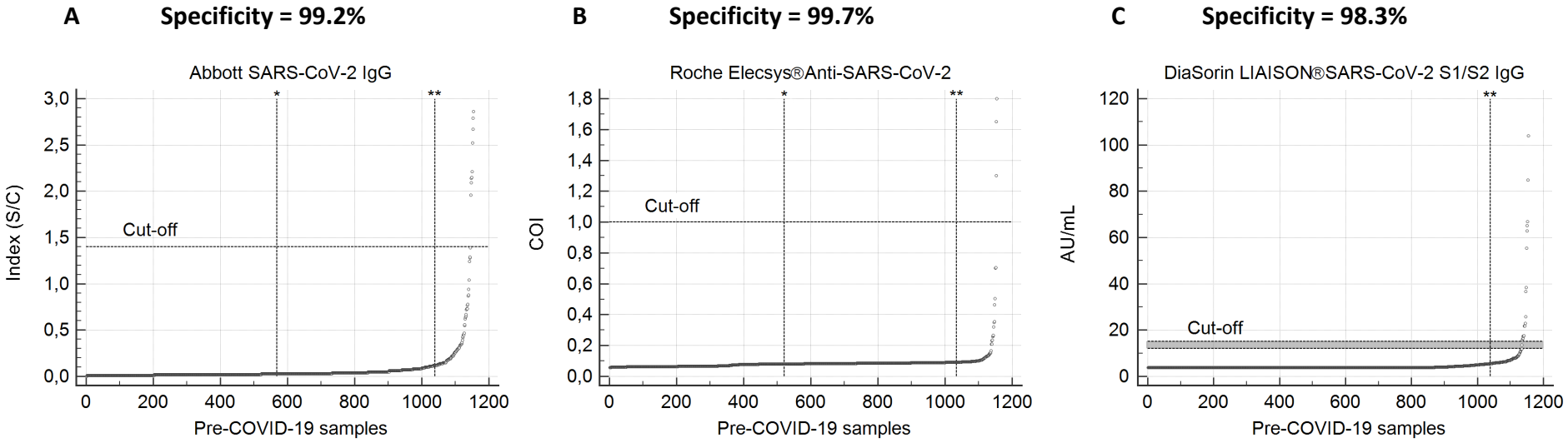
25. Jääskeläinen AJ, Kuivanen S, Kekäläinen E, Ahava MJ, Loginov R, Kallio-Kokko H, et al. Performance of six SARS-CoV-2 immunoassays in comparison with microneutralisation. medRxiv. Cold Spring Harbor Laboratory Press; 2020;:2020.05.18.20101618.

26. Wu F, Wang A, Liu M, Wang Q, Chen J, Xia S, et al. Neutralizing Antibody Responses to SARS-CoV-2 in a COVID-19 Recovered Patient Cohort and Their Implications. SSRN Journal. 2020.

27. Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. International Federation of Clinical Chemistry and Laboratory Medicine; 2009;19:203–11.

28. US Food Drug Administration. EUA Authorized Serology Test Performance [Internet]. fda.gov. 2020 [cited 2020 Jun 2]. Available from: https://www.fda.gov/medical-devices/emergency-situations-medical-devices/eua-authorized-serology-test-performance

Figure 1A-C: Specificity was determined using 1,154 serum samples taken before the circulation of SARS-CoV-2. For SARS-CoV-2 antibody tests Abbott SARS-CoV-2 IgG (**A**), Roche Elecsys® Anti-SARS-CoV-2 (**B**), and DiaSorin LIAISON® SARS-CoV-2 S1/S2 IgG (**C**), values of specificity samples are shown in rank order. Horizontal dotted lines mark the respective cut-offs recommended by the manufacturer and, in the case of DiaSorin, a gray zone for equivocal results. Vertical dotted lines indicate the median (*) and the 90th percentile values (**). Median and 90th percentile values of negative samples were 0.025 and 0.115 for Abbott, 0.0815 and 0.0927 for Roche, below LOD and 5.52 for DiaSorin.



Figures 2A and 2B: Quality criteria of the SARS-CoV-2 antibody tests Abbott SARS-CoV-2 IgG, Roche Elecsys® Anti-SARS-CoV-2, and DiaSorin LIAISON® SARS-CoV-2 S1/S2 IgG. Columns represent values of sensitivity, specificity (**A**), PPV, and NPV (**B**) at 1%, 5%, and 10% assumed seroprevalence; bars indicate the 95% CI.

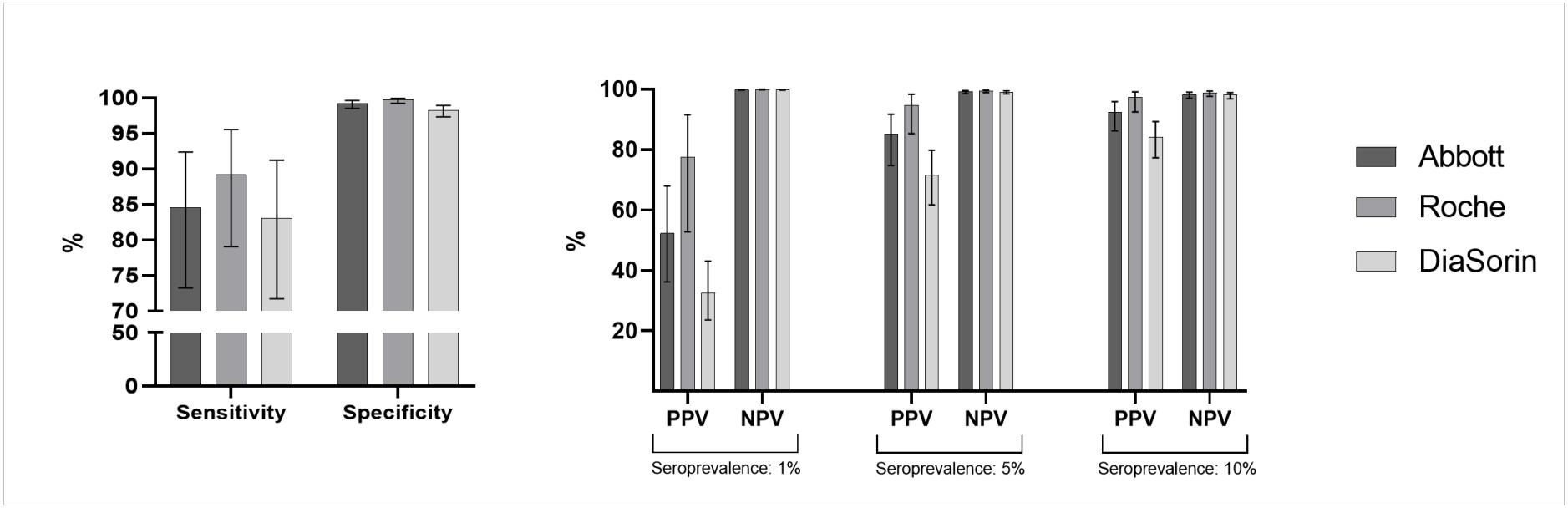


Figure 3A-C: Receiver Operating Characteristic (ROC) curves for Abbott (**A**), Roche (**B**), and DiaSorin (**C**) are shown. The Area Under the Curve (AUC) indicates the test accuracy, 95% confidence intervals are represented by gray dotted lines.

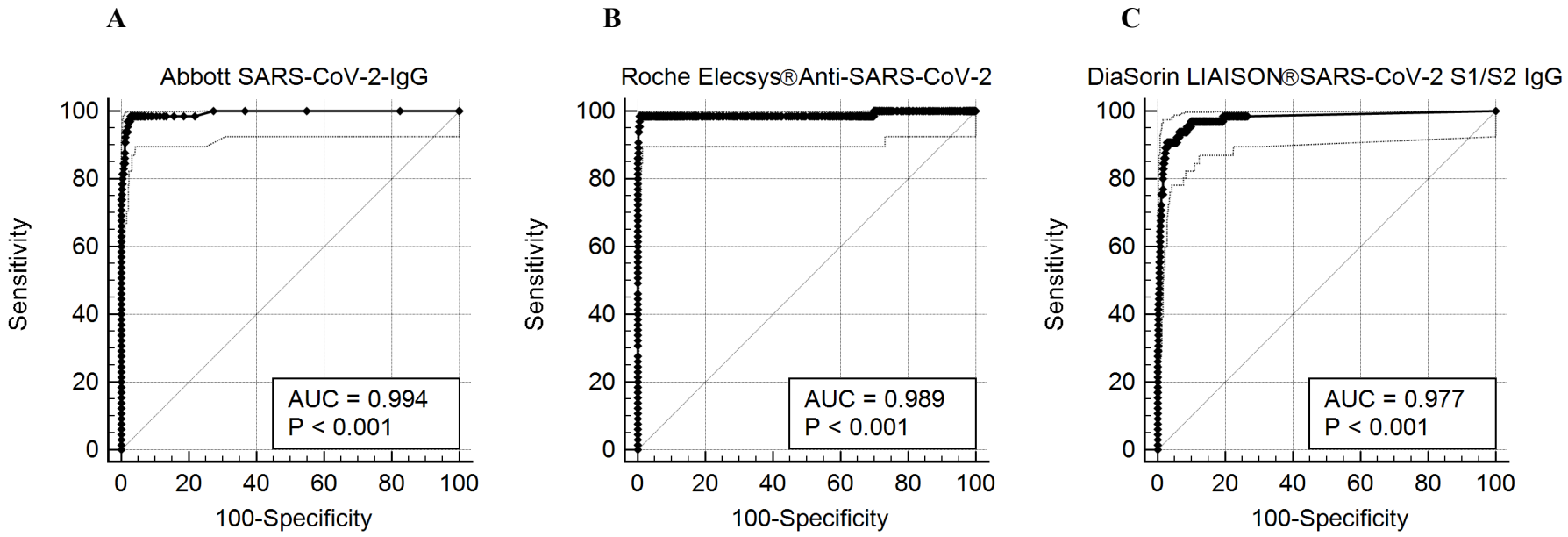


Table 1. Numbers and percentages of false positive SARS-CoV-2 antibody reactivities in three different specificity cohorts: Cohort A (LEAD-Study), Cohort B (Healthy donor collective), and Cohort C (Rheumatic diseases cohort).

	COHORT A	**COHORT B**	**COHORT C**	**TOTAL**
	n=494	*n=302*	*n=358*	***n=1,154***
Roche Elecsys® Anti-SARS-CoV-2	0 (0.0%)	1 (0.3%)	2 (0.6%)	**3** (0.3%)
Abbott SARS-CoV-2 IgG	4 (0.8%)	3 (1.0%)	2 (0.6%)	**9** (0.8%)
DiaSorin LIAISON® SARS-CoV-2 S1/S2 IgG	6 (1.2%)	5 (1.7%)	9 (2.5%)	**20** (1.7%)

Table 2. Values for Specificity, Sensitivity, Positive-Predictive-Value (PPV) and Negative-Predictive-Value (NPV) at 1%, 5% and 10% SARS-CoV-2 seroprevalence (SP) with 95% confidence intervals (95% CI).

	Roche Elecsys® Anti-SARS-CoV-2		**Abbott SARS-CoV-2 IgG**		**DiaSorin LIAISON® SARS-CoV-2 S1/S2 IgG**	
Statistic	Value	95% CI	Value	95% CI	Value	95% CI
Sensitivity	**89.2%**	79.1-95.6	**84.6%**	73.6-92.4	**83.1%**	71.3-91.2
Specificity	**99.7%**	99.2-100	**99.2%**	98.6-99.7	**98.3%**	97.3-98.9
1% Seroprevalence						
PPV	**77.6%**	52.8-91.5	**52.3%**	36.2-67.9	**32.6%**	23.6-43.1
NPV	**99.9%**	99.8-100	**99.9%**	99.7-99.9	**99.8%**	99.7-99.9
5% Seroprevalence						
PPV	**94.8%**	85.3-98.3	**85.1%**	74.7-91.7	**71.6%**	61.7-79.8
NPV	**99.4%**	98.9-99.7	**99.2%**	98.6-99.5	**99.1%**	98.5-99.5
10 % Seroprevalence						
PPV	**97.4%**	92.5-99.2	**92.3%**	86.2-95.9	**84.2%**	77.3-89.3
NPV	**98.8%**	97.6-99.4	**98.3%**	97.0-99.0	**98.1%**	96.8-98.9

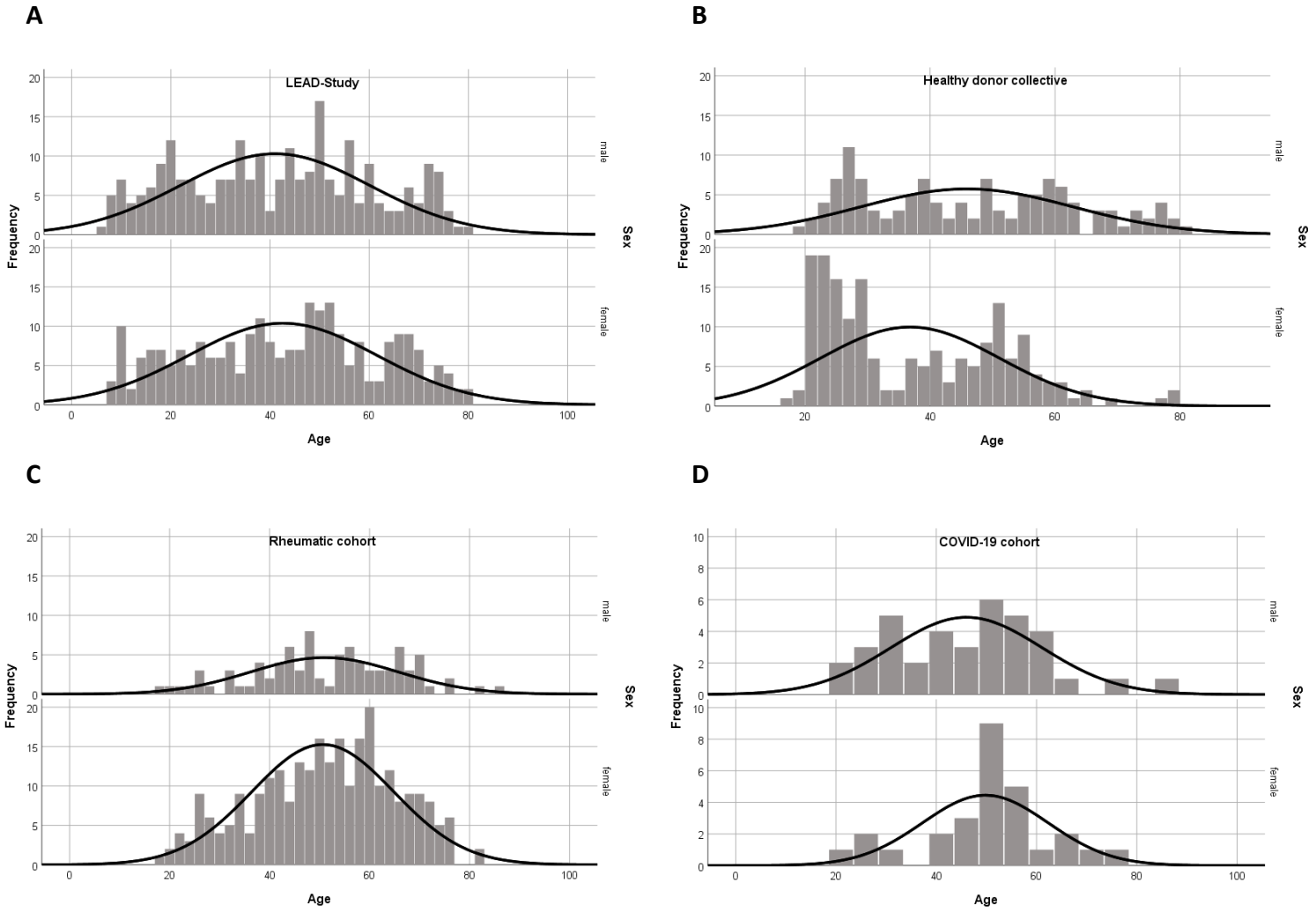
Table 3. Inter-rater agreement (Cohen´s kappa) with linear weights. Value of K <0.20 poor agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 good agreement, and 0.81-1.00 very good agreement.

	Abbott			Kappa (95% CI)
Roche	NEG	POS		**0.87 (0.81-0.94)**
NEG	1149	9	1158 (95.0%)	Standard Error
POS	6	55	61 (5.0%)	0.032
	1155 (94.7%)	64 (5.3%)	1219	

	Abbott			Kappa (95% CI)
DiaSorin	NEG	POS		**0.71 (0.62-0,79)**
NEG	1131	14	1145 (93.9%)	Standard Error
POS	24	50	74 (6.1%)	0.045
	1155 (94.7%)	64 (5.3%)	1219	

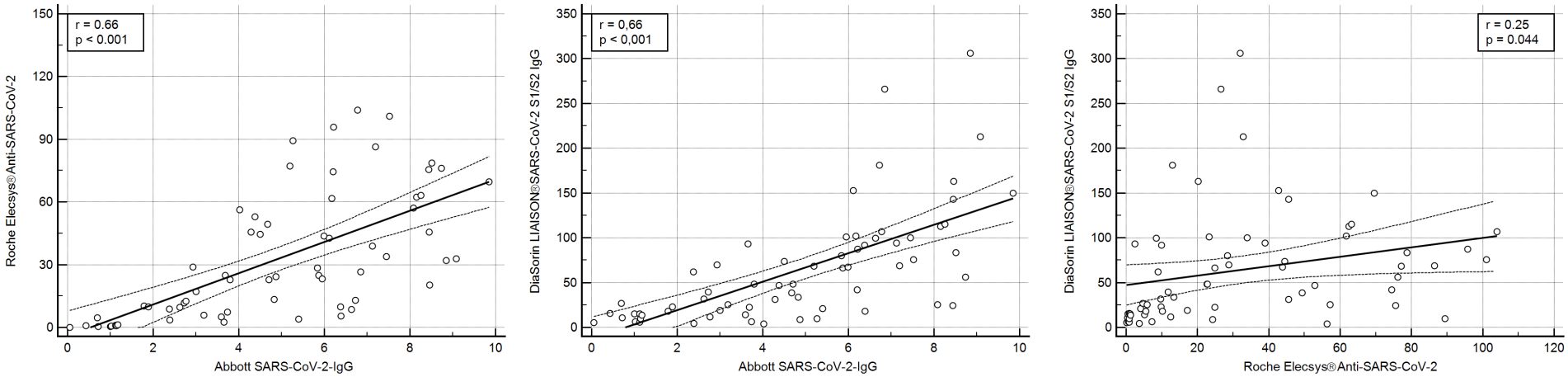
	DiaSorin			Kappa (95% CI)
Roche	NEG	POS		**0.76 (0.67-0.84)**
NEG	1136	22	1158 (95.0%)	Standard Error
POS	9	52	61 (5.0%)	0.042
	1145 (93.9%)	74 (6.1%)		

Supplementary Figure 1: Age distribution of SARS-CoV-2-negative and positive sample cohorts shown by sex. LEAD-Study (**A**), Healthy donor collective (**B**), Rheumatic disease cohort (**C**), and COVID-19 cohort (**D**).



Supplementary Figure 2: Pairwise correlational analysis of measurement values between all three test systems: Abbott/Roche (**A**), Abbott/DiaSorin (**B**), and Roche/DiaSorin (**C**). Pearson correlation coefficient and p value are indicated.

A



Supplementary Table 1. Cohort characteristics

	SARS-CoV-2 Serology Specificity cohorts (all collected before 2020)			**SARS-CoV-2 Serology Sensitivity Cohorts (COVID-19 cohort)**		
Cohort	COHORT A LEAD-study ({BreyerKohansal:2019jw}) (N=494)	COHORT B MedUni Wien Biobank healthy donor collective (N=302)	COHORT C Cohort of rheumatic diseases (N=358)	Post-COVID-19 donors at the MedUni Wien Biobank (N=39)	COVID-19-Convalescent plasma donors at the Department of Transfusion Medicine (N=7)	Diagnostic excess serum samples from COVID-19 patients sent to Department of Laboratory Medicine (N=19)
Cohort description	Population-based cohort, representing a cross-section of the Viennese and surrounding area population; balanced for sex, age, and health status	Population- based cohort: samples from healthy donors collected at the MedUni Wien Biobank;	Disease-specific cohort: patients with the full spectrum of rheumatological diseases from the Division of Rheumatology, Medical University of Vienna	Participants with a history of COVID-19, either confirmed by a positive SARS-CoV-2 RT-PCR (N=29) or symptomatic patients with close contact to PCR-positive COVID-19 patients (N=10)	Plasma donors with a history of SARS-CoV-2 RT-PCR confirmed COVID-19	Samples from in- or outpatients with SARS-CoV-2 RT-PCR confirmed COVID-19 Intensive care unit patients (N=13)
Inclusion criteria	Age 8 – 80 years Only material collected in the months November – April (enrichment of possible cross-reactive antibodies for seasonal respiratory infections) was used for this study	Age >18 years Self-assessment: healthy	Age >18 years Rheumatological disease	Age >18 years positive SARS-CoV-2 RT-PCR OR close contact to patient with positive SARS-CoV-2 PCR considered healthy at the time of sample donation written informed consent	Age >18 years positive SARS-CoV-2 RT-PCR followed by 2 negative SARS-CoV-2 RT-PCRs Considered healthy at the time of plasma donation Written informed consent	Age >18 years Positive SARS-CoV-2 RT-PCR
Exclusion criteria	Unavailability of biomaterial	Unavailability of biomaterial	Unavailability of biomaterial		Inconclusive history of COVID-19 Unavailability of biomaterial	Unavailability of biomaterial

Supplementary Table 2. Demographical data of SARS-CoV-2 serology specificity cohorts. Data are presented as median and interquartile range or counts and percentages. The term sample age describes the time (in days) from sample collection to sample access for SARS-CoV-2 antibody detection.

	LEAD-Study	**Healthy donor collective**	**Rheumatic cohort**
Number (% of total)	494 (43%)	302 (26%)	358 (31%)
Age (y)	43 (26 – 56)	38 (26 – 52)	52 (41 – 61)
Female sex (% of cohort)	247 (50%)	179 (59%)	272 (86%)
Sample age (d)	1.946 (1.615 – 2.331)	1.841 (960 – 2.294)	4.076 (2.994 – 4.846)
Rheumatic disease (% of cohort)			
Rheumatoid Arthritis			142 (40%)
Systemic lupus erythematosus			98 (27%)
Systemic sclerosis			49 (14%)
Psoriatic arthritis			25 (7%)
Spondyloarthritis			25 (7%)
Sjögren's Syndrome			19 (5%)

Supplementary Table 3. Characteristics of the COVID-19 cohort, comprising voluntary post-COVID-19 sample donors (N=39), convalescent plasma donors (N=7), and diagnostic excess serum samples from COVID-19 patients sent to Department of Laboratory Medicine (N=19). Data are presented as median and interquartile range or counts and percentages.*

	COVID-19 cohort
Number	65
Age (y)	49 (40 – 55)
Female sex (% of cohort)	28 (43%)
Symptom severity	
Asymptomatic	6 (9%)
Mild	27 (42%)
Moderate	15 (23%)
Severe	4 (6%)
Intensive care unit	13 (20%)
Symptom onset* (d before sampling)	41 (28 – 49)
Antibody response values	
Abbott [Index (S/C)], Cut-off 1.40	4.87 (2.77 – 6.78)
Roche [COI], Cut-off 1.00	24.20 (7.22 – 52.90)
DiaSorin [Au/ml], Cut-off 15.0 (borderline: 12-15)	47.0 (18.3 – 93.1)

Supplementary Table 4. ROC-Analysis for all three test systems. Sensitivities, specificities, and predictive values for three different seroprevalence levels (1%, 5%, and 10%) are given for the cut-off value associated with the Youden's index.

						Seroprevalence: 1%		**Seroprevalence: 5%**		**Seroprevalence: 10%**	
	ROC-AUC	Youden's Index	Associated criterion	Sensitivity	Specificity	PPV	NPV	PPV	NPV	PPV	NPV
Abbott	*0.994 (0.987 – 0.997)*	*0.957*	*>0.42*	*98.5 (91.7 – 100)*	*97.2 (96.1 – 98.1)*	*26.4 (20.3 – 33.6)*	*100.0 (99.9 – 100.0)*	*65.1 (57.0 – 72.5)*	*99.9 (99.4 – 100.0)*	*79.8 (73.7 – 84.8)*	*99.8 (98.8 – 100.0)*
Roche	*0.989 (0.981 – 0.994)*	*0.979*	*>0.355*	*98.5 (91.7 – 100)*	*99.4 (98.8 – 99.8)*	*62.1 (43.9 – 77.4)*	*100.0 (99.9 – 100.0)*	*89.5 (80.3 – 94.7)*	*99.9 (99.4 – 100.0)*	*94.7 (89.6 – 97.4)*	*99.8 (98.8 – 100.0)*
DiaSorin	*0.977 (0.967 – 0.985)*	*0.880*	*>8.76*	*90.8 (81.0 – 96.5)*	*97.2 (96.1 – 98.1)*	*24.8 (18.9 – 31.9)*	*99.9 (99.8 – 100.0)*	*63.3 (54.8 – 71.0)*	*99.5 (98.9 – 99.8)*	*78.4 (71.9 – 83.8)*	*99.0 (97.8 – 99.5)*

Supplementary Table 5. McNemar´s statistic to test rater disagreement. P<0.05 is statistically significant.

	Roche			Difference
Abbott	NEG	POS		**-0.25%**
NEG	1149	6	1155 (94.7%)	95% CI
POS	9	55	64 (5.3%)	-0.87-0.38
	1155 (95.0%)	61 (5.0%)	1219	P=0.6072

	DiaSorin			Difference
Abbott	NEG	POS		**0.82%**
NEG	1131	24	1155 (94.7%)	95% CI
POS	14	50	64 (5.3%)	-0.17-1,81
	1145 (93.9%)	74 (6.1%)	1219	P=0.1433

	Roche			Difference
DiaSorin	NEG	POS		**-1.07%**
NEG	1136	9	1158 (95.0%)	95% CI
POS	22	52	61 (5.0%)	-1.96-0.17
	1158 (93.9%)	61 (6.1%)	1219	***P=0.0294***