American College of Rheumatology
Education • Treatment • Research
Arthritis Care & Research

An International, Multi-Specialty Validation Study of the IgG4-Related Disease Responder Index

Journal:	*Arthritis Care and Research*
Manuscript ID	ACR-17-0360.R1
Wiley - Manuscript type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Wallace, Zachary; Massachusetts General Hospital, Rheumatology Unit Khosroshahi, Arezou; Emory University, Rheumatology Division Carruthers, Mollie; University of Britich Columbia, Rheumatology Perugino, Cory; Mass General Hospital, Medicine, Division of Rheumatology Choi, Hyon; Massachusetts General Hospital , Rheumatology; Massachusetts General Hospital, Campochiaro, Corrado; Ospedale San Raffaele, Immunology Culver, Emma ; John Radcliffe Hospital Cortazar, Frank; Massachusetts General Hospital, Nephrology Della-Torre, Emanuel; San Raffaele Scientific Institute, Unit of Medicine and Clinical Immunology Ebbo, Mikael; Hopital de la Timone, AP-HM, Department of Internal Medicine; Aix-Marseille Université, Fernandes, Ana; Massachusetts General Hospital Frulloni, Luca; University of Verona, Medicine - Pancreas Center Hart, Phil; Ohio State University Medical College, Gastroenterology Karadag, Omer; Hacettepe University, School of Medicine, Rheumatology Kawa, Shigeyuki; Shinshu University, Center for Health, Safety and Environmental Kawano, Mitsuhiro; Kanazawa University Hospital, Rheumatology Kim, Myung-Hwam; University of Ulsan College of Medicine, Asan Medical Center, Medicine Lanzillotta, Marco; IRCCS -San Raffaele Scientific Institute, Università Vita-Salute San Raffaele Matsui, Shoko; University of Toyama, Health Administration Center Okazaki, Kazuichi; Kansai Medical University, Gastroenterology and Hepatology Ryu, Jay; Mayo Clinic, Division of Pulmonary and Critical Care Medicine Saeki, Takako; Nagaoka Red Cross Hospital, Department of Internal Medicine schleinitz, nicolas; Aix Marseille Universite, AP-HM, internal medicine; Tanasa, Paula; Emory University School of Medicine Umehara, Hisanore; Shiritsu Nagahama Byoin, Division of Rheumatology and Immunology Webster, George; University College London Medical School Zhang, Wen; Chinese Academy of Medical Science, Peking Union Medical College Hospital, Rheumatology Stone, John H; Massachusetts General Hospital, Clinical Rheumatology

Key Words: Please choose key words for your manuscript. Choosing the best match of the journal's keywords to your work will likely result in better match to reviewers with expertise in your interest area. It may also help to speed the review process.:	IgG4-related Disease, Response Criteria, Disease Activity, Damage Index, Clinical Trial
Optional Terms: You have the option to add additional key words that are not on the journal's list to more specifically categorize your submission.:	

An International, Multi-Specialty Validation Study of the IgG4-Related Disease Responder Index

Zachary S Wallace, MD[1], Arezou Khosroshahi, MD[2], Mollie D Carruthers, MD[3], Cory A Perugino, DO[1], Hyon Choi, MD, Dr PH[1], Corrado Campochiaro, MD[4], Emma L Culver, MD, PhD[5], Frank Cortazar, MD[6], Emanuel Dellatorre, MD[4], Mikael Ebbo, MD[7], Ana Fernandes, BA[1], Luca Frulloni, MD[4], Philip A Hart, MD[8], Omer Karadag, MD[9], Shigeyuki Kawa, MD, PhD[10], Mitsuhiro Kawano, MD, PhD[11], Myung-Hwan Kim, MD, PhD[12], Marco Lanzillotta, MD[13], Shoko Matsui, MD, PhD[14], Kazuichi Okazaki, MD, PhD[15], Jay H Ryu, MD[16], Takako Saeki, MD, PhD[17], Nicolas Schleinitz, MD, PhD[18], Paula Tanasa, MD[2], Hisanori Umehara, MD, PhD[19], George Webster, MD, PhD[20], Wen Zhang[21], John H Stone, MD, MPH[1]

[1]Rheumatology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA
[2]Division of Rheumatology, Emory University School of Medicine, Atlanta, GA
[3]Division of Rheumatology, University of British Columbia, British Columbia, Canada
[4] Unit of Medicine and Clinical Immunology, IRCCS San Raffaele Scientific Institute, Università Vita-Salute San Raffaele, Milan, Italy
[5]Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK
[6]Renal Division, Massachusetts General Hospital and Harvard Medical School, Boston, MA
[7]Department of Internal Medicine, La Timone University Hospital, Marseille
[8]Department of Gastroenterology, University of Verona, Verona, Italy.
[9]Division of Gastroenterology, Hepatology, and Nutrition, The Ohio State University Wexner Medical Center, Columbus, Ohio
[10]Hacettepe University Faculty of Medicine, Ankara, Turkey
[11]Center for Health, Safety, and Environmental Management, Shinshu University, Matsumoto, Japan
[12]Department of Human Pathology, Kanazawa University Graduate School of Medicine , Kanazawa , Japan
[13]Department of Internal Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, South Korea
[14]Health Administration Center, University of Toyama, Toyama, Japan
[15]The Third Department of Internal Medicine, Division of Gastroenterology and Hepatology, Kansai Medical University, Osaka, Japan
[16]Division of Pulmonary and Critical Care, Mayo Clinic, Rochester, MN, USA
[17]Division of Clinical Rheumatology and Nephrology, Department of Internal Medicine, Nagaoka Red Cross Hospital
[18]Aix-Marseille Université, Assistance Publique Hôpitaux de Marseille, Marseille, France
[19]Northern County Center for RA and Autoimmune Diseases, Hayashi General Hospital, Fukui, Japan
[20]Dept of Gastroenterology, University College London Hospitals, London, UK
[21]Department of Rheumatology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China

Abstract: 250 (excluded headers) Text: 3,302 (with Tables)

Corresponding Author:
John H. Stone, MD, MPH
55 Fruit Street / Yawkey 2C

Massachusetts General Hospital
Boston, MA 02114

This work was funded by a Scientist Development Award from the Rheumatology Research Foundation (ZSW), a National Institutes of Health Loan Repayment Award (ZSW), National Institute of Arthritis, Musculoskeletal, and Skin Diseases Training Grant (T32- AR007258) (ZSW), and Xencor.

Abstract

Objective:

IgG4-related disease (IgG4-RD) can cause fibro-inflammatory lesions in nearly any organ, leading to organ dysfunction and failure. The IgG4-RD Responder Index (RI) was developed to help investigators assess the efficacy of treatment in a structured manner. We sought to validate the RI in a multi-national investigation.

Methods:

The RI guides investigators through assessments of disease activity and damage in 25 domains, incorporating higher weights for disease manifestations that require treatment urgently or that worsen despite treatment. After a training exercise, investigators reviewed 12 written IgG4-RD vignettes (mean length: 279 words, range: 76-511 words) based upon real patients. Investigators calculated both an RI score as well as a physician global assessment (PGA) for each vignette. Three investigators used the RI on fifteen patients followed over serial visits after treatment. We assessed inter- and intra-rater reliability, precision, validity, and responsiveness.

Results:

Twenty-six physician-investigators included representatives from 6 specialties and 9 countries. The inter-rater and intra-rater reliabilities of the RI were strong (0.88 and 0.69, respectively) and superior to those of the PGA. Correlations (construct validity) between the RI and PGA were high (Spearman's $r=0.9$, $P<0.0001$). The RI was sensitive to change (discriminant validity). Following treatment, there was significant improvement in the RI (mean change 10.5 (95% CI 5.4-12), $P<0.001$) which correlated with the change in the PGA. Urgent disease and damage were captured effectively.

Discussion:

In this international, multi-specialty study, we found that the RI is a valid, and reliable disease activity assessment tool that can be used to measure response to therapy.

Significance & Innovation

IgG4-related disease (IgG4-RD) is an emerging multi-organ inflammatory condition now recogized by the American College of Rheumatology (ACR) as a unique disease. A Classification Criteria effort funded by the ACR the European League Against Rheumatism (EULAR) is now in the validation stage. IgG4-RD is diagnosed all over the world now, and international collaborations have led to consensus publications on nomenclature, pathology findings, and management approach. The stage is set for multi-center clinical trials that will likely be international in scope. The IgG4-RD Responder Index has been developed as a clinical trials assessment tool designed for use in evaluating disease activity in a systematic manner. This validation study has engaged investigators from North America, Asia, Europe, and South America in the interest of facilitating international collaboration on treatment outcomes in this disease.

Introduction

IgG4-related disease (IgG4-RD) is a fibroinflammatory condition that can affect nearly any organ.[1] Common manifestations include dacryoadenitis, chronic sclerosing sialoadenitis, autoimmune pancreatitis, tubulointerstitial nephritis, and retroperitoneal fibrosis.[2] Untreated disease can lead to organ dysfunction, permanent organ injury (i.e., damage), and even death.[2,3]

Disease activity in IgG4-RD is typically assessed using a combination of factors including findings in the history and on physical examination, the results of laboratory investigations, and radiology studies.[4] None of these factors alone, however, is sufficiently specific and sensitive

from patient to patient (and across organ systems within individual patients) to permit reliance upon a single factor alone as a reflection of overall disease activity. As treatment options evolve, it is critical to establish a standardized instrument for measuring disease activity and damage that can be used in clinical trials. A useful instrument would be one capable of distinguishing disease activity from damage (e.g., changes unlikely to respond to treatment) which is essential to assessing treatment response. No widely validated activity index for IgG4-RD exists, although an earlier prototype was developed and partially validated at a single center.[5]

The concept of the IgG4-RD RI is based upon an instrument developed to assess disease activity in another multi-organ inflammatory condition, granulomatosis with polyangiitis (formerly known as Wegener's). That instrument, known as the Birmingham Vasculitis Activity Score for Wegener's Granulomatosis[6], has been used as a disease activity assessment measure in multiple international clinical trials in antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis.[7,8]

Given the protean manifestations of IgG4-RD and its prevalence around the world, a tool understood and adopted by many types of specialists from all over the world is necessary. Moreover, given the variations in disease activity associated with IgG4-RD, an instrument capable of capturing ranges of activity with good precision is necessary. Thus, we developed the IgG4-RD responder index (RI) and assessed its validity in this study. In the interest of unifying disease status indices for IgG4-RD into a single index for both disease activity and disease-associated damage, we also incorporated assessments of organ-based damage.

Methods

Construction of the IgG4-RD RI: The IgG4-RD RI concept was based on that of the BVAS-WG, in which investigators assess disease activity organ by organ, with the sum of organ assessments summing to a total score. Disease activity (over the preceding 28 days) is determined by the investigator and reflects a patient's symptoms attributable to active IgG4-RD as well as significant findings from the physical examination, imaging studies, and laboratory evaluations. *Organ Involvement:* Investigators are guided through the scoring of disease activity and damage in twenty-four standard organs/sites (**Table 1**) but can also enter additional sites of involvement as free text. Constitutional symptoms (weight loss, fever, fatigue) comprise a 25th domain of disease activity.

Scoring Disease Activity: In the prototypical version of the instrument, disease activity in each organ or site was scored on a scale of 0 to 4, where 4 reflected the most severe disease activity ("Worsened or new disease despite treatment") and 0 reflected no disease activity (Unaffected or "resolved"). A score of 1 represented "Improved but still persistent" disease activity, a score of 2 represented "Persistent/Unchanged from last visit" disease activity," and a score of 3 represented "New or recurrent disease while off of treatment." The online exercise, which emphasized scoring patients only at one point in time, employed this scoring scheme.

Experience during this validation exercise, however, led to the realization that the original scoring scheme could suggest improvement in disease activity even if, in fact, the disease activity was unchanged. More specifically, in the event that a patient's score within an individual organ went from 3 "New or recurrent disease while off treatment" to 2

"Persistent/unchanged from last visit," the overall disease activity score would decline in the absence of true clinical improvement. Therefore, for the longitudinal exercise using real patients, the scoring levels were modified such that each level reflected a unique disease activity status within a given organ. The final scoring levels are as follows:

0 = Unaffected or resolved

1 = Improved but persistent

2 = New or recurrence (while off of treatment) or unchanged

3 = Worse or new (despite treatment)

This updated system was studied in the final stage of this study when discriminant validity was assessed across longitudinal visits in real clinic patients. A patient whose disease is scored as a 1 "Improved but persistent" continues to receive this score on subsequent visits if the disease persists (e.g., unresolved) but remains improved when compared to their pre-treatment baseline. This is to contrast it with a score of 2 where "unchanged" refers to no response to treatment.

In certain situations, IgG4-RD may necessitate urgent treatment to prevent serious or irreversible organ dysfunction[4]. In such cases, the score for the organ or site is weighted higher by doubling it. This is described further in the IgG4-RI Manual of Operations (**Online Supplement**).

A common scoring scheme was used for each disease site, derived using an empiric approach. IgG4-RD can lead to myriad manifestations and within an individual organ disease severity can vary substantially. For instance, pulmonary nodules may be asymptomatic but a large

pseudotumor could lead to dyspnea and other symptoms. Similarly, cervical lymphadenopathy may be asymptomatic or lead to significant discomfort (both physically and cosmetically) for a patient. As such, it is difficult to assign varying weights for each organ and therefore we elected to account for varying severity across organs by doubling the score for disease requiring "urgent" treatment.

Capturing Damage due to Disease: Organ damage refers to irreversible organ dysfunction (e.g., exocrine pancreatic insufficiency) or failure (e.g., chronic kidney disease) caused by IgG4-RD. Damage can also occur as a consequence of surgical interventions performed to diagnose or treat IgG4-RD (e.g., modified Whipple procedures, submandibular gland excisions)[2]. The presence or absence of damaged is assessed at each site in the RI. Damage caused by IgG4-RD itself must be distinguished from damage caused by IgG4-RD treatment which, in the context of a clinical trial, would be recorded separately as adverse events.

Capturing Symptomatic Disease: Although some IgG4-RD manifestations (e.g., submandibular gland sialoadenitis) are often symptomatic, others frequently occur in the absence of symptomatology (e.g., pulmonary nodules, lymphadenopathy). The RI permits the investigator to assess whether organ system disease is symptomatic or not within each organ system. Given that symptoms may be due to active disease or damage, the RI differentiates symptoms attributed to each.

Investigators: Forty clinicians representing diverse specialties (rheumatology, nephrology, immunology, pulmonology, general internal medicine, and gastroenterology) and with expertise

in the diagnosis and treatment of IgG4-RD were invited to participate in the study. To ensure that this tool could be used by investigators around the world and for whom English is a second language, we invited experts from the USA, Japan, the UK, Canada, Italy, France, Turkey, China, and South Korea to participate.

Case Vignettes: We (JHS, ZSW, AK, MC, CAP) prepared 15 written cases describing patients with diverse manifestations of IgG4-RD. These vignettes included photographs of physical examination findings as well as images from radiology studies and biopsy specimens (**Online Supplement**). Three of the vignettes used in Phase 3 (see below) described a vignette patient's follow-up after treatment to permit an assessment of responsiveness. Four cases described patients in remission. Five cases described patients with damage as a result of IgG4-RD (e.g., aortic dissection requiring repair) and five described patients with disease manifestations appropriately described as urgent. Cases with damage rather than disease activity were used to preliminarily assess discriminant validity.

Study Design: The study occurred in five phases. In the first four phases, all investigators received clinical vignettes and completed the RI as well as a physician global assessment (PGA) of disease activity on a 100mm scale for each case. The PGA was measured so that it could be used as a comparison for the RI as an assessment of disease activity. In phase five, the updated scoring version of the RI (see *Scoring Disease Activity* above) was employed along with the PGA and patient global assessment (PtGA) in a longitudinal manner in fifteen patients with newly-active IgG4-RD. The first four phases of this study were exempt from the Partners

HealthCare Institutional Review Board (IRB). The fifth phases of this study were approved by the Partners HealthCare IRB.

Tutorial Exercise (Phase 1): All investigators received a Manual of Operations (available in both English and Japanese; **Online Supplement**) describing the use of the RI (Translations of English versions into Japanese were performed by TransPerfect Life Sciences, Irvine, CA). The investigators were also invited to join an online Web-Ex for further instruction. All investigators received three practice clinical vignettes and completed an RI and PGA for each vignette. The clinical vignettes were also available in both English and Japanese (Translations of English versions into Japanese were performed by TransPerfect Life Sciences, Irvine, CA). Scoring of these three cases was reviewed by two authors (ZSW and JHS) and investigators were given feedback regarding their performance.

Inter-Rater Reliability Validation Exercise (Phase 2): Once the three practice clinical vignettes had been completed and reviewed, investigators received 12 new clinical vignettes and scored an RI and PGA for each one.

Responsiveness Exercise Using Vignettes (Phase 3): Responsiveness was evaluated by all investigators using six (of the original 12 cases) written cases describing patients before (3 cases) and after (3 cases) treatment.

Intra-Rater Reliability (Test-Retest) Validation Exercise (Phase 4): Three months after Phase 2, all investigators received three of the same clinical vignettes from Phase 2 and were asked to

repeat their RI and PGA assessments. They were instructed to do so without referencing any notes from Phase 2.

Longitudinal Assessment of Real Patients (Phase 5): Finally, three investigators employed the RI and PGA in 15 consecutive patients with newly-active disease who were started on treatment and followed longitudinally. The RI, PGA, and patient global assessment (PtGA) were assessed prospectively over six months.

Statistical Analysis:

The intra- and inter-observer reliabilities of the RI and PGA were assessed using intra-class correlation coefficients (ICCs) using a previously described methodology that uses ANOVA to determine mean squares which are then used to calculate the ICC.[6] The inter-observer variation (precision) was evaluated by applying the signed rank test to the differences between the coefficients of variation for the RI and PGA of each case.[6] Using the responses to the paper cases, we evaluated construct validity by determining the correlation between the RI and the PGA using the Spearman's rank correlation coefficient. Additionally, construct validity was assessed prospectively using repeated measures correlation to assess the longitudinal relationship of changes in the RI with changes in the PGA and PtGA.[9] Responsiveness was assessed in two ways. First, for real patients followed longitudinally, RI scores before and 6 months after treatment were compared using a paired T-test. Second, for the clinical vignettes that required an investigator to score a patient's RI and PGA before and after treatment, a paired T-test and correlation coefficient were measured, as above. The proportion of investigators reporting urgent disease and damage were tabulated. The modified Wald method was used to determine 95%

confidence intervals for proportions. SAS Version 9.3 (For all analyses unless otherwise noted), R Version 3.4.1 (Repeated measures correlation), and SPSS Version 24 (ICC determination) were used for all analyses.

Results:

Investigators: Forty investigators were invited to participate in the study. Twenty-six physician-investigators participated in Phases 1, 2, and 3 and 20 participated in Phase 4. The discriminant validity using real patients with longitudinal follow up (Phase 5) was completed by three investigators (JHS, ZSW, CP), all of whom are rheumatologists. In terms of investigators in Phases 1-3, there were 11 rheumatologists, 6 gastroenterologists, 4 immunologists, 2 pulmonologists, 2 nephrologists, and 1 internist. Investigators represented 9 different countries, including the USA, United Kingdom, Canada, Italy, France, Turkey, Japan, China, and South Korea.

Clinical Vignettes: The written case vignettes captured a variety of organ involvement, disease activity, and damage (**Table 2**). The average RI assessment for each case ranged from 0.07 (± 0.4) to 14.6 (± 2.8). Disease activity ranged from remission (e.g., history of retroperitoneal, submandibular, and parotid disease in case 12, RI=0), to mild/moderate (e.g., submandibular gland and lymph node involvement in case 4, RI=8), to severe (e.g., aortitis and pancreatitis in case 8, RI=14). In some cases, investigators were asked to properly distinguish damage from disease activity (e.g., case 2). The average PGA for each case ranged from 0.5 (± 2.1) to 79.7 (± 21.6).

Reliability: The RI had similar but higher inter-rater reliability to the PGA (0.89, 95% CI 0.80-0.96, vs. 0.88, 95% CI 0.77-0.96). The intra-rater (test-retest) reliability of the RI was and PGA

were similar. The median ICC for the RI and PGA were 0.73 (range 0.32-0.92) and 0.74 (range 0.44-0.79), respectively.

Precision: To assess precision (inter-rater variation) of the RI and PGA, the coefficients of variation (CV) for the RI and PGA in each case were calculated and the differences in the CVs (DCVs) were determined for each case by subtracting the RI CV from the PGA CV (**Table 3**). The CV represents the level of agreement between raters for each case. Whereas a DCV value of zero would imply that the RI and PGA were equally precise, a positive value suggests lower variability (i.e., greater precision) of the RI compared with the PGA. The mean DCV for all cases was -0.4 (± 0.8, P=0.5), suggesting that the RI and PGA had approximately equal precision. For cases with very low or absent disease activity, the CV tended to be higher, suggesting that some investigators equated complete remission with a low RI or PGA rather than a zero

Correlation (Convergent Validity): We evaluated the correlation between the RI and PGA using the RIs and PGAs calculated in Phase 1. When all cases were included, the RI and PGA had high correlation (Spearman's r=0.9, P< 0.0001). When cases with no disease activity (Cases 2, 9, 10, and 12) as well as very low disease activity (Case 7) were excluded – because of the potential for inflated correlations in such cases – correlation remained high (Spearman's r=0.6, P<0.0001). We also assessed the correlation between the RI, PGA, and PtGA prospectively in treated patients. There was strong correlation (r=0.81, 95% CI: 0.74-0.86, P<0.001) between the RI and PGA over repeated assessments. Similarly, there was significant correlation (r=0.26, 95% CI: 0.09-0.42, P=0.003) between the RI and the PtGA (**Table 4**).

Responsiveness: Based on a review of six vignettes that described three unique cases before and after treatment, both the RI and PGA showed good discriminant validity. There were significant differences by paired T-tests and correlation between changes in the RI and PGA before and after treatment (**Table 4**). In clinic, three investigators assessed fifteen patients before and 6 months after treatment. The PGA and the RI showed good responsiveness in the clinical setting (**Table 4**).

Urgent Disease and Damage: In the clinical vignettes, there were five cases in which organ damage had occurred as a result of IgG4-RD and five cases which required urgent treatment. Damage was correctly identified, on average, 86% of the time. Urgent disease was correctly identified, on average, 76% of the time, indicating that the RI is able to discriminate between active disease and damage (discriminant validity, **Table 5**).

Discussion

In this international, multi-specialty validation study, we demonstrated that the IgG4-RD RI is a practical, reliable, and responsive means of assessing and recording disease activity and damage. Our findings also support the validity of the RI. The RI, the first tool of its kind in IgG4-RD, will be instrumental in future clinical trials and other types of studies in this disease. The RI demonstrated strong inter- and intra-rater reliabilities. In addition, the precision of the RI was similar to that of the PGA and the two types of assessments were highly correlated both cross-sectionally and prospectively, supporting the instrument's validity as an assessment of disease activity. In longitudinal assessments of patients, we also demonstrated that the RI has good responsiveness, indicating sensitivity to changes in disease activity over serial visits following

treatment. The RI was able to appropriately differentiate disease activity from damage (discriminant validity).

As in other systemic illnesses, there are many potential ways of measuring disease activity, including findings on the history and physical examination, the results of laboratory studies, and the interpretation of imaging abnormalities. The ideal instrument for assessing disease activity, however, accommodates a broad diversity of organ system involvement, variation in resource availability, and other factors specific to both IgG4-RD and to individual patients. Although several types of biomarkers have been proposed in IgG4-RD (e.g., serum IgG4 concentrations,[2,10,11] circulating plasmablast levels,[10,12] and complement concentrations[2]), none of these measures is sufficiently sensitive or specific for disease activity. The clinical context of these measurements must be interpreted by an investigator in order to make proper attributions of their implications for disease activity and treatment decisions. Thus, a cognitive tool such as the RI that allows the investigator to consider information from a variety of sources and distill these parts into a sum of disease activity and damage reflected in individual organ manifestations is critical to IgG4-RD, as in other multi-organ conditions. An earlier version of the RI[5] included a scoring domain for the serum IgG4 concentration, but greater experience with IgG4-RD led to the removal of this domain because many patients in remission never achieve a normal serum IgG4 concentration or do not do so within a timeframe that is appropriate for clinical trials.[13] However, the serum IgG4 concentration may be an important reflection of disease activity for an individual patient; this may be considered by a provider when assessing disease activity.

Longitudinal use of the RI in real patients in this study led to practical insights on the appropriate application of the instrument in clinical trials. We deleted one scoring level from the initial version of the RI because its inclusion had the potential to indicate falsely that a patient's disease activity had improved over the baseline assessment, regardless of whether or not true clinical improvement had actually occurred. Phase 5 of the study, application of the RI in real patients on a longitudinal basis, employed the updated scoring system.

Successful application of the RI, which may appear deceptively simple, requires substantial clinical experience and judgment in order to address both the protean nature of IgG4-RD and the RI's subtleties. It is crucial, for example, to distinguish active IgG4-RD within a specific organ from damage that occurred to that same organ from previously active but now quiescent disease. It is also possible that both active disease and damage can co-exist at the same time in a given organ, a fact that requires clinical acumen to discern and record appropriately. The findings from this validation study indicate that following appropriate training, investigators from many different countries, speaking many different primary languages, and representing an array of medical specialties can all use the RI successfully. When using the RI in the context of a clinical trial, thorough pre-trial training and assessments of the investigators will be required, as performed in the context of this validation study.

The most common challenge faced by investigators during the training phase of this study was distinguishing disease activity and damage due to IgG4-RD. This distinction is critical because damage is not expected to respond to treatment. The erroneous attribution of clinical manifestations resulting from damage to active IgG4-RD leads inevitably to incorrect

conclusions regarding treatment efficacy. The investigators reported damage correctly in 86% of the scenarios described in this study, including damage related to surgical procedures required to establish the diagnosis of IgG4-RD. The design of the RI includes the concept that any surgical intervention beyond a fine needle aspiration should be considered damage, given that such procedures pose a risk to patients and, at the least, leave patients with scars. Future studies will focus on defining, assessing, and reporting damage due to IgG4-RD.

The RI assigns a higher weight (two-fold) to urgent disease to reflect the greater severity of certain manifestations of IgG4-RD. Urgent disease refers specifically to the need to begin treatment immediately for certain manifestations in order to prevent irreversible damage of an organ or site. For example, a patient with an aortic dissection due to IgG4-RD requires urgent management of their disease. Given that investigators identified urgent disease correctly only 76% of the time, future studies will address sources of disagreement to improve guidelines. Despite this, the RI was found to be a reliable and precise tool for the assessment of disease activity. To assist providers using the RI, we have provided additional details regarding damage and urgent disease in an online supplement.

IgG4-RD is a protean disease with wide variations in disease activity which we sought to capture in this study. To maximize participation among investigators we had to balance the number of clinical vignettes we asked them to review with the reality that asking investigators to review too many cases would discourage participation. We chose cases that were representative of the IgG4-RD spectrum of disease, including various organ sites and combinations of disease activity, symptoms, urgent disease, and damage. Further, we chose clinical scenarios that required

investigators to use a variety of tools to assess disease activity. Less commonly affected disease sites (e.g., pituitary, meninges) were not included in the clinical vignettes but we have no reason to suspect that investigators would have difficulty assessing disease activity in these sites given their ability to do so in other sites (e.g., aorta, biliary, lung) which also rely on imaging along with other factors (e.g., physical exam) to assess disease activity.

Our study has potential weaknesses. These relate primarily to the challenges of recording subtle gradations of disease activity in a multi-organ condition in which degrees of activity do not necessarily fall into discrete levels from visit to visit. Some disease manifestations of IgG4-RD require imaging to gauge the level of improvement or worsening. If a disease manifestation cannot be assessed with certainty without follow-up imaging, then that manifestation should be scored a "2" to reflect that concept that as far as the investigator knows – while awaiting imaging – the manifestation is unchanged from the previous visit. Because of the need in some cases to await imaging, the recording of improvement or worsening on the RI may lag behind the true clinical state in these situations. This fact, however, reflects the realities of clinical practice. An additional limitation was that cases in which disease improved but did not resolve were under-represented in clinical vignettes. However, we have no reason to believe that clinicians would be unable to distinguish improvement (but persistent disease) from remission and worsening disease activity. Additionally, the ability to distinguish remission, damage, and disease activity, as demonstrated in this study, is critical for the use of the RI in an IgG4-RD clinical trial and was a priority in this study. Finally, intra-rater reliability was lower than the inter-rater reliability which was unexpected. However, we suspect that this is related to suboptimal power given that

investigators were asked to only re-analyze three cases. Future studies in the clinical setting will be able to address these limitations.

Despite these challenges, changes in the RI over time should correspond either to disease flares – clear worsenings of disease activity that lead to increases in treatment – or to improvement, corresponding to less need for therapy. In this way, the RI should justify alterations in therapy that occur over the course of a clinical trial, and the instrument offers a means of checking investigators' decisions to escalate therapy.

In summary, in this international validation study of an RI for IgG4-RD, we found the RI to be a reliable, responsive, and valid instrument with which to measure disease activity and record disease-associated damage, regardless of the manifestation or specialist managing the case. The RI will be an important tool in monitoring disease activity in clinical trials.

1. Kamisawa T, Zen Y, Pillai S, Stone JH. IgG4-related disease. *Lancet*. 2015;385(9976):1460-1471.

2. Wallace ZS, Deshpande V, Mattoo H, et al. IgG4-related disease: Clinical and laboratory features in one hundred twenty-five patients. *Arthritis Rheumatol*. 2015;67(9):2466-2475.

3. Huggett MT, Culver EL, Kumar M, et al. Type 1 autoimmune pancreatitis and IgG4-related sclerosing cholangitis is associated with extrapancreatic organ failure, malignancy, and mortality in a prospective UK cohort. *Am J Gastroenterol*. 2014;109(10):1675-1683.

4. Khosroshahi A, Wallace ZS, Crowe JL, et al. International consensus guidance statement on the management and treatment of IgG4-related disease. *Arthritis Rheumatol*. 2015.

5. Carruthers MN, Stone JH, Deshpande V, Khosroshahi A. Development of an IgG4-RD responder index. *Int J Rheumatol*. 2012;2012:259408.

6. Stone JH, Hoffman GS, Merkel PA, et al. A disease-specific activity index for wegener's granulomatosis: Modification of the birmingham vasculitis activity score. international network for the study of the systemic vasculitides (INSSYS). *Arthritis Rheum*. 2001;44(4):912-920.

7. Stone JH, Merkel PA, Spiera R, et al. Rituximab versus cyclophosphamide for ANCA-associated vasculitis. *N Engl J Med*. 2010;363(3):221-232.

8. Wegener's Granulomatosis Etanercept Trial (WGET) Research Group. Etanercept plus standard therapy for wegener's granulomatosis. *N Engl J Med*. 2005;352(4):351-361.

9. Bakdash JZ, Marusich LR. Repeated measures correlation. *Front Psychol*. 2017;8:456.

10. Wallace ZS, Mattoo H, Mahajan VS, et al. Predictors of disease relapse in IgG4-related disease following rituximab. *Rheumatology (Oxford)*. 2016;55(6):1000-1008.

11. Carruthers MN, Khosroshahi A, Augustin T, Deshpande V, Stone JH. The diagnostic utility of serum IgG4 concentrations in IgG4-related disease. *Ann Rheum Dis*. 2015;74(1):14-18.

12. Wallace ZS, Mattoo H, Carruthers M, et al. Plasmablasts as a biomarker for IgG4-related disease, independent of serum IgG4 concentrations. *Ann Rheum Dis*. 2014.

13. Carruthers MN, Topazian MD, Khosroshahi A, et al. Rituximab for IgG4-related disease: A prospective, open-label trial. *Ann Rheum Dis*. 2015.

Table 1: Potential Disease Activity Captured in the IgG4-RD Responder Index (RI)

Meninges	Pituitary Gland
Orbital Lesion	Lacrimal Gland
Parotid Gland	Submandibular Gland
Other Salivary Gland*	Mastoiditis/Middle Ear Disease
Nasal Cavity Lesion	Sinusitis
Other ENT Lesion*	Thyroid
Lung	Lymph Node^
Aorta/Large Blood Vessel	Heart/Pericardium
Retroperitoneal Fibrosis	Sclerosing Mediastinitis
Sclerosing Mesenteritis	Pancreas
Liver	Bile Duct
Kidney	Skin
Constitutional Symptoms (Weight Loss, Fever, Fatigue due to IgG4-RD)	Other*

*Provides free-text space for investigator to capture disease activity not captured elsewhere (e.g., breast, prostate); ^Asks investigator to specify region of lymphadenopathy (e.g., mediastinal)

Table 2: Clinical Vignette Descriptions

Case	Organs/Sites of Involvement	Constitutional Symptoms	Active Disease	Organs/Sites of Active Disease	Urgent Disease	Damage
1	Biliary, Lung, Orbit, Renal	Yes	Yes	Renal	Yes	No
2	Aorta, LAD	No	No	N/A	No	Yes
3	Orbital, Lacrimal, Parotid, Skin	No	Yes	Orbital, Lacrimal, Skin	Yes	No
4	LAD, SG	No	Yes	LAD, SG	No	No
5	Lacrimal, Orbit, LAD, Lung, Pre-splenic mass	No	Yes	Lacrimal, LAD, Lung, Pre-splenic mass	Yes	No
6	Orbit, Lacrimal	No	Yes	Orbit, Lacrimal	No	No
7	Follow up of case 6	No	Yes*	Orbit, Lacrimal	No	No
8	Aorta, LAD, Pancreas	No	Yes	Aorta, LAD, Pancreas	Yes	Yes
9	Follow up of case 8	No	No	N/A	No	Yes
10	Orbit, Parotid	No	No	N/A	No	Yes
11	RP, Parotid, SG	No	Yes	RP, Parotid, SG	Yes	No
12	Follow up of case 11	No	No	N/A	No	Yes

N/A = Not applicable; RP = Retroperitoneum; SG = Submandibular gland; LAD = Lymphadenopathy; *Disease was improved

Table 3: Precision of the IgG4-RD Responder Index and Physician Global Assessment

Case	Mean (SD) RI	RI CV	Mean (SD) PGA	PGA CV	DCV^
1	9.0 (1.2)	13.3	68.0 (15.9)	23.4	0.1
2	0.2 (0.5)	301.7	4.8 (11)	230.5	-0.7
3	11.4 (2.8)	24.7	63.9 (18)	28.2	0.0
4	8.2 (2.4)	28.6	55.2 (16.6)	30.0	0.0
5	14.6 (2.8)	19.0	76.8 (14.7)	19.1	0.0
6	8.5 (3.4)	39.5	56.8 (22.3)	39.2	0.0
7	0.1 (0.4)	509.9	0.5 (2.1)	386.8	-1.2
8	13.8 (4.6)	33.1	79.7 (21.6)	27.1	-0.1
9	0.1 (0.4)	509.9	2.6 (6.6)	255.3	-2.5
10	0.5 (1.0)	205.9	5.8 (11.1)	191.9	-0.1
11	14.9 (2.2)	15.0	72.0 (15.9)	22.1	0.1
12	0.2 (0.4)	186.2	5.3 (10.2)	193.3	0.1
				Mean (SD)	-0.4 (0.8)*

SD = Standard Deviation; CV = Coefficient of Variation; DCV = Difference of CV (PGA CV – RI CV); ^CV was divided by 100 to calculate DCV; *P=0.5

Table 4: Discriminant Validity of the RI and PGA

Case	Mean Difference (SD) in RI	P-Value*	Mean Difference (SD) in PGA	P-Value*	Correlation^	P-Value
1	8.4 (7.0-9.9)	<0.0001	56.2 (46.7-65.7)	<0.0001	0.6 (0.2-0.8)	0.0003
2	13.8 (11.8-15.7)	<0.0001	77 (67.7-86.1)	<0.0001	0.5 (0.2-0.8)	0.005
3	14.6 (13.6-15.6)	<0.0001	65.9 (57.9-73.9)	<0.0001	0.5 (0.07-0.7)	0.02
Clinical Series	10.5 (6.5-14.6)	<0.0001	41.4 (31.1-51.7)	<0.0001	0.81 (0.7-0.9)	<0.0001

*Paired T-test; ^ Tested the correlation of the difference in the RI and PGA before and after treatment for the paper cases and used repeated measure correlation analysis for the longitudinal assessment; SD=Standard Deviation

Table 5: Proportion Correctly Identifying Damage and Urgent Disease

Case	Proportion (95% CI) Correctly Classifying Damage	Proportion (95% CI) Correctly Classifying Urgent Disease
1	N/A	Renal: 96% (80%-99.9%)
2	Lymph Node: 54% (35%-71%) Aorta: 96% (80%-99.9%)	N/A
3	N/A	Orbit: 57.7% (39%-75%)
5	N/A	Lung: 62% (43%-78%)
8	Aorta: 85% (66%-95%)	Aorta: 85% (66%-95%)
9	Aorta: 96% (80%-99.9%) Pancreas: 92% (75%-99%)	N/A
10	Orbit: 92% (75%-99%)	N/A
11	N/A	RPF: 81% (62%-92%)
12	RPF: 88% (70%-97%)	N/A

An International, Multi-Specialty Validation Study of the IgG4-Related Disease Responder Index

ZS Wallace[1], A Khosroshahi[2], M Carruthers[3], C Perugino[1], H Choi[1], C Campochiaro[4], EL Culver[5], F Cortazar[6], E Dellatorre[4], M Ebbo[7], A Fernandes[1], L Frulloni[4], P Hart[8], O Karadag[9], S Kawa[10], M Kawano[11], MH Kim[12], M Lanzillotta[13], S Matsui[14], K Okazaki[15], JH Ryu[16], T Saeki[17], N Schleinitz[18], P Tanasa[2], H Umehara[19], G Webster[20], W Zhang[21], JH Stone[1]

[1]Rheumatology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA
[2]Division of Rheumatology, Emory University School of Medicine, Atlanta, GA
[3]Division of Rheumatology, University of British Columbia, British Columbia, Canada
[4] Unit of Medicine and Clinical Immunology, IRCCS San Raffaele Scientific Institute, Università Vita-Salute San Raffaele, Milan, Italy
[5]Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK
[6]Renal Division, Massachusetts General Hospital and Harvard Medical School, Boston, MA
[7]Department of Internal Medicine, La Timone University Hospital, Marseille
[8]Department of Gastroenterology, University of Verona, Verona, Italy.
[9]Division of Gastroenterology, Hepatology, and Nutrition, The Ohio State University Wexner Medical Center, Columbus, Ohio
[10]Hacettepe University Faculty of Medicine, Ankara, Turkey
[11]Center for Health, Safety, and Environmental Management, Shinshu University, Matsumoto, Japan
[12]Department of Human Pathology, Kanazawa University Graduate School of Medicine , Kanazawa , Japan
[13]Department of Internal Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, South Korea
[14]Health Administration Center, University of Toyama, Toyama, Japan
[15]The Third Department of Internal Medicine, Division of Gastroenterology and Hepatology, Kansai Medical University, Osaka, Japan
[16]Division of Pulmonary and Critical Care, Mayo Clinic, Rochester, MN, USA
[17]Division of Clinical Rheumatology and Nephrology, Department of Internal Medicine, Nagaoka Red Cross Hospital
[18]Aix-Marseille Université, Assistance Publique Hôpitaux de Marseille, Marseille, France
[19]Northern County Center for RA and Autoimmune Diseases, Hayashi General Hospital, Fukui, Japan
[20]Dept of Gastroenterology, University College London Hospitals, London, UK
[21]Department of Rheumatology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China

Abstract: 250 (excluded headers) Text: 3,302 (with Tables)

Corresponding Author:
John H. Stone, MD, MPH
55 Fruit Street / Yawkey 2C
Massachusetts General Hospital
Boston, MA 02114

This work was funded by a Scientist Development Award from the Rheumatology Research Foundation (ZSW), a National Institutes of Health Loan Repayment Award (ZSW), National Institute of Arthritis, Musculoskeletal, and Skin Diseases Training Grant (T32- AR007258) (ZSW), and Xencor.

Abstract

Objective:

IgG4-related disease (IgG4-RD) can cause fibro-inflammatory lesions in nearly any organ, leading to organ dysfunction and failure. The IgG4-RD Responder Index (RI) was developed to help investigators assess the efficacy of treatment in a structured manner. We sought to validate the RI in a multi-national investigation.

Methods:

The RI guides investigators through assessments of disease activity and damage in 25 domains, incorporating higher weights for disease manifestations that require treatment urgently or that worsen despite treatment. After a training exercise, ~~participants~~ investigators reviewed 12 written IgG4-RD vignettes (mean length: 279 words, range: 76-511 words) based upon real patients. Investigators calculated both an RI score as well as a physician global assessment (PGA) for each vignette. Three investigators used the RI on ~~nine~~ fifteen patients followed over serial visits after treatment. We assessed inter- and intra-rater reliability, precision, ~~construct validity, and discriminant validity~~validity, and responsiveness.

Results:

Twenty-six physician-investigators included representatives from 6 specialties and 9 countries. The inter-rater and intra-rater reliabilities of the RI were strong (0.88 and 0.69, respectively) and superior to those of the PGA. Correlations (construct validity) between the RI and PGA were high (Spearman's r=0.9, P<0.0001). The RI was sensitive to change (discriminant validity). Following treatment, there was significant improvement in the RI (mean change ~~11.8~~10.5 (95% CI 5.4-~~18~~12), P<0.001) which correlated with the change in the PGA. Urgent disease and damage were captured effectively.

Discussion:

In this international, multi-specialty study, we found that the RI is a valid, responsive, and reliable disease activity assessment tool that can be used to test the efficacy of treatment.

IgG4-related disease (IgG4-RD) is a fibroinflammatory condition that can affect nearly any organ.[1] Common manifestations include dacryoadenitis, chronic sclerosing sialoadenitis, autoimmune pancreatitis, tubulointerstitial nephritis, and retroperitoneal fibrosis.[2] Untreated disease can lead to organ dysfunction, permanent organ injury (i.e., damage), and even death.[2,3]

Disease activity in IgG4-RD is typically assessed using a combination of factors including findings in the history and on physical examination, the results of laboratory investigations, and radiology studies.[4] None of these factors alone, however, is sufficiently specific and sensitive from patient to patient (and across organ systems within individual patients) to permit reliance upon a single factor alone as a reflection of overall disease activity. As treatment options evolve, it is critical to establish a standardized instrument for measuring disease activity and damage that can be used in clinical trials. A useful instrument would be one capable of distinguishing disease activity from damage (e.g., changes unlikely to respond to treatment) which is essential to assessing treatment response. No widely validated activity index for IgG4-RD exists, although an earlier prototype was developed and partially validated at a single center.[5]

The concept of the IgG4-RD RI is based upon an instrument developed to assess disease activity in another multi-organ inflammatory condition, granulomatosis with polyangiitis (formerly known as Wegener's). That instrument, known as the Birmingham Vasculitis Activity Score for Wegener's Granulomatosis[6], has been used as a disease activity assessment measure in multiple international clinical trials in antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis.[7,8]

Given the protean manifestations of IgG4-RD and its prevalence around the world, a tool understood and adopted by many types of specialists from all over the world is necessary. Moreover, given the variations in disease activity associated with IgG4-RD, an instrument capable of capturing ranges of activity with good precision is necessary. Thus, we developed the IgG4-RD responder index (RI) and assessed its validity in this study. In the interest of unifying disease status indices for IgG4-RD into a single index for both disease activity and disease-associated damage, we also incorporated assessments of organ-based damage.

Methods

Construction of the IgG4-RD RI: The IgG4-RD RI concept was based on that of the BVAS-WG, in which investigators assess disease activity organ by organ, with the sum of organ assessments summing to a total score. Disease activity (over the preceding 28 days) is determined by the investigator and reflects a patient's symptoms attributable to active IgG4-RD as well as significant findings from the physical examination, imaging studies, and laboratory evaluations.

Organ Involvement: Investigators are guided through the scoring of disease activity and damage in twenty-four standard organs/sites (**Table 1**) but can also enter additional sites of involvement as free text. Constitutional symptoms (weight loss, fever, fatigue) comprise a 25th domain of disease activity.

Scoring Disease Activity: In the prototypical version of the instrument, disease activity in each organ or site was scored on a scale of 0 to 4, where 4 reflected the most severe disease activity ("Worsened or new disease despite treatment") and 0 reflected no disease activity (Unaffected or "resolved"). A score of 1 represented "Improved but still persistent" disease activity, a score of 2

represented "Persistent/Unchanged from last visit" disease activity," and a score of 3 represented "New or recurrent disease while off of treatment." The online exercise, which emphasized scoring patients only at one point in time, employed this scoring scheme.

Experience during this validation exercise, however, led to the realization that the original scoring scheme could suggest improvement in disease activity even if, in fact, the disease activity was unchanged. More specifically, in the event that a patient's score within an individual organ went from 3 "New or recurrent disease while off treatment" to 2 "Persistent/unchanged from last visit," the overall disease activity score would decline in the absence of true clinical improvement. Therefore, for the longitudinal exercise using real patients, the scoring levels were modified such that each level reflected a unique disease activity status within a given organ. The final scoring levels are as follows:

0 = Unaffected or resolved

1 = Improved but persistent

2 = New or recurrence (while off of treatment) or unchanged

3 = Worse or new (despite treatment)

This updated system was studied in the final stage of this study when discriminant validity was assessed across longitudinal visits in real clinic patients. A patient whose disease is scored as a 1 "Improved but persistent" continues to receive this score on subsequent visits if the disease persists (e.g., unresolved) but remains improved when compared to their pre-treatment baseline. This is to contrast it with a score of 2 where "unchanged" refers to no response to treatment.

In certain situations, IgG4-RD may necessitate urgent treatment to prevent serious or irreversible organ dysfunction[4]. In such cases, the score for the organ or site is weighted higher by doubling it. This is described further in the IgG4-RI Manual of Operations (**Online Supplement**).

A common scoring schemea was used for each disease site, derived using an empiric approach. IgG4-RD can lead to myriad manifestations and within an individual organ disease severity can vary substantially. For instance, pulmonary nodules may be asymptomatic but a large pseudotumor could lead to dyspnea and other symptoms. Similarly, cervical lymphadenopathy may be asymptomatic or lead to significant discomfort (both physically and cosmetically) for a patient. As such, it is difficult to assign varying weights for each organ and therefore we elected to account for varying severity across organs by doubling the score for disease requiring "urgent" treatment.

Capturing Damage due to Disease: Organ damage refers to irreversible organ dysfunction (e.g., exocrine pancreatic insufficiency) or failure (e.g., chronic kidney disease) caused by IgG4-RD. Damage can also occur as a consequence of surgical interventions performed to diagnose or treat IgG4-RD (e.g., modified Whipple procedures, submandibular gland excisions)[2]. The presence or absence of damaged is assessed at each site in the RI. Damage caused by IgG4-RD itself must be distinguished from damage caused by IgG4-RD treatment which, in the context of a clinical trial, would be recorded separately as adverse events.

Capturing Symptomatic Disease: Although some IgG4-RD manifestations (e.g., submandibular gland sialoadenitis) are often symptomatic, others frequently occur in the absence of symptomatology (e.g., pulmonary nodules, lymphadenopathy). The RI permits the investigator to assess whether organ system disease is symptomatic or not within each organ system. Given that symptoms may be due to active disease or damage, the RI differentiates symptoms attributed to each.

~~Validation Study Participants~~Investigators: Forty clinicians representing diverse specialties (rheumatology, nephrology, immunology, pulmonology, general internal medicine, and gastroenterology) and with expertise in the diagnosis and treatment of IgG4-RD were invited to participate in the study. To ensure ~~cross-cultural validity~~that this tool could be used by investigators around the world and ~~the ability of those~~ for whom English is a second language ~~to use the tool~~, we invited experts from the USA, Japan, the UK, Canada, Italy, France, Turkey, China, and South Korea to participate.

Case Vignettes: We (JHS, ZSW, AK, MC, CAP) prepared 15 written cases describing patients with diverse manifestations of IgG4-RD. These vignettes included photographs of physical examination findings as well as images from radiology studies and biopsy specimens (**Online Supplement**). Three of the vignettes used in Phase 3 (see below) described a vignette patient's follow-up after treatment to permit an assessment of ~~discriminant validity~~responsiveness. Four cases described patients in remission. Five cases described patients with damage as a result of IgG4-RD (e.g., aortic dissection requiring repair) and five described patients with disease

manifestations appropriately described as urgent. Cases with damage rather than disease activity were used to preliminarily assess discriminant validity.

Study Design: The study occurred in five phases. In the first four phases, all ~~participants~~ investigators received clinical vignettes and completed the RI as well as a physician global assessment (PGA) of disease activity on a 100mm scale for each case. The PGA was measured so that it could be used as a comparison for the RI as an assessment of disease activity. In phase five, the updated scoring version of the RI (see *Scoring Disease Activity* above) was employed along with the PGA and patient global assessment (PtGA) in a longitudinal manner in ~~nine~~ fifteen patients with newly-active IgG4-RD. The first four phases of this study were exempt from the Partners HealthCare Institutional Review Board (IRB). The fifth phases of this study were approved by the Partners HealthCare IRB.

Tutorial Exercise (Phase 1): All investigators received a Manual of Operations (available in both English and Japanese; **Online Supplement**) describing the use of the RI (Translations of English versions into Japanese were performed by TransPerfect Life Sciences, Irvine, CA). The investigators were also invited to join an online Web-Ex for further instruction. All investigators received three practice clinical vignettes and completed an RI and PGA for each vignette. The clinical vignettes were also available in both English and Japanese (Translations of English versions into Japanese were performed by TransPerfect Life Sciences, Irvine, CA). Scoring of these three cases was reviewed by two authors (ZSW and JHS) and investigators were given feedback regarding their performance.

Formatted: Font: Not Italic

Inter-Rater Reliability Validation Exercise (Phase 2): Once the three practice clinical vignettes had been completed and reviewed, investigators received 12 new clinical vignettes and scored an RI and PGA for each one.

~~*Discriminant Validity*~~*Responsiveness Exercise Using Vignettes (Phase 3)*: ~~Discriminant validity~~Responsiveness was evaluated by all investigators using six (of the original 12 cases) written cases describing patients before (3 cases) and after (3 cases) treatment.

Intra-Rater Reliability (Test-Retest) Validation Exercise (Phase 4): Three months after Phase 2, all investigators received three of the same clinical vignettes from Phase 2 and were asked to repeat their RI and PGA assessments. They were instructed to do so without referencing any notes from Phase 2.

~~*Discriminant Validity*~~*Responsiveness ~~Exercise Using~~ Longitudinal Assessment of Real Patients (Phase 5):* Finally, three investigators employed the RI and PGA in ~~9~~ 15 consecutive patients with newly-active disease who were started on treatment and followed longitudinally. The RI, PGA, and patient global assessment (PtGA) were assessed prospectively over six months. ~~and PGA were compared at baseline (prior to treatment) and the last date of follow up for an additional assessment of discriminant validity.~~

Statistical Analysis:

The intra- and inter-observer reliabilities of the RI and PGA were assessed using intra-class correlation coefficients (ICCs) using a previously described methodology that uses ANOVA to

determine mean squares which are then used to calculate the ICC.[6] The inter-observer variation (precision) was evaluated by applying the signed rank test to the differences between the coefficients of variation for the RI and PGA of each case.[6] Using the responses to the paper cases, we evaluated ~~Construct~~ construct validity by determining the correlation between the RI and the PGA using the Spearman's rank correlation coefficient. Additionally, construct validity was assessed prospectively using repeated measures correlation to assess the longitudinal relationship of changes in the RI with changes in the PGA and PtGA.[9] ~~Discriminant validity~~Responsiveness was assessed in two ways. First, for real patients followed longitudinally, RI scores before and 6 months after treatment were compared using a paired T-test ~~and the differences in the RI and PGA over the study period were compared using the Spearman's rank correlation coefficient~~. Second, for the clinical vignettes that required an investigator to score a patient's RI and PGA before and after treatment, a paired T-test and correlation coefficient were measured, as above. The proportion of investigators reporting urgent disease and damage were tabulated. The modified Wald method was used to determine 95% confidence intervals for proportions. SAS Version 9.3 (For all analyses unless otherwise noted), R Version 3.4.1 (Repeated measures correlation), ~~was~~ and SPSS Version 24 (ICC determination) were used for all analyses.

Results:

Investigators: Forty investigators were invited to participate in the study. Twenty-six physician-investigators participated in Phases 1, 2, and 3 and 20 participated in Phase 4. The discriminant validity using real patients with longitudinal follow up (Phase 5) was completed by three investigators (JHS, ZSW, CP), all of whom are rheumatologists. In terms of ~~participants~~ investigators in Phases 1-3, there were 11 rheumatologists, 6 gastroenterologists, 4

immunologists, 2 pulmonologists, 2 nephrologists, and 1 internist. Investigators represented 9 different countries, including the USA, United Kingdom, Canada, Italy, France, Turkey, Japan, China, and South Korea.

Clinical Vignettes: The written case vignettes captured a variety of organ involvement, disease activity, and damage (**Table 2**). The average RI assessment for each case ranged from 0.07 (± 0.4) to 14.6 (± 2.8). Disease activity ranged from remission (e.g., history of retroperitoneal, submandibular, and parotid disease in case 12, RI=0), to mild/moderate (e.g., submandibular gland and lymph node involvement in case 4, RI=8), to severe (e.g., aortitis and pancreatitis in case 8, RI=14). In some cases, ~~participants~~ investigators were asked to properly distinguish damage from disease activity (e.g., case 2). The average PGA for each case ranged from 0.5 (± 2.1) to 79.7 (± 21.6).

Reliability: The RI had ~~greater~~ similar but higher inter-rater reliability ~~than~~ to the PGA (0.89, 95% CI 0.80-0.96,~~8~~ vs. 0.~~83~~88, 95% CI 0.77-0.96). The intra-rater (test-retest) reliability of the RI was ~~also superior to that of the~~and PGA were similar ~~(0.69 vs. 0.20)~~. The median ICC for the RI and PGA were 0.73 (range 0.32-0.92) and 0.74 (range 0.44-0.79), respectively.

Precision: To assess precision (inter-rater variation) of the RI and PGA, the coefficients of variation (CV) for the RI and PGA in each case were calculated and the differences in the CVs (DCVs) were determined for each case by subtracting the RI CV from the PGA CV (**Table 3**). The CV represents the level of agreement between raters for each case. Whereas a DCV value of zero would imply that the RI and PGA were equally precise, a positive value suggests lower variability (i.e., greater precision) of the RI compared with the PGA. The mean DCV for all cases was -0.4 (± 0.8, P=0.5), suggesting that the RI and PGA had approximately equal precision. For cases with very low or absent disease activity, the CV tended to be higher,

suggesting that some investigators equated complete remission with a low RI or PGA rather than a zero

Correlation (Convergent Validity): We evaluated the correlation between the RI and PGA using the RIs and PGAs calculated in Phase 1. When all cases were included, the RI and PGA had high correlation (Spearman's r=0.9, P< 0.0001). When cases with no disease activity (Cases 2, 9, 10, and 12) as well as very low disease activity (Case 7) were excluded – because of the potential for inflated correlations in such cases – correlation remained high (Spearman's r=0.6, P<0.0001). We also assessed the correlation between the RI, PGA, and PtGA prospectively in treated patients. There was strong correlation (r=0.81, 95% CI: 0.74-0.86, P<0.001) between the RI and PGA over repeated assessments. Similarly, there was significant correlation (r=0.26, 95% CI: 0.09-0.42, P=0.003) between the RI and the PtGA (**Table 4**).

~~Discriminant Validity~~Responsiveness: Based on a review of six vignettes that described three unique cases before and after treatment, both the RI and PGA showed good discriminant validity. There were significant differences by paired T-tests and correlation between changes in the RI and PGA before and after treatment (**Table 4**). In clinic, three investigators assessed ~~nine~~ fifteen patients before and 6 months after treatment. ~~The mean number of days between baseline and re-assessment was 128.7 days (range 48-188 days). Both t~~The PGA and the RI showed good ~~discriminant validity~~responsiveness in the clinical setting (**Table 4**). ~~In the clinical setting, the correlation approached but did not achieve significance (P=0.088), perhaps a function of the small number of cases (N=9). Notably, the estimated Spearman's correlation (r=0.6) was similar to that obtained in other assessments of discriminant validity during the study (**Table 4**).~~

Urgent Disease and Damage: In the clinical vignettes, there were five cases in which organ damage had occurred as a result of IgG4-RD and five cases which required urgent treatment. Damage was correctly identified, on average, 86% of the time. Urgent disease was correctly identified, on average, 76% of the time, indicating that the RI is able to discriminate between active disease and damage (discriminant validity, **Table 5**).

Discussion

In this international, multi-specialty validation study, we demonstrated that the IgG4-RD RI is a practical~~,~~ ~~valid~~reliable, and responsive~~, and useful~~ means of assessing and recording disease activity and damage. Our findings also support the validity of the RI. The RI, the first tool of its kind ~~to be validated~~ in IgG4-RD, will be instrumental in future clinical trials and other types of studies in this disease. The RI demonstrated strong inter- and intra-rater reliabilities ~~that were superior to those of the PGA~~. In addition, the precision of the RI was similar to that of the PGA and the two types of assessments were highly correlated both cross-sectionally and prospectively, ~~further~~ supporting the instrument's validity as an assessment of disease activity. In longitudinal assessments of patients, we also demonstrated that the RI has good ~~discriminant validity~~responsiveness, indicating sensitivity to changes in disease activity over serial visits following treatment. The RI was able to appropriately differentiate disease activity from damage (discriminant validity).

As in other systemic illnesses, there are many potential ways of measuring disease activity, including findings on the history and physical examination, the results of laboratory studies, and

the interpretation of imaging abnormalities. The ideal instrument for assessing disease activity, however, accommodates a broad diversity of organ system involvement, variation in resource availability, and other factors specific to both IgG4-RD and to individual patients. Although several types of biomarkers have been proposed in IgG4-RD (e.g., serum IgG4 concentrations[2,10,11], circulating plasmablast levels[10,12], and complement concentrations[2]), none of these measures is sufficiently sensitive or specific for disease activity. The clinical context of these measurements must be interpreted by an investigator in order to make proper attributions of their implications for disease activity and treatment decisions. Thus, a cognitive tool such as the RI that allows the investigator to consider information from a variety of sources and distill these parts into a sum of disease activity and damage reflected in individual organ manifestations is critical to IgG4-RD, as in other multi-organ conditions. An earlier version of the RI[5] included a scoring domain for the serum IgG4 concentration, but greater experience with IgG4-RD led to the removal of this domain because many patients in remission never achieve a normal serum IgG4 concentration or do not do so within a timeframe that is appropriate for clinical trials[13]. However, the serum IgG4 concentration may be an important reflection of disease activity for an individual patient; this may be considered by a provider when assessing disease activity.

Longitudinal use of the RI in real patients in this study led to practical insights on the appropriate application of the instrument in clinical trials. We deleted one scoring level from the initial version of the RI because its inclusion had the potential to indicate falsely that a patient's disease activity had improved over the baseline assessment, regardless of whether or not true clinical improvement had actually occurred. Phase 5 of the study, application of the RI in real patients on a longitudinal basis, employed the updated scoring system.

Successful application of the RI, which may appear deceptively simple, requires substantial clinical experience and judgment in order to address both the protean nature of IgG4-RD and the RI's subtleties. It is crucial, for example, to distinguish active IgG4-RD within a specific organ from damage that occurred to that same organ from previously active but now quiescent disease. It is also possible that both active disease and damage can co-exist at the same time in a given organ, a fact that requires clinical acumen to discern and record appropriately. The findings from this validation study indicate that following appropriate training, investigators from many different countries, speaking many different primary languages, and representing an array of medical specialties can all use the RI successfully. When using the RI in the context of a clinical trial, thorough pre-trial training and assessments of the investigators will be required, as performed in the context of this validation study.

The most common challenge faced by investigators during the training phase of this study was distinguishing disease activity and damage due to IgG4-RD. This distinction is critical because damage is not expected to respond to treatment. The erroneous attribution of clinical manifestations resulting from damage to active IgG4-RD leads inevitably to incorrect conclusions regarding treatment efficacy. The investigators reported damage correctly in 86% of the scenarios described in this study, including damage related to surgical procedures required to establish the diagnosis of IgG4-RD. The design of the RI includes the concept that any surgical intervention beyond a fine needle aspiration should be considered damage, given that such procedures pose a risk to patients and, at the least, leave patients with scars. Future studies will focus on defining, assessing, and reporting damage due to IgG4-RD.

The RI assigns a higher weight (two-fold) to urgent disease to reflect the greater severity of certain manifestations of IgG4-RD. Urgent disease refers specifically to the need to begin treatment immediately for certain manifestations in order to prevent irreversible damage of an organ or site. For example, a patient with an aortic dissection due to IgG4-RD requires urgent management of their disease. Given that investigators identified urgent disease correctly only 76% of the time, future studies ~~might~~ will address sources of disagreement to improve guidelines. Despite this, the RI was found to be a reliable and precise tool for the assessment of disease activity. To assist providers using the RI, we have provided additional details regarding damage and urgent disease in an online supplement.

IgG4-RD is a protean disease with wide variations in disease activity which we sought to capture in this study. To maximize participation among investigators we had to balance the number of clinical vignettes we asked them to review with the reality that asking ~~participants~~ investigators to review too many cases would discourage participation. We chose cases that were representative of the IgG4-RD spectrum of disease, including various organ sites and combinations of disease activity, symptoms, urgent disease, and damage. Further, we chose clinical scenarios that required investigators to use a variety of tools to assess disease activity. Less commonly affected disease sites (e.g., pituitary, meninges) were not included in the clinical vignettes but we have no reason to suspect that investigators would have difficulty assessing disease activity in these sites given their ability to do so in other sites (e.g., aorta, biliary, lung) which also rely on imaging along with other factors (e.g., physical exam) to assess disease activity.

Our study has potential weaknesses. These relate primarily to the challenges of recording subtle gradations of disease activity in a multi-organ condition in which degrees of activity do not necessarily fall into discrete levels from visit to visit. Some disease manifestations of IgG4-RD require imaging to gauge the level of improvement or worsening. If a disease manifestation cannot be assessed with certainty without follow-up imaging, then that manifestation should be scored a "2" to reflect that concept that as far as the investigator knows – while awaiting imaging – the manifestation is unchanged from the previous visit. Because of the need in some cases to await imaging, the recording of improvement or worsening on the RI may lag behind the true clinical state in these situations. This fact, however, reflects the realities of clinical practice. An additional limitation was that cases in which disease improved but did not resolve were under-represented in clinical vignettes. However, we have no reason to believe that clinicians would be unable to distinguish improvement (but persistent disease) from remission and worsening disease activity. Additionally, the ability to distinguish remission, damage, and disease activity, as demonstrated in this study, is critical for the use of the RI in an IgG4-RD clinical trial and was a priority in this study. Finally, intra-rater reliability was lower than the inter-rater reliability which was unexpected. However, we suspect that this is related to ~~insufficient~~ suboptimal power given that investigators were asked to only re-analyze three cases ~~to minimize the burden~~. Future studies in the clinical setting will be able to address these limitations.

Despite these challenges, changes in the RI over time should correspond either to disease flares – clear worsenings of disease activity that lead to increases in treatment – or to improvement, corresponding to less need for therapy. In this way, the RI should justify alterations in therapy

that occur over the course of a clinical trial, and the instrument offers a means of checking investigators' decisions to escalate therapy.

In summary, in this international validation study of an RI for IgG4-RD, we found the RI to be a reliable, responsive, and valid ~~and useful~~ instrument with which to measure disease activity and record disease-associated damage, regardless of the manifestation or specialist managing the case. The RI will be an important tool in monitoring disease activity in clinical trials.

1. Kamisawa T, Zen Y, Pillai S, Stone JH. IgG4-related disease. *Lancet*. 2015;385(9976):1460-1471.

2. Wallace ZS, Deshpande V, Mattoo H, et al. IgG4-related disease: Clinical and laboratory features in one hundred twenty-five patients. *Arthritis Rheumatol*. 2015;67(9):2466-2475.

3. Huggett MT, Culver EL, Kumar M, et al. Type 1 autoimmune pancreatitis and IgG4-related sclerosing cholangitis is associated with extrapancreatic organ failure, malignancy, and mortality in a prospective UK cohort. *Am J Gastroenterol*. 2014;109(10):1675-1683.

4. Khosroshahi A, Wallace ZS, Crowe JL, et al. International consensus guidance statement on the management and treatment of IgG4-related disease. *Arthritis Rheumatol*. 2015.

5. Carruthers MN, Stone JH, Deshpande V, Khosroshahi A. Development of an IgG4-RD responder index. *Int J Rheumatol*. 2012;2012:259408.

6. Stone JH, Hoffman GS, Merkel PA, et al. A disease-specific activity index for wegener's granulomatosis: Modification of the birmingham vasculitis activity score. international network for the study of the systemic vasculitides (INSSYS). *Arthritis Rheum*. 2001;44(4):912-920.

7. Stone JH, Merkel PA, Spiera R, et al. Rituximab versus cyclophosphamide for ANCA-associated vasculitis. *N Engl J Med*. 2010;363(3):221-232.

8. Wegener's Granulomatosis Etanercept Trial (WGET) Research Group. Etanercept plus standard therapy for wegener's granulomatosis. *N Engl J Med*. 2005;352(4):351-361.

9. Bakdash JZ, Marusich LR. Repeated measures correlation. *Front Psychol*. 2017;8:456.

10. Wallace ZS, Mattoo H, Mahajan VS, et al. Predictors of disease relapse in IgG4-related disease following rituximab. *Rheumatology (Oxford)*. 2016;55(6):1000-1008.

11. Carruthers MN, Khosroshahi A, Augustin T, Deshpande V, Stone JH. The diagnostic utility of serum IgG4 concentrations in IgG4-related disease. *Ann Rheum Dis*. 2015;74(1):14-18.

12. Wallace ZS, Mattoo H, Carruthers M, et al. Plasmablasts as a biomarker for IgG4-related disease, independent of serum IgG4 concentrations. *Ann Rheum Dis*. 2014.

13. Carruthers MN, Topazian MD, Khosroshahi A, et al. Rituximab for IgG4-related disease: A prospective, open-label trial. *Ann Rheum Dis*. 2015.

Table 1: Potential Disease Activity Captured in the IgG4-RD Responder Index (RI)

Meninges	Pituitary Gland
Orbital Lesion	Lacrimal Gland
Parotid Gland	Submandibular Gland
Other Salivary Gland*	Mastoiditis/Middle Ear Disease
Nasal Cavity Lesion	Sinusitis
Other ENT Lesion*	Thyroid
Lung	Lymph Node^
Aorta/Large Blood Vessel	Heart/Pericardium
Retroperitoneal Fibrosis	Sclerosing Mediastinitis
Sclerosing Mesenteritis	Pancreas
Liver	Bile Duct
Kidney	Skin
Constitutional Symptoms (Weight Loss, Fever, Fatigue due to IgG4-RD)	Other*

*Provides free-text space for investigator to capture disease activity not captured elsewhere (e.g., breast, prostate); ^Asks investigator to specify region of lymphadenopathy (e.g., mediastinal)

Table 2: Clinical Vignette Descriptions

Case	Organs/Sites of Involvement	Constitutional Symptoms	Active Disease	Organs/Sites of Active Disease	Urgent Disease	Damage
1	Biliary, Lung, Orbit, Renal	Yes	Yes	Renal	Yes	No
2	Aorta, LAD	No	No	N/A	No	Yes
3	Orbital, Lacrimal, Parotid, Skin	No	Yes	Orbital, Lacrimal, Skin	Yes	No
4	LAD, SG	No	Yes	LAD, SG	No	No
5	Lacrimal, Orbit, LAD, Lung, Pre-splenic mass	No	Yes	Lacrimal, LAD, Lung, Pre-splenic mass	Yes	No
6	Orbit, Lacrimal	No	Yes	Orbit, Lacrimal	No	No
7	Follow up of case 6	No	Yes*	Orbit, Lacrimal	No	No
8	Aorta, LAD, Pancreas	No	Yes	Aorta, LAD, Pancreas	Yes	Yes
9	Follow up of case 8	No	No	N/A	No	Yes
10	Orbit, Parotid	No	No	N/A	No	Yes
11	RP, Parotid, SG	No	Yes	RP, Parotid, SG	Yes	No
12	Follow up of case 11	No	No	N/A	No	Yes

N/A = Not applicable; RP = Retroperitoneum; SG = Submandibular gland; LAD = Lymphadenopathy; *Disease was improved

Table 3: Precision of the IgG4-RD Responder Index and Physician Global Assessment

Case	Mean (SD) RI	RI CV	Mean (SD) PGA	PGA CV	DCV^
1	9.0 (1.2)	13.3	68.0 (15.9)	23.4	0.1
2	0.2 (0.5)	301.7	4.8 (11)	230.5	-0.7
3	11.4 (2.8)	24.7	63.9 (18)	28.2	0.0
4	8.2 (2.4)	28.6	55.2 (16.6)	30.0	0.0
5	14.6 (2.8)	19.0	76.8 (14.7)	19.1	0.0
6	8.5 (3.4)	39.5	56.8 (22.3)	39.2	0.0
7	0.1 (0.4)	509.9	0.5 (2.1)	386.8	-1.2
8	13.8 (4.6)	33.1	79.7 (21.6)	27.1	-0.1
9	0.1 (0.4)	509.9	2.6 (6.6)	255.3	-2.5
10	0.5 (1.0)	205.9	5.8 (11.1)	191.9	-0.1
11	14.9 (2.2)	15.0	72.0 (15.9)	22.1	0.1
12	0.2 (0.4)	186.2	5.3 (10.2)	193.3	0.1
				Mean (SD)	-0.4 (0.8)*

SD = Standard Deviation; CV = Coefficient of Variation; DCV = Difference of CV (PGA CV – RI CV); ^CV was divided by 100 to calculate DCV; *P=0.5

Table 4: Discriminant Validity of the RI and PGA

Case	Mean Difference (SD) in RI	P-Value*	Mean Difference (SD) in PGA	P-Value*	~~Spearman~~ Correlation^	P-Value
1	8.4 (7.0-9.9)	<0.0001	56.2 (46.7-65.7)	<0.0001	0.6 (0.2-0.8)	0.0003
2	13.8 (11.8-15.7)	<0.0001	77 (67.7-86.1)	<0.0001	0.5 (0.2-0.8)	0.005
3	14.6 (13.6-15.6)	<0.0001	65.9 (57.9-73.9)	<0.0001	0.5 (0.07-0.7)	0.02
Clinical Series	~~11.8 (5.4-18.1)~~10.5 (6.5-14.6)	~~0.003~~<0.0001	41.4 (31.1-51.7)~~40.8 (25-56.6)~~	~~0.0003~~<0.0001	0.81 (0.7-0.9)~~0.6 (0.1-0.9)~~	<0.0001~~0.088~~

*Paired T-test; ^ Tested the correlation of the difference in the RI and PGA before and after treatment for the paper cases and used repeated measure correlation analysis for the longitudinal assessment; ~~&Not statistically significant but likely related to small sample size;~~ SD=Standard Deviation

Formatted Table

Table 5: Proportion Correctly Identifying Damage and Urgent Disease

Case	Proportion (95% CI) Correctly Classifying Damage	Proportion (95% CI) Correctly Classifying Urgent Disease
1	N/A	Renal: 96% (80%-99.9%)
2	Lymph Node: 54% (35%-71%) Aorta: 96% (80%-99.9%)	N/A
3	N/A	Orbit: 57.7% (39%-75%)
5	N/A	Lung: 62% (43%-78%)
8	Aorta: 85% (66%-95%)	Aorta: 85% (66%-95%)
9	Aorta: 96% (80%-99.9%) Pancreas: 92% (75%-99%)	N/A
10	Orbit: 92% (75%-99%)	N/A
11	N/A	RPF: 81% (62%-92%)

12	RPF: 88% (70%-97%)	N/A

Version: July 25, 2016

IgG4-RD Responder Index Validation Study

Scoring Rules

Scoring refers to manifestations of disease activity present in the last 28 days

Scoring: 0 Normal or resolved
1 Improved but still present
2 New / Recurrence while patient is off treatment or unchanged from the previous visit*
3 Worsened or new disease manifestation despite treatment

*Unchanged from previous visit will often refer to disease manifestations that require follow-up imaging to assess accurately

Definitions

Organ/Site score: The overall level of IgG4-RD activity within a specific organ system

Symptomatic: Is the disease manifestation in a particular organ system symptomatic? (Y = yes; N = no)

Urgent disease: Disease that requires treatment immediately to prevent serious organ dysfunction (Y = yes; N = no)
(*Presence of **urgent disease** within an organ leads to **DOUBLING** of that organ system score*)

Damage: Organ dysfunction that has occurred as a result of IgG4-RD and is considered permanent (Y = yes; N = no)

	Activity			Damage	
Organ/Site	Organ/Site Score (0-3)	Symptomatic (Yes/No)	Urgent (Yes/No)	Yes/No	Symptomatic (Yes/No)
Meninges					
Pituitary Gland					
Orbital lesion (specify location): _____					
Lacrimal Glands					
Parotid Glands					
Submandibular Glands					
Other Salivary Glands (specify): _____					
Mastoiditis / Middle ear disease					
Nasal Cavity Lesions					
Sinusitis					
Other ENT Lesions, e.g., tonsillitis, pharyngitis (specify): _____					
Thyroid					
Lungs					
Lymph Nodes (please circle site of involvement, below):					
Submental Submandibular Cervical Axillary Mediastinal Hilar Abdominal/Pelvic Inguinal Other lymph node chains:					

	Activity			Damage	
Organ/Site	Organ/Site Score (0-3)	Symptomatic (Yes/No)	Urgent (Yes/No)	Yes/No	Symptomatic (Yes/No)
Aorta / Large Blood Vessels					
Heart/Pericardium					
Retroperitoneal Fibrosis					
Sclerosing Mediastinitis					
Sclerosing Mesenteritis					
Pancreas					
Liver					
Bile ducts					
Kidney					
Skin					
Constitutional symptoms not attributable to involvement of a particular organ (weight loss, fever, fatigue caused by active IgG4-RD)					
Other involvement - specify: (Consider prostate, breast, gallbladder involvement; and other. Each "Other" item is counted separately.) _____ _____	_____ _____	_____ _____	_____ _____	_____ _____	_____ _____

Total Activity Score

Organ/sites (x 2 if urgent): _____

Total **urgent** organs: _____

Total **symptomatic (active)** organs: _____

Total **damaged** organs: _____

Total **symptomatic (damage)** organs: _____

Version: October 21, 2017

Manual of Operations (Instruction Guide)

IgG4-Related Disease Responder Index (IgG4-RD RI)

Assessing Disease Activity, Disease Damage, & Treatment Response

I. Introduction

The IgG4-RD-RI is a two-page instrument that is designed to record both disease activity and damage.

The IgG4-RD-RI is based upon the Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS/WG)[1]. The BVAS/WG has been used as an outcome measure in multiple international clinical trials in vasculitis. The underlying principle of both the BVAS/WG and the IgG4-RD RI is that the investigator rates the degree of disease activity or the presence of damage in each organ, based on the available clinical, laboratory, radiology, and pathology evidence. The individual organ scores are then summed to provide an overall score.

The IgG4-RD RI will be most useful in following individual patients longitudinally over time, as a record of changes in disease activity and the accrual of damage.

Some of the major principles of IgG4-RD RI scoring are:

- If the patient's disease activity has improved since the previous visit, the IgG4-RD RI should be lower than the previous recording.
- If there is no active disease in any organ, the overall IgG4-RD RI Total Activity Score should be zero.
- If damage has occurred but there is no active IgG4-RD, the Total Activity Score (see page 2 of the scoring sheet) should be zero.
- When the Total Activity Score for the IgG4-RD RI is zero, the corresponding Physician Global Assessment of disease activity should also be zero.
- If the patient is being evaluated for the first time and has received no treatment, all manifestations of active disease are considered to be New, even if present for months.

These principles will be reinforced through a series of case-based paper exercises. Important definitions are shown on the following page.

II. Definitions

Disease activity refers to manifestations due to inflammation within an organ related to active IgG4-RD. In contrast, **damage** refers to organ system dysfunction or permanent defects (e.g., organ enlargement) that has occurred as a result of active IgG4-RD or its surgical management. It is possible to have disease activity and disease-associated damage simultaneously within the same organ, but activity and damage should be recorded separately on the IgG4-RD RI.

Disease damage due to IgG4-RD: damage can stem from disease that is currently active or has been active previously, is generally regarded as irreversible and causes organ dysfunction or organ failure. In contrast to disease activity, disease damage is NOT expected to improve with immunosupressive treatment or with time.

Disease damage due to surgical intervention for IgG4-RD: damage can also result from surgical intervention to diagnose or manage IgG4-RD. Major surgical procedures such as partial pancreatectomy, thyroidectomy, or pulmonary lobectomy should be regarded as damage. In contrast, fine needle aspirations are usually tolerated well and should not be scored as damage.

Damage is recorded in the right-sided columns of the IgG4-RD RI. Clinical symptoms and signs that are associated with damage as opposed to active disease are recorded in the damage column, not in the disease activity section.

Patients may have both damage and activity in the same organ at the same time. Both damage and activity should be recorded on the IgG4-RD RI if both are present simultaneously, but damage should be scored in the damage section of the RI and activity scored in the activity section.

III. Impact of Serological Markers

Serological markers such as the serum IgG4 concentration, the IgE concentration, the erythrocyte sedimentation rate, and complement levels are not considered directly in the IgG4-RD RI. These measures may influence the investigator's thinking about whether or not the overall disease is active in a particular organ system, however, and might be recorded separately as additional outcome measures, depending on the study. As an example, serum complement levels are often low in patients with IgG4-related tubulointerstitial nephritis. If the C3 and C4 concentrations have normalized at a particular visit,

this may influence the investigator's belief that the degree of kidney involvement has improved or resolved.

IV. Use of Imaging Studies

Imaging studies are often helpful in the diagnosis of IgG4-RD and in the longitudinal assessment of disease activity and damage. The results of imaging studies can inform an investigator's judgment about whether or not the disease is active in a particular organ, or if damage has occurred in that organ.

Imaging studies must be used when clinically appropriate in the evaluation of patients. Imaging studies are also performed at certain prescribed intervals in clinical trials. For example, positron emission tomography (PET) studies might be performed at the time of enrollment and at the time of primary outcome assessment or other major trial milestones in a clinical trial. When evaluating the meaning of imaging study findings, the investigator must always use careful clinical judgment to determine if the findings on an imaging study represent active disease, damage, or another process that is unrelated to IgG4-RD.

For reasons of cost, radiation exposure, and convenience, it is not possible to obtain imaging studies on every patient at every visit. In scoring the IgG4-RD-RI, therefore, investigators must sometimes consider that if a patient's overall clinical status has improved following treatment, it is also reasonable to assume that any potentially reversible imaging findings have also improved. Obviously, in the context of a clinical trial, this assumption must be confirmed at the next time of protocolized imaging.

V. Recording Disease Activity

Three columns in the IgG4-RD RI describe different aspects of disease activity. These columns are the:

- "Organ/Site score" column: Is IgG4-RD active at this site/organ now?
- "Symptomatic" column: Is the IgG4-RD activity symptomatic for the patient?
- "Urgent" column: Is the need to institute treatment for this manifestation of IgG4-RD activity urgent?

We define the "Organ/Site score" and the "Symptomatic" and "Urgent" columns in further detail, below:

Organ/Site score: The overall level of IgG4-RD activity within a specific organ system or body site is recorded in this column. Each organ or site is given a separate score (see description of the individual scores, Section VI). Individual

scores within a total of approximately 25 different organs & sites are summed to give the total IgG4-RD Activity Score.

Symptomatic: Is the disease manifestation in a particular organ system symptomatic? (Y = yes; N = no)

Even if IgG4-RD is active, not all organ system features lead to symptoms. As examples, some cases of lymphadenopathy, pulmonary nodules, and proteinuria generally to not cause patients to have symptoms. Even if a disease feature is not symptomatic, however, it may be important: progression of asymptomatic IgG4-RD activity can lead to significant organ damage or organ failure over time.

Urgent disease: Does the disease in this organ/site require treatment immediately to prevent serious organ dysfunction (Y = yes; N = no)?

Some features of IgG4-RD require treatment urgently to prevent serious organ dysfunction. Examples include IgG4-related renal disease, biliary tract disease, aortitis, and pachymeningitis. If an IgG4-RD feature is sufficiently severe as to require immediate treatment to prevent serious organ dysfunction, that organ/site disease is considered to be urgent.

Urgent disease at an organ or site is weighted by a factor of two within that organ system. For example, if the kidney is assessed to have active disease and given a score of "3" to indicate the presence of New/Worse Disease and the investigator believes that it is important to institute treatment urgently, then the score for the kidney item is doubled to "6".

Urgent disease does not reflect the extent of the disease. In some patients, IgG4-RD may be widespread and involve multiple organ systems or sites, yet not be worrisome enough in any body area to require treatment immediately. In contrast, disease that is apparently isolated to one organ system (e.g., the liver/biliary tract) may be severe enough to require treatment urgently.

VI. Scoring System for Disease Activity at Each Individual Organ/Site

The score should be determined based on disease activity over the preceding 28 days (or the most recent visit if less than 28 days ago). Here is the scoring system for each overall organ or site:

0 – If the organ/site is normal, or if previous IgG4-RD disease activity has resolved, then the score in that organ is zero. Zero signifies no active IgG4-RD in this particular organ or site. Damage from previously active IgG4-RD, in contrast, can be recorded in this organ or at this site, in the appropriate column.

1 – If the disease activity in the organ/site has improved (but not resolved)

2 – If the disease activity is persistent (Unchanged; still active)

3 – If there is new disease activity at an organ or site where there was previously none (either because disease was in remission or never present in that organ), or if there is recurrent disease at a site when the patient is off treatment

4 – If the disease activity is worse in that organ/site *despite treatment*, or if it is a new manifestation that has occurred *despite treatment*.

To determine whether disease activity is worse despite treatment with oral medications, we consider "on treatment" to include all days on which the patient takes the medication. For instance, if they take prednisone daily and then flare four days after completing treatment, they would receive a score of 3. By contrast, if they flare the last day they take prednisone, they would receive a score of 4.

The standard is different for infusion medications. If the patient is being treated with rituximab, he or she is considered "on treatment" so long as B cells remain depleted (as measured by flow cytometry) or 6 months after the first infusion, whichever comes first.

VII. Recording of Damage

One column in the IgG4-RD RI relates to damage as opposed to activity. This is the "Damage" column, separated on the right from the activity columns on the left.

Damage: "Damage" refers to organ system dysfunction or other changes in the organ or anatomic site that have occurred as a result of active IgG4-RD, whether or not the disease is still active in that organ system. Damage is regarded as a permanent organ/site "scar" that will not improve, even with appropriate treatment for IgG4-RD.

As is true with disease activity, the effects of disease damage can be either symptomatic or asymptomatic. The investigator should also indicate in the final column on the right whether or not the damage from IgG4-RD is symptomatic. For example, if IgG4-related retroperitoneal fibrosis leads to chronic flank or abdominal pain for the patient, this would be considered damage. In contrast, the damage associated with IgG4-related renal disease is generally asymptomatic, even in the setting of substantial renal dysfunction.

Three important points related to the scoring of damage:

1) Patients who have undergone a surgical procedure (aside from a fine-needle aspirate) for the purpose of either diagnosis or treatment are considered to have sustained damage at that particular site.

2) Damage that has occurred primarily as a result of **treatment** (for example, osteoporosis, cataracts, or infection) is not scored in the IgG4-RD RI.

3) Sometimes it can be difficult to be certain if a particular type of organ dysfunction is related to IgG4-RD (in which case it should be scored) or if the organ dysfunction is related to treatment (in which case it should not be scored). An example of this is diabetes mellitus: both type 1 IgG4-related autoimmune pancreatitis and treatment with glucocorticoids can lead to diabetes. The investigator must often make a clinical decision regarding the relative contribution of disease as opposed to treatment in leading to organ dysfunction.

We provide some examples of IgG4-RD-associated damage, below.

Examples of damage to organs or sites affected by IgG4-RD:

- Pancreatic insufficiency requiring enzyme replacement
- Biopsy of the meninges for the purpose of establishing the diagnosis
- The need for hormone replacement therapy because of pituitary involvement
- Proptosis that persists after treatment
- Vision loss
- Dry eyes that persists after treatment
- Thyroid surgery to treat IgG4-related Riedel's thyroiditis
- Aortic dissection
- Hydronephrosis requiring permanent ureteral stents
- Chronic pain from retroperitoneal fibrosis
- Superior vena cava syndrome resulting from IgG4-related sclerosing (fibrosing) mediastinitis
- Diabetes mellitus caused by type 1 (IgG4-related autoimmune pancreatitis)
- Permanent decline in creatinine clearance or end-stage renal disease

Version: October 21, 2017

VIII. Totaling the Scores

On page 2 of the Scoring Sheet, please total the Activity and Damage scores as shown below. An example is provided:

Suppose the patient is a newly-diagnosed, previously untreated 62 year-old man just recognized to have:

- Type 1 (IgG4-related) autoimmune pancreatitis (AIP)
- "Mikulicz disease" affecting the lacrimal, parotid, and submandibular glands
- IgG4-related pachymeningitis

Both the AIP and the pachymeningitis are symptomatic (causing abdominal pain and headache, respectively). You believe that the need to treat both the AIP and the pachymeningitis is urgent. The lacrimal, parotid, and submandibular gland disease, in contrast, are not symptomatic and do not require treatment urgently.

The patient's total activity score is therefore:

	Organ/Site Score (0-4)	Activity Symptomatic (Yes/No)	Urgent (Yes/No)	Damage Yes/No	Damage Symptomatic (Yes/No)
AIP	3*	Yes	Yes	No	No
Lacrimal	3	No	No	No	No
Parotid	3	No	No	No	No
Submandibular	3	No	No	No	No
Pachymeninges	3*	Yes	Yes	No	No
TOTAL	**21**				

***Multiplied times two, because urgent. Therefore, both of these 3 scores become 6 scores.**

Total Activity Score

Organ/sites (x 2 if urgent): **21**

Total **urgent** organs: **2**

Total **symptomatic (active)** organs: **2**

Total **damaged** organs: **0**

Total **symptomatic (damage)** organs: **0**

Version: October 21, 2017

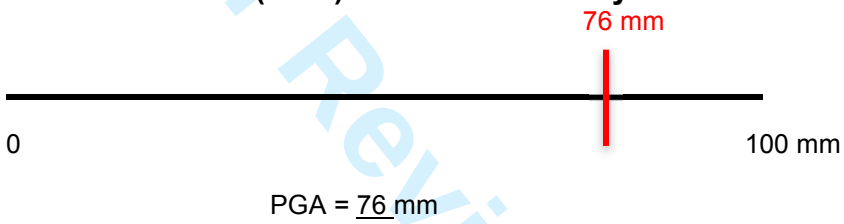
IX. Physician Global Assessment of Disease Activity

Finally, record your global assessment of the patient's disease activity. This is done by marking a point on the 100 mm line that you believe represents the degree of the patient's IgG4-RD activity at this assessment.

On this 0-100 mm scale, 0 mm represents NO DISEASE ACTIVITY (i.e., remission), and 100 mm represents extremely active, multi-organ disease.

You mark a point on the line and measure the distance of that point from 0. Your assessment of the degree of disease activity of this patient is 76 (out of a possible 100). Record this measurement at the appropriate site: PGA = ___mm.

Physician Global Assessment (PGA) of disease activity



PGA = 76 mm

Please proceed now to the three practice cases. Please refer back to this Manual of Operations/Instructions as necessary.

Version: October 21, 2017

X. REFERENCE

Stone JH, Hoffman GS, Merkel PA, Min YI, Uhlfelder ML, Hellmann DB, Specks U, Allen NB, Spiera RF, Calabrese LH, Wigley FM, Davis JC, Maiden N, Valente RM, Niles JL, Fye KH, McCune JW, St. Clair EW, Luqmani RA. A disease-specific activity index for Wegener's granulomatosis: Modification of the Birmingham Vasculitis Activity Score. Arthritis & Rheumatism 2001; 44: 912-920.