Arthritis & Rheumatology
Vol. 72, No. 1, January 2020, pp 7–19
DOI 10.1002/art.41120

American College of Rheumatology
Empowering Rheumatology Professionals

SPECIAL ARTICLE

The 2019 American College of Rheumatology/European League Against Rheumatism Classification Criteria for IgG4-Related Disease

Zachary S. Wallace,[1] Ray P. Naden,[2] Suresh Chari,[3] Hyon Choi,[1] Emanuel Della-Torre,[4] Jean-Francois Dicaire,[5] Phil A. Hart,[6] Dai Inoue,[7] Mitsuhiro Kawano,[8] Arezou Khosroshahi,[9] Kensuke Kubota,[10] Marco Lanzillotta,[11] Kazuichi Okazaki,[12] Cory A. Perugino,[1] Amita Sharma,[1] Takako Saeki,[13] Hiroshi Sekiguchi,[3] Nicolas Schleinitz,[14] James R. Stone,[1] Naoki Takahashi,[3] Hisanori Umehara,[15] George Webster,[16] Yoh Zen,[17] and John H. Stone,[1] for the American College of Rheumatology/European League Against Rheumatism IgG4-Related Disease Classification Criteria Working Group

This criteria set has been approved by the European League Against Rheumatism (EULAR) Executive Committee and the American College of Rheumatology (ACR) Board of Directors. This signifies that the criteria set has been quantitatively validated using patient data, and it has undergone validation based on an independent data set. All ACR/EULAR-approved criteria sets are expected to undergo intermittent updates.

The ACR is an independent, professional, medical and scientific society that does not guarantee, warrant, or endorse any commercial product or service.

Objective. IgG4-related disease (IgG4-RD) can cause fibroinflammatory lesions in nearly any organ. Correlation among clinical, serologic, radiologic, and pathologic data is required for diagnosis. This work was undertaken to develop and validate an international set of classification criteria for IgG4-RD.

Methods. An international multispecialty group of 86 physicians was assembled by the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR). Investigators used consensus exercises, existing literature, derivation and validation cohorts of 1,879 subjects (1,086 cases, 793 mimickers), and multicriterion decision analysis to identify, weight, and test potential classification criteria. Two independent validation cohorts were included.

Results. A 3-step classification process was developed. First, it must be demonstrated that a potential IgG4-RD case has involvement of at least 1 of 11 possible organs in a manner consistent with IgG4-RD. Second, exclusion criteria consisting of a total of 32 clinical, serologic, radiologic, and pathologic items must be applied; the presence of any of these criteria eliminates the patient from IgG4-RD classification. Third, 8 weighted inclusion criteria domains, addressing clinical findings, serologic results, radiology assessments, and pathology interpretations, are applied. In the first validation cohort, a threshold of 20 points had a specificity of 99.2% (95% confidence interval [95% CI] 97.2–99.8%) and a sensitivity of 85.5% (95% CI 81.9–88.5%). In the second, the specificity was 97.8% (95% CI 93.7–99.2%) and the sensitivity was 82.0% (95% CI 77.0–86.1%). The criteria were shown to have robust test characteristics over a wide range of thresholds.

Conclusion. ACR/EULAR classification criteria for IgG4-RD have been developed and validated in a large cohort of patients. These criteria demonstrate excellent test performance and should contribute substantially to future clinical, epidemiologic, and basic science investigations.

Introduction

IgG4-related disease (IgG4-RD) is an immune-mediated condition associated with fibroinflammatory lesions that can occur at nearly any anatomic site (1,2). It often presents as a multiorgan disease and may be confused with malignancy, infection, or other immune-mediated conditions, such as Sjögren's syndrome or vasculitis, associated with antineutrophil cytoplasmic antibodies (ANCAs). Rheumatologists, internists, gastroenterologists, nephrologists, pulmonologists, neurologists, radiologists, pathologists, and other practitioners are often involved in the evaluation of patients with this condition. IgG4-RD can lead to organ dysfunction, organ failure, and death. Its epidemiology remains poorly described because of its relatively recent recognition as a discrete condition, yet the disease is now seen by both generalists and specialists all across the world.

IgG4-RD was first recognized as a distinct disease in 2003 (3,4). Over the next decade, it became clear that although the disease could affect virtually any organ, there are strong predilections for certain organs (1,5). These include the major salivary glands (submandibular, parotid, sublingual), the orbits and lacrimal glands, the pancreas and biliary tree, the lungs, the kidneys, the aorta and retroperitoneum, the meninges, and the thyroid gland (Riedel's thyroiditis) (6–8). Many of the early diagnoses of IgG4-RD relied on pathologic assessment of surgical resection specimens (9). These discoveries were often incidental findings made following resections of lesions with suspected malignancy. The large pathologic samples available from such procedures generally permitted identification of a full range of findings considered characteristic of IgG4-RD: a lymphoplasmacytic infiltrate, storiform fibrosis, obliterative phlebitis, and dramatic IgG4+ plasma cell infiltrates, among others (9). With growing recognition of this condition, however, the diagnosis is now made using increasingly small biopsy samples that frequently do not demonstrate the full spectrum of pathologic findings (7,9,10). In a subset of patients with classic combinations of clinical, serologic, or radiologic findings, clinical diagnoses are sometimes made in the absence of biopsy, but the threshold to perform biopsies of accessible sites when there is significant concern about malignancy or infection remains appropriately low.

Other cases diagnosed early in the course of IgG4-RD were identified because of striking elevations in serum IgG4 concentrations (4). However, it is now recognized that serum IgG4 levels are normal in a substantial percentage of patients with clinicopathologic diagnoses of IgG4-RD (6,11,12). Although serum IgG4 concentrations can provide an important clue to the diagnosis and some guidance in the longitudinal assessment of disease activity, the centrality of IgG4 in the overall pathophysiology of this condition has been called into question (13). The presence of an elevated serum IgG4 level is no longer considered essential to the diagnosis of IgG4-RD. Indeed, certain organ systems and anatomic regions (e.g., the retroperitoneum) are less likely to be associated with a serum IgG4 elevation than are others (6).

Finally, the radiologic features of IgG4-RD have also been described with increasing thoroughness. Radiologic findings such as a sausage-shaped pancreas and periaortitis affecting the infrarenal aorta are now viewed as being strongly suggestive of IgG4-RD if detected in the proper clinical context (14,15). Nevertheless, radiologic findings in isolation—without reference to clinical, serologic, or pathologic data—are never sufficient for either clinical diagnosis or appropriate disease classification.

In short, although clinical, serologic, radiologic, and pathologic features all contribute to the classification of IgG4-RD, none of these approaches alone provides definitive evidence for the accurate classification of patients. The proper categorization of patients for both research studies and clinical purposes relies upon integration of data from all 4 domains of evidence. Given the recent recognition of IgG4-RD as a distinct condition, along with its multiorgan nature and the absence of a single diagnostic feature, classification criteria are now needed for the conduct of high-quality clinical and epidemiologic investigations in this disease.

Methods

This study was approved by the Partners HealthCare Institutional Review Board.

Study overview. The development and testing of the classification criteria for IgG4-RD was based on consensus-based and data-driven methods using prospectively collected data and decision analytics (16–19).

This article is published simultaneously in the January 2020 issue of *Annals of the Rheumatic Diseases*.

Supported by the American College of Rheumatology and the European League Against Rheumatism.

[1]Zachary S. Wallace, MD, Hyon Choi, MD, PhD, Cory A. Perugino, DO, Amita Sharma, MD, James R. Stone, MD, MPH, John H. Stone, MD, MPH: Massachusetts General Hospital, Boston; [2]Ray P. Naden, MD, FRACP: New Zealand Health Ministry, Auckland, New Zealand; [3]Suresh Chari, MD, Hiroshi Sekiguchi, MD, Naoki Takahashi, MD, PhD: Mayo Clinic, Rochester, Minnesota; [4]Emanuel Della-Torre, MD: IRCCS Ospedale San Raffaele, Milan, Italy; [5]Jean-Francois Dicaire: Pinnacle, Inc., Montreal, Quebec, Canada; [6]Phil A. Hart, MD: Ohio State University College of Medicine, Columbus; [7]Dai Inoue, MD, MPH: Kanazawa University, Kanazawa, Japan; [8]Mitsuhiro Kawano, MD, MPH: Kanazawa University Hospital, Kanazawa, Japan; [9]Arezou Khosroshahi, MD: Emory University, Atlanta, Georgia; [10]Kensuke Kubota, MD: Yokohama City University, Yokohama, Japan; [11]Marco Lanzillotta, MD: IRCCS, Milan, Italy; [12]Kazuichi Okazaki, MD, PhD: Kansai Medical University, Hirakata, Japan; [13]Takako Saeki, MD, PhD: Nagaoka Red Cross Hospital, Nagaoka, Japan; [14]Nicolas Schleinitz, MD: Aix-Marseille University, Marseille, France; [15]Hisanori Umehara, MD, PhD: Kanazawa Medical University, Uchinada, Japan, and Hayashi Hospital, Echizen, Japan; [16]George Webster, MD, FRCP: University College London, London, UK; [17]Yoh Zen, MD, PhD, FRCP: Kobe University, Kobe, Japan.

No potential conflicts of interest relevant to this article were reported.

Address correspondence to John H. Stone, MD, MPH, Massachusetts General Hospital, Rheumatology Unit, Yawkey 2, 55 Fruit Street, Boston, MA 02114. E-mail: jhstone@mgh.harvard.edu.

Submitted for publication August 18, 2018; accepted in revised form September 12, 2019.

Investigators. A Steering Committee composed of investigators from North America, Europe, and Asia was established. The Steering Committee directed the entire project and invited other investigators who were assigned to specific Advisory Groups addressing clinical, serologic, radiologic, and pathologic issues. In addition to members of the Steering Committee and the Advisory Groups, other investigators were invited to participate by submitting cases of IgG4-RD and of mimicking conditions to be used in the development and testing phases of the study. This full group of investigators is known as the American College of Rheumatology (ACR)/European League Against Rheumatism (EULAR) IgG4-RD Classification Criteria Working Group (Appendix A).

Item generation. Each Advisory Group consisted of a Steering Committee member and experts in the field being addressed by the specific Advisory Group. The Advisory Groups were tasked with using evidence- and consensus-based approaches to identify items that might be relevant to the classification of patients as having or not having IgG4-RD. These items comprised preliminary exclusion criteria and preliminary inclusion criteria. Preliminary exclusion criteria were defined as items that would lead to termination of consideration of the patient as an IgG4-RD case. In contrast, preliminary inclusion criteria could either increase or decrease the likelihood of classification of the patient as an IgG4-RD case. Preliminary inclusion criteria that demonstrated discriminatory ability to increase the likelihood of classification were later selected as inclusion criteria. A 24-member Steering Committee of the ACR/EULAR IgG4-RD Classification Criteria Development Group met in Boston in April 2016 to begin this process. At this initial Steering Committee meeting, 104 rounds of consensus-based decision-making were conducted. Consensus was achieved for 79 (76%) of these decisions, the process of which is described below. Item generation and the subsequent task of item reduction were continued through teleconferences and e-mail discussions.

Process of consensus. The rules regarding consensus were set out at the time of the first face-to-face meeting. Consensus was considered to have been reached when 80% of the members of the Steering Committee were in agreement on a given point. Discussion was permitted following achievement of the 80% threshold, however, if individuals in the minority wished to express the rationale behind their opinion. During discussions, evidence was presented by participants to support arguments. Discussants referred to the medical literature when relevant to illuminate a particular question. In some instances, in the setting of a persuasive argument by a member of the minority, discussion led to re-voting and occasionally to a change in the ultimate decision on a particular point.

Item reduction. Following item generation, the Steering Committee participated in 2 exercises to reduce the number of items. First, the Committee reviewed all proposed inclusion and exclusion criteria and reduced the potential criteria into 8–10 domains through the consensus process described above. Related items were clustered within domains that were independent of the other domains; for preliminary inclusion criteria, items contributed positive or negative weights toward classifying cases as IgG4-RD. For instance, biopsy immunohistochemistry results (e.g., IgG4+ plasma cells/high-power field [hpf] and IgG4+IgG+ plasma cells/hpf) were listed under an immunohistochemistry domain. Within each preliminary inclusion criteria domain, items were arranged by group members according to the degree to which they either increased or decreased the likelihood of classification as IgG4-RD (e.g., an infiltrate of ≥40 IgG4+ plasma cells/hpf was positioned above an infiltrate of 0–9 IgG4+ plasma cells/hpf). Definitions for each item were determined such that cases could be assigned clearly to only 1 item in a domain.

The Steering Committee then ranked each potential preliminary inclusion criteria item on a Likert scale from –5 ("Highly confident the patient does not have IgG4-RD if this item is present") to +5 ("Highly confident the patient has IgG4-RD if this item is present"). Items associated with an average confidence between –2.0 and +2.0 were deemed to have insufficient sensitivity or specificity and were excluded from further consideration.

Derivation case collection. Investigators were invited to submit cases of IgG4-RD or mimicking conditions that they had managed and to report the presence or absence of each preliminary item for each submitted case using standardized data collection forms. No identifying data on these patients were collected. Investigators were encouraged to submit data on a broad range of IgG4-RD cases, including cases in which they were highly confident in the diagnosis as well as those in which they were less confident. The investigator submitting the case proposed the initial classification of the case as IgG4-RD or as a mimicker of IgG4-RD. This initial classification of all cases was reviewed by a subset of the Steering Committee to confirm the appropriateness of the initial designation. Cases that appeared to be inappropriately classified by the investigator or cases with insufficient information on which to base a classification decision were discarded.

Approach to assigning relative weights to inclusion criteria items. Twenty of the submitted cases representing a combination of IgG4-RD and mimickers were selected for a Steering Committee exercise designed for 2 purposes. First, the exercise was used to assign preliminary weights to the inclusion criteria. Second, it fostered discussion and facilitated consensus on the definitions of individual items. Only cases that did not fulfill any of the exclusion criteria were selected for this exercise. The cases selected represented a broad range of manifestations in order to assess the performance of all potential criteria. Investigators were

asked to rank all cases in order from most likely to least likely to be classified as IgG4-RD. In addition, investigators were asked to indicate the point at which they would divide the cases into those that should be classified as IgG4-RD and those that were more likely to be mimickers.

The draft IgG4-RD classification criteria consisted of 8 domains and a total of 29 items. Once preliminary domains and items had been selected, the Steering Committee met in person for a 2-day session employing decision science theory and computer adaptive technology. A computer software program known as 1000minds (http://www.1000minds.com) was used. Investigators participated in a series of discrete, forced-choice experiments through pairwise rankings of alternatives that led to quantified weights for each item (20–22). During this exercise, investigators were presented with a series of paired scenarios (A and B), each of which contained the same 2 domains (e.g., serum IgG4 concentrations and salivary gland disease). Different combinations of the domains' items were grouped together in each scenario.

For each paired scenario choice, investigators selected the scenario they believed to contribute more toward classification of the patient as having IgG4-RD, assuming that all other aspects of the case were the same. The distribution of votes (percent who voted for A, B, or "equal probability") was presented for each pair of scenarios after each vote. Discussions and re-voting were pursued when necessary, using the same process of consensus described above. Consensus was considered to have been achieved when all participants either indicated complete agreement as to which scenario represented a higher probability of IgG4-RD or indicated that they could accept the majority opinion. During this phase of classification criteria development, 160 rounds of consensus-based decision-making were conducted. Based on this voting, the computer software assigned relative weights to each item. The specific weights assigned to each item were not revealed to investigators.

Scoring of weighted items. If >1 item was present within a given domain, only the highest-weighted item was scored. As an example from the Chest domain, if a patient had peribronchovascular and septal thickening evident on computed tomography of the chest (weighted 4 points) as well as a paravertebral band-like soft tissue mass in the thorax (weighted 10), only the weight of the paravertebral band-like soft tissue mass in the thorax would count in the patient's total classification criteria score.

Identifying a threshold for classifying IgG4-RD. Each derivation case that was not removed by an exclusion criterion was assigned a total score based on the aggregation of weighted inclusion criteria present. These cases were ranked and a preliminary threshold was identified based on targets of >90% for specificity and >80% for sensitivity. Cases around the threshold were selected for discussion among the investigators, who reached consensus on a cutoff point between the group of patients who should be classified as having IgG4-RD and those who could not be confidently classified as having IgG4-RD. A preliminary threshold of 20 was selected by 2 of the investigators (RPN and JHS) after an in-person review of cases around this threshold revealed a common point at which cases were more likely to be classified by investigators as not clearly being IgG4-RD. This preliminary threshold was then tested in the first of 2 validation phases, using newly submitted cases of IgG4-RD and IgG4-RD mimickers. This preliminary threshold was not revealed to other investigators as the cases for the validation phase were collected.

Collection of IgG4-RD cases and mimickers for the first validation phase. Investigators were invited to submit a second set of data from cases of IgG4-RD or mimicking conditions. None of the cases in this second set had been included in the derivation set. The investigators reported the presence or absence of each finalized item using standardized data collection forms. For each case, investigators reported their confidence in the diagnosis on a scale of 0–3 in which 0 = uncertain, 1 = slightly confident, 2 = confident, and 3 = very confident.

Testing of the IgG4-RD classification criteria and other statistical analyses. We evaluated the performance of the preliminary classification criteria among those cases that fulfilled the entry criteria. To determine the test performance, we only analyzed cases in which investigators were at least "confident" or "very confident" in the diagnosis (IgG4-RD or mimicker); thus, a "confident" or "very confident" diagnosis was considered the gold standard for the purpose of assessing test performance. The number of patients with "confident" or "very confident" designations as either IgG4-RD cases or IgG4-RD mimickers was 771, or 85% of all of the patients included in the first validation phase.

We assessed the test performance of the classification criteria at the preliminary threshold of 20 as well as at a range of thresholds above and below 20. To determine the optimal threshold, we considered the goal of our classification criteria for use in clinical trials (specificity >90% and sensitivity >80%). We also considered other measures such as area under the curve (AUC) (23), Youden's criteria (24), distance from (0,1) on a receiver operating characteristic curve (ROC), difference between sensitivity and specificity, and the diagnostic odds ratio (positive likelihood ratio/negative likelihood ratio) (25).

Sensitivity analyses. We performed several sensitivity analyses to test the performance of the criteria. These sensitivity analyses included the following considerations: 1) if all cases, regardless of confidence level were included; 2) if all of the exclusion criteria were removed; 3) if information on serum IgG4 concentrations was not available; 4) if biopsies were not available; and 5) if the mimickers without data on serum IgG4 concentrations or biopsies were assumed to have the highest values for each item. Chi-square tests, Fisher's exact tests,

t-tests, and Wilcoxon rank sum tests were used to compare subgroups, as appropriate.

Testing the final threshold in a second validation cohort. Investigators were invited to submit another set of data from cases of IgG4-RD or mimicking conditions that they had managed but had not yet contributed to the previous derivation or validation cohorts. This second validation cohort was collected because minor changes in the some of the definitions of inclusion and exclusion criteria had been made after the derivation set of patients had been collected, in the interest of clarifying definitions for investigators. However, the definitions of inclusion criteria and exclusion criteria used in the 2 validation cohorts were exactly the same. Using the same approach as above, we assessed the performance of the classification criteria at the identified threshold of 20. We used all cases and mimickers for whom the diagnosis was considered "confident" or "very confident" by the investigator as the gold standard (n = 402 [83%]).

Table 1. Exclusion criteria definitions

Clinical
Fever: Documented, recurrent temperature >38°C, with fever being a prominent part of the patient's overall presentation with the underlying disease, in the absence of any clinical features of infection.
No objective response to glucocorticoids: If the patient has been treated with prednisone at a minimum of 40 mg/day (~0.6 mg/kg/day) for a period of 4 weeks, it is assumed that the patient has not demonstrated an objective clinical response. An objective response includes unequivocal improvement of the clinical lesions, biochemical abnormalities, or radiologic findings. There are 2 additional points to consider with regard to glucocorticoid response: Improvement only in the serum IgG4 concentration should not be regarded as a clinical response without improvement in other aspects of the disease. Some forms of IgG4-related disease (IgG4-RD) associated with advanced fibrosis, e.g., some cases of retroperitoneal fibrosis or sclerosing mesenteritis, may not demonstrate obvious radiologic responses to glucocorticoids.
Serologic
Leukopenia and thrombocytopenia without alternative explanation: Reduction in the total white blood cell count and platelet count to levels below those normal for the reference laboratory, having no apparent explanation except for the underlying disease. Reductions in both the white blood cell count and platelet count are unusual in IgG4-RD but are typical of, for example, myelodysplastic syndromes, hematopoietic malignancies, and autoimmune conditions within the systemic lupus erythematosus spectrum.
Peripheral eosinophilia: To a concentration of >3,000 mm^3.
Positive antineutrophil cytoplasmic antibody (ANCA): Enzyme-linked immunosorbent assay results positive for ANCA targeted against proteinase 3 or myeloperoxidase.
Positive antibodies: Ro, La, double-stranded DNA, RNP, or Sm antibodies positive in titers greater than normal suggest an alternative diagnosis. Other autoantibody associated with high specificity for another immune-mediated condition that is a reasonable explanation for the patient's presentation. Such specific autoantibodies include antisynthetase antibodies (e.g., anti–Jo-1), anti–topoisomerase III (Scl-70), and anti–phospholipase A_2 receptor antibodies. This does not include autoantibodies of low specificity such as rheumatoid factor, antinuclear antibodies, antimitochondrial antibodies, anti–smooth muscle antibodies, and antiphospholipid antibodies.
Cryoglobulinemia: Cryoglobulinemia (type I, II, or III) occurring in a clinical context that provides a reasonable explanation for the patient's presentation.
Radiologic
Known radiologic findings suspicious for malignancy or infection that have not been investigated sufficiently: Such radiologic findings include mass lesions that have not been evaluated thoroughly, necrosis, cavitation, hypervascular or exophytic mass, bulky or matted lymphadenopathy, loculated abdominopelvic fluid collection, among others.
Rapid radiologic progression: Defined as significant worsening within a 4–6-week interval.
Long bone abnormalities consistent with Erdheim-Chester disease: Multifocal osteosclerotic lesions of the long bones, usually associated with bilateral diaphyseal involvement.
Splenomegaly: >14 cm in the absence of alternative explanation (e.g., portal hypertension).
Pathologic
Cellular infiltrates suspicious for malignancy that have not been investigated sufficiently: A high likelihood of malignancy may be suggested by cellular atypia, a monotypic nature of immunohistochemistry findings, or light chain restriction on in situ hybridization studies. If malignancy is suspected, this must be excluded by appropriate studies before inclusion.
Markers consistent with inflammatory myofibroblastic tumor: Known positivity for a marker suggestive of inflammatory myofibroblastic tumor, e.g., anaplastic lymphoma kinase 1 or ROS, a receptor tyrosine kinase that is encoded by the gene *ROS1*.
Prominent neutrophilic inflammation: Neutrophilic infiltrates are unusual in IgG4-RD, with the exception of occasional examples in the lung or near mucosal sites. Extensive neutrophilic infiltrates or neutrophilic abscesses strongly indicate the possibility of a non–IgG4-RD diagnosis.
Necrotizing vasculitis: Although vascular injury (e.g., obliterative phlebitis or arteritis) is a hallmark of IgG4-RD, the presence of fibrinoid necrosis within blood vessel walls provides strong evidence against IgG4-RD.
Prominent necrosis: Small foci of necrosis may rarely be present around the luminal surface of ductal organs, but zonal necrosis with no alternative explanation (e.g., stenting) provides strong evidence against IgG4-RD.
Primary granulomatous inflammation: Inflammation rich in epithelioid histiocytes, including multinucleated giant cell formation and granuloma formation, is highly atypical of IgG4-RD.
Pathologic features of a macrophage/histiocytic disorder: Example: known S100-positive macrophages demonstrating emperipolesis, a pathologic feature of Rosai-Dorfman disease.
Specific disease exclusions
Known diagnoses of the following diseases are exclusion criteria: Multicentric Castleman's disease Crohn's disease (if pancreatobiliary disease is present) Ulcerative colitis (if pancreatobiliary disease is present)
Hashimoto thyroiditis (if the thyroid is the only proposed disease manifestation). Patients with IgG4-RD can certainly have Hashimoto thyroiditis separately from IgG4-RD, but Hashimoto thyroiditis is part of the IgG4-RD spectrum.

Results

Research group. The Steering Committee consisted of investigators from North America, Europe, and Asia. There were 3 Advisory Groups: clinical and serologic, radiologic, and pathologic. A total of 86 investigators submitted cases for the derivation and/or validation sets.

Item generation and reduction. At the conclusion of item generation, definitions for the entry criteria, exclusion criteria, and inclusion criteria were established. The entry criteria were defined as 1) characteristic clinical or radiologic involvement of a typical organ (e.g., pancreas, bile ducts, orbits, lacrimal glands, major salivary glands, retroperitoneum, kidney, aorta, pachymeninges, or thyroid gland [Riedel's thyroiditis]) or 2) pathologic evidence of an inflammatory process accompanied by a lymphoplasmacytic infiltrate of uncertain etiology in one of these same organs. "Characteristic" involvement generally refers to enlargement of the organ or a tumor-like mass within an affected organ. It also includes 3 organ-specific features, with reference to 1) the bile ducts, where narrowing tends to occur, 2) the aorta, where wall thickening or aneurysmal dilatation is typical, and 3) the lungs, where thickening of the bronchovascular bundles is common.

Supplementary Tables 1 and 2 (on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.41120/abstract) list the preliminary exclusion criteria and the preliminary inclusion criteria. There was initially a total of 78 such criteria (51 preliminary exclusion criteria and 27 preliminary inclusion criteria). The preliminary exclusion criteria and preliminary inclusion criteria demonstrating the highest discrimination of IgG4-RD from disease mimickers were chosen as draft classification criteria. Complete definitions of the exclusion criteria and the inclusion criteria are shown in Table 1 and Table 2, respectively. Following the consensus exercises and the Likert scale rating of the preliminary inclusion criteria, refined lists of exclusion and positive and negative inclusion criteria were created (Supplementary Table 2).

Derivation and validation cohorts. Table 3 describes the derivation cohort and the first and second validation cohorts used to develop and assess the performance of the classification criteria. A total of 1,879 patients were included in the overall IgG4-RD classification criteria effort, including 486 in the derivation cohort (272 IgG4-RD cases, 214 mimickers), 908 in the first validation cohort (493 cases, 415 mimickers), and 485 in the second validation cohort (321 cases, 164 mimickers). The patients' status as a case or mimicker, proposed by the submitting investigator, was confirmed by members of the Steering Committee. In both the derivation and validation cohorts, the majority of cases were male patients and typically in their sixth decade of life, consistent with the demographics of IgG4-RD and many of its mimicking conditions.

Classification criteria (Table 4). The derivation cohort was used to assess the relative performance of each proposed exclusion and inclusion criterion. The exclusion criteria are not designed to be a "laundry list" of evaluations that must be checked off as negative before a patient can be classified as having IgG4-RD. Rather, they serve as a reminder to the investigator of evaluations that might be appropriate to consider in specific clinical scenarios.

Criteria that did not distinguish IgG4-RD cases from mimickers were eliminated, and those that helped distinguish IgG4-RD cases from mimickers were retained. The final entry criteria and items were modified through in-person discussion after completion of the 1000minds program and review of the derivation cases (n = 486) ranked in order of points accrued by totaling the weights associated with each inclusion criteria item after cases fulfilling exclusion criteria had been excluded. A preliminary score of 20 was identified

Table 2. Inclusion criteria definitions

Immunostaining
IgG+ cells can be identified using either IgG staining or CD138 staining.
Head and neck gland involvement
A "set" of glands refers to both lacrimal glands or both submandibular glands, etc. If a gland has been surgically removed for the purpose of diagnosis, it can be considered to have been involved if confirmed by pathology.
Involvement of the lacrimal glands and the major salivary glands in IgG4-related disease is bilateral (but can be asymmetric). Involvement of the glands can be determined either by clinical examination or by a radiology study (e.g., positron emission tomography scan or computed tomography scan).
Chest
Peribronchovascular and septal thickening in the lung must be determined by a cross-sectional imaging study of the chest.
The paravertebral band-like soft tissue in the thorax is usually right-sided, located between T8 and T11, and does not encase the aorta.
Pancreas and biliary tree
Diffuse pancreas enlargement usually encompasses more than two-thirds of the pancreas.
The type of biliary involvement that is highly consistent with IgG4-related sclerosing cholangitis involves the proximal biliary tract (i.e., intrahepatic and extrapancreatic portions of the extrahepatic bile ducts). The bile duct walls often have smooth thickening.
Kidney
Hypocomplementemia pertains to low serum levels of C3, C4, or both.
Renal pelvic wall thickening can be either unilateral or bilateral, usually without severe stenosis or luminal irregularity.
Low-density areas in both renal cortices can be seen only on contrast-enhanced computed tomography and are usually patchy or round-shaped in appearance.
Retroperitoneum
The location of IgG4-related retroperitoneal fibrosis or periaortitis is typically circumferential or on the anterolateral sides of the aorta. The segment of aorta involved tends to be the infrarenal aorta, often extending to include the iliac vessels.

Table 3. Demographic and disease characteristics of the derivation and validation cohorts*

	Derivation cohort (n = 486)†	Validation cohort 1 (n = 908)†	Validation cohort 2 (n = 485)†
IgG4-related disease	272 (56)	493 (54)	321 (66)
Mimickers‡	214 (44)	415 (46)	164 (34)
Vasculitis	26 (12)	106 (26)	34 (21)
Malignancy	51 (24)	31 (7)	36 (22)
Sjögren's syndrome	13 (6)	59 (14)	8 (5)
Other pancreatitis	5 (2)	15 (4)	7 (4)
Other	119 (56)	204 (49)	79 (48)
Male sex	319 (66)	503 (55)	288 (59)
Age at diagnosis, mean ± SD years	58.2 ± 14.5	55.5 ± 16.5	56.4 ± 16.8
Select organ involvement			
Salivary glands	153 (31)	278 (31)	151 (31)
Orbit	101 (21)	188 (21)	146 (30)
Pulmonary	128 (26)	173 (19)	75 (15)
Lymph nodes	176 (36)	262 (29)	95 (20)
Aorta	52 (11)	97 (11)	37 (8)
Retroperitoneal fibrosis	78 (16)	108 (12)	50 (10)
Pancreas	132 (27)	269 (30)	160 (33)
Biliary	75 (15)	149 (16)	91 (19)
Renal	90 (19)	137 (15)	74 (15)
No. of organs involved, median (interquartile range)	2 (1–4)	2 (1–3)	2 (1–3)

* In validation cohort 1, the judgment of a case as being IgG4-related disease (IgG4-RD) or as being an IgG4-RD mimicker was "confident" or "very confident" in 771 cases (84.9% of all cases and mimickers included in that cohort). In validation cohort 2, the judgment of a case as being IgG4-RD or as being an IgG4-RD mimicker was "confident" or "very confident" in 431 cases (88.9% of all cases and mimickers included in that cohort). Except where indicated otherwise, values are the number (%).
† Includes all submitted cases and mimickers.
‡ Mimicker conditions are listed in Supplementary Table 5, on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.41120/abstract.

as the cutoff point at or above which the majority of investigators considered the patient to have IgG4-RD; with this threshold, a sensitivity of >80% and high specificity were also achieved.

Validating the classification criteria. We then tested the performance of the classification criteria in the first validation cohort (n = 908). To determine the optimal cutoff, we assessed the test performance of criteria at various thresholds (Table 5). Given that the purpose of the criteria was to identify patients with IgG4-RD for enrollment in research studies, the ideal threshold would have excellent specificity while retaining good sensitivity (>80%). The preliminary threshold of 20 had a specificity of 99.2% (95% confidence interval [95% CI] 97.2–99.8%) and a sensitivity of 85.5% (95% CI 81.9–88.5%). Moreover, the threshold of 20 had excellent discrimination, with an AUC of 0.924 (95% CI 0.906–0.941). A threshold of either 21 or 22 had a specificity identical to that obtained with the threshold of 20, but sensitivity decreased at those thresholds, as reflected in other measures of threshold performance, including the AUC. A threshold of 20 also had the highest diagnostic odds ratio compared to other thresholds.

Because of the emphasis placed upon specificity, we considered the test characteristics obtained with a threshold of 20 superior to those of other potential thresholds. Of note, however, a threshold of 16 performed better in certain measures, including sensitivity (88.6%), Youden's criteria, distance from (0,1) on the ROC curve (0.12), and AUC (0.933 [95% CI 0.916–0.950]). The threshold of 16 was associated with a slightly lower specificity: 98.1% versus 99.2%. When comparing a threshold of 20 to a threshold of 16 with regard to the diagnostic odds ratio, a threshold of 20 was associated with superior test performance (761.5 versus 394.5). The consistent performance of these classification criteria across a range of thresholds suggests that the criteria will be robust when used in the clinic for purposes of research.

Analyses were then performed using the second validation cohort (n = 485). In this group the classification criteria had a specificity of 97.8% (95% CI 93.7–99.2%) and a sensitivity of 82.0% (95% CI 77.0–86.1%).

Sensitivity analyses with a threshold of 20. We performed a number of sensitivity analyses to assess the robustness of the classification criteria at a threshold of 20 in the first validation cohort. If all cases, regardless of confidence in the diagnosis, were included, the classification criteria performed very well, with a sensitivity of 83% and a specificity of 98.9%. The IgG4-RD classification criteria are the first of its kind in any rheumatic disease to incorporate absolute exclusion criteria. In a sensitivity analysis that removed exclusion criteria from the classification algorithm, we found that the specificity of the criteria decreased from 99.2%

Table 4. The 2019 American College of Rheumatology/European League Against Rheumatism classification criteria for IgG4-related disease

Step	Categorical assessment or numeric weight
Step 1. Entry criteria	**Yes**† or No
Characteristic* clinical or radiologic involvement of a typical organ (e.g., pancreas, salivary glands, bile ducts, orbits, kidney, lung, aorta, retroperitoneum, pachymeninges, or thyroid gland [Riedel's thyroiditis]) OR pathologic evidence of an inflammatory process accompanied by a lymphoplasmacytic infiltrate of uncertain etiology in one of these same organs	
Step 2. Exclusion criteria: domains and items‡	Yes or **No**§
Clinical	
Fever	
No objective response to glucocorticoids	
Serologic	
Leukopenia and thrombocytopenia with no explanation	
Peripheral eosinophilia	
Positive antineutrophil cytoplasmic antibody (specifically against proteinase 3 or myeloperoxidase)	
Positive SSA/Ro or SSB/La antibody	
Positive double-stranded DNA, RNP, or Sm antibody	
Other disease-specific autoantibody	
Cryoglobulinemia	
Radiologic	
Known radiologic findings suspicious for malignancy or infection that have not been sufficiently investigated	
Rapid radiologic progression	
Long bone abnormalities consistent with Erdheim-Chester disease	
Splenomegaly	
Pathologic	
Cellular infiltrates suggesting malignancy that have not been sufficiently evaluated	
Markers consistent with inflammatory myofibroblastic tumor	
Prominent neutrophilic inflammation	
Necrotizing vasculitis	
Prominent necrosis	
Primarily granulomatous inflammation	
Pathologic features of macrophage/histiocytic disorder	
Known diagnosis of the following:	
Multicentric Castleman's disease	
Crohn's disease or ulcerative colitis (if only pancreatobiliary disease is present)	
Hashimoto thyroiditis (if only the thyroid is affected)	
If case meets entry criteria and does not meet any exclusion criteria, proceed to step 3.	
Step 3. Inclusion criteria: domains and items¶	
Histopathology	
Uninformative biopsy	0
Dense lymphocytic infiltrate	+4
Dense lymphocytic infiltrate and obliterative phlebitis	+6
Dense lymphocytic infiltrate and storiform fibrosis with or without obliterative phlebitis	+13
Immunostaining#	0–16, as follows: Assigned weight is 0 if the IgG4+:IgG+ ratio is 0–40% or indeterminate and the number of IgG4+ cells/hpf is 0–9.** Assigned weight is 7 if 1) the IgG4+:IgG+ ratio is ≥41% and the number of IgG4+ cells/hpf is 0–9 or indeterminate; or 2) the IgG4+:IgG+ ratio is 0–40% or indeterminate and the number of IgG4+ cells/hpf is ≥10 or indeterminate. Assigned weight is 14 if 1) the IgG4+:IgG+ ratio is 41–70% and the number of IgG4+ cells/hpf is ≥10; or 2) the IgG4+:IgG+ ratio is ≥71% and the number of IgG4+ cells/hpf is 10–50. Assigned weight is 16 if the IgG4+:IgG+ ratio is ≥71% and the number of IgG4+ cells/hpf is ≥51.

(Continued)

Table 4. *(Cont'd)*

Step	Categorical assessment or numeric weight
Serum IgG4 concentration	
Normal or not checked	0
> Normal but <2× upper limit of normal	+4
2–5× upper limit of normal	+6
>5× upper limit of normal	+11
Bilateral lacrimal, parotid, sublingual, and submandibular glands	
No set of glands involved	0
One set of glands involved	+6
Two or more sets of glands involved	+14
Chest	
Not checked or neither of the items listed is present	0
Peribronchovascular and septal thickening	+4
Paravertebral band-like soft tissue in the thorax	+10
Pancreas and biliary tree	
Not checked or none of the items listed is present	0
Diffuse pancreas enlargement (loss of lobulations)	+8
Diffuse pancreas enlargement and capsule-like rim with decreased enhancement	+11
Pancreas (either of above) and biliary tree involvement	+19
Kidney	
Not checked or none of the items listed is present	0
Hypocomplementemia	+6
Renal pelvis thickening/soft tissue	+8
Bilateral renal cortex low-density areas	+10
Retroperitoneum	
Not checked or neither of the items listed is present	0
Diffuse thickening of the abdominal aortic wall	+4
Circumferential or anterolateral soft tissue around the infrarenal aorta or iliac arteries	+8
Step 4: Total inclusion points	
A case meets the classification criteria for IgG4-RD if the entry criteria are met, no exclusion criteria are present, and the total points is ≥20.	

* Refers to enlargement or tumor-like mass in an affected organ except in 1) the bile ducts, where narrowing tends to occur, 2) the aorta, where wall thickening or aneurysmal dilatation is typical, and 3) the lungs, where thickening of the bronchovascular bundles is common.
† If entry criteria are not fulfilled, the patient cannot be further considered for classification as having IgG4-related disease (IgG4-RD).
‡ Assessment for the presence of exclusion criteria should be individualized depending on a patient's clinical scenario.
§ If exclusion criteria are met, the patient cannot be further considered for classification as having IgG4-RD.
¶ Only the highest-weighted item in each domain is scored.
# Biopsies from lymph nodes, mucosal surfaces of the gastrointestinal tract, and skin are not acceptable for use in weighting the immunostaining domain.
** "Indeterminate" refers to a situation in which the pathologist is unable to clearly quantify the number of positively staining cells within an infiltrate yet can still ascertain that the number of cells is at least 10/high-power field (hpf). For a number of reasons, most often pertaining to the quality of the immunostain, pathologists are sometimes unable to count the number of IgG4+ plasma cells with precision yet even so, can be confident in grouping cases into the appropriate immunostaining result category.

to 89.2%, while the sensitivity increased from 85.5% to 90.0%. As is typical of clinical practice, serum IgG4 concentrations were not measured, or biopsies not performed, in some cases of IgG4-RD (3% and 15%, respectively) and mimickers (36% and 16%, respectively). When exclusion and inclusion criteria related to biopsy results or serum IgG4 concentrations were removed from the classification algorithm, the classification criteria maintained excellent specificity in both scenarios (98.9% when biopsy criteria were removed, 99.3% when serum IgG4 concentrations were removed). The sensitivity decreased substantially in the absence of pathology data or serum IgG4 concentrations, to 48.6% and 75.0%, respectively. When we assumed the worst-case scenario in which all the mimickers without biopsy or serum IgG4 concentration data were assigned the highest weights for each (e.g., IgG4 concentrations >5 times the upper limit of normal), the specificity of the classification criteria remained high (92.7%).

Reasons for cases not achieving a classification of IgG4-RD. Of the 428 and 267 IgG4-RD cases from the first and second validation cohorts used to test the classification criteria, 62 (14%) and 48 (18%), respectively, did not fulfill the classification criteria. In both the first and second validation cohorts, the majority of these false-negative cases (43 [69%] and 39 [81%], respectively) did not achieve sufficient inclusion criteria points (Table 6), partly because they were less likely to have had biopsies compared to true-positive cases (65% versus 91% [$P < 0.001$] and 73% versus 88% [$P = 0.007$], respectively). Twenty false-negative cases in the first validation cohort

Table 5. Performance of various thresholds of the 2019 American College of Rheumatology/European League Against Rheumatism classification criteria for IgG4-related disease using validation cohort 1*

Threshold	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)	Youden index	Distance to (0,1)	Specificity − sensitivity	Diagnostic odds ratio
14	0.89 (0.86–0.92)	0.95 (0.91–0.97)	0.92 (0.90–0.94)	0.84	0.12	0.06	142.4
15	0.89 (0.85–0.91)	0.97 (0.95–0.99)	0.93 (0.91–0.95)	0.86	0.12	0.09	286.1
16	0.89 (0.85–0.91)	0.98 (0.96–0.99)	0.93 (0.92–0.95)	0.87	0.12	0.10	394.5
17	0.88 (0.85–0.91)	0.98 (0.96–0.99)	0.93 (0.92–0.95)	0.86	0.12	0.10	385.6
18	0.88 (0.84–0.90)	0.98 (0.96–0.99)	0.93 (0.91–0.95)	0.86	0.13	0.11	360.8
19	0.86 (0.83–0.89)	0.99 (0.92–0.99)	0.93 (0.91–0.94)	0.85	0.14	0.12	408.3
20	0.86 (0.82–0.89)	0.99 (0.97–100.0)	0.92 (0.91–0.94)	0.85	0.15	0.14	761.5
21	0.83 (0.79–0.86)	0.99 (0.97–0.99)	0.91 (0.89–0.93)	0.82	0.18	0.17	607.2
22	0.82 (0.78–0.85)	0.99 (0.97–0.99)	0.91 (0.89–0.92)	0.81	0.18	0.18	578.8

* 95% CI = 95% confidence interval; AUC = area under the curve.

(32%) and 9 in the second validation cohort (19%) met at least 1 exclusion criterion. Of all of the IgG4-RD cases submitted in the first and second validation cohorts, 24 (4.9%) and 42 (8.7%), respectively, did not meet the initial entry criterion (characteristic organ involvement). In addition, 23 (5%) and 13 (4%) of the submitted IgG4-RD cases in the first and second validation cohorts, respectively, fulfilled at least 1 exclusion criterion, most often a clinical or serologic exclusion criterion (Table 7).

In the first validation cohort, 64 (20%) of 324 mimickers considered when deriving thresholds for the classification criteria did not meet entry criteria. Similarly, in the second validation cohort, 17 (10%) of the 164 mimickers did not meet entry criteria. Of those who met entry criteria in each validation cohort (260 and 147, respectively), 258 (99%) and 144 (98%), respectively, did not fulfill the classification criteria (true-negatives). The majority of mimickers in both cohorts (201 [77%] and 93 [65%], respectively) were eliminated at the exclusion criteria stage (Table 7). Supplementary Tables 3 and 4 (on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.41120/abstract) list the inclusion criteria fulfilled by the cases classified as IgG4-RD and cases submitted as mimickers in the first and second validation cohorts.

Discussion

The 2019 ACR/EULAR IgG4-RD criteria represent a significant milestone in IgG4-RD, a multiorgan condition with myriad

Table 6. Comparison of differences in false-negative and true-positive IgG4-related disease cases from the validation cohorts*

	Validation cohort 1			Validation cohort 2		
	False-negatives (n = 62)	True-positives (n = 366)	*P*	False-negatives (n = 48)	True-positives (n = 219)	*P*
Male sex	38 (61)	244 (67)	0.4	29 (60)	150 (69)	0.3
Age at diagnosis, mean ± SD years	57.5 ± 14.9	60.5 ± 13.4	0.1	60.4 ± 15.9	58.8 ± 14.8	0.5
Age at symptom onset, mean ± SD years	55.6 ± 15.0	58.6 ± 14.0	0.1	57.9 ± 16.2	56.7 ± 15.3	0.6
No. of organs involved, median (interquartile range)	2 (1–3)	3 (2–4)	0.002	2 (1–3)	2 (2–4)	0.01
Biopsy performed	40 (65)	332 (91)	<0.001	35 (73)	193 (88)	0.007
Reason criteria not met						
Exclusion criteria present	20 (32)	–		9 (19)	–	
Clinical	7 (11)	–		4 (8)	–	
Serologic	7 (11)	–		3 (6)	–	
Radiologic	5 (8)	–		2 (4)	–	
Pathologic	2 (3)	–		0 (0)	–	
Inclusion criteria score <20	43 (69)	–		39 (81)	–	
Total points toward inclusion criteria, mean ± SD	22.9 ± 17.1	38.9 ± 12.2	<0.001	18.6 ± 11.9	37.9 ± 12.7	<0.001

* "Gold standard" cases and mimickers were used in this analysis. Except where indicated otherwise, values are the number (%).

Table 7. Percentage of validation cohort cases and mimickers fulfilling exclusion criteria*

	Validation cohort 1		Validation cohort 2	
Exclusion criteria met†	IgG4-RD	Mimicker	IgG4-RD	Mimicker
Clinical exclusion criteria	7 (2)	81 (31)	5 (2)	25 (17)
Fever	1 (<1)	44 (17)	4 (1)	15 (10)
No response to glucocorticoids	1 (<1)	23 (9)	0 (0)	9 (6)
Leukopenia and thrombocytopenia	1 (<1)	19 (7)	0 (0)	2 (1)
Peripheral eosinophilia (>3,000 mm^3)	4 (1)	9 (4)	1 (<1)	4 (3)
Serologic exclusion criteria	7 (2)	108 (42)	5 (2)	32 (22)
Positive PR3- or MPO-ANCA	2 (1)	48 (19)	1 (<1)	26 (18)
Positive anti-Ro or anti-La	5 (1)	51 (20)	2 (1)	6 (4)
Positive extractable nuclear antigen (e.g., anti-Sm antibody)	0 (0)	6 (2)	1 (<1)	2 (1)
Other specific antibody positive	0 (0)	0 (0)	0 (0)	0 (0)
Cryoglobulins	0 (0)	10 (4)	1 (<1)	1 (1)
Radiologic exclusion criteria	5 (1)	24 (9)	2 (1)	20 (14)
Rapid radiographic progression	0 (0)	5 (2)	0 (0)	3 (2)
Long bone abnormalities (e.g., Erdheim-Chester disease)	0 (0)	3 (1)	0 (0)	1 (1)
Splenomegaly	3 (1)	14 (5)	0 (0)	3 (2)
Infectious/malignancy radiographic concern	2 (1)	4 (2)	2 (1)	13 (9)
Pathologic exclusion criteria	2 (1)	110 (42)	2 (1)	66 (45)
Malignant infiltrate on biopsy	1 (<1)	26 (10)	0 (0)	30 (20)
Inflammatory pseudotumor pathology	0 (0)	2 (1)	0 (0)	1 (1)
Prominent neutrophilic infiltrate	0 (0)	6 (2)	1 (<1)	9 (6)
Necrotizing vasculitis	0 (0)	36 (14)	0 (0)	11 (8)
Prominent necrosis	0 (0)	2 (1)	0 (0)	7 (5)
Primarily granulomatous inflammation	0 (0)	39 (15)	0 (0)	21 (14)
Prominent histiocytic infiltrate	1 (<1)	7 (3)	0 (0)	7 (5)
Multicentric Castleman's pathology	0 (0)	6 (2)	1 (<1)	2 (1)

* Includes all cases and mimickers fulfilling entry criteria. Values are the number (%). IgG4-RD = IgG4-related disease; PR3 = proteinase 3; MPO = myeloperoxidase; ANCA = antineutrophil cytoplasmic antibody.
† Total will sum to >100% because cases and mimickers could meet >1 exclusion criterion.

clinical presentations (3,4). Our approach reflects the fact that in clinical practice, information from clinical, serologic, radiologic, and pathologic evaluations must be integrated to arrive at a confident decision about whether to classify a patient as having IgG4-RD. The excellent sensitivity and specificity of these criteria will assist in the conduct of clinical trials and other studies of IgG4-RD. The purpose of these classification criteria is to facilitate the identification of more homogeneous groups of subjects for inclusion into clinical trials and observational studies (26–28).

No set of classification criteria can be constructed so as to include all patients within the spectrum of a disease. Accordingly, attempts to include all conceivable patients with clinical diagnoses of IgG4-RD would inevitably involve major sacrifices in specificity that would lead to the unacceptable inclusion of a significant percentage of false-positive cases. Our principal goal in constructing these classification criteria was to create a criteria set with the highest possible specificity while retaining moderately high sensitivity. The specificity of 97.8% achieved at a threshold of ≥20 points will include few false-positive cases: a highly desirable performance measure for clinical trials and other investigations. The sensitivity of 82.0% at this threshold also captures a broad spectrum of the patient population about whose IgG4-RD classification investigators are confident. The classification criteria for IgG4-RD that we have developed demonstrate robust test characteristics across a range of thresholds, suggesting that they will have broad relevance to the field of IgG4-RD investigation.

These criteria are not intended for use in clinical practice as the basis of establishing the diagnosis of IgG4-RD (29). If the appropriate clinical diagnosis for a patient is IgG4-RD, then failure to fulfill the ACR/EULAR classification criteria should not prevent the management of that patient's condition accordingly. There might be a substantial likelihood of this when, for example, a representative biopsy sample is difficult to obtain (30). These criteria provide a useful framework for clinicians considering the diagnosis of IgG4-RD in a patient. They highlight findings such as bilateral salivary gland enlargement, common features of IgG4-related kidney disease, and typical pancreas abnormalities that increase the likelihood that a patient has IgG4-RD. They also describe findings that suggest alternative diagnoses are more likely, such as primary granulomatous inflammation, ANCA positivity, and fevers. However, the exclusion criteria should not be interpreted as a list of studies or tests a clinician must obtain on every patient.

An important strength of this criteria set is that a patient may be classified accurately as having IgG4-RD in many cases even in the absence of a biopsy. Although biopsies

are essential in many settings to establish the diagnosis of IgG4-RD and exclude mimickers, we aimed to develop criteria in which biopsy is not required when the diagnosis of IgG4-RD is straightforward on the basis of clinical, serologic, and radiologic findings. Such criteria are consistent with clinical practice (7,31), compatible with research, and essential to the appropriate diagnosis of patients in both clinical and research settings. The fact that the 2019 ACR/EULAR IgG4-RD classification criteria require neither a biopsy nor an elevated serum IgG4 level reflects important changes in the approaches whereby classifications of this disease are now assigned (and clinical diagnoses rendered). Nearly 20% of cases classified as IgG4-RD had a normal serum IgG4 concentration or did not have a serum IgG4 value available. Moreover, 9% of the IgG4-RD cases did not have a biopsy, 37% lacked the classic histopathologic findings, and >40% did not meet previously defined cutoffs for IgG4+ plasma cell infiltrates (9). These criteria reflect the reality of clinical care and clinical investigation in IgG4-RD; clinicians consider a combination of factors when determining whether to classify a patient as having this disease (10).

The 2019 IgG4-RD classification criteria are one of the first sets of classification criteria in rheumatology to include absolute exclusion criteria that are not based solely on having an alternative diagnosis, but rather focus on clinical, serologic, radiologic, and pathologic features. This approach has strong appeal, particularly when the common mimickers of IgG4-RD themselves pose challenges in classification because of their multiorgan nature. Our sensitivity analysis indicated that in the absence of exclusion criteria, the specificity of the classification criteria decreased by nearly 10%, yet was accompanied by only a small improvement in sensitivity.

Some patients with clinical diagnoses of IgG4-RD will not fulfill these classification criteria. There are several explanations for this. First, we excluded patients with disease that affected only organs or sites that are involved only infrequently in IgG4-RD (e.g., patients with pituitary, breast, skin, or prostate disease). We focused our classification criteria development efforts on patients with more typical and common manifestations because of the desire to enroll relatively homogeneous populations in clinical trials. Second, some patients were excluded because their clinical evaluations identified exclusion criteria. Again, for the purposes of clinical trials, the exclusion of exceptional cases is usually prudent. Third, some patients met the entry criteria and did not meet exclusion criteria but still failed to accrue sufficient inclusion points to be classified as having IgG4-RD. Patients considered with confidence by their investigators to have IgG4-RD who did not fulfill the classification criteria were significantly less likely to have had a biopsy. It is possible that in some of these cases, a biopsy showing typical features of IgG4-RD might be useful for achieving sufficient points for the patient to be classified as having IgG4-RD.

Our study has a number of strengths. First, a cohort of nearly 1,900 patients with either IgG4-RD or a mimicking condition was assembled by an international group of investigators. Second, the experts involved in the consensus exercises, decision analysis, and cohort development represented investigators from a variety of specialties (e.g., rheumatology, gastroenterology, pathology, and radiology) and from around the world, including the Americas, Europe, Asia, and Australia. Moreover, many investigators involved in cohort development were not involved in other aspects of the classification criteria development, minimizing any influence of circularity of reasoning. Such a bias can occur when the same investigators who define criteria also develop derivation and validation cohorts (22). Our design prevented this potential bias. Third, we applied multicriteria decision analysis to derive the weights for each inclusion criteria item. These weights can be adjusted easily if or when other tests or information relevant to diagnosis become available.

Despite these strengths, our study has certain limitations. First, although the derivation and validation sets included a wide range of IgG4-RD mimickers, the performance of these classification criteria might be further evaluated in specific populations enriched for malignant conditions, non–IgG4-RD pancreatobiliary diseases, and infections. Because of the specific exclusion criteria intended to address these groups of mimickers, however, the 2019 ACR/EULAR criteria should perform well under such circumstances. Second, the laboratory, imaging, and pathology findings were not assessed centrally. Although the sensitivity and specificity of certain results may consequently have varied between investigator sites, this is unlikely to have affected our results significantly because of the expertise of the research group overall.

In summary, these are the first classification criteria for IgG4-RD, developed and tested using a data-driven approach and multicriterion decision analysis. The criteria perform well over a wide range of thresholds. They represent a significant advance in this rapidly evolving field and should be used in future clinical trials and epidemiologic studies of IgG4-RD.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. J. H. Stone had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Wallace, Naden, Chari, Choi, Della-Torre, Dicaire, Hart, Inoue, Kawano, Khosroshahi, Kubota, Lanzillotta, Okazaki, Perugino, Sharma, Saeki, Sekiguchi, Schleinitz, J. R. Stone, Takahashi, Umehara, Webster, Zen, J. H. Stone.

Acquisition of data. Wallace, Naden, Chari, Choi, Della-Torre, Dicaire, Hart, Inoue, Kawano, Khosroshahi, Kubota, Lanzillotta, Okazaki, Perugino, Sharma, Saeki, Sekiguchi, Schleinitz, J. R. Stone, Takahashi, Umehara, Webster, Zen, J. H. Stone.

Analysis and interpretation of data. Wallace, Naden, Chari, Choi, Della-Torre, Dicaire, Hart, Inoue, Kawano, Khosroshahi, Kubota, Lanzillotta, Okazaki, Perugino, Sharma, Saeki, Sekiguchi, Schleinitz, J. R. Stone, Takahashi, Umehara, Webster, Zen, J. H. Stone.

REFERENCES

1. Stone JH, Zen Y, Deshpande V. IgG4-related disease [review]. N Engl J Med 2012;366:539–51.
2. Kamisawa T, Zen Y, Pillai S, Stone JH. IgG4-related disease [review]. Lancet 2015;385:1460–71.
3. Kamisawa T, Funata N, Hayashi Y, Eishi Y, Koike M, Tsuruta K, et al. A new clinicopathological entity of IgG4-related autoimmune disease. J Gastroenterol 2003;38:982–4.
4. Hamano H, Kawa S, Horiuchi A, Unno H, Furuya N, Akamatsu T, et al. High serum IgG4 concentrations in patients with sclerosing pancreatitis. N Engl J Med 2001;344:732–8.
5. Stone JH, Khosroshahi A, Deshpande V, Chan JK, Heathcote JG, Aalberse R, et al. Recommendations for the nomenclature of IgG4-related disease and its individual organ system manifestations. Arthritis Rheum 2012;64:3061–7.
6. Wallace ZS, Deshpande V, Mattoo H, Mahajan VS, Kulikova M, Pillai S, et al. IgG4-related disease: clinical and laboratory features in one hundred twenty-five patients. Arthritis Rheumatol 2015;67:2466–75.
7. Sekiguchi H, Horie R, Kanai M, Suzuki R, Yi ES, Ryu JH. IgG4-related disease: retrospective analysis of one hundred sixty-six patients. Arthritis Rheumatol 2016;68:2290–9.
8. Kuroda N, Nao T, Fukuhara H, Karashima T, Inoue K, Taniguchi Y, et al. IgG4-related renal disease: clinical and pathological characteristics. Int J Clin Exp Pathol 2014;7:6379–85.
9. Deshpande V, Zen Y, Chan JK, Yi EE, Sato Y, Yoshino T, et al. Consensus statement on the pathology of IgG4-related disease. Mod Pathol 2012;25:1181–92.
10. Brito-Zerón P, Bosch X, Ramos-Casals M, Stone JH. IgG4-related disease: advances in the diagnosis and treatment. Best Pract Res Clin Rheumatol 2016;30:261–78.
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;74:14–8.
12. Hao M, Liu M, Fan G, Yang X, Li J. Diagnostic value of serum IgG4 for IgG4-related disease: a PRISMA-compliant systematic review and meta-analysis. Medicine (Baltimore) 2016;95:e3785.
13. Mahajan VS, Mattoo H, Deshpande V, Pillai SS, Stone JH. IgG4-related disease. Annu Rev Pathol 2014;9:315–47.
14. Chari ST. Diagnosis of autoimmune pancreatitis using its five cardinal features: introducing the Mayo Clinic's HISORt criteria. J Gastroenterol 2007;42 Suppl 18:39–41.
15. Perugino CA, Wallace ZS, Meyersohn N, Oliveira G, Stone JR, Stone JH. Large vessel involvement by IgG4-related disease. Medicine (Baltimore) 2016;95:e3344.
16. Fransen J, Johnson SR, van den Hoogen F, Baron M, Allanore Y, Carreira PE, et al. Items for revised classification criteria in systemic sclerosis: results of a consensus exercise. Arthritis Care Res (Hoboken) 2012;64:351–7.
17. Johnson SR, Khanna D, Daikh D, Cervera R, Costedoat-Chalumeau N, Gladman DD, et al. Use of consensus methodology to determine candidate items for systemic lupus erythematosus classification criteria. J Rheumatol 2019;46:721–6.
18. Johnson SR, Naden RP, Fransen J, van den Hoogen F, Pope JE, Baron M, et al. Multicriteria decision analysis methods with 1000Minds for developing systemic sclerosis classification criteria. J Clin Epidemiol 2014;67:706–14.
19. Tedeschi SK, Johnson SR, Boumpas DT, Daikh D, Dörner T, Diamond B, et al. Multicriteria decision analysis process to develop new classification criteria for systemic lupus erythematosus. Ann Rheum Dis 2019;78:634–40.
20. Neogi T, Jansen TL, Dalbeth N, Fransen J, Schumacher HR, Berendsen D, et al. 2015 gout classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheumatol 2015;67:2557–68.
21. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO III, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.
22. Van den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2013;65:2737–47.
23. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.
24. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5.
25. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56:1129–35.
26. Classification and Response Criteria Subcommittee of the American College of Rheumatology Committee on Quality Measures. Development of classification and response criteria for rheumatic diseases [editorial]. Arthritis Rheum 2006;55:348–52.
27. Johnson SR, Goek ON, Singh-Grewal D, Vlad SC, Feldman BM, Felson DT, et al. Classification criteria in rheumatic diseases: a review of methodologic properties. Arthritis Rheum 2007;57:1119–33.
28. Dougados M, Gossec L. Classification criteria for rheumatic diseases: why and how? [editorial]. Arthritis Rheum 2007;57:1112–5.
29. Aggarwal R, Ringold S, Khanna D, Neogi T, Johnson SR, Miller A, et al. Distinctions between diagnostic and classification criteria? [review]. Arthritis Care Res (Hoboken) 2015;67:891–7.
30. Felson DT, Anderson JJ. Methodological and statistical approaches to criteria development in rheumatic diseases. Baillieres Clin Rheumatol 1995;9:253–66.
31. Chari ST, Takahashi N, Levy MJ, Smyrk TC, Clain JE, Pearson RK, et al. A diagnostic strategy to distinguish autoimmune pancreatitis from pancreatic cancer. Clin Gastroenterol Hepatol 2009;7:1097–103.

APPENDIX A: THE AMERICAN COLLEGE OF RHEUMATOLOGY/EUROPEAN LEAGUE AGAINST RHEUMATISM IgG4-RELATED DISEASE CLASSIFICATION CRITERIA WORKING GROUP

Members of the ACR/EULAR IgG4-RD Classification Criteria Working Group are as follows: Drs. Takashi Akamizu, Mitsuhiro Akiyama, Lillian Barra, Adrian Bateman, Daniel Blockmans, Pilar Brito-Zeron, Corrado Campochiaro, Mollie Carruthers, Suresh Chari, Tsutomu Chiba, Hyon Choi, Lynn Cornell, Emma Culver, Saman Darabian, Emanuel Della-Torre, Vikram Deshpande, Mr. Jean-Francois Dicaire, Drs. Lingli Dong, Mikael Ebbo, Andreu Fernández-Codina, Judith A. Ferry, George Fragkoulis, Fabian Frost, Luca Frulloni, Phil A. Hart, Gabriela Hernandez-Molina, Dai Inoue, Haihan Ji, Karuna Keat, Terumi Kamisawa, Shigeyuki Kawa, Mitsuhiro Kawano, Arezou Khosroshahi, Hiroshi Kobayashi, Yuzo Kodama, Satoshi Kubo, Kensuke Kubota, Marco Lanzillotta, Haiyang Leng, Markus Lerch, Yanying Liu, Zhifu Liu, Matthias Löhr, Eduardo Martin-Nares, Ferran Martinez-Valle, Chiara Marvisi, Yasufumi Masaki, Shoko Matsui, Ichiro Mizushima, Ray P. Naden, Seiji Nakamura, Jan Nordeide, Kenji Notohara, Kazuichi Okazaki, Sergio Paira, Cory A. Perugino, Jovan Popovic, Manel Ramos-Casals, James Rosenbaum, Jay Ryu, Takako Saeki, Yasuharu Sato, Nicolas Schleinitz, Hiroshi Sekiguchi, Amita Sharma, Evgeniya V. Sokol, James R. Stone, John H. Stone, Wenwu Sun, Hiroki Takahashi, Naoki Takahashi, Masayuki Takahira, Yoshiya Tanaka, Hisanori Umehara, Augusto Vaglio, Alejandra Villamil, Yoko Wada, Zachary S. Wallace, George Webster, Kazunori Yamada, Motohisa Yamamoto, Joanne Yi, Yinlan Yi, Giuseppe Zamboni, Yoh Zen, Wen Zhang.