ACR Open Rheumatology
Vol. 4, No. 4, April 2022, pp 371–377
DOI 10.1002/acr2.11405

American College of Rheumatology
Empowering Rheumatology Professionals

Derivation and Validation of Algorithms to Identify Patients With Immunoglobulin-G4-Related Disease Using Administrative Claims Data

Zachary S. Wallace,[1] Xiaoqing Fu,[2] Claire Cook,[2] Cory A. Perugino,[1] Yuqing Zhang,[1] John H. Stone,[1] and Hyon K. Choi[1]

Objective. Immunoglobulin-G4-related disease (IgG4-RD) is a systemic autoimmune disease that can affect nearly any organ, but its epidemiology remains poorly understood. Validated algorithms to identify cases in claims data will enable studies to describe IgG4-RD epidemiology in the general population.

Methods. Potential claims-based algorithms were developed by IgG4-RD experts using a combination of *International Classification of Diseases, Ninth Revision* (ICD-9) and *International Classification of Diseases, 10th Revision* (ICD-10) codes, dispensed medications, and procedure codes for immunoglobulin G (IgG) subclass testing. Algorithms were tested using Medicare Parts A, B, and D linked to medical records (2007-2017). Classification of cases as IgG4-RD was determined using the American College of Rheumatology (ACR) and European League Against Rheumatism (EULAR) classification criteria for IgG4-RD. We estimated the positive predictive value (PPV) of each algorithm; sensitivity was determined using a cohort of patients with IgG4-RD also enrolled in Medicare Parts A, B, and D during the study period.

Results. We identified seven algorithms that used a combination of ICD-9 and ICD-10 codes, medication prescriptions, and/or IgG subclass tests to identify patients with IgG4-RD. The PPV of algorithms in the derivation cohort ranged from 57% to 100%, and sensitivity ranged from 0% to 58%. The best performing algorithm in the validation cohort had a PPV of 81% and a sensitivity of 58%. Typical IgG4-RD manifestations were observed in the cohort (n = 36) assembled by this algorithm, including 50% with sialadenitis, 64% with pancreatic disease, 31% with renal disease, and 59% with an elevated IgG4 concentration.

Conclusion. We derived and validated a well-performing algorithm to identify IgG4-RD cases with typical manifestations of the disease. The claims-based algorithm can be used in research studies of IgG4-RD.

INTRODUCTION

Immunoglobulin-G4-related disease (IgG4-RD) is a systemic rheumatic disease that causes fibrosing inflammatory lesions in nearly any organ (1,2). Common manifestations include autoimmune pancreatitis, submandibular and/or parotid gland sialadenitis, dacryoadenitis, sclerosing cholangitis, and retroperitoneal fibrosis (3). Serum immunoglobulin G4 (IgG4) concentrations are elevated in the majority of patients with IgG4-RD (3) and are often tested serially as part of the diagnostic evaluation and management of patients (4). IgG4-RD is most often treated with glucocorticoids, but other common treatments include azathioprine, methotrexate, mycophenolate mofetil, and rituximab (2).

Our understanding of the epidemiology of IgG4-RD is largely derived from cohort studies describing patients assembled from referral centers, which may be biased toward more severe or difficult to treat cases. These studies suggest that IgG4-RD tends to affect men more than women, most often in their fifth to seventh decades of life (3). There are no population-based estimates of IgG4-RD incidence and prevalence, in part because there are no validated methods for identifying patients with IgG4-RD using administrative claims data. A prior study estimated the incidence

Dr. Wallace's work was supported by the NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases (K23-AR-073334), Bristol Myers Squibb, and Principia/Sanofi.

[1]Zachary S. Wallace, MD, MSc, Cory A. Perugino, DO, Yuqing Zhang, ScD, John H. Stone, MD, MPH, Hyon K. Choi, MD, DrPH: Massachusetts General Hospital, Boston; [2]Xiaoqing Fu, MS, Claire Cook, MPH: Harvard Medical School, Boston, Massachusetts.

Dr. Wallace reports consulting fees from Viela Bio and Medpace. No other disclosures relevant to this article were reported.

Author disclosures are available at https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Facr2.11405&file=acr211405-sup-0001-Disclosureform.pdf.

Address correspondence to Zachary S. Wallace, MD, MSc, Massachusetts General Hospital, Division of Rheumatology, Allergy, and Immunology, Clinical Epidemiology Program, 16th Floor, 100 Cambridge Street, Boston, MA 02114. Email: zswallace@mgh.harvard.edu.

Submitted for publication December 3, 2021; accepted December 7, 2021.

SIGNIFICANCE & INNOVATION

- Immunoglobulin-G4-related disease (IgG4-RD) is a recently recognized multiorgan fibroinflammatory disease for which population-based studies of its epidemiology, the comparative effectiveness of interventions, and other outcomes are needed.
- There is no specific *International Classification of Diseases, Ninth Revision* (ICD-9) or *International Classification of Diseases, 10th Revision* (ICD-10) code for IgG4-RD, limiting our ability to conduct studies using administrative claims and other data sources.
- Here, we derived and validated an algorithm to accurately identify cases of IgG4-RD in administrative claims data using a combination of ICD-9 and ICD-10 codes associated with IgG4-RD manifestations, dispensed medications, and testing for immunoglobulin G subclass concentrations.
- This algorithm will enable the conduct of subsequent studies using administrative claims data and represents a novel approach for identifying cases of a disease with no specific ICD-9 or ICD-10 codes.

of autoimmune pancreatitis, a common manifestation of IgG4-RD, to be 0.9 cases per 100,000 people in Japan by random sampling of hospitals (5). A major obstacle to using administrative claims data to study IgG4-RD is that there is no specific *International Classification of Diseases, Ninth Revision* (ICD-9) or *International Classification of Diseases, 10th Revision* (ICD-10) code for IgG4-RD, so providers often bill for encounters using nonspecific codes reflecting an autoimmune disease or a specific manifestation of IgG4-RD.

We used Medicare Parts A, B, and D data linked to medical records to develop algorithms for use in administrative claims databases to identify patients with IgG4-RD for epidemiologic and comparative effectiveness studies.

MATERIALS AND METHODS

Data source. We used longitudinal Medicare claims data from Parts A (inpatient coverage), B (outpatient coverage), and D (prescription benefits) for patients with Medicare Fee-for-Service from January 1, 2007, to December 31, 2017. These data were linked to the Mass General Brigham (MGB) Research Patient Data Registry (RPDR), which is a centralized data warehouse that contains unstructured and structured data for all patients seen in the MGB health care system. MGB includes two tertiary care hospitals (Massachusetts General Hospital [MGH] and Brigham and Women's Hospital) as well as community hospitals, primary care clinics, and subspecialty clinics in the greater Boston, MA, area (6). For algorithm development, we randomly divided the entire Medicare–RPDR linked cohort in half, with n = 71,393 used for algorithm development and n = 71,393 for validation testing. Random selection of cases for the derivation and validation sets was performed using PROC SURVEYSELECT in SAS (SAS Institute, Inc.).

To assess sensitivity of algorithms, we identified patients from the MGH Center for IgG4-RD registry who were enrolled in Medicare Fee-for-Service with Part A, B, and D coverage during the study period after the initial date of their IgG4-RD diagnosis. The MGH Center for IgG4-RD maintains a registry of all patients evaluated who have IgG4-RD confirmed to be definite, probable, or atypical according to the American College of Rheumatology (ACR) and European League Against Rheumatism (EULAR) classification criteria for IgG4-RD and previously described methods (7–9). This study was approved by the MGB Institutional Review Board.

Potential algorithm criteria. Three authors (ZSW, CAP, and JHS) with expertise in IgG4-RD developed potential algorithms to be used to identify patients with IgG4-RD in claims data. The development of these algorithms was informed, in part, by similar algorithms for identification of antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis and systemic lupus erythematosus cases in claims data because, like IgG4-RD, these are multiorgan rheumatic diseases with protean manifestations and specific diagnostic tests (10,11). Algorithms were developed using a combination of ICD-9-Clinical Modification and ICD-10-Clinical Modification codes (referred to as ICD-9 and ICD-10 codes throughout this study), dispensed medications, and procedure codes for IgG subclass testing.

ICD-9 and ICD-10 code inclusion criteria. There are no specific ICD-9 or ICD-10 codes for IgG4-RD or its manifestations. The authors therefore identified an initial list of 821 ICD-9 and ICD-10 codes using two approaches. First, we identified ICD-9 and ICD-10 codes with description keywords potentially related to IgG4-RD on the basis of the ACR/EULAR classification criteria: Orbit, Lacrima* (eg, lacrimal, lacrimation), Subman* (eg, submandibular, submandib, submand), Salivary gland, Parotid, Retropriton*, Pseudotumor, Pancrea* (eg, pancreatitis, pancreas, pancreatic), Thyroiditis, Nephritis, Cholangitis, Bile Duct, Biliary, Aortitis, Tubulointerstitial nephritis, Mesenteritis, Mediastinitis, and Meningitis. Second, we reviewed all ICD-9 and ICD-10 codes used to bill for clinical encounters of patients with IgG4-RD in the MGH Center for IgG4-RD registry, which is linked to the RPDR. From these codes, we identified codes not identified in the keyword search that might be relevant to a diagnosis of IgG4-RD. From this list of 821 codes, two authors (ZSW and JHS) excluded 622 ICD-9 and ICD-10 codes because they were specific to malignancy, infection, trauma, or other non-IgG4-related conditions. From the remaining 199 codes, two authors (ZSW and JHS) classified 54 ICD-9 and ICD-10 codes that were thought to have a high likelihood of identifying IgG4-RD on the basis of their expertise (Supplementary Table 1). Four ICD-9 and ICD-10 codes commonly used to bill for IgG4-RD clinical encounters were also identified (ICD-9: 279.49

and 279.9; ICD-10: D89.89 and D89.9). In total, there were 216 ICD-9 and ICD-10 codes of interest. Where possible, we categorized ICD-9 and ICD-10 codes as being associated with a common anatomic site of IgG4-RD, as identified by the entry criteria included in the ACR/EULAR classification criteria for IgG4-RD.

ICD-9 and ICD-10 code exclusion criteria. Because immunoglobulin G (IgG) subclass testing is commonly used to screen for immunodeficiency and because some ICD-9 and ICD-10 codes that may be used in encounters for IgG4-RD may also be used in encounters for immunodeficiency, we identified ICD-9 and ICD-10 codes associated with immunodeficiency for potential exclusion (Supplementary Table 2). Because some manifestations of pancreatitis or cholangitis are manifestations of inflammatory bowel disease and are treated with similar medications as those for IgG4-RD, we also identified ICD-9 and ICD-10 codes for ulcerative colitis and Crohn disease for potential exclusion.

IgG subclass testing criteria. The *Current Procedural Terminology* (CPT) code for IgG subclass testing is 82787. We required that patients needed to have at least two procedure codes for IgG subclass testing as one criterion in these algorithms. We required at least two procedure codes because our preliminary work demonstrated that requiring only one code was associated with a substantially lower positive predictive value (PPV) across algorithms. For instance, the PPV of each of the seven algorithms reported in this study was substantially lower when only one CPT code for IgG subclass testing was required, as opposed to at least two CPT codes. This decrease in PPV ranged from 16% to 46%. Sensitivity was not affected.

Medication criteria. Medications commonly used to treat IgG4-RD were identified by their national drug code (NDC) and included glucocorticoids (ie, prednisone, prednisolone, methylprednisolone), methotrexate, mycophenolate mofetil or mycophenolic acid, mercaptopurine, and azathioprine (Supplementary Table 3). Treatment with rituximab was identified by either NDC or CPT codes.

Algorithm development. In iterative testing using the derivation cohort, we determined that algorithms including only ICD-9 and ICD-10 codes had poor performance. For instance, requiring only one high-likelihood code identified 12,707 individuals, the vast majority of whom do not have IgG4-RD. Additionally, we found that requiring two CPT codes for IgG subclass testing enhanced the performance of algorithms, as discussed previously. Through this iterative process of trial and error, we found that the algorithms performed better after incorporating criteria for medication prescriptions and IgG subclass testing. On the basis of this work and the previous algorithms developed for other multiorgan rheumatic diseases, we identified seven algorithms for additional testing.

TABLE 1. Characteristics of the Medicare cohort and IgG4-RD cohort used to assess algorithm performance

	Medicare cohort	IgG4-RD cohort
n	142,786	30
Age, mean (SD)[a]	70 (11)	71 (11)
Female sex, n (%)	81,339 (57)	8 (27)
Race, n (%)		
White	123,369 (86)	26 (87)
Black	7735 (5)	0 (0)
Asian	2195 (2)	4 (13)
American Indian or Alaska Native	99 (<1)	0 (0)
Other or not reported	9388 (7)	0 (0)
Hispanic	3402 (2)	1 (3)
ACR/EULAR classification, n (%)		
Definite IgG4-RD	—	29 (97)
Probable IgG4-RD	—	0 (0)
Atypical IgG4-RD	—	1 (3)
IgG4-RD features, n (%)		
Orbit	—	3 (10)
Lacrimal gland	—	8 (27)
Salivary gland	—	18 (60)
Pancreas	—	15 (50)
Biliary	—	6 (20)
Kidney	—	11 (37)
RPF	—	6 (20)
Single organ	—	6 (20)
Multiorgan	—	24 (80)
Elevated IgG4 levels	—	25 (83)

Abbreviations: ACR, American College of Rheumatology; EULAR, European League Against Rheumatism; IgG4, immunoglobulin G4; IgG4-RD, immunoglobulin-G4-related disease; RPF, retroperitoneal fibrosis.
[a] For the Medicare cohort, age represents the age at the start of their Medicare data availability during the study period; for the IgG4-RD cohort, age represents the age at IgG4-RD diagnosis.

Establishing the IgG4-RD diagnosis in chart review. Two authors (ZSW and JHS) reviewed cases identified by each algorithm and determined whether they were IgG4-RD on the basis of the classification criteria approved by the EULAR and ACR (7). We included patients who fell into one of three categories: 1) definite IgG4-RD, 2) probable IgG4-RD, and 3) atypical IgG4-RD (8,9). Patients in the definite category fulfilled the published classification criteria. Patients who were considered probable fulfilled two parts of the ACR/EULAR classification criteria (ie, had clinical involvement of a typical organ and were rigorously evaluated to ensure that they did not meet any exclusion criteria) but did not reach the threshold of 20 inclusion points according to the criteria. Patients with probable IgG4-RD often had lesions (eg, retroperitoneal fibrosis) inaccessible to perform a safe biopsy or were seen or evaluated prior to the recognition of IgG4-RD and might have had incomplete data collected, but sufficient evidence was available to classify them as having IgG4-RD. Patients who were considered atypical met the previously established pathological and immunostaining criteria for diagnosing IgG4-RD but presented with involvement of an atypical organ (eg, breast, prostate) that was not considered in the ACR/EULAR classification criteria (7). Patients identified in Medicare claims as fulfilling an algorithm's criteria but lacking records in the RPDR from the same time period were excluded.

Statistical analysis. Descriptive statistics were used to describe the demographics and clinical characteristics of patients identified by each algorithm. The PPV for each algorithm in the derivation and validation cohorts was based on the proportion of cases classified as IgG4-RD from all of the cases identified by each algorithm. The sensitivity of each algorithm in the derivation and validation cohorts was based on the proportion of cases from the MGH Center for IgG4-RD linked to Medicare that were identified by each algorithm. The 95% confidence intervals of the PPV and sensitivity estimates were calculated using the normal approximation of the binomial distribution. SAS version 9.4 was used for all analyses.

TABLE 2. Algorithms chosen for testing in derivation and validation cohorts

Algorithm	ICD-9 and ICD-10 codes[a]	Medications[a] used for IgG4-RD dispensed ±365 days of earliest ICD-9 or ICD-10 code	At least two CPT codes for IgG subclass tests	No use of ICD-9 or ICD-10 codes associated with immunodeficiency or inflammatory bowel disease
1	≥1 high-likelihood ICD-9 or ICD-10 code from >1 entry criteria category	Yes	Yes	Yes
2	≥1 high-likelihood ICD-9 or ICD-10 code from >1 entry criteria category used twice ≥30 days apart	Yes	Yes	No
3	≥1 high-likelihood ICD-9 or ICD-10 code from >1 entry criteria category used twice ≥30 days apart	Yes	Yes	Yes
4	ICD-10 code D89.89 or D89.9 used twice ≥30 days apart	Yes	Yes	Yes
5	Algorithm 1 but includes ICD-10 codes commonly used to bill for IgG4-RD[b]	Yes	Yes	Yes
6	Algorithm 3 but includes ICD-10 codes commonly used to bill for IgG4-RD[b]	Yes	Yes	No
7	Algorithms 2 and 4 combined together	Yes	Yes	Algorithm 2 (no), algorithm 4 (yes)

Note: These algorithms require fulfillment of ICD-9 or ICD-10 code criteria, medication exposure, and CPT codes for IgG subclass testing; some also require that other ICD-9 or ICD-10 codes are not used (indicated by "Yes" in the last column).

Abbreviations: CPT, *Current Procedural Terminology*; ICD-9, *International Classification of Diseases, Ninth Revision*; ICD-10, *International Classification of Diseases, 10th Revision*; IgG, immunoglobulin G; IgG4-RD, immunoglobulin-G4-related disease.

[a] A complete list of ICD-9 and ICD-10 codes and medications used in each algorithm is available in the Supplementary Material.

[b] ICD-10 code refers to D89.89 or D89.9.

TABLE 3. Algorithm performance for identification of IgG4-RD cases in derivation and validation cohorts

	PPV				Sensitivity			
	Derivation cohort		Validation cohort		Derivation cohort (n = 12)		Validation cohort (n = 18)	
Algorithm	n[a]	PPV (95% CI)	n[a]	PPV (95% CI)	n[a]	Sensitivity (95% CI)	n[a]	Sensitivity (95% CI)
1	7	57% (18%-90%)	6	100% (54%-100%)[b]	0	0% (0%-26%)[c]	4	22% (6%-48%)
2	12	83% (52%-98%)	10	80% (44%-97%)	4	33% (9%-65%)	4	22% (6%-48%)
3	4	100% (40%-100%)[b]	3	100% (29%-100%)[b]	0	0% (0%-26%)[c]	2	11% (1%-35%)
4	12	83% (52%-98%)	16	81% (54%-96%)	3	25% (5%-57%)	8	44% (22%-69%)
5	9	67% (30%-93%)	14	71% (42%-92%)	0	0% (0%-26%)[c]	8	44% (22%-69%)
6	5	100% (48%-100%)[b]	10	90% (56%-100%)	0	0% (0%-26%)[c]	7	39% (17%-64%)
7	21	81% (58%-94%)	24	79% (58%-93%)	7	58% (28%-85%)	10	56% (31%-78%)

Abbreviations: CI, confidence interval; IgG4-RD, immunoglobulin-G4-related disease; PPV, positive predictive value.
[a] The total number of cases identified by each algorithm.
[b] Approach higher boundary, one-sided 97.5% CI.
[c] Approach lower boundary, one-sided 97.5% CI.

RESULTS

Study population. In the linked cohort of MGB patients with Medicare records (Part A, B, and D), there were a total of 142,786 individuals with any potential ICD-9 or ICD-10 code of interest. The mean age was 70 (SD 11), 61,447 (43%) were male, and 123,369 (86%) were White. This cohort was randomly divided in half, with one half (n = 71,393) reserved for use as the derivation cohort and the other half (n = 71,393) reserved for the validation cohort (Table 1). Of the 364 patients with IgG4-RD in the MGH Center for IgG4-RD registry, 30 (8%) were identified in the available Medicare data (12 in the derivation cohort and 18 in the validation cohort) and used to estimate the sensitivity of algorithms (Table 1).

TABLE 4. Characteristics of Patients with IgG4-RD identified by the high-performing algorithm

Characteristic	Distribution
N (%)	36 (100)
Age, mean (SD)	70.7 (6.7)
Male sex, n (%)	30 (83)
Race, n (%)	
White	32 (89)
Black	1 (3)
Asian	2 (6)
Hispanic	1 (3)
ACR/EULAR classification criteria, n (%)	
Definite	31 (86)
Probable	4 (11)
Atypical	1 (3)
IgG4-RD features	
Orbit	2 (6)
Lacrimal gland	6 (17)
Salivary gland	18 (50)
Pancreas	23 (64)
Biliary	11 (31)
Kidney	11 (31)
RPF	5 (14)
Single organ	12 (33)
Multiorgan	24 (67)
Elevated IgG4 concentration	20 (59)
Specialist establishing diagnosis	
Rheumatology	16 (44)
Gastroenterology	13 (36)
Other	6 (17)
Unknown	1 (3)

Abbreviations: ACR, American College of Rheumatology; EULAR, European League Against Rheumatism; IgG4, immunoglobulin G4; IgG4-RD, immunoglobulin-G4-related disease; RPF, retroperitoneal fibrosis.

Algorithms tested and performance. On the basis of preliminary work, we identified seven algorithms for additional testing in the derivation and validation cohorts (Table 2). The PPV of each algorithm in the derivation cohort ranged from 57% to 100% (Table 3). In the validation cohort, the PPV of each algorithm ranged from 71% to 100%. The sensitivity of each cohort in the derivation and validation cohorts ranged from 0% to 58% and from 11% to 56%, respectively. When we considered the PPV and specificity of each algorithm, as measured in the validation cohort, the best performing algorithm was algorithm 7.

Characteristics of patients with IgG4-RD identified by high-performing algorithm. The demographic and disease-specific characteristics of patients identified by algorithm 7 in the derivation and validation cohorts are included in Table 4. Overall, 45 patients with IgG4-RD were identified. The mean age was 71 (SD 7) years, and the majority of patients were male (30, 83%) and White (32, 89%). The median of organs affected was 3 (interquartile range 1-5), and 20 (59%) had an elevated IgG4 concentration. The most common locations of IgG4-RD manifestations in this cohort were the pancreas (23, 64%), salivary gland (18, 50%), and kidney (11, 31%). Of these patients, 31 (86%) fulfilled the ACR/EULAR classification criteria, and one (3%) and four (11%) were considered to have atypical or probable IgG4-RD, respectively. The majority of patients were diagnosed with IgG4-RD by a rheumatologist or gastroenterologist (29, 81%); the remainder were diagnosed by other providers, including surgeons and nephrologists. The cases falsely identified

as IgG4-RD by this algorithm included two cases of rheumatoid arthritis and a case each of systemic lupus erythematosus, recurrent infection, chronic pancreatitis, multiple myeloma, ANCA-associated vasculitis, idiopathic serositis, and idiopathic fibrotic disease resembling Erdheim–Chester disease (but a biopsy was not performed).

DISCUSSION

We derived and validated an algorithm to identify IgG4-RD cases in claims data with high PPV and good sensitivity. The best performing algorithm used a combination of ICD-9 and ICD-10 codes, dispensed medications for common IgG4-RD treatments, and IgG subclass testing procedure codes. This algorithm identified patients with diverse manifestations of IgG4-RD with characteristics similar to those in other described cohorts. This algorithm may be used in claims data as well as electronic health record (EHR) data to accurately identify cases of IgG4-RD in the general population to describe the epidemiology of this recently recognized disease.

This study represents a novel addition to the literature, expanding on prior work in IgG4-RD, which established the first classification criteria, common terminology for reporting IgG4-RD, the pathologic definition of the disease, and a consensus approach to treatment. Like these prior studies, the algorithm validated here will enable important future studies that will advance our understanding of IgG4-RD by defining its epidemiology and the comparative effectiveness of treatments in general population cohorts.

Like other systemic rheumatic diseases, such as systemic lupus erythematosus and ANCA-associated vasculitis, IgG4-RD is a heterogeneous condition that can affect nearly any organ or anatomic site. However, unlike systemic lupus erythematosus and ANCA-associated vasculitis (10,11), IgG4-RD does not have any specific ICD-9 or ICD-10 codes associated with the diagnosis. As such, we developed algorithms that would identify patients with a variety of manifestations using a variety of ICD-9 and ICD-10 codes. Reassuringly, the best performing algorithm in this study identified IgG4-RD cases with diverse IgG4-RD features, including hepatopancreatobiliary disease, salivary gland disease, renal disease, and orbital disease. Additionally, the algorithm identified patients with single-organ disease and multiorgan disease. Indeed, the distribution of single-organ (33% vs 24%) and multiorgan involvement (67% vs 76%), the proportion with an elevated IgG4 concentration (59% vs 79%), and the proportion with specific organ manifestations were similar in the cohort assembled by this algorithm as in the large international cohort (n = 493) assembled to derive the ACR/EULAR classification criteria for IgG4-RD (3). Patients identified by the high-performing algorithm identified in this study had been diagnosed with IgG4-RD by a variety of specialists, including both rheumatologists and gastroenterologists, as is typical in practice. Collectively, these findings suggest that this algorithm may be used to identify an IgG4-RD cohort representative of the usual spectrum of disease.

We used structured data in this study to develop rule-based algorithms for IgG4-RD case identification. We focused on structured data available in claims data, as opposed to unstructured data (eg, provider notes) available in EHRs, so that the algorithm could be used in administrative claims data, which are not typically linked to EHRs. However, because there is no specific ICD-9 or ICD-10 code for IgG4-RD, it may be that leveraging unstructured data with natural language processing to identify free-text references to the diagnosis of IgG4-RD improves the sensitivity of an algorithm in contexts in which EHR and claims data are linked. Additionally, although we evaluated many potential algorithms for case identification, we were limited in the number and type of criteria incorporated into each algorithm because these algorithms were developed manually. It may be that machine learning methods are useful for identifying algorithms with better sensitivity. Additional studies are needed to evaluate the potential role of these advanced bioinformatics methods to identify IgG4-RD cases in claims and EHR data.

Our study has important strengths. First, we leveraged a unique data linkage of Medicare claims with a large health care system's EHR to develop and validate these algorithms. Second, we applied the recent ACR/EULAR classification criteria for IgG4-RD to validate cases. Despite these strengths, our study has certain limitations. First, we validated these algorithms using Medicare claims data, which typically include patients at least 65 years old. Therefore, the performance of these algorithms in other age groups or claims databases is uncertain. However, we do not have reason to suspect that our algorithms, which require a combination of ICD-9 and ICD-10 codes, medication prescriptions, and IgG subclass testing, would perform differently in other age groups or databases. Once the *International Classification of Diseases, 11th Revision* is introduced, the PPV of a specific code for IgG4-RD (4A43.0) will require testing. It is important to note that the PPV is related to baseline prevalence, so there is potential for overestimation when testing in a Medicare population because IgG4-RD may be particularly common in this age group. We were unable to estimate specificity, but this may be further evaluated in future studies. Second, the sensitivity of these algorithms was moderate, suggesting that a portion of patients with IgG4-RD will not be captured by these algorithms. In particular, patients not treated with medications (eg, mild salivary gland, resected lesions) or those in whom normal serum IgG4 subclass testing at baseline led providers to not repeat testing are unlikely to be identified by these algorithms. However, the majority of patients with IgG4-RD are treated at some point in their disease course, and our best performing algorithm still identified a large portion of patients with normal serum IgG4 concentrations. Third, these algorithms were derived and tested within a health care system that included a center for IgG4-RD, which may limit the

generalizability of our findings. However, the cases identified by these algorithms included many not managed in our center, which suggests that these algorithms will be generalizable beyond specialized centers. Indeed, the diagnosis of IgG4-RD was established by rheumatologists, gastroenterologists, and other clinicians in the cohort assembled by the best performing algorithm.

In conclusion, the high-performing algorithm identified in this study to identify IgG4-RD cases can be leveraged in future studies to characterize the epidemiology of IgG4-RD and enable comparative effectiveness analyses. This algorithm will enable investigators to leverage administrative claims data to study this recently recognized disease in the general population.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual contact, and all authors approved the final version to be published. Dr. Wallace 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, Fu, Cook, Perugino, Zhang, Stone, Choi.

Acquisition of data. Wallace, Fu, Cook, Choi.

Analysis and interpretation of data. Wallace, Fu, Zhang.

ROLE OF THE STUDY SPONSOR

Principia/Sanofi had no role in the study design or in the collection, analysis, or interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication. Publication of this article was not contingent upon approval by Principia/Sanofi.

REFERENCES

1. Stone JH, Zen Y, Deshpande V. IgG4-related disease. N Engl J Med 2012;366:539-51.
2. Zhang W, Stone JH. Management of IgG4-related disease. Lancet Rheumatol 2019;1:e55-65.
3. Wallace ZS, Zhang Y, Perugino CA, Naden RL, Choi HK, Stone JH, et al. Clinical phenotypes of IgG4-related disease: an analysis of two international cross-sectional cohorts. Ann Rheum Dis 2019;78: 406-12.
4. Khosroshahi A, Wallace ZS, Crowe JL, Akamizu T, Azumi A, Carruthers MN, et al. International consensus guidance statement on the management and treatment of IgG4-related disease. Arthritis Rheumatol 2015;67:1688-99.
5. Kanno A, Nishimori I, Masamune A, Kikuta K, Hirota M, Kuriyama S, et al. Nationwide epidemiological survey of autoimmune pancreatitis in Japan. Pancreas 2012;41:835-9.
6. Nalichowski R, Keough D, Chueh HC, Murphy SN. Calculating the benefits of a research patient data repository. AMIA Annu Symp Proc 2006;2006:1044.
7. Wallace ZS, Naden RP, Chari S, Choi H, Della-Torre E, Dicaire JF, et al. The 2019 American College of Rheumatology/European League Against Rheumatism classification criteria for IgG4-related disease. Arthritis Rheumatol 2020;72:7-19.
8. Wallwork R, Perugino CA, Fu X, Harkness T, Zhang Y, Choi HK, et al. The association of smoking with immunoglobulin G4-related disease: a case-control study. Rheumatology (Oxford) 2021;60:5310-7.
9. Sanders S, Fu X, Zhang Y, Perugino CA, Wallwork R, Della-Torre E, et al. Lifetime allergy symptoms in IgG4-related disease: a case-control study. Arthritis Care Res (Hoboken) 2020. E-pub ahead of print.
10. Sreih AG, Annapureddy N, Springer J, Casey G, Byram K, Cruz A, et al. Development and validation of case-finding algorithms for the identification of patients with anti-neutrophil cytoplasmic antibody-associated vasculitis in large healthcare administrative databases. Pharmacoepidemiol Drug Saf 2016;25:1368-74.
11. Jorge A, Castro VM, Barnado A, Gainer V, Hong C, Cai T, et al. Identifying lupus patients in electronic health records: development and validation of machine learning algorithms and application of rule-based algorithms. Semin Arthritis Rheum 2019;49:84-90.