Associations of Autoantibodies, Autoimmune Risk Alleles, and Clinical Diagnoses From the Electronic Medical Records in Rheumatoid Arthritis Cases and Non–Rheumatoid Arthritis Controls

Katherine P. Liao,1 Fina Kurreeman,2 Gang Li,1 Grant Duclos,1 Shawn Murphy,3 Raul Guzman,4 Tianxi Cai,5 Namrata Gupta,6 Vivian Gainer,4 Peter Schur,1 Jing Cui,1 Joshua C. Denny,7 Peter Szolovits,8 Susanne Churchill,9 Isaac Kohane,10 Elizabeth W. Karlson,1 and Robert M. Plenge11

Objective. The significance of non–rheumatoid arthritis (RA) autoantibodies in patients with RA is unclear. The aim of this study was to assess associations of autoantibodies with autoimmune risk alleles and with clinical diagnoses from the electronic medical records (EMRs) among RA cases and non-RA controls.

Supported by NIH grant U54-LM-008748. Dr. Liao’s work was supported by NIH grant K08-AR-060257 (the Katherine Swan Ginsburg Fund). Dr. Karlson’s work was supported by NIH grants K24-AR-052403, R01-AR-049880, and P60-AR-047782. Dr. Plenge’s work was supported by NIH grants R01-AR-057108, R01-AR-056768, U01-GM-092691, and R01-AR-059648. Dr. Plenge holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund.

Methods. Data on 1,290 RA cases and 1,236 non-RA controls of European genetic ancestry were obtained from the EMRs of 2 large academic centers. The levels of anti–citrullinated protein antibodies (ACPAs), antinuclear antibodies (ANAs), anti–tissue transglutaminase antibodies (AGTAs), and anti–thyroid peroxidase (anti-TPO) antibodies were measured. All subjects were genotyped for autoimmune risk alleles, and the association between number of autoimmune risk alleles present and number of types of autoantibodies present was studied. A phenome-wide association study (PheWAS) was conducted to study potential associations between autoantibodies and clinical diagnoses among RA cases and non-RA controls.

Results. The mean ages were 60.7 years in RA cases and 64.6 years in non-RA controls. The proportion of female subjects was 79% in each group. The prevalence of ACPAs and ANAs was higher in RA cases compared to controls (each P < 0.0001); there were no differences in the prevalence of anti-TPO antibodies and AGTAs. Carriage of higher numbers of autoimmune risk alleles was associated with increasing numbers of autoantibody types in RA cases (P = 2.1 ¥ 10−5) and non-RA controls (P = 5.0 ¥ 10−3). From the PheWAS, the presence of ANAs was significantly associated with a diagnosis of Sjögren’s/sicca syndrome in RA cases.

Conclusion. The increased frequency of autoantibodies in RA cases and non-RA controls was associated with the number of autoimmune risk alleles carried by an individual. PheWAS of EMR data, with linkage to laboratory data obtained from blood samples, provide a novel method to test for the clinical significance of biomarkers in disease.
Presence of anti–citrullinated protein antibodies (ACPAs) is a component of the rheumatoid arthritis (RA) classification criteria (1). ACPAs can be found in patient sera more than 10 years prior to the diagnosis of disease (2–5), and the presence of ACPAs portends higher disease severity (6,7). Other autoantibodies, such as antinuclear antibodies (ANAs) (which are associated with systemic lupus erythematosus [SLE]), are found more frequently in RA patients compared to controls (8). However, the prevalence and clinical significance of these autoantibodies in RA is not well established. Moreover, little is known about the factors associated with autoantibody production (9).

The majority of studies in which autoantibodies were assessed focused on the association between genetic risk alleles and the risk of having a disease (e.g., HLA genes and ACPA-positive RA versus controls) (10,11), rather than on autoantibody formation itself as the outcome of interest. Genetics studies of individuals with ANAs, anti–thyroid peroxidase (anti-TPO) antibodies, and anti–tissue transglutaminase antibodies (AGTAs) are conducted within the associated autoimmune diseases of SLE, autoimmune thyroid disease (AITD), and celiac disease, respectively. SLE risk alleles are associated with ANA autoantibodies in patients with SLE (12), but the effect of SLE risk alleles on ANA formation in non-SLE patients has not been investigated. The diagnoses of AITD and celiac disease are tightly linked to the presence of anti–TPO antibodies and AGTAs, respectively, but no study has investigated whether AITD and celiac disease risk alleles are associated with the formation of these autoantibodies.

In this study, we hypothesized that there is a genetic basis for autoantibody formation not only in patients with RA, but also in control subjects without clinical evidence of a rheumatic disease. Furthermore, integrating clinical, genetic, and autoantibody data may provide additional information regarding clinical subsets of RA. Specifically, we hypothesized that 1) RA cases will carry more types of autoantibodies than controls, and 2) in both RA cases and controls, disease-specific risk alleles will be associated with disease-specific autoantibodies (e.g., association of SLE risk alleles with ANAs) even when the associated autoimmune disease is not present. In addition, we conducted a phenome-wide association study (PheWAS) (13) as an exploratory analysis to investigate associations of autoantibodies with autoimmune risk alleles and with clinical diagnosis codes ascertained through the electronic medical records (EMRs).

SUBJECTS AND METHODS

Study population. Our study was conducted using data for 1,290 RA cases and 1,236 non-RA controls of European ancestry identified from the EMRs of Brigham and Women’s Hospital (BWH) and Massachusetts General Hospital; these subjects have been characterized previously in published studies from our group (10,14). RA cases were identified using a previously described EMR RA phenotype algorithm, which has a positive predictive value (PPV) of 94% (14). The algorithm was trained on a gold standard set of subjects who were classified as either RA cases or not RA cases by 3 board-certified rheumatologists (KPL, EWK, and RMP), using the American College of Rheumatology 1987 revised classification criteria for RA (15) as the benchmark. This algorithm was further validated at 2 other academic institutions (16) (see ref. 14 for details on development, training, and validation of this RA phenotype algorithm).

The non-RA control group was created from the EMR by excluding all subjects with an International Classification of Diseases, Ninth Revision (ICD-9) code for any rheumatic disease in the EMR (this excluded all subjects in the RA cohort) (see ref. 10 for details). The remaining subjects were matched to RA cases (3:1) by age, sex, self-reported ethnicity, and level of health care utilization (represented by the number of “facts,” or contacts with the health care system [i.e., office visits, blood withdrawals for laboratory tests]) (17). For both RA cases and non-RA controls, information regarding age, sex, ICD-9 codes, laboratory test results, and electronic prescriptions for medications were extracted from the structured EMR data. Information on bone erosions was obtained from the bone radiology reports of the subjects, using natural language processing in the Health Information Text Extraction system (14,18).

Discarded blood samples from 5 Partners Healthcare clinical laboratories were collected in the Clinical Specimen Bank at BWH from 2009 to 2010, using an Institutional Review Board–approved process as described in the study by Kurreeman et al (10). Thus, the final population of RA cases and non-RA controls analyzed for this study comprised subjects for whom blood samples were available and whose European ancestry could be confirmed by testing for ancestry-informative markers (AIMs). As a result, the RA cases and non-RA controls were no longer perfectly matched.

Genotyping. The detailed methods used for genotyping and assigning genetic ancestry of the subjects in the RA case and non-RA control groups were previously described in the study by Kurreeman et al (10). Briefly, processing and genotyping of the discarded blood samples were performed at The Broad Institute (Cambridge, MA). We genotyped 192 AIMs, 28 single-nucleotide polymorphisms (SNPs) associated with RA, 33 SNPs associated with SLE, and 16 SNPs associated with celiac disease (19–24) (disease-specific SNPs are listed in Supplementary Table 1, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). For quality control, we removed SNPs with a missing genotype rate of >10% and a minor allele frequency of <1%. Genetic ancestry based on the AIMs was determined using the Bayes classifier and principal components analysis.
**Calculation of the aggregate genetic risk score (GRS).** We calculated a cumulative aggregate GRS for RA, SLE, and celiac disease for each individual using the following formula (10,25,26): $GRS = \sum_{i=1}^{n} X_i$, where $n$ is the number of SNPs for the particular disease (RA, SLE, or celiac disease) (see Supplementary Table 1, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract), $i$ is the SNP, and $X_i$ is the number of risk alleles (0, 1, or 2).

The RA GRS excludes the HLA–DRB1*04 tag SNP because we were interested in understanding the effects of non-HLA risk alleles and production of ACPAs in RA. In addition, the associations in the HLA region are complex and require dense genotyping (27), an approach that was not available for this study. We also created a combined autoimmune GRS (AI GRS), which consisted of all risk alleles in the study with the exception of SNPs in linkage disequilibrium with another SNP (see Supplementary Table 2, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). All GRS were unweighted, due to absence of information on the strength of association for any individual risk allele and autoantibody outcome. The literature on AITD is less definitive (28), and we therefore did not construct a GRS for AITD.

**Measurement of autoantibodies.** We measured autoantibodies in blood samples using the following enzyme-linked immunosorbent assay (ELISA) kits (all from Inova): for ACPAs, the CCP3 IgG ELISA; for ANAs, the Quanta-Lite ANA ELISA; for anti-TPO antibodies, the Quanta-Lite TPO ELISA; and for AGTAs, the Quanta-Lite IgA tissue transglutaminase ELISA. We determined positivity for an autoantibody based on the manufacturer’s recommended cutoff levels: for ACPAs, a titer of ≥20 units; for ANAs, a titer of ≥20 units (high-titer ANAs defined as a titer of >60 units); for anti-TPO antibodies, a titer of >100 World Health Organization units; and for AGTAs, a titer of ≥20 units. These autoantibodies were selected because of the relationship between each relevant autoimmune disease and RA, as has been identified in both epidemiologic studies (29,30) and genetic studies (31–33). ANAs, anti-TPO antibodies, and AGTAs were measured in all RA cases and all non-RA controls. ACPAs were measured in all RA cases and 202 controls, which provided adequate power to detect a difference in ACPA prevalence between the 2 groups (see Supplementary Table 2, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract).

**Statistical analysis.** Determination of autoantibody prevalence. We first determined the distribution and overlap of ACPAs, ANAs, anti-TPO antibodies, and AGTAs among RA cases and non-RA controls. We then constructed separate logistic regression models for each of the autoantibodies to study the association between each autoantibody (ACPAs, ANAs, anti-TPO antibodies, and AGTAs) and RA case or control status, adjusted for age, sex, and health care utilization (represented by the number of facts). We then conducted an RA case-only analysis to examine whether the prevalence of ANAs, anti-TPO antibodies, and AGTAs differed according to the ACPA status of the RA cases, using logistic regression models that were adjusted for age, sex, and health care utilization.

Logistic regression analyses for associations between GRS and autoantibody positivity. We constructed separate logistic regression models to determine the association between the disease-specific GRS (for RA, SLE, and celiac disease) and their related autoantibodies (ACPAs, ANAs, and AGTAs, respectively) among RA cases. In controls, we tested only the association between the SLE GRS and ANAs, due to the low prevalence of ACPAs and AGTAs in the control subjects (results of post hoc power calculations are shown in Supplementary Table 2, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). Linear regression was used to assess the association between the AI GRS and the number of types of autoantibodies (ANAs, anti-TPO, and AGTAs) present in both RA cases and controls.

We utilized multivariate logistic regression models to determine the relative importance of age, sex, and disease-specific GRS in predicting autoantibody positivity among RA cases. In controls, we tested only the association between the SLE GRS and ANAs, due to the power issues stated above.

To test whether SLE risk alleles have the same effect on ANA production in both RA cases and controls, we categorized all subjects into tertiles of the SLE GRS (tertile 1 = SLE GRS ≤0 and ≤33.36, tertile 2 = SLE GRS >33.36 and ≤40.9, tertile 3 = SLE GRS >40.9). Within each tertile, we fitted logistic regression models to estimate the odds of ANA positivity in RA cases compared to non-RA controls, given a similar number of SLE risk alleles. Each model was adjusted for age, sex, and health care utilization. A subgroup analysis was conducted to assess only subjects who had not received a prescription for a tumor necrosis factor inhibitor (TNFi), because treatment with TNFi agents has been associated with conversion to ANA positivity (34).

**PheWAS for associations of autoantibodies and autoimmune risk alleles with clinical diagnoses.** To discover potential associations between non-RA autoantibody positivity and a clinical diagnosis, and between the AI GRS and a clinical diagnosis, we conducted a PheWAS. Clinical diagnoses were defined using ICD-9 codes that were manually grouped into clinically relevant diseases (referred to as PheWAS codes); e.g., ICD-9 codes 401–405 represent different types of hypertension, and thus complications are grouped into a single PheWAS code for hypertension. The PheWAS codes were assigned by a team of physicians in a previously published study (13). We included all PheWAS codes that had a prevalence of ≥1% in each group, yielding 512 clinical diagnosis codes for analysis in RA cases and 698 clinical diagnosis codes for analysis in non-RA controls.

We tested for potential associations between an autoantibody and PheWAS clinical diagnosis code in RA cases and non-RA controls separately, by fitting logistic regression models with ACPA, ANA, or anti-TPO antibody status as the predictor (adjusted for age and sex) and the presence or absence (designated 0 or 1, respectively) of the PheWAS code as the outcome. In these analyses, we did not study AGTAs in cases and controls and ACPAs in controls because of the low prevalence of these autoantibodies (≤1%), which would lead to unstable estimates. To study the associations between the AI GRS and PheWAS codes, we fitted logistic regression models...
with the AI GRS as the predictor (adjusted for age and sex) and the presence or absence of the PheWAS code as the outcome. Since multiple comparisons are applied in the PheWAS, suggestive associations could be attributable to chance. Therefore, we applied the Bonferroni correction for multiple comparisons, and set the cutoff level for a significant association as \( P \leq 9.76 \times 10^{-5} \) (0.05/512) in RA cases and \( P \leq 7.16 \times 10^{-5} \) (0.05/698) in non-RA controls.

When a significant association between an autoantibody and PheWAS clinical diagnosis code was found after Bonferroni correction, we reviewed the medical records to ascertain the diagnostic accuracy of the code. For example, in the RA case group, if a specific autoantibody was significantly associated with a specific PheWAS code, we randomly selected 20 RA cases with \( \geq 1 \) PheWAS code in their EMR, and a single investigator (KPL) reviewed the medical records for clinical documentation of this diagnosis by the treating physician. This method was applied for each PheWAS code that showed a significant association (after Bonferroni correction) with an autoantibody. The percentage of confirmed diagnoses represents the PPV of the PheWAS code.

This study was approved by the Partners Healthcare Institutional Review Board. Statistical analyses were conducted using the SAS software package (version 4.2; SAS Institute) and the R Project for Statistical Computing (http://www.r-project.org).

**RESULTS**

**Characteristics of the study population and prevalence of autoantibodies.** We studied 1,290 RA cases and 1,236 non-RA controls of European ancestry, using clinical data from the EMRs and laboratory data from DNA/plasma extracted from discarded blood samples. The mean age of the RA cases was 60.7 years, while the mean age of non-RA controls was 64.6 years. The proportion of women in both groups was 79%. The clinical characteristics of the RA cases included the following: 70.3% positive for ACPAs, 61.6% with bone erosions, 62.3% with \( \geq 1 \) electronic prescription for methotrexate, and 40.9% with \( \geq 1 \) electronic prescription for a TNFi at some point during their treatment (Table 1). ACPAs were tested in all of the RA cases and in 202 controls. The distribution of antibodies to tissue transglutaminase is not shown (14 RA cases [1.1%] and 8 non-RA controls [0.6%]).

![Figure 1](image_url) Distribution of autoantibodies (antibodies to citrullinated proteins [ACPAs], antinuclear antibodies [ANAs], and anti–thyroid peroxidase [anti-TPO]) antibodies among rheumatoid arthritis (RA) cases (n = 1,290) (A) and non-RA controls (n = 1,236) (B) of European ancestry. ACPAs were tested in all of the RA cases and in 202 controls. The distribution of antibodies to tissue transglutaminase is not shown (14 RA cases [1.1%] and 8 non-RA controls [0.6%]).

**Table 1.** Characteristics of the rheumatoid arthritis (RA) cases and non-RA controls*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>RA cases (n = 1,290)</th>
<th>Controls (n = 1,236)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean ± SD years</td>
<td>60.7 ± 13.6</td>
<td>64.6 ± 13.4</td>
</tr>
<tr>
<td>Female, no. (%)</td>
<td>1,024 (79.4)</td>
<td>965 (78.7)</td>
</tr>
<tr>
<td>Methotrexate, ever use, no. (%)</td>
<td>803 (62.3)</td>
<td>15 (1.2)</td>
</tr>
<tr>
<td>TNFi, ever use, no. (%)</td>
<td>528 (40.9)</td>
<td>8 (0.65)</td>
</tr>
<tr>
<td>Bone erosions, no. (%)</td>
<td>795 (61.6)</td>
<td>69 (5.6)</td>
</tr>
<tr>
<td>Measured autoantibodies, no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACPAs</td>
<td>901 (70.3)</td>
<td>1 (0.5)†</td>
</tr>
<tr>
<td>ANAs</td>
<td>445 (34.5)</td>
<td>193 (15.6)</td>
</tr>
<tr>
<td>High-titer ANAs‡</td>
<td>111 (8.6)</td>
<td>51 (4.1)</td>
</tr>
<tr>
<td>Anti-TPO antibodies</td>
<td>202 (15.6)</td>
<td>196 (15.9)</td>
</tr>
<tr>
<td>AGTAs</td>
<td>14 (1.1)</td>
<td>8 (0.7)</td>
</tr>
</tbody>
</table>

* TNFi = tumor necrosis factor inhibitor; ACPAs = antibodies to citrullinated proteins; anti-TPO = anti–thyroid peroxidase; AGTAs = anti–tissue transglutaminase antibodies.
† Of 202 controls.
‡ High-titer antinuclear antibodies (ANAs) were defined by the manufacturer as a titer of >60 units.
RA cases were more likely to be ANA positive compared to non-RA controls (Figure 2A), whereas there were no significant differences in the prevalence of anti-TPO and AGTAs between RA cases and controls. In an RA case–only subset analysis, we found that ACPA-positive RA cases were more likely to be ANA positive compared to ACPA-negative RA cases, but no significant differences in the prevalence of anti-TPO antibodies and AGTAs were found between ACPA-positive and ACPA-negative RA cases (Figure 2B).

Association between GRS and autoantibody positivity. To determine the genetic basis of autoantibodies in RA cases, we tested RA (23), SLE (19,35), and celiac disease (24,36) genetic risk alleles for association with the disease-specific autoantibodies (ACPAs, ANAs, and AGTAs, respectively). We found that the disease-specific GRS were predictive of the presence of the disease-specific autoantibodies among RA cases (Figures 3A–C). As we have shown previously in ACPA-positive patients with RA (10), the mean GRS for RA was significantly higher in ACPA-positive cases than in ACPA-negative cases ($P = 2.5 \times 10^{-9}$) (Figure 3A). Notably, calculation of the RA GRS did not include the HLA shared epitope alleles.

Moreover, as shown in Figure 3B, we found that ANA-positive RA cases had a significantly higher mean GRS for SLE compared to ANA-negative RA cases ($P = 8.0 \times 10^{-4}$). Finally, as shown in Figure 3C, AGTA-positive RA cases had a significantly higher mean GRS for celiac disease compared to AGTA-negative RA cases ($P = 0.03$).

Among RA cases, the associations between disease-specific GRS and autoantibody status did not change after adjustment for age, sex, and health care utilization in a multivariable logistic regression model. Furthermore, neither age nor sex was significantly associated with autoantibody status in this model (results not shown).

In non-RA controls, there was no significant association between the SLE GRS and ANA status, although the trend was similar to that in the RA patients. ANA-positive controls showed a trend toward a significantly higher mean GRS for SLE when compared to ANA-negative controls ($P = 0.31$) (see Supplementary Figure 1A, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). However, in contrast to RA cases, both age and sex were significantly associated with ANA positivity in non-RA controls (for age per year, odds ratio [OR] 1.01, 95% confidence interval [95% CI] 1.002–1.03, $P = 0.02$; for female sex, OR 1.7, 95% CI 1.2–2.75, $P = 0.007$).

We also combined all non-RA autoimmune risk alleles into a single GRS (the AI GRS) and tested for associations of the AI GRS with the number of autoantibodies present (with or without ACPAs). In RA cases, the increasing count of any autoimmune risk allele (i.e., a higher AI GRS) was significantly associated with increasing numbers of autoantibody types (range 0–4, comprising ACPAs, ANAs, anti-TPO antibodies, and AGTAs) ($P = 3.2 \times 10^{-8}$). In RA cases, the association remained after the removal of ACPAs as a potential
outcome; an increasing AI GRS was still associated with increasing numbers of autoantibody types (range 0–3, comprising ANAs, anti-TPO antibodies, and AGTAs) ($P = 2.6 \times 10^{-5}$) (Figure 3D).

In non-RA controls, increasing numbers of autoantibody types (range 0–3, comprising ANAs, anti-TPO antibodies, and AGTAs) were also significantly associated with having more autoimmune risk alleles (i.e., a higher AI GRS) ($P = 4.0 \times 10^{-4}$) (see Supplementary Figure 1B, available on the Arthritis & Rheumatism website at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). The association between the AI GRS and numbers of autoantibody types (range 0–3, comprising ANAs, anti-TPO antibodies, and AGTAs) remained significant in RA cases ($P = 2.1 \times 10^{-5}$) and in non-RA controls ($P = 5.0 \times 10^{-3}$) after adjustment for age, sex, and health care utilization.

One possible explanation for the association between ANA positivity and RA case status (Figure 2A) is that RA cases carry more SLE risk alleles than non-RA controls. Although the RA cases carried significantly more SLE risk alleles than the non-RA controls (mean ± SD SLE GRS 38.6 ± 5.3 in RA cases versus 37.8 ± 5.4 in controls; $P = 4.0 \times 10^{-4}$), we found that RA cases still had a higher rate of ANA positivity compared to non-RA controls after controlling for the effect of SLE risk alleles (Figure 4). More specifically, when RA cases and controls with similar numbers of SLE risk alleles (grouped by tertile of SLE GRS) were compared, we found that RA cases were still significantly more likely to be ANA positive.

A subgroup analysis comparing RA cases and non-RA controls who did not have a TNFi prescription electronically prescribed through the EMR (RA cases $n = 762$, controls $n = 1,228$) yielded similar findings. RA cases compared to non-RA controls, given the same number of SLE risk alleles, were still significantly more likely to be ANA positive ($P < 0.0001$ for all tertiles of the SLE GRS).

PheWAS findings of associations of autoantibodies and autoimmune risk alleles with clinical diagnoses. To determine whether the presence of autoantibodies or a higher number of autoimmune risk alleles could be associated with codified clinical diagnoses, we performed an exploratory PheWAS using our EMR data. From the PheWAS findings, we observed that in both RA cases and non-RA controls, the presence of anti-TPO antibodies was significantly associated with hypo-
thorothyroidism ($P = 1.22 \times 10^{-16}$ and $P = 9.22 \times 10^{-10}$, respectively) (Table 2), which was expected given the known association between anti-TPO antibodies and autoimmune thyroid disease. The PPV of the PheWAS code for acquired hypothyroidism was 75% in RA cases and 100% in non-RA controls, based on medical record review.

Two PheWAS codes were significantly associated with high-titer ANA positivity in RA cases and non-RA controls. In RA cases, the PheWAS code for a diagnosis of Sjögren’s/sicca syndrome was associated with high-titer ANA positivity ($P = 8.59 \times 10^{-6}$) (Table 2). The PPV of the PheWAS code for Sjögren’s/sicca syndrome was 70%. In non-RA controls, we observed a significant association between high-titer ANA positivity and chronic nonalcoholic liver disease ($P = 2.9 \times 10^{-5}$) (Table 2). The PPV of the PheWAS code for chronic nonalcoholic liver disease was 75%, based on medical record review for documentation of liver disease not associated with alcohol.

We found no significant association between ACPA positivity and any PheWAS outcome category among RA cases (see Supplementary Table 3, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). Furthermore, we observed no significant associations between carriage of higher numbers of autoimmune risk alleles in aggregate (the AI GRS) and a specific PheWAS code corresponding to a clinical diagnosis in the RA cases or controls.

**DISCUSSION**

RA diagnosis and classification are based on a distinct clinical phenotype (1). Underlying the clinical phenotype of RA are variations in the numbers of autoimmune risk alleles and differential production of RA and non-RA autoantibodies. In this study, we observed that 1) RA cases had a higher prevalence of ANAs, with a trend toward a higher prevalence of AGTAs (but not anti-TPO antibodies), when compared to non-RA controls, 2) SLE and celiac disease genetic risk alleles influenced the production of the SLE- and celiac disease–related autoantibodies (ANAs and AGTAs, respectively) among RA cases, 3) the number of autoimmune risk alleles carried by a patient influenced the number of types of autoantibodies carried in both RA cases and non-RA controls, 4) given a similar number of SLE risk alleles, RA cases were more likely to be ANA positive compared to non-RA controls, and 5) ANA positivity was associated with a diagnosis of Sjögren’s/sicca syndrome among RA cases, according to the results of PheWAS.

To our knowledge, this is the largest study to date that systematically tested and compared the prevalence of clinical autoantibodies (ACPAs, ANAs, anti-TPO antibodies, and AGTAs) in RA cases and non-RA controls. For the purposes of our study, we defined autoantibodies as being disease-specific, linking ACPAs with RA and linking other non-RA autoantibodies with specific autoimmune diseases (ANAs in SLE, AGTAs in celiac disease, and anti-TPO antibodies in AITD). However, our results show that this definition is not precise, since we found that non-RA autoantibodies (i.e., ANAs) were actually more common in RA cases compared to non-RA controls.

We propose 2 hypotheses to explain why RA patients are at an increased risk of having disease-specific autoantibodies compared to non-RA controls.

**Table 2.** Results of the phenome-wide association study for significant associations between autoantibodies and electronic medical record–based clinical diagnoses among rheumatoid arthritis (RA) cases ($n = 1,265$) and non-RA controls ($n = 1,225$) of European ancestry

<table>
<thead>
<tr>
<th>Group, autoantibody</th>
<th>Outcome/diagnosis</th>
<th>Prevalence of diagnosis, %</th>
<th>OR (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Autoantibody positive</td>
<td>Autoantibody negative</td>
<td></td>
</tr>
<tr>
<td>RA cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anti-TPO</td>
<td>Acquired hypothyroidism</td>
<td>40.1</td>
<td>14.3</td>
<td>4.2 (3.0–5.9)</td>
</tr>
<tr>
<td>High-titer ANAs†</td>
<td>Sjögren’s/sicca syndrome</td>
<td>13.5</td>
<td>3.5</td>
<td>4.2 (2.2–7.9)</td>
</tr>
<tr>
<td>Non-RA controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anti-TPO</td>
<td>Acquired hypothyroidism</td>
<td>48.5</td>
<td>24.1</td>
<td>2.8 (2.0–3.8)</td>
</tr>
<tr>
<td>Anti-TPO</td>
<td>Thyroiditis</td>
<td>24.1</td>
<td>2.7</td>
<td>4.7 (2.7–8.4)</td>
</tr>
<tr>
<td>High-titer ANAs†</td>
<td>Other chronic nonalcoholic liver disease</td>
<td>11.7</td>
<td>8.0</td>
<td>8.0 (3.0–21.1)</td>
</tr>
</tbody>
</table>

*The Bonferroni-corrected P values for a significant association were as follows: $P \leq 9.76 \times 10^{-5}$ in RA cases, and $P \leq 7.16 \times 10^{-5}$ in non-RA controls. OR = odds ratio; 95% CI = 95% confidence interval; anti-TPO = anti–thyroid peroxidase antibodies.

† High-titer antinuclear antibodies (ANAs) were defined by the manufacturer as a titer of >60 units.
Under a purely genetic model, RA patients have a higher prevalence of autoantibodies because they have a higher burden of autoimmune risk alleles. The increased frequency of ANAs in RA may be due to the fact that the same risk alleles are shared between RA and SLE (31,33,37), leading to a larger burden of SLE risk alleles in RA patients. Supporting this concept, we found that RA patients had a significantly higher mean GRS for SLE alleles compared to controls. A consequence of the increased burden of autoimmune risk alleles is that these risk alleles may also have general effects on self-tolerance. For example, the PTPN22 missense allele (R620W), a shared genetic risk factor for RA, SLE, and AITD, has been shown to cause defects in the counter-selection of autoreactive B cells (31,38,39), which leads to production of autoantibodies.

The second explanation, not mutually exclusive from the genetic model, is that differential environmental and endogenous (immune dysregulation) exposures are associated with RA (40). These exposures may render RA patients more susceptible to the downstream effects of carrying additional autoimmune risk alleles, since the threshold for producing autoantibodies is lower in RA patients than in non-RA controls. However, if general immune dysregulation alone were responsible for autoantibody formation, we would not expect to see the specific associations that were observed between the SLE and celiac disease risk alleles and the presence of ANAs and AGTAs, respectively, in RA cases. In addition, we would expect to see an increased frequency of all autoantibodies, but the prevalence of anti-TPO antibodies was the same in RA cases and non-RA controls.

Integrating the 2 explanations above, we interpret the similarities and differences in autoantibody prevalence between RA cases and non-RA controls in the following way. The common theme shared by RA cases and controls was that an increasing burden of autoimmune risk alleles was associated with carriage of more types of any autoantibody (Figure 3D) (see also Supplementary Figure 1B, available on the Arthritis & Rheumatism web site at http://onlinelibrary.wiley.com/doi/10.1002/art.37801/abstract). Autoimmune risk alleles, therefore, provide the substrate for autoantibody production. However, it is clear that additional factors contribute to increased autoantibody production. For example, among subjects carrying a similar number of SLE risk alleles, RA cases were more likely than controls to be ANA positive (Figure 4). We hypothesize that in patients with RA, the effect of inherited and environmental factors that leads to a breakdown in self-tolerance and the onset of clinical symptoms of RA also predisposes individuals to the development of additional autoantibodies, such as ANAs. In non-RA controls, there is less background immune dysregulation, and the same autoimmune risk alleles are less important for autoantibody production. Thus, we posit that autoimmune risk alleles predispose both RA patients and non-RA control subjects to the development of autoantibodies, but that environmental factors and endogenous exposures (immune dysregulation) are also important contributors.

A prediction of our integrated model is that carriage of autoimmune risk alleles leading to immune dysregulation will have phenotypic consequences beyond RA case-versus-control status. To investigate this hypothesis systematically, we utilized EMR clinical data to determine which phenotypes may be associated with specific autoantibodies or higher numbers of autoimmune risk alleles. This approach, termed PheWAS, has been previously applied to test associations between SNPs and diagnosis codes (13,41). However, this approach has not yet been used to test for clinical associations with autoantibodies or GRS. From our exploratory PheWAS, we observed, as expected, that the presence of anti-TPO antibodies was strongly associated with hypothyroidism in both RA cases and non-RA controls. We also observed associations between ANAs and Sjögren's/sicca syndrome in RA cases. Although this finding is consistent with existing clinical observations regarding the association between RA and Sjögren's syndrome (42), this association has not been definitively...
shown until now. Among RA cases, ACPA status was not associated with RA, because virtually all of the patients had a diagnosis code for RA.

Our study design represents a novel approach to translational research. The RA case and non-RA control study populations were created with anonymized clinical data from the EMRs (14), with linkage to a biospecimen repository of discarded blood samples. Clinical information (i.e., age, sex, diagnosis codes) was obtained from the structured EMR. This EMR platform afforded us the opportunity to understand the relationship between clinical features, genetic factors, and autoantibodies in RA, and to compare these findings with a control group. Moreover, we had the ability to integrate clinical EMR data and our research laboratory–based autoantibody and genetic data for thousands of patients, to determine significant associations between autoimmune risk alleles and autoantibodies and identify the potential clinical relevance of autoantibodies or a high number of autoimmune risk alleles.

There are important limitations to our study. We focused our study on a subset of autoimmune diseases, which limits the generalizability of our conclusions. In the selection of our controls, we excluded those with any ICD-9 code for a rheumatic disease (10). Therefore, our controls may have had a lower prevalence of autoantibodies than might be found in general population controls; however, the prevalence of ACPAs, ANAs, and anti-TPO antibodies that we observed in our controls was comparable to that observed in control populations in other studies (8,43–45). Because we do not have temporal information with regard to the point at which a patient developed an autoantibody, we cannot state that an autoantibody was predictive of a particular outcome. Similarly, we were not able to study whether an environmental exposure (i.e., smoking) might have contributed to autoantibody development.

We chose to use ACPAs rather than rheumatoid factor (RF) to stratify the RA cases. The correlation between these 2 autoantibodies is high, but discrepancies are known (46–48). Therefore, whether these results would also directly apply in a comparison of RF-positive and RF-negative RA cases is unclear. There is concern that the presence of RF can lead to false-positive results in multiplex antibody assays (49); however, we found no evidence that the presence of RF resulted in false-positive readings for the standard commercial ELISAs used in this study. Our findings suggest that RF interference was not a major issue; among RA cases, the prevalence of autoantibodies was similar between ACPA-positive and ACPA-negative patients, with the exception of ANAs. Finally, the results stemming from the PheWAS approach are exploratory and, therefore, require further refinement in terms of validation of outcomes (i.e., ICD-9 diagnosis codes) and covariates. Moreover, the metrics of what constitutes a true association (e.g., P value threshold, effect size) remain to be determined.

In conclusion, we utilized a novel EMR-based approach in this study to test genetic risk factors for associations with clinical autoantibodies. Our results provide insight into the immunologic heterogeneity underlying the clinical entity of RA. In the context of a linked clinical EMR–research laboratory database, we believe that the PheWAS approach can be a powerful hypothesis-generating tool for uncovering associations that are not readily apparent from our current knowledge of the pathways and mechanisms of disease.

ACKNOWLEDGMENTS

We thank Dr. Kurt Bloch (Massachusetts General Hospital) for his advice on the selection and testing of autoantibodies in this study, and Dr. Lynn Bry (Director, BWH Clinical Specimen Bank) for her help with coordinating the sample collection and testing.

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. Liao 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. Liao, Murphy, Cai, Denny, Szolovits, Churchill, Karlson, Plenge.

Acquisition of data. Liao, Li, Duclos, Murphy, Guzman, Gupta, Gainer, Kohane, Karlson, Plenge.

Analysis and interpretation of data. Liao, Kurreeman, Cai, Schur, Cui, Denny, Szolovits, Karlson, Plenge.

REFERENCES


5. Majka DS, Deane KD, Parrish LA, Lazar AA, Baron AE, Walker...


