EXTENDED REPORT

A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk

Annie Yarwood,[1] Buhm Han,[2] Soumya Raychaudhuri,[1,2] John Bowes,[1] Mark Lunt,[1] Dimitrios A Pappas,[3] Joel Kremer,[5] Jeffrey D Greenberg,[4] Robert Plenge,[6,7,8] Rheumatoid Arthritis Consortium International (RACI), Jane Worthington,[1,9] Anne Barton,[1,9] Steve Eyre[1]

Handling editor Tore K Kvien

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/annrheumdis-2013-204133).

For numbered affiliations see end of article.

Correspondence to
Dr Steve Eyre,
Arthritis Research UK
Epidemiology Unit, Stopford
Building, The University of
Manchester, Oxford Road,
Manchester M13 9PT, UK;
steve.eyre@manchester.ac.uk

Received 17 June 2013
Revised 9 August 2013
Accepted 10 September 2013
Published Online First
3 October 2013

To cite: Yarwood A, Han B, Raychaudhuri S, *et al*. *Ann Rheum Dis* 2015;**74**:170–176.

ABSTRACT

Background There is currently great interest in the incorporation of genetic susceptibility loci into screening models to identify individuals at high risk of disease. Here, we present the first risk prediction model including all 46 known genetic loci associated with rheumatoid arthritis (RA).

Methods A weighted genetic risk score (wGRS) was created using 45 RA non-human leucocyte antigen (HLA) susceptibility loci, imputed amino acids at HLA-DRB1 (11, 71 and 74), HLA-DPB1 (position 9) HLA-B (position 9) and gender. The wGRS was tested in 11 366 RA cases and 15 489 healthy controls. The risk of developing RA was estimated using logistic regression by dividing the wGRS into quintiles. The ability of the wGRS to discriminate between cases and controls was assessed by receiver operator characteristic analysis and discrimination improvement tests.

Results Individuals in the highest risk group showed significantly increased odds of developing anti-cyclic citrullinated peptide-positive RA compared to the lowest risk group (OR 27.13, 95% CI 23.70 to 31.05). The wGRS was validated in an independent cohort that showed similar results (area under the curve 0.78, OR 18.00, 95% CI 13.67 to 23.71). Comparison of the full wGRS with a wGRS in which HLA amino acids were replaced by a HLA tag single-nucleotide polymorphism showed a significant loss of sensitivity and specificity.

Conclusions Our study suggests that in RA, even when using all known genetic susceptibility variants, prediction performance remains modest; while this is insufficiently accurate for general population screening, it may prove of more use in targeted studies. Our study has also highlighted the importance of including HLA variation in risk prediction models.

INTRODUCTION

The recent advances in complex disease genetics has prompted discussion as to how this information could be used in personalised medicine and risk prediction. Predicting the development of disease, or progression to severe disease is one of the ultimate aims of genetic and epidemiological research. Accurate risk prediction would allow targeted preventive treatments, such as lifestyle or even pharmacological intervention.

Rheumatoid arthritis (RA) is a typical complex genetic disease, with the strongest genetic association lying within the major histocompatibility complex (MHC). In 2012 a study in this region identified five amino acids in three human leucocyte antigen (HLA) proteins (HLA-DRB1, HLA-DPB1 and HLA-B) explaining most of the association between the MHC and RA.[1] In addition, the most recent study to identify RA-associated loci comprehensively (Immunochip), has identified novel loci associated with disease, bringing the total number of Caucasian RA susceptibility loci to 48 (45 non HLA loci plus three HLA loci).[2]

Therefore, it is now timely to examine the predictive ability of these genetic data. The aims of this study were, first, to extend previous analysis[3–6] by creating a weighted genetic risk score (wGRS) including all known Caucasian RA susceptibility loci and epidemiological risk factors when possible (smoking, gender); second, to compare models to identify the most informative and economical method for inclusion into a screening model; third, to assess the performance of the wGRS in specific patient subgroups stratified by the presence of autoantibodies.

METHODS

Samples

Cases and controls were selected from the UK, USA, Sweden, The Netherlands and Spain as described previously.[2] All cases were Caucasian, over the age of 18 years and satisfied 1987 American College of Rheumatology criteria for RA modified for genetic studies.[7] For further information regarding the sample collections used in this study see the supplementary note (available online only).

Genotyping

Genotyping was carried out using a custom designed Illumina Infinium array, 'immunochip', and Illumina iScan technology, and quality control of the data was carried out as described previously.[2]

Statistical analysis

A wGRS was generated using a combination of RA susceptibility loci. Analysis was carried out using STATA V.11.0 (http://www.stata.com). wGRS were created using the 45 non-HLA loci confirmed to be associated with RA (wGRS45_G),[2] and then including either a HLA-DRB1 tag single-nucleotide polymorphism (SNP) (rs660895) or imputed HLA-DRB1 amino acids 11, 71, 74 and imputed amino acids nine at HLA-DPB1 and HLA-B,[1] in order to identify the most informative and economical method to incorporate HLA into a risk screening model. SNP-only and HLA-only models were also generated to determine the contribution of the disease-associated SNPs, which have modest effect sizes, to the risk score.

We assessed the accuracy of the imputation by comparing the imputed HLA-DRB1 alleles to the genotyped HLA-DRB1 alleles in a subset of 1894 samples.

Each wGRS was generated by weighting each allele using the natural log of the published OR reported in Eyre *et al*[2] or, in the case of the imputed HLA amino acids, were weighted by published OR for each haplotype as shown in Raychaudhuri *et al*.[1] The SNP rs10739580, rs39984 and rs78560100 were not available in the discovery data, therefore the following proxies were selected rs10739579, rs434816 and rs62323881, respectively. Individuals with missing SNP genotypes were assigned a value that was equal to two times the minor allele frequency.[6] Logistic regression was used to calculate OR and p values for each wGRS.

Environmental factors also contribute to the risk of RA, therefore gender and smoking were also incorporated into the model when data were available. Gender was available for all individuals and was weighted by the natural log of the OR in our cohort (OR 2.00) as the association varies by age.

Smoking data were available for controls from the 1958 birth cohort and a subset of RA patients from the UK, specifically recruited as part of the Norfolk Arthritis Register (NOAR). Smoking status was coded as 0 for never smokers and 1 for ever smokers.

To determine how well the wGRS discriminates between cases and controls we generated receiver operator characteristic (ROC) curves, defining the sensitivity and specificity of each wGRS, and calculated the area under the curve (AUC). It has previously been shown that susceptibility factors would need an exceptionally large OR to have a significant impact on the AUC, and that it may be more appropriate to assess reclassification rather than discrimination of a model.[8] [9] Therefore, integrated discrimination improvement (IDI) tests were carried out using STATA to determine the change in probability between the models. IDI is a measure of overall improvement in sensitivity and specificity, which allows the comparison of models and assessment of how much improvement is made by the addition of other variables.[10] In our case–control study the probability of RA is increased compared to the general population due to a large number of cases; however, we have used IDI tests to compare models within our dataset and not to generalise to the population as a whole.

Continuous net reclassification improvement tests[11] were used to quantify the overall improvement of reclassification between models, that is, the amount of correct reclassification among cases and controls.

Finally the wGRS scores were split into quintiles. Individuals in the 5th quintile were classed as high risk, allowing us to calculate sensitivity and specificity for each model.

Validation of wGRS

As several of the RA susceptibility loci included in the wGRS were identified in the cohort used in this study, the wGRS were tested in an independent European cohort of 2206 RA cases and 1863 healthy controls for validation. The SNP rs10683701 and rs13397 were not available in the validation dataset, therefore rs7979246 was used as a proxy ($r^2=0.98$) for rs10683701, no proxy was available for rs13397.

The samples were from the Consortium of Rheumatology Researchers of North America (CORRONA) registry and the Informatics for Integrating Biology and the Bedside (I2B2) centre. For the CORRONA samples, RA patients from the registry were recruited and consented for a genetics substudy from 40 rheumatology office sites.

Samples were genotyped using the custom designed Illumina Infinium genotyping chip 'immunochip'. Genotype clustering was carried out using Illumina's Beadstudio custom cluster files. SNP and sample quality control thresholds were applied in PLINK (SNP missingness >0.02, sample call rate <0.9). To identify related individuals' identity by descent analysis was carried out using a set of high quality (missingness <0.002, minor allele frequency >0.1) and linkage disequilibrium pruned ($r^2<0.2$) SNPs. Samples with a PI-HAT greater than 0.2 were removed. Principal components analysis was carried in EIGENSOFT, using HapMap3 samples as reference.

RESULTS

Data quality control

Sample and genotype quality control was carried out as described previously.[2] HLA imputation accuracy was shown to be 97% at two digit resolution and 91% at four digit resolution. Quality control of the imputed HLA data excluded 439 samples due to missing data or multiple haplotypes being assigned, leaving 11 366 cases and 15 489 controls for analysis (described previously).[2] Cohort characteristics can be seen in table 1.

Table 1 Cohort characteristic for the discovery cohort (11 366 cases and 15 489 controls)

	Cases				Controls	
Collection	All	Women (%)	ACPA+	ACPA−	All	Women (%)
UK	3768	2772 (73.57)	2360	954	8051	3504 (43.52)
Swedish EIRA	2759	1943 (70.42)	960	562	1939	1415 (72.98)
USA	2534	1908 (75.30)	1801	2726	2133	1383 (64.84)
Dutch	647	428 (66.15)	329	301	2004	844 (42.12)
Swedish UMEA	852	594 (69.72)	524	242	963	659 (68.43)
Spanish	806	599 (74.32)	396	216	399	260 (65.16)
Total	11 366	8244 (72.53)	6370	2868	15 489	8065 (52.07)

Rheumatoid arthritis (RA) cases and controls were assembled from a number of different studies from six centres across five countries. RA cases were classified as anti-citrullinated protein antibody (ACPA) positive (ACPA+) or ACPA negative (ACPA−). Swedish EIRA, Swedish Epidemiological Investigation of Rheumatoid Arthritis; Swedish Umea, Swedish.

Association of smoking with RA

Smoking data were available for 1978 anti-cyclic citrullinated peptide (CCP) positive cases and 1224 controls. We tested the association of smoking with RA in our dataset and found that smoking was not associated with an increased risk of RA (OR 0.78, 95% CI 0.67 to 0.91, p=0.002) in our data.

Analysis of wGRS

Logistic regression showed that all calculated wGRS scores were good predictors of RA (table 2). The ability of the model to discriminate correctly between individuals with RA and those without RA was measured by the AUC and was shown to be 0.74 (wGRSfull_G) (table 2) this was improved in anti-CCP-positive individuals to AUC 0.79 (table 2).

The AUC was improved by replacing the HLA-DRB1 tag SNP with imputed HLA variation at HLA-DRB1, HLA-DPB1 and HLA-B (wGRStag_G AUC 0.70, wGRSfull_G AUC 0.79) (in CCP-positive individuals) (table 2, figure 1). IDI tests showed an increase in sensitivity and specificity of 11%, and continuous net reclassification tests showed that including all variation in the HLA over a HLA tag SNP improves the probability of correctly identifying a high-risk individual to 70.09% from 64.21%, and showed an overall reclassification improvement of 66.65% between models.

Comparing the wGRSfull_G (in CCP-positive individuals) to a model containing only imputed variation in the HLA and gender (wGRSHLA_G) showed that the addition of the susceptibility SNPs to the model only slightly improved the ROC AUC (0.76 to 0.79, respectively) (table 2, figure 1).

As the wGRS performed best in anti-CCP-positive individuals, the remainder of the analysis was restricted to this subgroup.

The comparison of full model wGRSfull_G and wGRSHLA_G including HLA variation and gender using likelihood ratio tests showed a significant difference between the two (LR χ^2 1130.72, χ_{dist} $p = 1.91 \times 10^{-207}$). The overall discrimination improvement was 5% and continuous net reclassification improvement tests, which consider the reclassification of cases and controls separately showed a 44% improvement in reclassification between models ($p=4.16 \times 10^{-189}$).

The wGRSfull_G was split into five quintiles; individuals in the 5th quintile were classed as high risk (figure 2). Classifying everyone with a wGRS greater than 9.75 as high risk, 9.7% of our control population were classified into this high-risk group.

Logistic regression was used to compare individuals in the high-risk group to individuals in the lowest risk group (quintile 1) and individuals in the median-risk group (quintile 3). The results showed that the odds of developing RA were significantly increased in individuals in the highest risk group compared to individuals in the lowest risk group (OR 27.13, 95% CI 23.70 to 31.05) and individuals in the median-risk group (OR 6.38, 95% CI 5.81 to 7.02) (table 3).

The sensitivity and specificity of the wGRSfull_G was calculated (table 2), showing that, although the model has a high AUC, it still lacks in sensitivity (44%), resulting in a high proportion of individuals with RA being misclassified as non-RA.

Addition of smoking to the model

The addition of smoking to the wGRS slightly improved the AUC to 0.80 in anti-CCP-positive individuals (1978 cases (1269 smokers) and 1224 controls (851 smokers)). IDI tests showed no improvement in sensitivity and specificity, and total continuous net reclassification (by adding smoking) showed a small increase of 10.74% (p=0.0016). This minimal improvement is not surprising given the lack of association between smoking and RA in our data.

Validation

SNP genotypes for the 45 RA associated loci and a HLA tag SNP were available for 2206 RA cases and 1863 healthy controls, independent from the test cohort.

Eleven samples were excluded after HLA imputation due to multiple ambiguous alleles being assigned, two samples were excluded due to missing wGRS; 2200 cases and 1856 controls remained for analysis.

All wGRS were tested in an independent cohort of samples (2200 RA cases and 1856 controls). The full model including gender (wGRSfull_G) showed good discrimination between individuals with RA and healthy individuals (ROC AUC 0.72); however, this was less than was seen in the discovery dataset (AUC 0.73).

This was further increased in CCP-positive individuals (1406 cases, 1856 controls) (AUC 0.78). These results were similar to those seen in the discovery dataset of 11 366 RA cases and 15 489 healthy controls.

Again, comparison of the full wGRS containing all 45 SNPs, HLA variation and gender to the HLA-only model showed little difference between the models (AUC 0.78, AUC 0.76, respectively) (figure 3, table 2), showing that, as expected, the contribution of each individual susceptibility variant is minimal.

As before, the wGRS was split into quintiles and comparison of the high-risk group to the low-risk and median-risk groups showed a significantly increased risk of RA (in CCP-positive cases and healthy controls: OR 18.00, 95% CI 13.67 to 23.71, OR 4.92, 95% CI 3.87 to 6.26, respectively) (table 3).

DISCUSSION

We present the first weighted genetic risk prediction score for RA using recently published genetic loci and in the largest cohort to date. We have shown a significant association between all models tested, with the odds of RA being increased to 27.13 (95% CI 23.70 to 31.05) in CCP-positive individuals when compared to the lowest risk group (OR 6.38 when compared to the median-risk group, 95% CI 5.81 to 7.02). However, 9.70% of healthy controls were classed as high risk, showing that a substantial proportion of the general population would be classified as being at high risk of developing RA, whereas only 1% will develop the disease. The wGRSfull_G shows increased discrimination when in anti-CCP-positive cases compared to controls (AUC 0.79), and improvement on models previously tested using 31 loci and HLA-DRB1 in seropositive RA (AUC 0.65).[6] However, it must be noted that the majority of loci in the wGRS have been associated with CCP-positive disease only, particularly HLA-DRB1 alleles that predispose to autoantibody-positive erosive disease.

Replacing the single HLA-DRB1 tag SNP with the five HLA amino acids (wGRSfull_G) not only showed a significant improvement in AUC (from 0.70 to 0.79), but also showed that the inclusion of imputed amino acids at HLA-DRB1, HLA-DPB1 and HLA-B provided significantly more information and increased the reclassification of individuals by 66.65%, suggesting that the inclusion of HLA variation in risk prediction models will be essential.

The ultimate aim of genetic screening models is that they can be used in the general population to identify at-risk individuals for monitoring or intervention. In order to be of use, a risk prediction model needs to be sensitive, specific and cost efficient, and will therefore include as few markers as possible to lower

Table 2 The association of each wGRS model with RA

													Sample size	
wGRS	45 Susceptibility SNPs	HLA-DRB1 tag SNP	5 Amino acids	Smoking	Gender	OR	95% CI		AUC	Threshold	Sensitivity (%)	Specificity (%)	Cases	Controls
Discovery dataset														
wGRSfull_G	+	–	+	–	+	2.01	1.96	2.05	0.7375	>9.85	35	91	11 366	15 489
Anti-CCP-positive cases and healthy controls														
wGRS45_G	+	–	–	–	+	2.41	2.30	2.52	0.6667	>7.79	33	85	6370	15 489
wGRStag_G	+	+	–	–	+	2.52	2.42	2.62	0.7039	>8.49	37	87	6370	15 489
wGRSfull_G	+	–	+	–	+	2.42	2.35	2.49	0.7875	>9.51	44	90	6370	15 489
wGRSHLA_G		–	+	–	+	2.40	2.33	2.48	0.7563	>1.54	40	88	6370	15 489
wGRSfull_SG	+	–	+	+	+	2.64	2.45	2.85	0.8013	>10.81	28	95	1978 (1269 smoke)	1224 (851 smoke)
Anti-CCP-negative cases and healthy controls														
wGRSfull_G	+	–	+	–	+	1.43	1.38	1.48	0.6205	>9.27	32	82	2868	15 489
Validation														
All cases and controls														
wGRS_full_G	+	–	+	–	+	1.98	1.87	2.11	0.7175	>8.70	30	92	2200	1856
Anti-CCP-positive cases and healthy controls														
wGRStag_G	+	+	–	–	+	2.50	2.26	2.76	0.6937	>7.39	32	89	1406	1856
wGRSfull_G	+	–	+	–	+	2.50	2.31	2.69	0.7774	>8.73	36	92	1406	1856
wGRSHLA_G	–	–	+	–	+	2.61	2.41	2.83	0.7632	>3.53	36	92	1406	1856

The OR shown are from the logistic regression testing the association of each weighted genetic risk score (wGRS) with rheumatoid arthritis (RA). Sensitivity and specificity values were based on the high-risk threshold being set at the 5th quintile and are shown in the threshold column.

AUC, area under ROC curve; CCP, cyclic citrullinated peptide; ROC, receiver operator characteristic; SNP, single-nucleotide polymorphism; wGRSfull_G, wGRS including all 45 susceptibility SNPs, five HLA amino acids and gender; wGRStag_G, wGRS including all 45 susceptibility SNPs, HLA tag SNP and gender; wGRSHLA_G, wGRS containing five HLA amino acids and gender; wGRSfull_SG, wGRS containing 45 susceptibility SNPs, five HLA amino acids, smoking and gender.

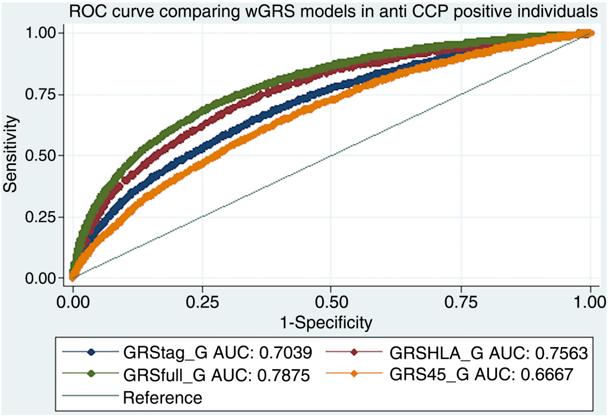

Figure 1 GRStag_G: wGRS containing 45 SNPs, a single HLA-DRB1 tag SNP and gender. GRSfull_G: wGRS containing 45 SNPs, all variation at the HLA region (HLA-DRB1, HLA-DPB1, HLA-B) and gender. GRSHLA_G: wGRS containing all variation at the HLA (HLA-DRB1, HLA-DPB1, HLA-B) and gender. GRS45_G: wGRS containing 45 susceptibility SNPs and gender. AUC, area under the curve; CCP, cyclic citrullinated peptide; ROC, receiver operator characteristic; SNP, single-nucleotide polymorphism; wGRS, weighted genetic risk score.

cost without compromising the predictive accuracy. Comparison of our full wGRS to a more practical wGRS containing only imputed HLA amino acids at HLA-DRB1, HLA-DPB1, HLA-B and gender showed that removing the 45 SNPs from the model decreased sensitivity from 44% to 40% and specificity from 90% to 88% (AUC 0.79, 0.76, respectively, table 2). This translates to 269 fewer patients being classed as high risk in the reduced model. Although a significant increase in information was gained by including the additional susceptibility markers (LR χ^2 1130.72, using 45° of freedom, $p = 1.91 \times 10^{-207}$), this statistically significant improvement in the model translates to a minimal clinical improvement, showing that each new susceptibility variant identified for RA is likely to improve the model; however, this improvement will be small. This is shown by the SNP-only model, incorporating 45 SNPs (wGRS45_G), which shows poor discriminatory ability (AUC 0.66) (figure 1).

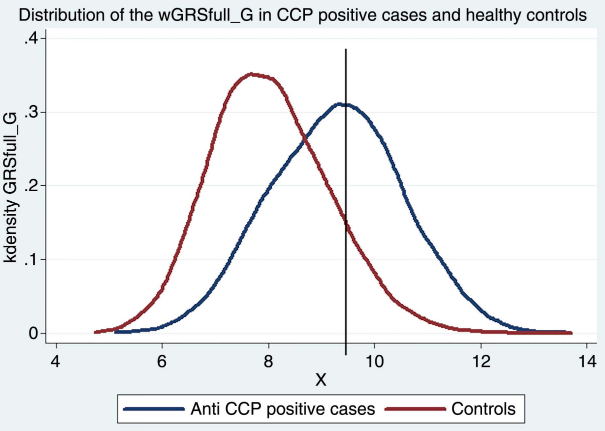

Figure 2 Distribution of the wGRS containing 45 SNPs, all variation at the HLA region (HLA-DRB1, HLA-DPB1, HLA-B) and gender in anti-CCP-positive cases (blue) and healthy controls (red) from the discovery dataset. Individuals were classed as high risk if they had a risk score greater than 9.51, indicated by the solid line. CCP, cyclic citrullinated peptide; SNP, single-nucleotide polymorphism; wGRS, weighted genetic risk score.

It should also be noted that determining the sensitivity and specificity of each model is threshold dependent. The choice of threshold by which individuals can be classed as high risk is arbitrary and can be altered to the requirements of the test.

Although ROC curves are useful to estimate the sensitivity and specificity of a model they do not take into account the prevalence of the disease in the population; therefore, even a high AUC may be of little use when the disease in question is rare in the general population. RA has a prevalence of only 1%, therefore general population screening with an AUC of approximately 80% is unlikely to prove informative.[12] Screening could be more informative in identifying individuals at risk of developing disease in groups already deemed to be at high risk due to environmental factors, such as family history and smoking. The selection of individuals for primary prevention trials is currently of great interest in RA research. Identifying individuals with a family history of RA who can then be stratified by other environmental and genetic risk factors for long-term follow-up may help identify pre-disease risk factors (serological or immunological), which will help to refine risk prediction models.

Our wGRS including all known RA susceptibility markers to date shows an improvement over previously published studies including fewer genetic susceptibility variants (Karlson *et al*[3] AUC 0.66 in an NHS cohort, 0.75 in a Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA) cohort), Kurreeman *et al*[5] AUC 0.71 and Chibnik *et al*[6] AUC 0.66). Although the improvement of our model is not substantial compared to that shown by Karlson *et al*[3] this may be due to the inclusion of clinical risk factors such as smoking and age in their model. It should also be noted that European samples used in Kurreeman *et al*[5] will have some overlap with those used in the validation stage of our study. In comparison to these earlier studies we have shown a greater increased risk of RA when we compare individuals in the highest risk group to those in the median risk group (OR 6.38, 95% CI 5.81 to 7.01) (Karlson *et al*[3] NHS cohort, OR 2.85, 95% CI 1.75 to 4.64; EIRA 3.36, 95% CI 2.27 to 4.97; Chibnik *et al*[6] OR 3.0, 95% CI 1.7 to 4.7). The results of our study (AUC 0.80) are comparable to genetic risk score analysis in other autoimmune diseases, for example coeliac disease (AUC 0.85)[13] and type one diabetes (AUC ~0.90).[14 15]

A limitation of our study is that several of the RA susceptibility SNPs were originally identified in our large discovery dataset of 11 366 cases and 15 489 controls. Therefore, the results of our modelling could be inflated due to over-fitting. However, we addressed this issue by testing our models in an independent validation dataset of 2200 cases and 1856 controls, which showed similar results.

Although we were able to include gender in our model, age is also a risk factor for RA; however, these data were not available for all subjects in our cohort and therefore were not included. It is possible that predictive accuracy would have been improved had age been included. A further limitation in the current study is that smoking status could only be classified as 'ever or never', whereas it is likely that pack-years or the smoking habits before diagnosis may be more important.

It should be noted that the OR reported in genome-wide association studies may not accurately represent the true relative risks as many of these studies include RA patients with long-term established disease and are largely seropositive. Therefore, care must be taken when evaluating the predictive accuracy of genetic models as they often perform better in the populations in which they are developed, and the performance of a model can be affected by differences in populations, genotype frequencies, phenotypic effect sizes and disease incidence.

Table 3 Logistic regression results after splitting the wGRSfull_G into quintiles (6370 anti-CCP-positive cases and 15 489 healthy controls)

Group		OR	p Value	95% CI
Compare each group to the lowest risk group (1st quintile)				
Quintile 1	Reference group (lowest risk group)	1	NA	NA
Quintile 2		2.25	<0.00001	1.94 to 2.61
Quintile 3	Median-risk group	4.25	<0.00001	3.70 to 4.88
Quintile 4		8.78	<0.00001	7.67 to 10.04
Quintile 5	High-risk group	27.13	<0.00001	23.70 to 31.05
Validation				
Quintile 5	High-risk group	18.00	<0.00001	13.67 to 23.71
Compare each quintile to the median-risk group (3rd quintile)				
Quintile 1	Lowest risk group	0.24	<0.00001	0.20 to 0.27
Quintile 2		0.53	<0.00001	0.47 to 0.59
Quintile 3	Reference group (median-risk group)	1	NA	NA
Quintile 4		2.07	<0.00001	1.88 to 2.27
Quintile 5	High-risk group	6.38	<0.00001	5.80 to 7.02
Validation				
Quintile 5	High-risk group	4.92	<0.00001	3.87 to 6.26

CCP, cyclic citrullinated peptide; wGRSfull_G, wGRS including all 45 susceptibility SNPs, five HLA amino acids and gender.

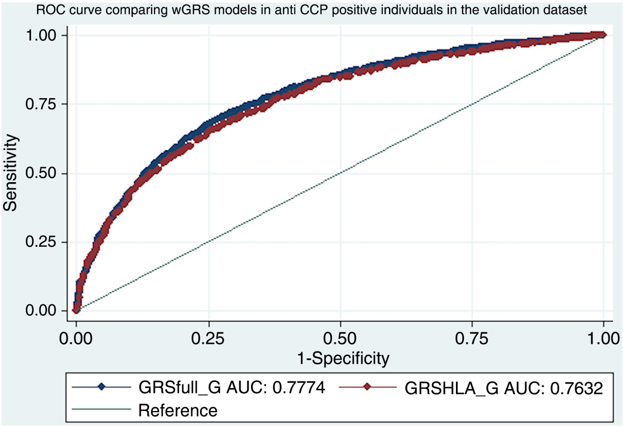

Figure 3 Comparison of wGRS in the validation dataset. GRSfull_G: wGRS containing 45 SNPs, all variation at the HLA region (HLA-DRB1, HLA-DPB1, HLA-B) and gender. GRSHLA_G: wGRS containing all variation at the HLA (HLA-DRB1, HLA-DPB1, HLA-B) and gender. AUC, area under the curve; CCP, cyclic citrullinated peptide; ROC, receiver operator characteristic; SNP, single-nucleotide polymorphism; wGRS, weighted genetic risk score.

The loci used in this study currently account for 51% of the heritability of RA,[2] and further genetic studies will account for a significantly larger proportion of heritability, although we have shown that the addition of susceptibility loci with modest effects is unlikely to improve classification significantly.

In summary, we have shown that in RA, even with all known genetic susceptibility variants (45 loci and amino acids at HLA-DRB1, HLA-DPB1 and HLA-B), prediction performance remains modest for the general population but could be used to identify at-risk individuals in suitably designed targeted screens.

Author affiliations

[1]Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, The University of Manchester, Manchester, UK
[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[3]Division of Rheumatology, Department of Medicine, New York Presbyterian Hospital, College of Physicians and Surgeons, Columbia University, New York, New York, USA
[4]Department of Rheumatology, New York University Hospital for Joint Diseases, New York, New York, USA
[5]Department of Medicine, Albany Medical College, New York, New York, USA
[6]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[7]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[8]Medical and Population Genetics Program, Broad Institute, Cambridge, Massachusetts, USA
[9]NIHR Manchester Musculoskeletal Biomedical Research Unit, Manchester Academic Health Science Centre, Central Manchester Foundation Trust, Manchester, UK

Acknowledgements The authors would like to thank Arthritis Research UK for their support (grant ref 17552). They acknowledge support from the National Institute of Health Research (NIHR) Manchester Musculoskeletal Biomedical Research Unit and from the innovative medicines initiative joint undertaking (IMI JU) funded project BeTheCure, (contract no. 115142-2). They also thank RACI, CORRONA and I2B2 for the use of samples and data.

Collaborators Rheumatoid Arthritis Consortium International.

Contributors Conception and design of study: AY, SE, AB, JW; Acquisition and analysis of data: AY, ML, JB, BH, SR, JDG, DAP, JK, RP; Interpretation of data: AY, ML, JB, SE, AB, BH, SR, RP; Manuscript preparation: AY, SE, AB, ML, JB, BH, SR, JDG, DAP, RP, JK, JW. AB and SE contributed equally to this manuscript.

Funding This study received funding from Arthritis Research UK (grant reference: 17552). AY is funded by BeTheCure (grant number: 115142-2).

Competing interests None.

Ethics approval The study was approved by the North West Ethics Committee (MREC 99/8/84).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Raychaudhuri S, Sandor C, Stahl EA, *et al*. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* 2012;44:291–6.
2 Eyre S, Bowes J, Diogo D, *et al*. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* 2012;44:1336–40.
3 Karlson EW, Chibnik LB, Kraft P, *et al*. Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk. *Ann Rheum Dis* 2010;69:1077–85.

4 van der Helm-van Mil A, Toes REM, *et al*. Genetic variants in the prediction of rheumatoid arthritis. *Ann Rheum Dis* 2010;69:1694–6.

5 Kurreeman F, Liao K, Chibnik L, *et al*. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet* 2011;88:57–69.

6 Chibnik LB, Keenan BT, Cui J, *et al*. Genetic risk score predicting risk of rheumatoid arthritis phenotypes and age of symptom onset. *PLoS One* 2011; 6:e24380.

7 Arnett FC, Edworthy SM, Bloch DA, *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988;31:315–24.

8 Pepe MS, Janes H, Longton G, *et al*. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;159:882–90.

9 Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008;54:17–23.

10 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, *et al*. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.

11 Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;30:11–21.

12 Jakobsdottir J, Gorin MB, Conley YP, *et al*. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* 2009;5:e1000337.

13 Romanos J, Rosen A, Kumar V, *et al*. Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. *Gut* 2014;63: 415–22.

14 Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet* 2009;5:e1000540.

15 Jostins L, Barrett JC. Genetic risk prediction in complex disease. *Hum Mol Genet* 2011;20:R182–8.