Genetic risk loadings influence the susceptibility and severity of systemic lupus erythematosus

Lingyan Chen[1,2], Phil Tombleson[1], Adrianna Bielowka[1], Christopher A Odhams[1], Amy L Roberts[1], Deborah S Cunninghame Graham[1], Timothy J Vyse[1,*] and David L Morris[1,*]

1. Department of Medical & Molecular Genetics, King's College London, London, UK.

2. MRC/BHF Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK.

*These are corresponding authors and joint last authors.

Abstract

Systemic lupus erythematosus (SLE) is a chronic autoimmune disease with a wide range of clinical manifestations. Of these, kidney involvement is the most common cause of morbidity and mortality. Genome-wide association studies (GWAS) have identified more than 80 loci that are associated with SLE. We calculated the genetic risk score (GRS) using SNPs that are associated with SLE. We studied three GWAS sets and found that the best GRS in the prediction of SLE generated an area under the ROC curve of 0.745 (95%CI 0.735-0.754). However, it is not known whether genetic factors affect the clinical features. We further showed a significant correlation between a GRS and renal involvement in two independent European GWAS: cohort 1 (N_{Renal+} = 1,152, N_{Renal-} = 1,949) and cohort 2 (N_{Renal+} = 146, N_{Renal-} = 378) – the higher the GRS, the higher risk of renal disease ($P_{cohort1}$ = 2.44e-08; $P_{cohort2}$ = 0.00205) and the younger age of SLE onset ($P_{cohort1}$ = 1.76e-12; $P_{cohort2}$ = 0.00384). When partitioning the patients according to the age of SLE onset, we found that the GRS performed better in the prediction of renal disease in the 'late onset' group comparing to the 'early onset' group. In conclusion, age of onset incorporating a GRS may assist prediction of lupus nephritis in a clinical setting.

Key words

Genetic risk score (GRS), Systemic lupus erythematosus (SLE), lupus nephritis, age of onset, severity

Introduction

Systemic lupus erythematosus (SLE) is a chronic inflammatory autoimmune disease characterized by a wide spectrum of signs and symptoms varying among affected individuals and can involve many organs and systems, including the skin, joints, kidneys, lungs, central nervous system, and haematopoietic system [1]. A recent report underscores that SLE is among the leading causes of death in young females, particular females among ages 15-24 years, in which SLE ranked tenth in the leading causes of death in all populations and fifth for African American and Hispanic females [2]. Lupus nephritis is the most common cause of morbidity and mortality. Patients with kidney disease are likely to have more severe clinical outcomes and a shorter lifespan. 30-60% of adults and up to 70% of children with SLE have renal disease, characterized by the glomerular deposition of immune complexes and an ensuring inflammatory response [3]. Genetic ancestry influences the incidence and prevalence of SLE and kidney involvement, being more frequent in Hispanics, Africans and Asians than in European [4-7]. Currently, kidney disease in SLE is diagnosed by use of light microscopy, which drives therapeutic decision-making. However, not all patients will respond to therapy, indicating that additional information focusing on the mechanism of tissue injury is required. Moreover, early detection of kidney involvement in SLE is important because early treatment can be applied to reduce the accumulation of renal disability.

Although the exact aetiology of lupus is not fully understood, a strong genetic link has been identified through the application of family [8, 9] and twins studies [10]. SLE does not follow a Mendelian pattern of inheritance, and so it is termed a non-Mendelian disease or complex trait. Complex traits are multi-factorial with both genetic and environmental contributions. Genome-wide association studies (GWAS) have been successfully used to investigate the genetic basis of a disease and this has dramatically advanced knowledge of the genetic aetiology of SLE. Our recent review summarized a total of 84 genetic loci that are implicated as SLE risk [11]. Despite the advances in the genetics of SLE, it is not clear how to utilise genetic information for the prediction of SLE risk or severity.

A genetic risk score (GRS) summarizes risk-associated variations by aggregating information from multiple risk single nucleotide polymorphisms (SNPs). The approach to calculate the GRS is to simply count disease-associated alleles or weighting the summed alleles by log Odds Ratios. Recent studies [12, 13] have proposed methods which select SNPs from GWAS by LD (linkage disequilibrium) pruning and clumping and thresholding for GRS calculation. As the number of SNPs included in a GRS

increases, the distribution approaches normality, even when individual risk alleles are relatively uncommon. Therefore, a GRS can be an effective means of constructing a genome-wide risk measurement that summarises an individual's genetic predisposition to SLE. Moreover, as GRSs pool information from multiple SNPs, each individual SNP does not strongly influence the summary measurement. Thus, the GRS is more robust to imperfect linkage for any tag SNP and causal SNP, and less sensitive to minor allele frequencies for individual SNPs [14-17].

In this study, we firstly tested whether a quantitative model - a GRS derived from SLE GWAS applying a range of methods, was an effective way to distinguish SLE patients and controls in three independent cohorts. Next, we classified SLE patients into two groups: SLE renal+ (patients with renal disease) and SLE renal- (patients without renal disease), and performed a case-case genome-wide association study (GWAS) in two independent SLE cohorts with available renal data for the identification of SLE renal susceptibility loci. However, no genome-wide significant genetic risk loci were identified in the SLE renal GWASs. We then tested whether a GRS derived from SLE GWAS was an effective way to distinguish SLE patients with or without renal disease in two independent cohorts.

Methods

Samples source

Samples were from three previously published SLE genome-wide association studies (GWASs) in the European population – the SLE main cohort [18], the SLEGEN cohort [19], and the Genentech cohort [20]. The SLE main cohort [18] was the biggest SLE GWAS, which consisted of 4,036 SLE patients and 6,959 healthy controls. A total number of 603,208 SNPs were available post quality control. The SLEGEN cohort [19] was carried out by The International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN) on women of European ancestry, which comprised 283,211 SNPs genotyped for 2,542 controls and 533 SLE patients. The Genentech cohort [20] was performed by Genentech on North American individuals of European descent, which comprised 487,208 SNPs genotyped for 1,165 cases and 2,107 controls.

Clinical sub-phenotypes were available for the SLE main cohort and SLEGEN cohort, which were documented according to the standard American College of Rheumatology (ACR) classification criteria.

87 Subgroups of patients with renal disease or without renal disease were identified according to the sub-
88 phenotype data using ACR classification. Following quality control, the sample size of patients with
89 renal disease, lupus nephritis (LN+) were 1,152 and 146 and patients without renal disease (LN-) were
90 1,949 and 378 in the SLE main cohort and SLEGEN cohort, respectively. More details are presented
91 in **Supplementary Table 1**.
92

93 Genome-wide association study (GWAS)

94 SLE GWAS

95 SLE GWASs were performed in genotyped SNPs including principal components consistent with the
96 original publications in all three independent cohorts.
97

98 SLE Renal GWAS within SLE cases

99 The SLE Renal GWASs were performed within SLE cases, i.e., genome-wide associations of patients
100 with renal disease (SLE Renal+, cases) and patients without renal disease (SLE Renal-, controls) in
101 two independent cohorts, i.e., the SLE main cohort and the SLEGEN cohort. For Renal GWASs, we
102 pre-phased the genotyped data using the SHAPEIT algorithm [21] and then used IMPUTE2 [22] to
103 impute to the density of the 1000 Genome reference data (phase 3 integrated set, release 20130502)
104 [23] (data unpublished). All case-control analysis was carried out using the SNPTEST algorithm [24].
105 SNPs with imputation INFO scores of < 0.7 and MAF (minor allele frequency) < 0.001 were removed.
106 After quality control (QC), there were 21,431,070 SNPs left for further analysis. Moreover, a genome-
107 wide association meta-analysis of the SLE main cohort and SLEGEN cohort was performed using the
108 summary statistics derived from the two Renal GWASs. A standard threshold of $P \leq 5e\text{-}08$ was used
109 to report genome-wide significance and a $P \leq 1e\text{-}05$ was used to report suggestive associated signals.
110

111 Polygenic analysis

112 We tested for non-zero standardised effect sizes (Z scores) for SLE association in the Genentech data
113 for groups of SNPs stratified by their *P* values in the SLE main cohort. The Z scores in the Genentech
114 data were polarized with respect to the SLE main cohort in that the effect allele was set to be the same
115 risk allele as in the SLE main cohort. Under the null hypothesis the Z scores will have zero mean, while
116 under the alternative the mean will be positive. SNPs were stratified by *P* value intervals of 1-0.9, 0.9-

117 0.8, 0.8-0.7, 0.7-0.6, 0.6-0.5, 0.5-0.4, 0.4-0.3, 0.3-0.2, 0.2-0.1, 0.1-0.00. We would expect a positive
118 mean for SNPs with very small *P* values in the main SLE data as these will be enriched for true positives,
119 while the same is not necessarily true over other *P* values ranges unless there are more widespread
120 true associations with very weak effects. We also ran this analysis on renal association standardised
121 effect sizes (Z scores) again polarised with respect to SLE association and stratified by SLE *P* values.
122 In all analyses, we used an LD clumped set of SNPs with an R^2 threshold of 0.1. When comparing the
123 SLE main cohort to the Genentech cohort or the SLEGEN cohort, we limited the clumping to SNPs that
124 overlap the GWASs.
125

126 ## Genetic risk score derivation

127 A Genetic risk score (GRS) is a quantitative trait of an individual's inherited risk based on the cumulative
128 impact of many genetic variants, which is calculated according to the method described by Hughes et
129 al [25], taking the number of risk alleles (i.e., 0, 1 or 2) for a given SNP and multiplying this by its
130 corresponding estimated effect - β coefficient, i.e. the natural log of its odds ratio (OR). The cumulative
131 risk score in each subject was calculated by summing the risk scores from the target risk loci:

132 $$\text{Genetic risk score} = \sum_{i}^{n} G_i \beta_i$$

133 where n represents the number of SLE risk loci, G_i is the number of risk alleles at a given SNP, and β_i
134 is the effect size of the risk SNP i.
135 We used two approaches to select SNPs for GRS calculation. The first approach – a weighted GRS
136 was derived from all published independent SLE risk SNPs (**Supplementary Table 2**) – including 78
137 SLE susceptibility loci (without the X chromosome), consisting of 93 SNPs outside of the MHC region
138 and 2 independent tag SNPs in the MHC region for two SLE associated HLA haplotypes. The risk allele
139 and its effect size for each SNP is derived from its original publication, which is summarized in a recent
140 review [11]. Each GRS for three SLE cohorts [18, 19, 26] was generated using R version 3.4.3.
141 The second approach – LD clumping and thresholding – was used to build 32 GRSs. Clumping and
142 thresholding scores were built using a *P* value and linkage disequilibrium (LD)-driven clumping
143 threshold in PLINK version 1.90b (www.cog-genomics.org/plink/1.9/) [27]. In brief, the algorithm forms
144 clumps around SNPs with association *P* values less than a provided threshold (Index SNPs). Each
145 clump contains all SNPs within a specified window of the index SNP that are also in LD with the index

146 SNP as determined by a provided pairwise correlation threshold (r^2) in the LD reference. The algorithm
147 loops through all index SNPs, beginning with the smallest *P* value and only allowing each SNP to appear
148 in one clump. The final output should contain the most significant disease-associated SNP for each
149 LD-based clump across the genome. Note that when performing LD clumping, we firstly removed the
150 X-chromosome and the MHC extended region (24-36MB) and kept all other autosomal SNPs. Then
151 we included the MHC region by using two tag SNPs for two well-known HLA haplotypes in SLE, i.e.
152 rs2187668 for HLA-DRB1*03:01 and rs9267992 for HLA-DRB1*15:01. A GRS was built using the
153 genotypes for the index SNPs weighted by the estimated effect sizes (β). Specifically, when training
154 the GRS in the SLE main cohort and testing in the SLEGEN cohort, we performed a GWAS on the
155 genotyped SNPs in the SLE main cohort and generated 32 lists of clumped SNPs over a set of P values
156 *(--clump-p1*: 0.1, 0.01, 1e-03, 1e-04, 1e-05, 1e-06, 1e-07,and 5e-08), r^2 *(--clump-r2:* 0.2 and 0,5) and
157 clumping radius *(--clump-kb*: 250 and 1000). The 32 list of SNPs were then used to generate 32 GRSs
158 by summing across all variants weighted by their respective effect size for samples in the SLEGEN
159 cohort. We performed this cross-validation in all three cohorts, generating six training-and-testing pairs.
160

161 **Receiver Operating Characteristic (ROC) curves for model evaluation**

162 The GRS with the best discriminative capacity was determined based on the maximal Area under the
163 ROC curve (AUC) with SLE or RENAL as the outcome and the candidate GRS as the predictor. AUC
164 confidence intervals were calculated using the '*pROC*' package within R and the difference between
165 the ROC curves was determined with the '*roc.test'* function, which used a non-parametric approach, as
166 described by De Long et al [28]. To assess the degree to which the age of SLE onset contributes to
167 the prediction of renal involvement within SLE cases, we generated ROCs as above with the GRS and
168 compared to ROC curves with SLE age onset as a single predictor and the ROC with both GRS and
169 age onset as predictor(s).
170

171 Partitioning the genetic risk of renal disease

172 Since a continuous score is difficult to interpret on an individual level when a physician needs to explain
173 the results of the GRS to a patient, we partitioned SLE patients into quintile according to genetic dosage
174 (SLE GRS). We used a chi-square test to study the association of the partitioned GRS and renal risk.

175 The odds ratios of renal risk were then calculated compared to the reference group - the first quintile
176 GRS group.
177 To test whether the GRS correlated with renal disease independently of age-of-onset, we partitioned
178 SLE patients into two groups according to their age of onset, with a cut-off at age of 30 - patients with
179 age above 30 were defined as 'Late age onset' and others as 'Early age onset'. A two-way ANOVA
180 test was then performed with the function '*aov*' in R, with *aov(GRS ~ age group * renal group)*. All
181 statistical analyses were conducted using R version 3.4.3 software (https://www.r-project.org/).
182
183

184 Results

185 The best GRS in SLE prediction

186 A GRS is a simple weighted sum of SLE risk alleles, however the choice of SNPs to use for the
187 calculation greatly effects its performance. We used three independent cohorts for cross-validation,
188 generating six training-and-testing pairs. For each training-and-testing pair of cohorts, we derived 32
189 predictors based on a clumping and thresholding method, and one additional predictor using the set of
190 SNPs that were previously reported to be associated with SLE (**Supplementary Table 2**). We then
191 evaluated the performance of the GRS as a predictor by its AUC. We found that the best GRS in the
192 prediction of SLE (with highest AUC) was the one derived from the published SLE SNPs (**Figure 1** &
193 **Supplementary Table 3**), with an AUC (95% CI) of 0.729 (0.706 - 0.753), 0.692 (0.673 - 0.71), and
194 0.745 (0.735 - 0.754) in SLE main cohort, SLEGEN cohort, and Genentech cohort, respectively. Among
195 the GRSs generated from LD clumping and thresholding, the predictor with the best discriminative
196 capacity was the one derived from SNPs clumping at *P* threshold (P_{th}) of 1e-05 in the SLE main cohort
197 and tested in both the SLEGEN and Genentech cohorts (**Figure 1** & **Supplementary Table 3**),
198 suggesting there may be more true positive signals than the genome-wide significant ones involved in
199 the risk of SLE. In fact, the predictive performance of the GRS using all pairs of training and test data
200 was maximised using SNPs below the standard genome-wide threshold (**Supplementary Table 3**).
201 This evidence for polygenicity was also seen in an analysis of the association statistics (Z scores) in
202 the Genentech GWAS polarised to the risk allele in the main GWAS, partitioned by their association *P*
203 value in the main GWAS (see methods). Here, we found evidence (**Figure 2** & **Supplementary Table**

204 **4**) against a zero mean (*P* = 3.91e-04) for the Z scores in Genentech data for SNPs with *P* values
205 between 0.3 and 0.2 in the main GWAS.
206

207 Lupus Nephritis GWAS within SLE cases

208 Lupus Nephritis (LN) occurs in approximately half of all SLE patients, and its frequency ranges from
209 25% to 75% depending on the population studied [29]. About one third of European SLE patients
210 experience renal disease [30]. Until recently, one of the most common causes of death in SLE patients
211 was kidney failure. According to the lupus severity index (LSI) using the ACR criteria developed by
212 Bello et al [31], renal involvement has the highest impact and particular strongly associated with disease
213 severity, hence we chose LN as a proxy of SLE severity in this study.
214 The within case LN GWAS in the SLE main cohort, which comprised 1152 SLE patients with renal
215 disease (LN+) and 1949 patients without renal disease (LN-), did not identify any genome-wide
216 significant associated loci ($P \leqslant$ 5e-08) (**Figure 3a**). Consistently, no inflation (genomic inflation factor:
217 λ = 1.014) was observed in the QQ plot (**Figure 3d**). Similarly, none of the SNPs reached genome-
218 wide significance in the SLEGEN cohort [19] (λ = 1.023) (**Figure 3b** & **3e**). In addition, no variant
219 passed genome-wide significance in the meta-analysis of the SLE main cohort and SLEGEN cohort for
220 Renal GWAS (λ = 0.9565) (**Figure 3c** & **3f**). Summary association statistics for SNPs with $P \leqslant$ 1e-05
221 are provided in **Supplementary tables 5 and 6**.
222 We did, however, see evidence that SNPs with very strong evidence for association with SLE (P $\leqslant$ 1e-
223 05) were associated with LN. This was evident from an analysis of the renal association statistics (Z
224 scores) polarised to the risk allele for SLE. There was strong evidence (**Figure 2** & **Supplementary**
225 **Table 4**, *P* = 8.72e-08) against a zero mean for the Renal Z scores for SNPs with P $\leqslant$ 1e-05 for SLE
226 in the main cohort. This result was replicated in the SLEGEN study with *P* = 2.42e-03 (**Figure 2** &
227 **Supplementary Table 4**). We only found evidence of renal association with SNPs showing very strong
228 evidence for association with SLE. This finding could be exploited for prediction of disease progression
229 and we explore this below.
230

231 ## Genetic risk loading of SLE is significantly higher in LN+ patients

232 While we observed that no individual SNPs were significantly associated with renal involvement in the
233 SLE cases, we did show that there was a deviation from zero mean for renal Z scores taken from SNPs
234 with very strong evidence for association with SLE. In view of this finding, we investigated the
235 correlation between the SLE GRS and renal disease in all SLE cases. To accomplish this, we used the
236 best GRS derived from a list of published SLE associated SNPs [11] for the comparison of the SLE
237 genetic risk burden in patients with and without renal disease. As expected, the GRS was higher in the
238 SLE patients compared to healthy controls in both independent cohorts (**Figure 4**).
239 A significantly higher GRS was observed in the group of patients with renal disease (LN+) compared to
240 patients without renal disease (LN-) (**Figure 4**). In the SLE main cohort, the mean (SD) of the GRS
241 was 18.1 (1.64) for LN+ patients and 17.8 (1.65) for LN- patients (P = 1.60e-07); the mean for the
242 SLEGEN cohort was 18.2 (1.66) for LN+ patients and 17.6 (1.69) for LN- patients (P = 0.0010).
243 Moreover, we saw a significant increasing trend of GRS over levels of diseases: Healthy control, LN-
244 patients, and LN+ patients, with a trend P < 1.0e-400 in the SLE main cohort and a P = 3.81e-73 in the
245 SLEGEN cohort (**Figure 4).**
246

247 ## Genetic risk of nephritis and age of onset in SLE

248 We partitioned the SLE cases into five groups according to quintiles for GRS to show the risk of renal
249 involvement. We observed over 1.5 folds higher risk of renal disease (OR = 1.58; 95% CI: 1.25 to 1.99;
250 P = 0.00015) between the top and bottom quintiles of GRS in the SLE main cohort (**Figure 5a**). This is
251 replicated in the SLEGEN cohort (**Figure 5b**), with odds ratios of 3.16 (95% CI: 1.62 to 6.13; P =
252 0.00091). A significantly earlier age of SLE onset was observed in those with renal disease compared
253 to those without renal disease. In the main cohort (**Figure 6a**), the mean (SD) for age of disease onset
254 was 29yrs (12) for LN+ patients and 35yrs (13) for LN- patients (P = 2.8e-27); the means for the
255 SLEGEN cohort (**Figure 6b**) were 28yrs (11) and 35yrs (13) for LN+ and LN-, respectively (P = 6.05e-
256 09). When testing the association of GRS with age of onset in the SLE main cohort, a significant
257 correlation was present – the higher the GRS, the earlier age of SLE onset (P = 2.4e-07). This
258 correlation was also detected in the SLEGEN cohort (P = 0.021).
259 To test whether the GRS correlated with renal disease independently of age-of-onset, we partitioned
260 SLE patients into two groups according to their age of onset, i.e. 'Late age onset' and 'Early age onset'

261 and performed a two-way ANOVA test (See Methods). The GRS was shown to positively correlate with
262 both renal disease and early age-of-onset (P_{Renal} = 7.64 × 10^{-5} and $P_{age\text{-}of\text{-}onset}$ = 1.06 × 10^{-9} in the SLE
263 main cohort; P_{Renal} = 0.0288 and $P_{age\text{-}of\text{-}onset}$ = 0.0513 in SLEGEN cohort), while we found that there was
264 no statistically significant interaction between renal and early age-of-onset in either the SLE main cohort
265 ($P_{Interaction}$ = 0.795) or the SLEGEN cohort ($P_{Interaction}$ = 0.0511) (**Supplementary Figure 1**). Notably, we
266 found that GRS was a better predictor of renal disease in the 'Late age onset' group (AUC = 0.621)
267 compared with the 'Early age onset' group (**Figure 7**).
268 Finally, we assessed the predictive ability of the partitioned SLE GRS (quintile GRS, see methods) over
269 the two age-of-onset groups. In the main SLE cohort there is a clear and significant risk effect for renal
270 involvement with increasing GRS in the 'Late age of onset' group, but no significant effect in the early
271 onset group. We observed over two fold higher risk of renal disease (OR = 2.33; 95% CI: 1.567 to
272 3.471; *P* = 3.762-05) between the upper fourth quintile and the bottom quintile in the 'Late age onset'
273 group in the SLE main cohort (**Figure 5a**). The results were similar in the SLEGEN cohort, with the risk
274 of renal disease between the top and bottom quintile of GRS being over five times (OR = 5.484; 95%
275 CI: 1.647 to 18.26; *P* = 0.006635) (**Figure 5b** & **Supplementary Table 7**) in patients of 'Late age onset'
276 but no significant differences in those with 'Early age onset'.
277
278

279 Discussion

280 GRS has been showed to be predictive for several diseases including cardiovascular disease
281 (AUC=0.81, 95%CI: 0.81-0.81) [12], inflammatory bowel disease (AUC=0.63, 95%CI: 0.62–0.64) [12]
282 and breast cancer (AUC=0.63, 95%CI: 0.63-0.65) [32]. However, in many of these applications the AUC
283 values are dependent on inclusion of age and sex for prediction and so the AUC due to genetics alone
284 would have been substantially lower [33]. We have shown that a SLE GRS using only SNPs has good
285 predictive power with AUC approaching 0.7. Our results, using three independent GWASs, shows that
286 a GRS using SNPs with association *P* values well below genome-wide levels of significance has the
287 best predictive performance. This is further evidence that SLE is a polygenic disease with many risk
288 variants as yet undiscovered, and that more powerful studies could lead to useful predictive models.
289 Genetic risk scores may also have utility in prediction of disease severity and we find evidence for this

290 to be so for SLE. Our data show that renal involvement is not related to specific genetic factors or
291 particular genes but simply to genetic load of risk alleles.
292 Until recently, the most common cause of death in SLE patients was kidney failure. Though the
293 frequency of death from kidney disease has decreased sharply due to better therapies (e.g. dialysis
294 and kidney transplantation), kidney failure is still potentially fatal in some people with SLE and causes
295 significant morbidity. According to the lupus severity index (LSI) using the ACR criteria developed by
296 Bello et al [31], renal involvement had the highest impact and particularly more strongly associated with
297 disease severity, hence we used renal involvement as a proxy of SLE severity in this study.
298 In the SLE within-case renal GWASs, we observed no genome-wide significant signals in either the
299 SLE main cohort or the SLEGEN cohort, or meta-analysis of these two. Both datasets had genetic
300 variants with less stringent *P* values ($P \leqslant$ 1e-05) for renal association, but none of them were replicated
301 in the other cohort. Considering the sample size of both cohorts are relatively small, we applied an
302 online genetic power calculator (http://zzz.bwh.harvard.edu/gpc/) to calculate the power of our current
303 sample size for the GWAS study (**Supplementary Table 8**). We assumed the effect sizes of SLE renal
304 risk alleles is similar to that seen in SLE GWAS, so the odds ratio (OR) of the risk allele would be
305 between 1.0 and 2.0. Therefore, we calculated power under a variety of parameters, including OR, risk
306 allele frequency (RAF) and alpha. As showed in **Supplementary Table 8**, we have a power of $\geqslant$ 0.8
307 to detect a genetic risk variant with an OR = 1.4 and RAF = 0.3 or an OR = 1.5 and RAF = 0.2 when
308 alpha = 5e-08. However, if we assume the renal associated variants are as weak as most of the SLE
309 associated variants (OR < 1.2), then we are under powered (< 0.8) to detect the true renal associations
310 at the GWAS significant threshold of *P* = 5e-08 in the current study.
311 We did however find evidence that SNPs most associated with SLE (*P* < 1e-05) were enriched for
312 associations with SLE renal and so we then tested the hypothesis that the genetic risk loading of SLE
313 may correlate with kidney involvement. Therefore, a genetic risk score (GRS) with the best performance
314 in SLE prediction was derived for the prediction of SLE renal disease. In both European cohorts, the
315 SLE main cohort and the SLEGEN cohort, the GRS was significantly higher in patients with renal
316 disease than patients without. In addition, patients with a higher GRS were more likely to have renal
317 involvement at a younger age, indicating the strong genetic background of SLE development. These
318 findings provide more evidence to support the opinion that younger-age onset lupus is generally more
319 severe than older-onset lupus as reported previously [34-36].

320 One may argue that if the severity of SLE is driven by multiple genes' contribution in a quantitative way,
321 the more risk alleles that are added to the model, the better the model would fit. In this study, we show
322 that a GRS is a useful tool for the classification of SLE renal+ and SLE renal- groups. The renal
323 association *P* values of the 95 SNPs (of 77 SLE risk loci) in the SLE main cohort and the SLEGEN
324 cohort are strongly inflated as shown in the QQ plots (**Supplementary Figure 2**), suggesting the
325 cumulative genetic burden from multiple SLE risk genes with modest effect.
326 Our analysis of Renal disease in SLE patients has shown that, while we find no SNPs significantly
327 associated with renal disease, the fact that SLE associated variants correlated with renal using a GRS
328 suggests that many SLE associated variants are also risk for renal involvement albeit with likely weaker
329 effects (Odds ratios). We find that the GRS and age-of-onset are correlated but the GRS is associated
330 with renal involvement independently of age-of-onset with no interaction observed. The GRS performs
331 better for predicting renal disease in patients with late age-of-onset (> 30 years). We also find that a
332 stratified GRS may be a more viable option for predicting renal disease, where we estimate significantly
333 high relative risks for those in the tails of the GRS distribution in both of our studies that had renal data.
334 This is the first study to investigate accumulated genetic risk and its relationship with the susceptibility
335 and severity of SLE. We found that the higher the GRS, the younger onset of SLE. In patients of late
336 onset, a higher GRS means patients are more likely to suffer from more severe disease. In brief, age
337 of onset incorporating a GRS may assist early prediction of lupus nephritis in a clinical setting.
338 Nevertheless, more clinical studies are needed to validate the usefulness of this application.
339
340
341 **Acknowledgements** The authors would like to thank Dr. Amy W.L Bulter and Mr. Akmal Droubi for
342 excellent clinical data administration on both SLE cohorts. This work was supported by the National
343 Institute for Health Research Biomedical Research Centre (NIHR BRC) at Guy's and St Thomas' NHS
344 Foundation and King's College London.
345 **Funding** This study is support by China Scholarship Council (CSC) and Medical Research Council
346 (MRC).
347 **Competing interests** None declared.
348 **Patients consent** Obtained.

Ethics approval Each participating centre has obtained approval from the local ethics committee for including a patient's data in the SLE cohort after the patient has given written informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

References

1. Kaul, A., et al., *Systemic lupus erythematosus.* Nat Rev Dis Primers, 2016. **2**: p. 16039.
2. Yen, E.Y. and R.R. Singh, *Brief Report: Lupus-An Unrecognized Leading Cause of Death in Young Females: A Population-Based Study Using Nationwide Death Certificates, 2000-2015.* Arthritis Rheumatol, 2018. **70**(8): p. 1251-1255.
3. Davidson, A., *What is damaging the kidney in lupus nephritis?* Nat Rev Rheumatol, 2016. **12**(3): p. 143-53.
4. Feldman, C.H., et al., *Epidemiology and sociodemographics of systemic lupus erythematosus and lupus nephritis among US adults with Medicaid coverage, 2000-2004.* Arthritis Rheum, 2013. **65**(3): p. 753-63.
5. Peschken, C.A., et al., *The 1000 Canadian faces of lupus: determinants of disease outcome in a large multiethnic cohort.* J Rheumatol, 2009. **36**(6): p. 1200-8.
6. Pons-Estel, B.A., et al., *The GLADEL multinational Latin American prospective inception cohort of 1,214 patients with systemic lupus erythematosus: ethnic and disease heterogeneity among "Hispanics".* Medicine (Baltimore), 2004. **83**(1): p. 1-17.
7. Alarcon, G.S., et al., *Baseline characteristics of a multiethnic lupus cohort: PROFILE.* Lupus, 2002. **11**(2): p. 95-101.
8. Lawrence, J.S., C.L. Martins, and G.L. Drake, *A family survey of lupus erythematosus. 1. Heritability.* J Rheumatol, 1987. **14**(5): p. 913-21.
9. Kuo, C.F., et al., *Familial Aggregation of Systemic Lupus Erythematosus and Coaggregation of Autoimmune Diseases in Affected Families.* JAMA Intern Med, 2015. **175**(9): p. 1518-26.
10. Deapen, D., et al., *A revised estimate of twin concordance in systemic lupus erythematosus.* Arthritis Rheum, 1992. **35**(3): p. 311-8.
11. Chen, L., D.L. Morris, and T.J. Vyse, *Genetic advances in systemic lupus erythematosus: an update.* Curr Opin Rheumatol, 2017.
12. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nat Genet, 2018. **50**(9): p. 1219-1224.
13. Euesden, J., C.M. Lewis, and P.F. O'Reilly, *PRSice: Polygenic Risk Score software.* Bioinformatics, 2015. **31**(9): p. 1466-8.
14. Belsky, D.W., et al., *Development and evaluation of a genetic risk score for obesity.* Biodemography Soc Biol, 2013. **59**(1): p. 85-100.
15. De Jager, P.L., et al., *Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score.* The Lancet Neurology, 2009. **8**(12): p. 1111-1119.
16. Karlson, E.W., et al., *Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk.* Ann Rheum Dis, 2010. **69**(6): p. 1077-85.
17. Yarwood, A., et al., *A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk.* Ann Rheum Dis, 2015. **74**(1): p. 170-6.
18. Bentham, J., et al., *Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus.* Nat Genet, 2015. **47**(12): p. 1457-1464.
19. International Consortium for Systemic Lupus Erythematosus, G., et al., *Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci.* Nat Genet, 2008. **40**(2): p. 204-10.
20. Hom, G., et al., *Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX.* N Engl J Med, 2008. **358**(9): p. 900-9.
21. O'Connell, J., et al., *Haplotype estimation for biobank-scale data sets.* Nat Genet, 2016. **48**(7): p. 817-20.
22. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
23. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
24. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
25. Hughes, T., et al., *Analysis of autosomal genes reveals gene-sex interactions and higher total genetic risk in men with systemic lupus erythematosus.* Ann Rheum Dis, 2012. **71**(5): p. 694-9.

411 26. Yang, W., et al., *Genome-wide association study in Asian populations identifies variants in*
412 *ETS1 and WDFY4 associated with systemic lupus erythematosus.* PLoS Genet, 2010. **6**(2): p.
413 e1000841.
414 27. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer*
415 *datasets.* Gigascience, 2015. **4**: p. 7.
416 28. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more*
417 *correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics,
418 1988. **44**(3): p. 837-45.
419 29. Cervera, R., et al., *Morbidity and mortality in systemic lupus erythematosus during a 10-year*
420 *period - A comparison of early and late manifestations in a cohort of 1,000 patients.* Medicine,
421 2003. **82**(5): p. 299-308.
422 30. Font, J., et al., *Clusters of clinical and immunologic features in systemic lupus erythematosus:*
423 *Analysis of 600 patients from a single center.* Seminars in Arthritis and Rheumatism, 2004.
424 **33**(4): p. 217-230.
425 31. Bello, G.A., et al., *Development and validation of a simple lupus severity index using ACR*
426 *criteria for classification of SLE.* Lupus Sci Med, 2016. **3**(1): p. e000136.
427 32. Mavaddat, N., et al., *Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer*
428 *Subtypes.* Am J Hum Genet, 2019. **104**(1): p. 21-34.
429 33. Curtis, D., *Clinical relevance of genome-wide polygenic score may be less than claimed.* Ann
430 Hum Genet, 2019: p. 1-4.
431 34. Lazaro, D., *Elderly-onset systemic lupus erythematosus: prevalence, clinical course and*
432 *treatment.* Drugs Aging, 2007. **24**(9): p. 701-15.
433 35. Janwityanujit, S., et al., *Age-related differences on clinical and immunological manifestations*
434 *of SLE.* Asian Pac J Allergy Immunol, 1995. **13**(2): p. 145-9.
435 36. Tomic-Lucic, A., et al., *Late-onset systemic lupus erythematosus: clinical features, course, and*
436 *prognosis.* Clin Rheumatol, 2013. **32**(7): p. 1053-8.
437

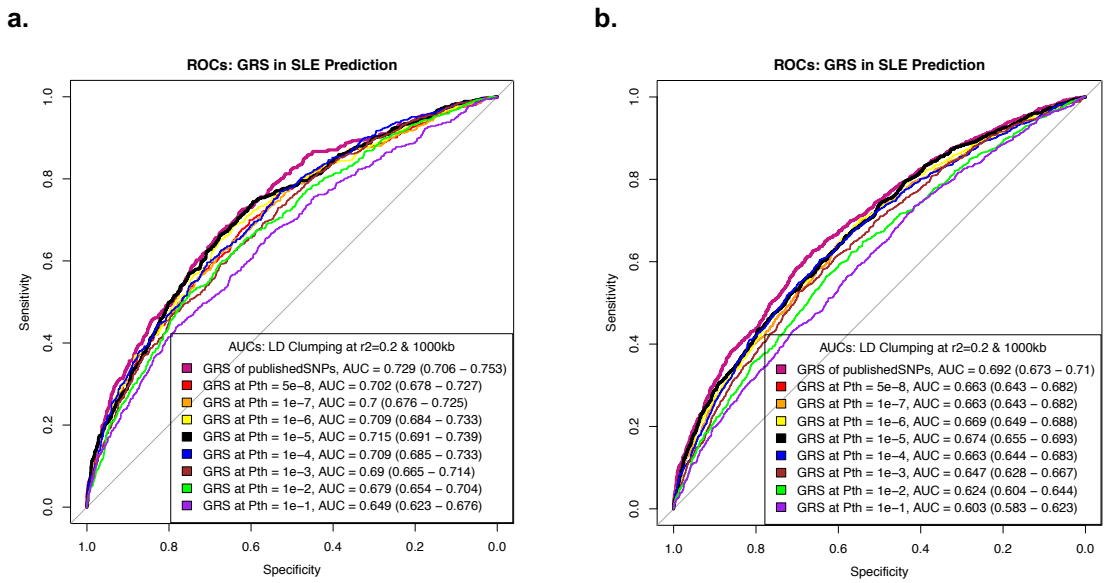


Figure 1. ROCs and AUCs of models in SLE prediction in European cohorts

GRSs for the prediction of SLE in the SLEGEN cohort (**a**) and Genentech cohort (**b**) were generated from SNPs of LD clumping and threshold derived from the SLE main cohort, and a list of published SLE risk SNPs (**Supplementary Table 2**). 'GRS at Pth' represented the GRS in the SLE prediction model was derived from the LD clumping at the according GWAS *P* value threshold.

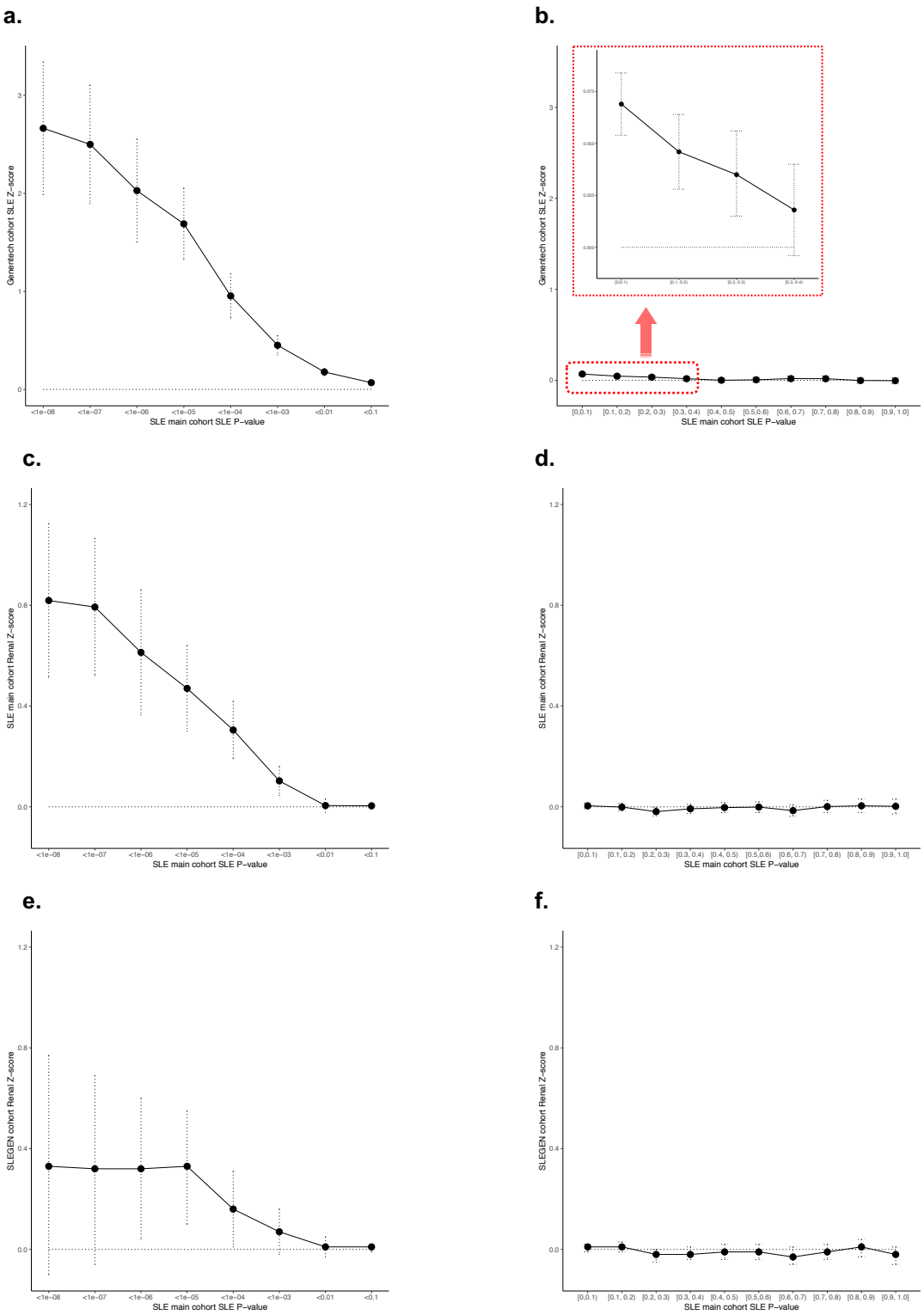


Figure 2. Polygenic test of SLE and Renal disease.

Polygenic test of SLE in Genentech cohort (**a** and **b**) and polygenic test of Renal disease in the SLE main cohort (**c** and **d**) and SLEGEN cohort (**e** and **f**). The SLE main cohort is used to generate *P* value for each SNP to stratify the SNPs into groups for the Z score calculation of SLE association or Renal association.

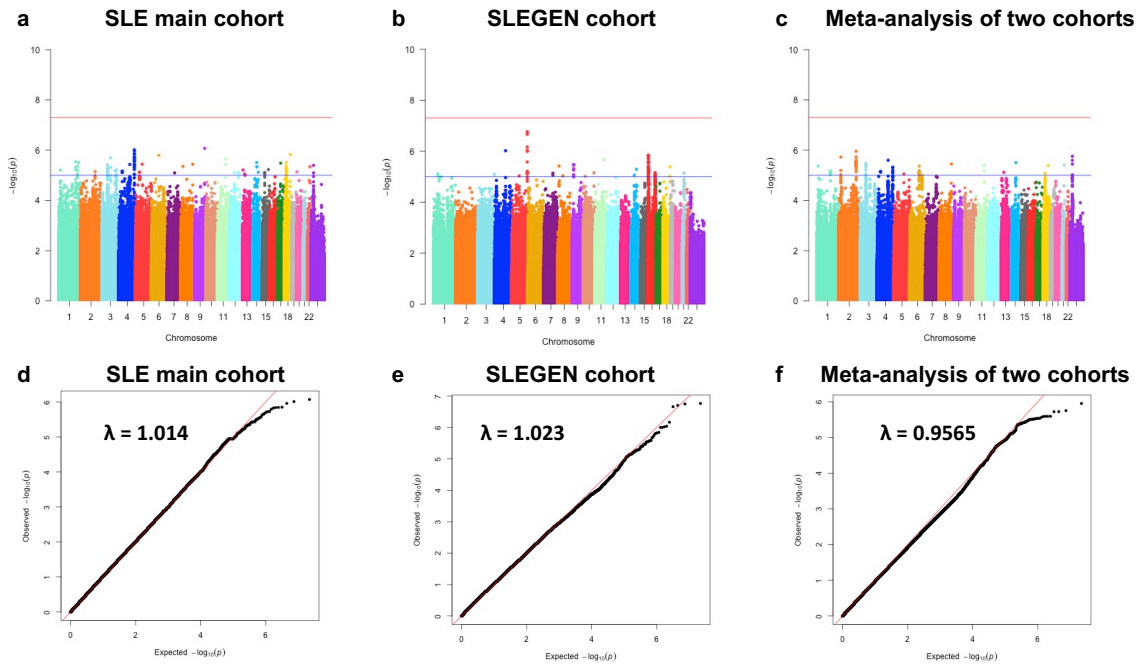


Figure 3. Genome-wide scans of LN associated variants.

(Upper) Manhattan plots showing the -log10-transformed p values (y axis) against physical genomic position (x axis) for each SNP in the SLE main cohort (**a**), the SLEGEN cohort (**b**), and the meta-analysis of these two cohorts (**c**). The red horizontal line represents the threshold for genome-wide significance (P ⩽ 5e-08) and the blue horizontal line represents the threshold for suggestive significance (P ⩽ 1e-05).

(Lower) Quantile-quantile plots showing the observed distribution of -log10-transformed p values (y axis) by the expected distribution (x axis) under the null hypothesis of no association (diagonal line) for the SLE main cohort (genomic inflation factor, λ = 1.014) (**d**), the SLEGEN cohort (λ = 1.023) (**e**), and the meta-analysis of these two cohorts (λ = 0.9565) (**f**).

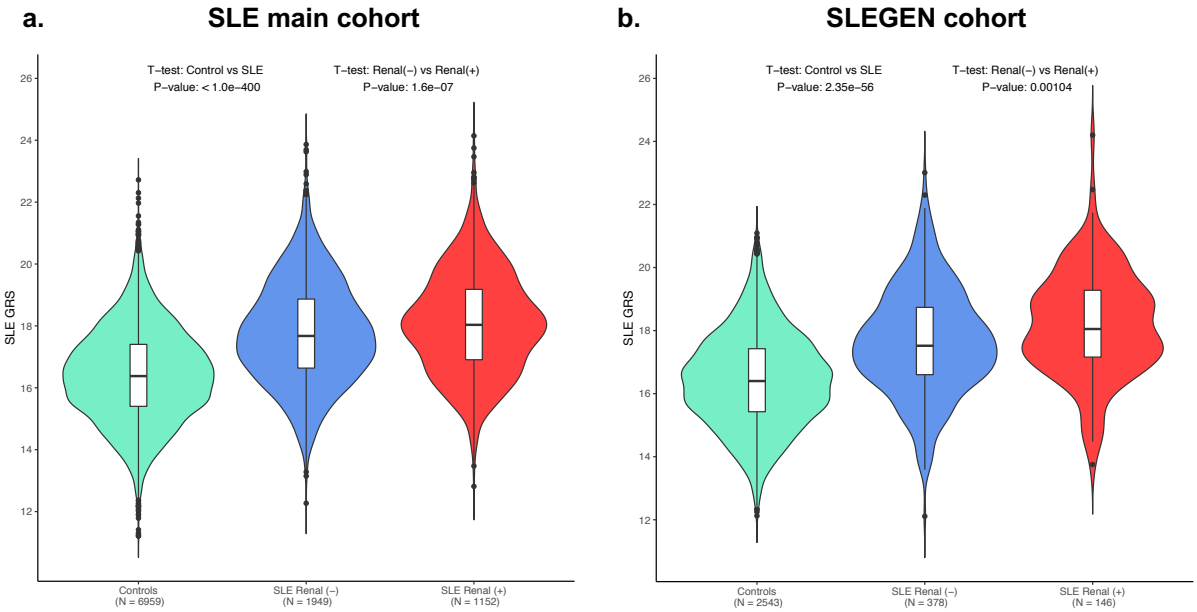


Figure 4. GRS over levels of disease: Controls / SLE Renal (-) / SLE Renal (+).
The violin-and-box plots show the summary GRS for each level of the disease in the SLE main cohort (**a**) and the SLEGEN cohort (**b**). The violins show the distribution of the GRS across each group. The bottom line of the box inside the violin is the 1st quantile, the top line is the 3rd quantile, and the box is divided at the median. Sample size (N) of each group is showed within brackets below the group name. Note that GRS for SLE main cohort and SLEGEN cohort are generated by 93 non-MHC SNPs and 2 MHC tag SNPs - a total of 95 SNPs.

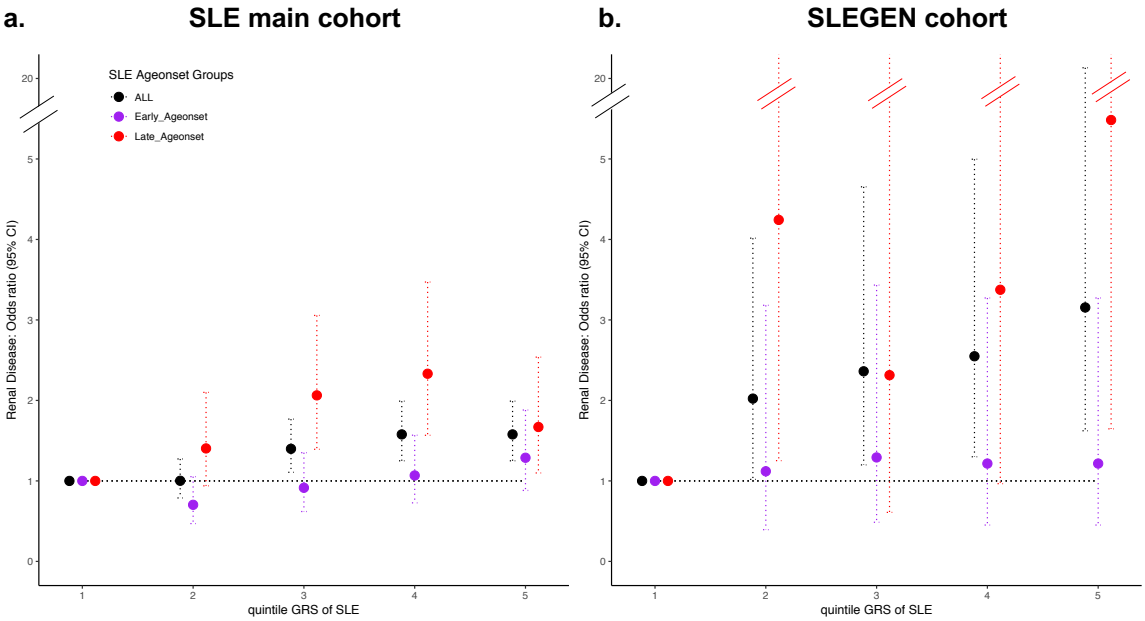


Figure 5. Relationship of quintiles of the GRS and risk of renal disease within SLE patients. Plots show the odds ratios of Renal disease for the SLE main cohort (**a**) and the SLEGEN cohort (**b**), comparing each of the upper four GRS quintiles with the lowest quintile; dotted lines represent the 95% confidence intervals; horizontal black dotted lines represent OR = 1.

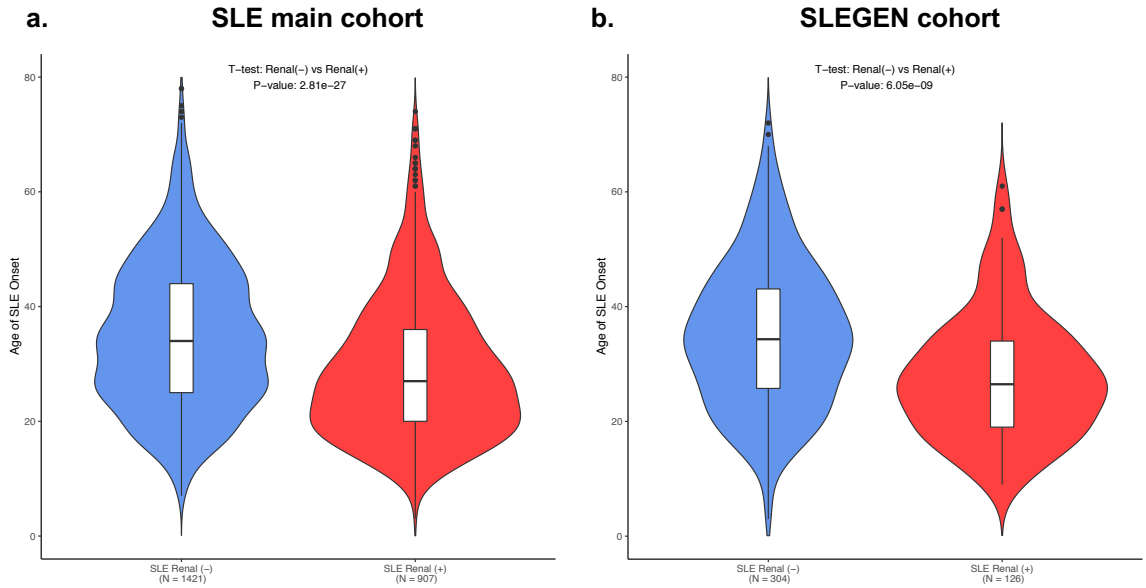


Figure 6. Age of SLE onset in patients of Renal(-) / Renal(+).

The violin-and-box plots show the age of SLE onset for each level of the disease in the SLE main cohort (**a**) and the SLEGEN cohort (**b**). The violins show the distribution of the Age of SLE onset across each group. The bottom line of the box inside the violin is the 1st quantile, the top line is the 3rd quantile, and the box is divided at the median. Sample size (N) of each group is showed within brackets below the group name.

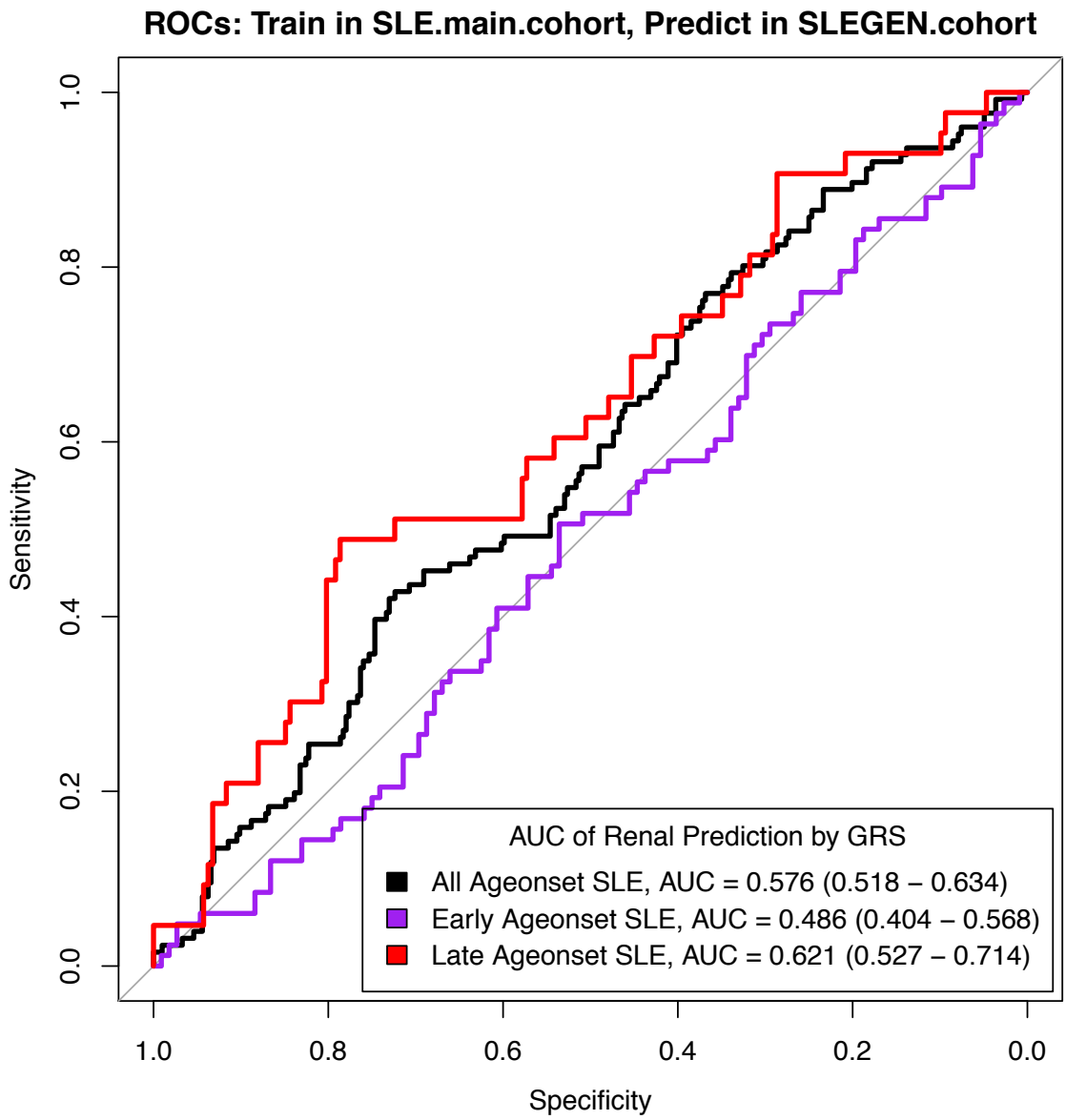
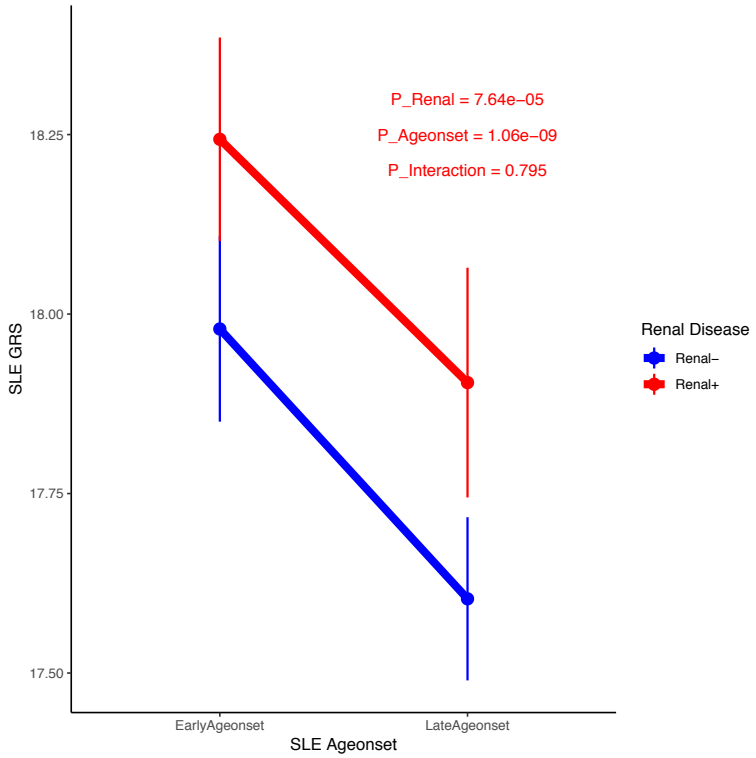
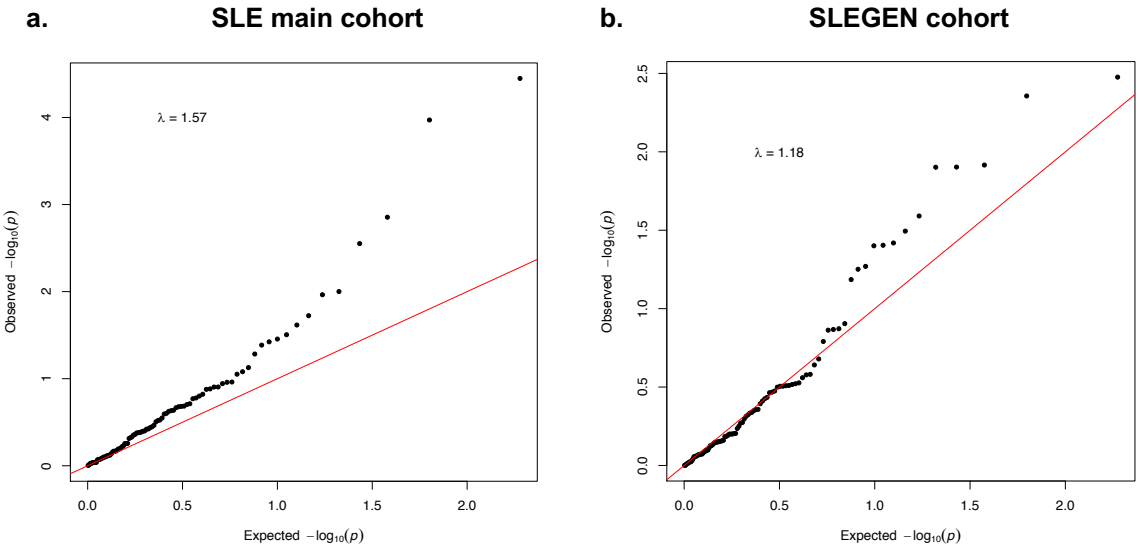


Figure 7. ROC Curves for models predicting a diagnosis of Renal disease in SLE patients using GRS, split by age-of-onset.

The models were trained in the SLE main cohort and tested in the SLEGEN cohort. The plots showed the ROC curves in the prediction of renal disease in SLE patients with GRS as a predictor, The ROC curve in black was trained and tested with all SLE samples, the purple curve was trained and tested in the 'Early age onset' patients, and the red curve was trained and tested in the 'Late age onset' group. AUC, area under the ROC curve is showed with 95% CI in brackets.



Supplementary Figure 1. Relationship of GRS and age onset in Renal disease.
The age of SLE onset <= 30 years was defined as “Early onset” and > 30 years was defined as “Late onset”. For each age onset and renal group, the GRS was plotted with mean and 95% CI for the SLE main cohort.



Supplementary Figure 2 . Quantile-quantile plots of Renal association results

QQ plots showed the observed distribution of -log10-transformed p values (y axis) by the expected distribution (x axis) under the null hypothesis of no association (diagonal line) for the SLE main cohort (**a**) and the SLEGEN cohort (**b**). The *P* values for the QQ plots were derived from Renal association test of the 95 SNPs (of 77 SLE risk loci) which used for the GRS calculation.