nature portfolio

Peer Review File

# Integrating electronic health records and GWAS summary statistics to predict the progression of autoimmune diseases from preclinical stages.

Corresponding Author: Dr Dajiang Liu

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript by Wang et al. developed a novel Genetic Progression Score (GPS) method to predict the progression from preclinical to disease states. GPS integrates the PRS constructed from case-control summary statistics by treating the PRS weights as priors with L2 penalty. Although preclinical data often has small sample sizes, the GPS can actually borrow strength from the large sample sizes of case-control studies, therefore improves prediction power over PRS methods. The authors have done very nice work in both simulations and real data analyses (SLE and RA) and have solidly demonstrated that GPS performs better than the competing methods. GPS can be more useful than PRS in clinical practice. Overall, the manuscript is well written and I only have minor comments for improving the presentation.

1) The simulation data assumes that progression phenotype and case-control phenotype share a common set of causal variants. Does GPS still perform well when progression phenotype and case-control phenotype only share a portion of causal variants? The L2 penalty to penalize the deviation between the PRS weight and the prior seems to suggest that GPS may only perform well under the common set of causal variants. In particular, the authors stated “Variants that separate healthy controls from diseases cases tend to differ from variants that separate preclinical individuals from disease cases” in Discussion.
2) The authors consider the PRS weights as priors. However, it seems GRS is nothing to do with a Bayesian method. Should a different name rather than “prior” be more reasonable? Perhaps the GPS is equivalent to a Bayesian approach. If so, some additional explanation is needed.
3) In the PheWAS analysis in the UK Biobank, the authors identified more significant PheCodes by CC-PRS than by GPS-PRS. Was UK Biobank included in the case-control GWAS of RA and SEL? If UK Biobank was a part of CC-PRS construction, it may result more bias or identifications for CC-PRS than for GPS-PRS.
4) In GPS model, log(yi=1|Xi.)=Xi.*beta. Log should be logit.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors propose a method to improve the prediction of the progression of autoimmune diseases from preclinical stages utilizing information from EHR-based biobanks while incorporating PRS weights for phenotype from large-scale case-control studies as prior via a penalized regression approach. Instead of jointly modeling SNP effect sizes in two different types of studies via penalized regression, it estimates the effect sizes in the case-control study (different PRS methods can be applied here) first, then incorporates them as prior information to estimate SNP effects for progression from preclinical stage to disease status, which is more flexible than the standard penalized regression approach. The method leads to some interesting results in two applications. But overall, the novelty, significance and impact of the study in terms of methodology and application seem to be not sufficient for publication in Nature Communications. The writing also needs to be improved.

Comments

1. Although the proposed method was compared with one multi-trait PRS method, MTAG, which was published in 2018, there are many more recent multi-trait PRS methods proposed that have shown to potentially perform better than MTAG (e.g., Xu et al., 2023, PMID: 37716346). In fact, the proposed method utilizes a penalized regression-based approach to integrating information from two types of studies, which is essentially very similar to some multi-trait methods, such as Bahda et al. (2023) (PMID: 37333772), and multi-ancestry PRS methods like PROSPER (Zhang et al., 2023). These methods need to be mentioned in the Introduction section. The author should also consider comparing the proposed method with at least one of the more recent multi-trait PRS methods, such as Bahda et al. (2023), to show that the proposed two-step method can outperform the standard penalized regression approach to combining information on two correlated traits. I'm also curious if training a weighted combination of the CC PRS and PROG PRS across different baseline models can lead to a similar or even higher R2.

2. Details about implementation of the different baseline PRS methods need to be provided. For example, what are the tuning parameter settings considered for LDpred2, lassosum, and PRS-CS? Are the software/packages of LDpred2, lassosum, and PRS-CS of the latest version? Was "PRS-CS" or "PRS-CS auto" used? The reason I ask about these questions is because it has been shown that different tuning parameter settings and different versions of the PRS algorithms (e.g., LDpred2) could lead to quite different results.

3. Based on the simulation results, GPS seems to have very notable improvement when Nprog is relatively small and the genetic correlation is relatively low. But as sample size of the biobank data increases from 1000 to 2000, such advantage appears to be a lot less notable. Does this mean the method will not be as needed in the near future?

4. It will be great if the authors can conduct PheWAS on the PRS calculated by the alternative methods (PROG, stacking, etc) and compare findings with the PheWAS findings for CC PRS and GPS PRS.

5. Tables 1 & 2: it seems like the best performing methods are all based on lassosum. For example, when using lassosum as the baseline method, PROC-lassosum, GPS-lassosum and STACKING-lassosum have very similar AUC for predicting RF positive to RA progression, and for predicting ANA positive to SLE progression, CC-lassosum, GPS-lassosum and STACKING-lassosum give very similar AUC. Does it mean that choosing the most appropriate baseline method is more important than the method used to combine information from case control study and biobank data?

6. How were the 95% confidence intervals calculated? Were they bootstrap CIs?

(Remarks on code availability)
The authors only provide code for the penalized regression step but not the step to estimate SNP weights based on data from case-control studies. The authors should also provide the code that can be used by readers to reproduce the results in the manuscript. Right now, it's unclear how the various baseline PRS models (lassosum, LDpred2, PRS-CS) were trained.

Reviewer #3

(Remarks to the Author)

Patients with autoimmune diseases may exhibit serological or other manifestations long before a full-blown disease develops and is diagnosed. However, only a portion of individuals who test positive for these features progress to full-blown disease. Predicting the risk of progression to a full-blown disease using biobank data is valuable, but the sample size is often small and the information incomplete, making it challenging to accurately model disease progress.

Case-control studies from GWAS have identified hundreds of loci associated with complex autoimmune diseases, such as RA and SLE. Polygenic Risk Scores (PRS) perform reasonably well in identifying high-risk individuals in the general population. In this study, Wang et al. developed a Genetic Progression Score (GPS) that incorporates information from biobanks and GWAS to predict progression from preclinical stages to the disease stage. GPS integrates PRS weights as prior via penalized regression, forcing the model to be similar to the prior if it helps improve prediction accuracy. Testing on both simulated data and real data from healthcare records, the model appears to perform better than other approaches, especially when the biobank data is small or when the correlation between biobank data and GWAS data is weak.

This study represents a commendable effort in using genetic information to predict progression from the preclinical stage to disease development, an area that holds significant value for future predictive and preventive medicine. The model seems to perform relatively well compared to other approaches. Interestingly, the model performs best in two scenarios: when biobank data is small, likely indicating that useful information primarily comes from GWAS, and when biobank data and GWAS data do not correlate well. In this case, it would be worthwhile to further examine the reasons behind this situation. Is it because one type of data is not as reliable? Does the progression from the preclinical stage towards the disease stage represent a very different mechanism? It may be some time before models like this see real-world application, but understanding the underlying mechanisms could be truly beneficial.

Some other minor points:

1. PRS typically stands for polygenic risk score, but the authors also introduced "predictive risk score." These two terms can easily be confused by readers, so it might be a good idea to find a way to clearly distinguish their uses.

2. I assume that the case-control data comes from European studies. As far as I know, there are many studies focused on East Asians for SLE, at least. Addressing the potential impact of population differences would be valuable.

3. Estimating Polygenic Risk Scores involves various methods. Is there a clear favorite in this case, or does it depend?

4. As acknowledged by the authors, using only RF and ANA as preclinical features may significantly impact the model's performance. Is it possible to change the selection criteria of preclinical cases to test the model's performance?

5. Each biobank may have its own biases and confounding factors, such as a preportions of different populations, age groups, genders, income levels, and environmental factors. Can the authors address the model's portability and the representativeness of BIOVU and ALL of US? I assume there could be a significant difference when using the model on a biobank with 25% African Americans versus one with 5%.

6. RA and glaucoma do not share genetic predisposition, while RA shares a strong genetic risk with other autoimmune diseases. Thus, the findings in the UKBiobank data may have different interpretations—shared genetic predisposition versus having one condition increase the risk of the other. False signals are always possible depending on the sample size, analysis method, and whether confounding factors are fully controlled for. Further analysis in this regard might be beneficial.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors addressed my comments well. I don't have additional concerns.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have thoroughly revised the manuscript. Overall, I'm satisfied with the responses to my comments. There are still some grammar issues (e.g., a total 1405 PheWAS codes" should be "a total of", "141 and 34 PheWAS codes respectively" should be "141 and 34 PheWAS codes, respectively", "softwares" should be "software", in "Code Availability", "PRS-CS(version XXX)", there should be a space between PRS-CS and "(", etc.

(Remarks on code availability)
The code and the R package look good to me.

Reviewer #3

(Remarks to the Author)
Thank you for addressing my questions and concerns. I have no further questions.

(Remarks on code availability)

**RESPONSE TO REVIEWERS**

Reviewer #1 (Remarks to the Author):

The manuscript by Wang et al. developed a novel Genetic Progression Score (GPS) method to predict the progression from preclinical to disease states. GPS integrates the PRS constructed from case-control summary statistics by treating the PRS weights as priors with L2 penalty. Although preclinical data often has small sample sizes, the GPS can actually borrow strength from the large sample sizes of case-control studies, therefore improves prediction power over PRS methods. The authors have done very nice work in both simulations and real data analyses (SLE and RA) and have solidly demonstrated that GPS performs better than the competing methods. GPS can be more useful than PRS in clinical practice. Overall, the manuscript is well written and I only have minor comments for improving the presentation.

1) The simulation data assumes that progression phenotype and case-control phenotype share a common set of causal variants. Does GPS still perform well when progression phenotype and case-control phenotype only share a portion of causal variants? The L2 penalty to penalize the deviation between the PRS weight and the prior seems to suggest that GPS may only perform well under the common set of causal variants. In particular, the authors stated "Variants that separate healthy controls from diseases cases tend to differ from variants that separate preclinical individuals from disease cases" in Discussion.

RESPONSE: Thank you for the comment and suggestion! In the revised manuscript, we further evaluated GPS's performance when only a portion of causal variants are shared between case-control and progression phenotypes. Specifically, we conducted additional simulations by setting the proportion of causal variants that are shared between two traits to be 0.25, 0.5, and 0.75 as presented in **Methods** - *Generating simulation data* (**page 13, paragraph 2**). In these new simulation scenarios, GPS remains to be the best method across different proportions of shared causal variants (Supplementary Figure S2). GPS-PRS-CS models yields the highest prediction $R^2$ which is consistent with the results where all causal variants are shared. This is not unexpected since genetic risk prediction does not require the identification of causal variants. Progression and case-control phenotype having different causal variants or having the same causal variants with different effects will both lead to different marginal effects which have similar consequences in risk predictions.

As the reviewer pointed out, we mentioned in **Discussion** that variants separating healthy control vs disease and those separating preclinical vs disease can be different. We were specifically referring to the difference in the marginal effects of variants. We modified the statement to "Variants that separate healthy controls from diseases cases tend to have different marginal effects compared to variants that separate preclinical individuals from disease cases"(**page 9, paragraph 6**). The difference in marginal effects may be due to different casual variants or same causal variants with different effects. At current sample sizes, we do not have adequate power to perform fine mapping or colocalization analysis for progression and case control traits, to definitively answer this question. Larger sample sizes may be needed.

2) The authors consider the PRS weights as priors. However, it seems GRS is nothing to do with a Bayesian method. Should a different name rather than "prior" be more reasonable? Perhaps the GPS is equivalent to a Bayesian approach. If so, some additional explanation is needed.

RESPONSE: Thank you for the comment! The reviewer correctly pointed out that our approach has Bayesian connections, just as Lasso and Ridge regression can be interpreted as a Bayesian model. We have made this point clearer in the revised manuscript (**page 11 paragraph 5**).

3) In the PheWAS analysis in the UK Biobank, the authors identified more significant PheCodes by CC-PRS than by GPS-PRS. Was UK Biobank included in the case-control GWAS of RA and SEL? If UK Biobank was a part of CC-PRS construction, it may result more bias or identifications for CC-PRS than for GPS-PRS.

RESPONSE: Thank you for the comment! UK Biobank was included in the case-control GWAS for RA and SLE. To further validate the results in an independent biobank, we have conducted PheWAS using CC-PRS and GPS-PRS in *All of Us*, which are not part of model training. Among marginally significant PheWAS results (i.e., with p-value < 0.05) in UK Biobank, the correlation with *All of Us* is strong (RA CC-PRS $r^2$=0.53, p-value<2.2x$10^{-16}$; RA GPS-PRS $r^2$=0.6, p-value<2.2x$10^{-16}$; SLE CC-PRS $r^2$=0.70, p-value=<2.2x$10^{-16}$; SLE GPS-PRS $r^2$=0.44, p-value=5x$10^{-9}$). Among the 5 PheWAS codes uniquely associated to RA GPS-PRS in UK Biobank, 4 were replicated in *All of Us* [i.e., nodular lymphoma (p-value = 0.0037), multiple sclerosis (p-value = 1.09 x10-22), other inflammatory spondylopathies (p-value = 0.018), and ankylosing spondylitis (p-value = 0.0002)]. These PheWAS codes remained insignificant for RA CC-PRS in *All of Us* (p-value > 0.05). There were no PheWAS codes uniquely associated with SLE GPS-PRS in UK Biobank. Among the 23 PheWAS codes associated with SLE GPS-PRS in UK Biobank, 17 were replicated for SLE GPS-PRS in *All of Us* (p-value < 0.05). Lastly, among the 41 PheWAS codes uniquely associated with SLE CC-PRS in UK Biobank, 27 remained insignificant in *All of Us* for SLE GPS-PRS (p-value ≥ 0.05). In summary, our results suggest PheWAS effect sizes estimated in UK Biobank can be replicated in *All of Us*, which suggests the validity of the results despite sample overlaps.

We have added the PheWAS results from *All of Us* in the Supplementary Table S6 and S8 for RA and SLE CC-PRS and GPS-PRS. We also added the comparison of effect sizes for both RA and SLE in UK Biobank and *All of Us* in Supplementary Figure S8-9. We have summarized our findings in “**Results -** PheWAS analysis in the UK Biobank and *All of US*” (**Page 9, paragraph 2**).

4) In GPS model, log(yi=1|Xi.)=Xi.*beta. Log should be logit.

RESPONSE: Thank you for pointing this out! We have fixed this in the **Methods –** *GPS model* (**page 10, paragraph 6**).

Reviewer #2 (Remarks to the Author):

The authors propose a method to improve the prediction of the progression of autoimmune diseases from preclinical stages utilizing information from EHR-based biobanks while incorporating PRS weights for phenotype from large-scale case-control studies as prior via a penalized regression approach. Instead of jointly modeling SNP effect sizes in two different types of studies via penalized regression, it estimates the effect sizes in the case-control study (different PRS methods can be applied here) first, then incorporates them as prior information to estimate SNP effects for progression from preclinical stage to disease status, which is more flexible than the standard penalized regression approach. The method leads to some interesting results in two applications. But overall, the novelty, significance and impact of the study in terms of methodology and application seem to be not sufficient for publication in Nature Communications. The writing also needs to be improved.

RESPONSE: Thank you for the comment! We appreciate the reviewer recognize that our methods lead to interesting applications. We did comprehensive revision and performed additional simulations to address the comments. Please allow us to elaborate on a few key advances brought by our method below.

1) GPS is the only method that consistently performs the best among all methods in both simulation and applied data analyses. We did comprehensive simulations studies considering different genetic architectures, sample sizes and genetic correlations and compared against 20 alternative PRS models. In all scenarios, particularly when genetic correlation is small or sample size of progression cohort is limited, GPS achieves sizable improvement over the second-best method. In real data applications, GPS models demonstrated highest accuracy as evaluated by $R^2$ and AUPRC as well as the strongest association between disease prevalence and risk score quantiles. On the other hand, some methods, while performing well in certain scenarios, yield much lower accuracy in others. For example, the model trained with case-control data only and the model trained by TL-PRS method perform well when genetic correlation is high and the sample size of progression cohort is limited (Figure 2A, gcor=0.8), but

these two models fail to yield comparable accuracies in other scenarios. Since the true model is unknown in practice, GPS stands out as the method of choice.

2) GPS can flexibly incorporate different baseline PRS methods. This feature allows GPS to outperform other transfer learning-based methods that stick with a fixed baseline PRS and have to jointly estimate weights for both case-control and progression cohorts. For example, in simulations, GPS-PRS-CS models outperform the multivariate Lassosum (MVL) in all scenarios (Figure 2). In real data analysis, none of the MVL models yields $R^2$ significantly greater than 0 (Table 1-2). Moreover, in certain scenarios, PRS-CS models consistently outperform lassosum-based models, which further underscores the importance of being able to incorporate different baseline methods as priors. For example, in simulations, when genetic correlation is high (Figure2 A-D, gcor: 0.6, 0.8), for all combination strategies, the models constructed using PRS-CS as baseline consistently yield higher prediction accuracy than the model constructed using Lassosum as the baseline method.

3) GPS model is also the only method that can effectively integrate information from case-control studies and biobanks and consistently outperforms methods using only case-control or biobank datasets. In applied data analyses, GPS is the only method that yields $R^2$ estimates that are significantly greater than zero regardless of the choice of baseline PRS methods. Prediction accuracy of GPS is always better than the PRSs trained in CC or progression cohort alone. In contrast, other methods may fail to incorporate information from case-control studies or biobanks. For example, for RA, the STACKING-Lassosum score is exactly the same as PROG-Lassosum score as the weight assigned to case-control risk scores is zero. It indicates that the stacking strategy fails to borrow strength from case-control study of RA (**page 7, paragraph 1**). In summary, our GPS method presents itself as a more effective approach to integrate prior to improve prediction accuracy.

Comments

1. Although the proposed method was compared with one multi-trait PRS method, MTAG, which was published in 2018, there are many more recent multi-trait PRS methods proposed that have shown to potentially perform better than MTAG (e.g., Xu et al., 2023, PMID: 37716346). In fact, the proposed method utilizes a penalized regression-based approach to integrating information from two types of studies, which is essentially very similar to some multi-trait methods, such as Bahda et al. (2023) (PMID: 37333772), and multi-ancestry PRS methods like PROSPER (Zhang et al., 2023). These methods need to be mentioned in the Introduction section. The author should also consider comparing the proposed method with at least one of the more recent multi-trait PRS methods, such as Bahda et al. (2023), to show that the proposed two-step method can outperform the standard penalized regression approach to combining information on two correlated traits. I'm also curious if training a weighted combination of the CC PRS and PROG PRS across different baseline models can lead to a similar or even higher R2.

RESPONSE: Thank you for the comments and suggestions! For the PROSPR method from Zhang et al., 2023, we have included in first version of our manuscript in **Results** - *Connections with other methods*. We cited the PROSPR method as an adaptation of fused lasso. In the revised manuscript, we further cited mtPGS method from Xu et.al 2023 and multivariate lassosum method from Bahda et al. (2023) in **Introduction** (**page 3**, **paragraph 4**).

As the reviewer suggested, we choose multivariate Lassosum as another multi-trait PRS method. For both simulations and real data applications, the multivariate Lassosum method consistently have lower prediction accuracy compared to GPS (Figure 2-4, Supplementary Figure S1-2 and Table 1-2).

Furthermore, we trained a stacking model combining all baseline CC PRS and PROG PRS, which we call ALL-BASE_stacking. The PRS from ALL-BASE_stacking integrates data from both CC and progression phenotype as well as three different baseline methods. Following the same principle, we also trained weighed combinations all GPS models (GPS_stacking), MTAG models (MTAG_stacking), and TL-PRS models (TL-PRS_stacking). Details can be found in **Results – *Overview of GPS*** (**page 4**, **paragraph 4**) and Supplementary Table S1. These four models are referred as super-stacking models in the revised manuscript. Among the super-stacking models, GPS_stacking yields the highest prediction accuracy in both simulations (Supplementary Figure S2-4) and real applications (Supplementary Table S3-4 and Supplementary Figure S5-6) while ALL-BASE_stacking is the second-best method. This new comparison shows that GPS continues to be a useful contributor among super-stacking models.

2. Details about implementation of the different baseline PRS methods need to be provided. For example, what are the tuning parameter settings considered for LDpred2, lassosum, and PRS-CS? Are the software/packages of LDpred2, lassosum, and PRS-CS of the latest version? Was "PRS-CS" or "PRSCS auto" used? The reason I ask about these questions is because it has been shown that different tuning parameter settings and different versions of the PRS algorithms (e.g., LDpred2) could lead to quite different results.

RESPONSE: Thank you for the comments! We provide the script for training the baseline PRS methods at https://github.com/wangc29/GPS_paper_script. The link is also included in **Methods –** *Code availability* in the revised manuscript (**Page 15, paragraph 2**). The tuning parameters for LDpred2, Lassosum and PRS-CS are selected as the ones that maximize the prediction $R^2$ in the validation cohort of progression phenotypes in BioVU (**Methods –** *Code availability*, **Page 15, paragraph 2**). We did not use PRS-CS auto, as the authors of the PRS-CS method mentioned that PRS-CS auto does not perform well when training sample is small (<10000)[2].

In terms of versions of LDpred2, Lassosum and PRS-CS, we previously provided version information in the Reporting Summary form of the original submission and they are all of the latest version. To make it clearer, we included the details in **Methods –** *Code availability* (**Page 15, paragraph 2)** in the revised manuscript.

We understand that different tuning parameters can lead to different PRS scores. **Importantly, when constructing PRS models using different combination strategies (e.g., GPS, MTAG, TL-PRS, STACKING), we use baseline PRS models trained with the same set of tuning parameters. As a result, while the methods of choosing the tuning parameters affect the accuracy of baseline PRS methods, they would not affect the comparison of different combination strategies**. The goal of our GPS method is to provide a more effective way to integrate information from case-control studies and biobanks to enhance the prediction of progression phenotypes.

3. Based on the simulation results, GPS seems to have very notable improvement when Nprog is relatively small and the genetic correlation is relatively low. But as sample size of the biobank data increases from 1000 to 2000, such advantage appears to be a lot less notable. Does this mean the method will not be as needed in the near future?

RESPONSE: Thank you for the comments! As indicated in the simulation results, the improvement achieved by GPS is still sizable even when Nprog is 2000, particularly when the genetic correlation is lower (Figure 2 and Supplementary Figure S1). In addition, for progression phenotypes from preclinical stages, the effective sample size is very small in Biobanks. For example, in *All of US*, there are only 126 patients with RA and 397 patients with preclinical RA resulting in an effective sample size of 4/(1/126+1/397)=382.6. Similarly, for SLE, the effective sample size of progression cohort in *All of US* is 603.1. Both BioVU and *All of Us* are among the largest biobanks, yet, the number of preclinical individuals and disease cases are still small. As such, our methods will be very useful for boosting the power and prediction accuracy.

This represents a general challenge of using population-based biobanks for disease studies. To define clinically meaningful phenotypes, the number of disease cases is often small even if the total sample sizes are large. Methods that can combine the strengths of the large sample sizes of case-control studies and the detailed phenotypic information from biobanks will continue to be very useful.

4. It will be great if the authors can conduct PheWAS on the PRS calculated by the alternative methods (PROG, stacking, etc) and compare findings with the PheWAS findings for CC PRS and GPS PRS.

RESPONSE: Thank you for the comment! As suggested, we have now conducted PheWAS in UK Biobank and *All of US* using the PRS calculated from remaining 21 alternative methods for RA and SLE. Of the 21 alternative methods, 7 were stacking based methods including 4 super-stacking methods (GPS_stacking, MTAG_stacking, TL-PRS_stacking, and ALL-BASE_stacking) and 3 ordinary stacking methods (STACKING-Lassosum, STACKING-LDpred2, and STACKING-PRS-CS). The number of significantly associated PheWAS codes (Bonferroni corrected p-value < 0.05) identified by each PRS method varied (UK Biobank mean 93 PheWAS codes, range 0-335; *All of Us* mean 16 PheWAS codes, range 0-129). Overall, for both RA and SLE, we observed GPS related PRS are associated with a restricted but more biologically relevant set of PheWAS codes specific to RA and SLE when compared to non-GPS PRS methods. This pattern was observed in both UK Biobank and *All of Us*, which provides further support of the specificity of GPS related PRS.

We have added the PheWAS results for all 23 PRS methods for RA and SLE in UK Biobank and *All of Us* in the Supplementary Tables S5-S8. We have summarized the PheWAS comparison for each method with GPS related PRS in Supplementary Materials section “PheWAS analysis in UK Biobank and *All of US* for 23 PRS methods” (**Page 1 and paragraph 1**) and included Supplementary Figures 10-13 to visualize the comparison using upset plots.

5. Tables 1 & 2: it seems like the best performing methods are all based on lassosum. For example, when using lassosum as the baseline method, PROC-lassosum, GPS-lassosum and STACKING-lassosum have very similar AUC for predicting RF positive to RA progression, and for predicting ANA positive to SLE progression, CC-lassosum, GPS-lassosum and STACKING-lassosum give very similar AUC. Does it mean that choosing the most appropriate baseline method is more important than the method used to combine information from case control study and biobank data?

RESPONSE: Thank you for the comments! We think both baseline methods and combination strategies are important. Specifically, we would like to stress the following points:

Lassosum is not always the best baseline methods when evaluated using the area under the precision and recall curve (AUPRC), which is more appropriate here since the sample is unbalanced. For example, for RA and when evaluating the AUPRC, the best baseline method is LDpred2 when used with GPS, with AUPRC 0.397, compared to the AUPRC of 0.376 for GPS-Lassosum. For SLE, comparing $R^2$ among alternative methods, the best baseline method is PRS-CS when used with STACKING, with $R^2$ 0.039 for STACKING-PRS-CS, compared to 0.034 for STACKING-Lassosum.

Besides, the strategies used for combining case-control and progression PRS matter a lot for the performance. For RA, as shown in Table 1, when LDpred2 is selected as the baseline method, only GPS yields significantly positive $R^2$. When Lassosum or PRS-CS is selected as the baseline method, only GPS and Stacking yields significantly positive $R^2$ and GPS models yield the highest $R^2$.

In many scenarios, GPS is the most effective method to integrate information from case-control studies, while other combining methods may end up using only the biobank datasets or yield lower prediction accuracy than using biobanks or case-control samples only. For example, we mentioned in our manuscript that for RA, the STACKING-Lassosum score is exactly the same as PROG-Lassosum score as the weights assigned to case-control risk scores is zero. It indicates that the stacking strategy fails to borrow strength from case-control study of RA (**page 7, paragraph 1**). We observe similar patterns in SLE. The $R^2$ and AUPRC for GPS-Lassosum are 0.044 and 0.124 respectively, which are higher than CC-Lassosum (0.033 and 0.112) and STACKING-Lassosum (0.034 and 0.111). GPS-Lassosum achieved sizable improvement over stacking.

6. How were the 95% confidence intervals calculated? Were they bootstrap CIs?

RESPONSE: Thank you for the comments! The 95% confidence intervals for Nagelkerke's $\boldsymbol{R}^2$ estimates are calculated using the *CI.Rsq* function in the psychometric R package (version 2.3). This function constructs confidence intervals for $\boldsymbol{R}^2$ based on an approximated standard error estimates[3]. The 95% confidence intervals for AUPRC and AUC are calculated by bootstrapping 1000 times on the testing dataset. We included these details in **Methods -** *Building PRS models for progression risk of autoimmune disorders* in the revised manuscript (**page 14, paragraph 4**).

Reviewer #2 (Remarks on code availability):

The authors only provide code for the penalized regression step but not the step to estimate SNP weights based on data from case-control studies. The authors should also provide the code that can be used by readers to reproduce the results in the manuscript. Right now, it's unclear how the various baseline PRS models (lassosum, LDpred2, PRS-CS) were trained.

RESPONSE: Thank you for the comments and suggestions! We have now provided the script for training the baseline PRS methods as well as other multi-source PRS methods. The scripts can be found at https://github.com/wangc29/GPS_paper_script.

Reviewer #3 (Remarks to the Author):

Patients with autoimmune diseases may exhibit serological or other manifestations long before a full-blown disease develops and is diagnosed. However, only a portion of individuals who test positive for these features progress to full-blown disease. Predicting the risk of progression to a full-blown disease using biobank data is valuable, but the sample size is often small and the information incomplete, making it challenging to accurately model disease progress.

Case-control studies from GWAS have identified hundreds of loci associated with complex autoimmune diseases, such as RA and SLE. Polygenic Risk Scores (PRS) perform reasonably well in identifying high risk individuals in the general population. In this study, Wang et al. developed a Genetic Progression Score (GPS) that incorporates information from biobanks and GWAS to predict progression from preclinical stages to the disease stage. GPS integrates PRS weights as prior via penalized regression, forcing the model to be similar to the prior if it helps improve prediction accuracy. Testing on both simulated data and real data from healthcare records, the model appears to perform better than other approaches, especially when the biobank data is small or when the correlation between biobank data and GWAS data is weak.

This study represents a commendable effort in using genetic information to predict progression from the preclinical stage to disease development, an area that holds significant value for future predictive and preventive medicine. The model seems to perform relatively well compared to other approaches. Interestingly, the model performs best in two scenarios: when biobank data is small, likely indicating that useful information primarily comes from GWAS, and when biobank data and GWAS data do not correlate well. In this case, it would be worthwhile to further examine the reasons behind this situation. Is it because one type of data is not as reliable? Does the progression from the preclinical stage towards the disease stage represent a very different mechanism? It may be some time before models like this see real-world application, but understanding the underlying mechanisms could be truly beneficial.

RESPONSE: Thank you for the comments and suggestions! We observed that variants that separate healthy controls from diseases cases tend to have different marginal effects compared to variants that separate preclinical individuals from disease cases (Figure 5). The difference in marginal effects may be due to different casual variants or different magnitude of effects. As other GWAS studies, distinguishing these scenarios requires colocalization analysis. Given the sample sizes of preclinical individuals, such analyses would be too underpowered. Future datasets with larger sample sizes are needed to better understand the genetic mechanisms affecting control to preclinical and those affecting preclinical to disease progressions.

Some other minor points:

1. PRS typically stands for polygenic risk score, but the authors also introduced "predictive risk score." These two terms can easily be confused by readers, so it might be a good idea to find a way to clearly distinguish their uses.

RESPONSE: Thank you for the comments. We didn't find the usage of "predictive risk score" in our manuscript. We introduced the concept of "Genetic Progression Score" and provide its definition as a score to predict disease progressions from preclinical stages (**Page 3, paragraph 6**).

2. I assume that the case-control data comes from European studies. As far as I know, there are many studies focused on East Asians for SLE, at least. Addressing the potential impact of population differences would be valuable.

RESPONSE: Thank you for the comments! Yes, publicly available GWAS summary statistics contains case-control studies from East Asian ancestry. Yet biobanks contain too few preclinical samples to train the model. Specifically, in BioVU, only 24 ANA+ samples are from East Asian ancestry, 89 ANA+ samples are from African ancestry and 18 ANA+ samples are from American ancestry. So we cannot yet evaluate the transferability issues in East Asian ancestry.

Instead, we investigated the transferability for samples of African ancestry and American ancestry, which have bigger sample sizes, in our response to comment 5. Yet, the risk scores trained in samples of European ancestry yield even higher $R^2$ in non-European population in the validation dataset. It indicates that inadequate sample sizes of non-European ancestry individuals in the validation dataset may lead to unstable results.

3. Estimating Polygenic Risk Scores involves various methods. Is there a clear favorite in this case, or does it depend?

RESPONSE: Thank you for the comments! In simulations, we found that GPS models trained with CC priors from PRS-CS always yields the highest prediction accuracy. In real data analysis of RA and SLE, when CC PRS weights trained by Lassosum and PRS-CS are used as priors for GPS, the resulting GPS models seem to work the best (Table 1-2). Lassosum and PRS-CS rely on different assumptions of the genetic architecture, with the former favoring sparse models and the latter favoring polygenic models. As in other genetic risk score studies, users may train priors using different baseline methods and when used with GPS, choose the GPS model that performs the best in a validation study.

4. As acknowledged by the authors, using only RF and ANA as preclinical features may significantly impact the model's performance. Is it possible to change the selection criteria of preclinical cases to test the model's performance?

RESPONSE: Thank you for the comment! We agree that preclinical features chosen may impact the model's performance. We chose positive test for presence of antinuclear antibodies (ANA) and rheumatoid factor (RF) because they are commonly measured in the clinic to screen for systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA). Sample sizes from other laboratory markers are much smaller. We only identified 47, 56, and 130 patients with positive measured double stranded DNA (dsDNA) antibodies, anti-smith antibodies, and cyclic citrullinated peptide (CCP) antibodies respectively in BioVU. Similarly, in *All of Us* data, we only identified 56, 0, and 17 patients with positive measured dsDNA antibodies, anti-smith antibodies, and CCP antibodies respectively in *All of US*. As we showed in the response to point 5 below, the small sample sizes of validation dataset introduce high variability in the $R^2$ estimates, making it infeasible to train or validate different models.

5. Each biobank may have its own biases and confounding factors, such as a preportions of different populations, age groups, genders, income levels, and environmental factors. Can the authors address the model's portability and the representativeness of BIOVU and ALL of US? I assume there could be a significant difference when using the model on a biobank with 25% African Americans versus one with 5%.

RESPONSE: Thank you for the comments! In our manuscript, we only focus on samples of European ancestry as we trained our models in BioVU and evaluated model performance in *All of Us*. As we will show below, the number of non-European pre-clinical samples from these biobanks are very small. The estimates of $R^2$ are noisy, making it infeasible to validate (and train) the model in non-European populations.

To assess the model portability, we have now assessed the PRS models trained with samples of European ancestry and evaluate their accuracy in non-European populations. Due to the small non-European sample sizes, the results may not be reliable and have high variances. For example, in *All of US*, only 11 RF+ samples came from East Asian ancestry.

We evaluated portability of different PRS models in non-European cohorts with sample size greater than 100 in *All of US*, including the African ancestry cohort of RF+ → RA and African and American ancestry cohort of ANA+ → SLE. Detailed results and sample size information are presented in Supplementary Table S9-10 in the revised manuscript.

In the African cohort of RF+ → RA, no PRS model has significant non-zero prediction $R^2$.

In the American cohort of ANA+ → SLE, GPS-PRS-CS ($R^2 = 0.054$) and GPS-Lassosum ($R^2 = 0.047$) yield top two $R^2$ estimates that are significantly greater than zero. Although the models are trained in European ancestry, they both yield higher $R^2$ estimates in the American cohort than in the European cohort (Table 2, $R^2$ estimates are 0.042 and 0,044, respectively ), suggesting the small sample sizes of the testing cohort may lead to high variability of $R^2$ values.

In the African cohort of ANA+ → SLE, CC-PRS-CS ($R^2 = 0.041$) and STACKING-PRS-CS ($R^2 = 0.037$) models yield top two $R^2$ estimates. Similar to above, the CC-PRS-CS trained with European samples yields bigger $R^2$ values in the African cohort than in the European cohort (Table 2, $R^2 = 0.032$ for CC-PRS-CS).

We also compare the basic demographics of the BioVU and *All of Us* biobanks and summarized the age of diagnosis of SLE, RA and ANA+ and RF+. Overall, we observed the distribution of age at RF+ and ANA+ overlaps between the two biobanks. The sex ratios are also comparable.

Table R1. Distribution of age at positive rheumatoid factor (RF+) and antinuclear antibody (ANA+) tests between BioVU and *All of US* datasets for European samples.

| | **BioVU** | ***All of US*** |
|---|---|---|
| **RF+/RA- female (controls)** | 60.6% | 76.07% |
| **RF+/RA+ female (cases)** | 70.2% | 76.98% |
| **Age at RF+** | | |
| **Mean** | 57 | 54 |

| | | |
|---|---|---|
| **Median** | 58 | 55 |
| **25th percentile** | 48 | 45 |
| **75th percentile** | 67 | 64 |
| | | |
| **ANA+/SLE- female (controls)** | 69.9% | 76.51% |
| **ANA+/SLE+ female (cases)** | 86.5% | 90.22% |
| **Age at ANA+** | | |
| **Mean** | 52 | 48 |
| **Median** | 53 | 51 |
| **25th percentile** | 41 | 40 |
| **75th percentile** | 64 | 59 |

6. RA and glaucoma do not share genetic predisposition, while RA shares a strong genetic risk with other autoimmune diseases. Thus, the findings in the UKBiobank data may have different interpretations—shared genetic predisposition versus having one condition increase the risk of the other. False signals are always possible depending on the sample size, analysis method, and whether confounding factors are fully controlled for. Further analysis in this regard might be beneficial.

RESPONSE: Thank you for the comment! Indeed, the overlapping PheWAS associations may be explained by the shared causal variants between two traits, or by the causality of one trait over the other. Additional analyses are needed to distinguish different underlying explanations. A recent Mendelian randomization study from Teng et al[4] suggests that RA may causally increase the risk of glaucoma, which helps clarify the mechanisms underlying our PheWAS associations.

REFERENCE:

1. PRS-CS github page: https://github.com/getian107/PRScs
2. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications, 10(1), 1776.
3. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.
4. Teng, M., Wang, J., Su, X., Tian, Y., Wang, J., & Zhang, Y. (2024). Causal associations between rheumatoid arthritis, cataract and glaucoma in European and East Asian populations: A bidirectional two-sample mendelian randomization study. PLOS ONE, 19(3), e0299192.

**RESPONSE TO REVIEWERS**

Reviewer #1 (Remarks to the Author):

The authors addressed my comments well. I don't have additional concerns.

RESPONSE:

We thank the reviewers for their constructive and helpful comments, which have significantly improved the quality and clarity of our manuscript.

Reviewer #2 (Remarks to the Author):

The authors have thoroughly revised the manuscript. Overall, I'm satisfied with the responses to my comments. There are still some grammar issues (e.g., a total 1405 PheWAS codes" should be "a total of", "141 and 34 PheWAS codes respectively" should be "141 and 34 PheWAS codes, respectively", "softwares" should be "software", in "Code Availability", "PRS-CS(version XXX)", there should be a space between PRS-CS and "(", etc.

RESPONSE:

We thank the reviewers for their constructive and helpful comments, which have significantly improved the quality and clarity of our manuscript.

We have also resolved the grammar issues in the manuscript including the ones that are pointed out by the reviewer.

Reviewer #3 (Remarks to the Author):

Thank you for addressing my questions and concerns. I have no further questions.

RESPONSE:

We thank the reviewers for their constructive and helpful comments, which have significantly improved the quality and clarity of our manuscript.