Benchmarking the Accuracy of Polygenic Risk Scores and their Generative Methods

Scott Kulm[1,2,3], **Jason Mezey**[4,5,*], **and Olivier Elemento**[2,3,*]

[1]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY
[2]Caryl and Israel Englander Institute of Precision Medicine, Weill Cornell Medicine, New York, NY
[3]Physiology, Biophysics and Systems Biology Graduate Program, Weill Cornell Medicine, New York, NY
[4]Department of Genetic Medicine, Weill Cornell Medicine, New York, NY
[5]Department of Computational Biology, Cornell University, Ithaca, NY
[*]corresponding authors

ABSTRACT

The estimate of an individual's genetic susceptibility to a disease can provide critical information when setting screening schedules, prescribing medication and making lifestyle change recommendations. The polygenic risk score is the predominant susceptibility metric, with many methods available to describe its construction. However, these methods have never been comprehensively compared or the predictive value of their outputs systematically assessed, leaving the clinical utility of polygenic risk scores uncertain. This study aims to resolve this uncertainty by deeply comparing the maximum possible, currently available, 15 polygenic risk scoring methods to 25 well-powered, UK Biobank derived, disease phenotypes. Our results show that simpler methods, which employ heuristics, bested complex, methods, which predominately model linkage disequilibrium. Accuracy was assessed with AUC improvement, the difference in area under the receiver operating curve generated by two logistic regression models, both of which share the covariates of age, sex, and principal components, while the second model also contains the polygenic risk score. To better determine the maximal utility of polygenic risk scores, straightforward score ensembles, which bested all methods across all traits in the training data-set, were evaluated in the withheld data-set. The score ensembles revealed that the accuracy gained by considering a polygenic risk score varied greatly, with AUC improvement greater than 0.05 for 9 traits. Many additional analyses revealed widespread pleiotropy across scores, large variations between assessment statistics, peculiar patterns amongst phenotype definitions, and wide ranges in the optimal number of variants used for scoring. If these many variable aspects of score creation can be well controlled and documented, simple methods can easily generate polygenic risk score that well predict an individual's future liability of certain diseases.

Introduction

Public health may be largely improved by accurately stratifying a population's disease risk, as the proper intensity of care can then be distributed accordingly[1]. Many clinical models incorporate age, sex and pre-existing conditions into predictions of such risk, generating useful preventative recommendations[2]. Genetic information is nearly universally excluded from these models, despite evidence that it may significantly increase prediction accuracy[3]. The few situations where genetics is considered involve specific gene mutations. For example, BRCA mutations are considered in relation to breast cancer as they significantly increase risk, and therefore those affected start mammogram screening earlier in life[4]. However, the heritable component for the vast majority of traits is not captured by a few mutations, but rather is determined by variants scattered across the entire genome[5]. The polygenic risk score condenses this diffuse information into one metric, by summing over multiple genetic locations the product between a genetic variant's externally determined effect size and the number of alleles at a variant. Polygenic risk scores have been proven to be as accurate as clinically relevant monogenic mutations[6]. Beyond direct risk prediction, polygenic risk

scores have identified sub-populations that respond well to statins[7] and lifestyle changes[8], revealed underlying cross-disease correlations of schizophrenia[9] and breast cancer[10], and exposed hidden structures within electrocardiograms[11] and magnetic resonance imaging[12]. Despite the deep and diverse upside of polygenic risk scores, various reporting inconsistencies may be preventing their maximal utility from currently being well appreciated.

The expansive array of cohorts, traits, covariates and metrics polygenic risk score may be applied within leads to a difficult assessment of their true worth. For example, one of the first polygenic risk scores was developed on a schizophrenia cohort with 6,090 individuals, generating an R^2 of 0.032[13]. A later analysis of coronary artery disease (CAD) in the 409,258 person UK Biobank, which included covariates of age and sex, led to an AUC value of 0.81[6]. From these two publications it is not immediately clear whether the CAD polygenic risk score is actually superior to the schizophrenia score, or if instead its performance reflects the large sample size or unrelated covariates. This pattern of difficult comparisons extends to numerous other scores, leaving each of their maximal, true utility unclear. In addition to the confusion produced from inconsistent reporting, various methods exist which select the component genetic variants within the polygenic risk score. The choice of which method to employ is a critical but currently difficult choice to make when developing an optimal, genetically-informed clinical model.

The selection of genetic variants which best predict a phenotype is a fundamental problem in the field of genetics. In non-human studies this problem is often resolved with mixed-model type approaches applied directly to genetic information[14]. In human studies, privacy concerns have motivated the safeguarding of genetic information. Therefore, only a small group of investigators with access to the full genetic information perform a genome wide association study. The resulting univariate associations are published in the form of summary statistics, which describe each variant's effect size and significance[15]. A naive method for creating a polygenic risk score would include every variant tested within a GWAS, denoted as the basic method in this study. However, a large proportion of these variants likely reflect, via linkage disequilibrium, a sparse true causal signal[16]. Without any actual relation to a causal mechanism, these false positive variants will increase the polygenic risk score without a proportional increase in disease risk. Alternatively, only genome-wide significant variants can be included. While vastly increasing the chance of only incorporating true positive signals, assessments have shown this approach is less accurate than using more liberal p-value thresholds[17], that is, including variants above classical statistical relevance thresholds . Further gradations between these two extremes are possible by modifying the effect sizes of variants, or including all variants within a certain statistical significance. To find the optimal set of variants and their effect sizes, multiple methods have been introduced that span from simple heuristics to complex models (Table 1).

The first variety of polygenic risk score methods developed were simple and heuristic-based. The clumping[18] method selects variants below a p-value limit and within a linkage disequilibrium range. The variations on this approach, WC-2d[19] and stackCT[20], use multiple limits simultaneously. Alternatively an empirical likelihood can be calculated and used to select variants, an approach utilized by WC-Likelihood[19], WC-Lasso[19], and Tweedie[21] methods. An entirely different branch of complex methods selects variants by attempting to approximate the results of a mixed model approach. LDpred[22] was the first such method, utilizing a MCMC algorithm for posterior effect estimation. It inspired the related methods LDpred-funct[23], PRScs[24] and AnnoPred[25]. The sBLUP[26] and SBayesR[27] specified the methodology even further by recreating a multiple regression framework. Lastly, the GraBLD[28] and lassosum[29] methods both utilized generally sophisticated algorithms, although did not attempt the same approximation as the other complex methods. Most of these methods perform a small comparison of methods within their respective publication, each concluding their own process is the most accurate. However, bias towards reporting positive results, along with unique computational environments that are difficult to replicate, may partially impact these claims.

A few studies have attempted to compare methods or assess wide-scale polygenic risk score accuracy, but all have been limited in the scope of their analyses. The majority of analyses focus on correlations between scores[30,31] rather than predictive performance. Studies which do assess accuracy either utilize only a limited number of variants[32], compare just a couple of methods[6] (often LDpred and clumping), or analyze a single trait[7,8]. Additionally, the majority of polygenic risk score studies

do not consider or provide supplementary analyses that enable easy comparisons among results within the current literature. For example, a disagreement on whether self-reported traits should be counted toward the total number of cases creates two types of analyses that cannot be reconciled[33]. Without a comprehensive polygenic risk score study these disparate analyses will likely continue, minimizing the contribution scores can make in improving clinical care.

To remedy the current lack of knowledge in the field this study unbiasedly ranks the performance of polygenic risk score methods and reports the best predictive accuracy for a maximal number of well-powered traits. A depth of context is further provided to cover all assessment metrics, unaccounted covariates, possible correlations and other potential caveats that facilitate reproducible or innovative analyses. Future polygenic risk score practitioners may reference these results to simply determine which method is best and if it can produce the accuracy needed in their work. In certain situations the interpretation of these results may warrant the consideration or even inclusion of polygenic risk scores into clinical decision making.

Results

Comparison of Methods

Polygenic risk scores were generated from the maximum available 25 sets of well-powered GWAS summary statistics and 15 scoring methods, applied to UK Biobank data[34]. The comparison of all polygenic risk scores revealed that simpler, heuristic-based methods, generated the most accurate polygenic risk scores, as they produced the highest area under the curve (AUC) improvement in 21 out 25 traits (Fig. 1). AUC improvement is defined as the difference in area under the receiver operating curve generated by two logistic regression models. Both models share the covariates of age, sex, and principal components, while the second model also contains the polygenic risk score. The methods which produced the highest AUC improvement in the largest number of traits was a tie between clumping and WC-2d, each chosen as the best method in 10 traits (Fig. 1b). However, the WC-2d method produced the highest average AUC improvement (mean = 0.046, 95% CI 0.029-0.063) when looking across best parameter performance for each trait (Fig. 1).

Various other statistics were also used to assess predictive performance, including: a p-Value extracted from the polygenic risk score term in the full logistic regression model; Nagelgerke's R^2 calculated from a logistic regression model with the polygenic risk score as the independent variable; and odds ratios calculated by setting the exposed group as individuals with a polygenic risk score above a specified threshold, non-exposed group as individuals with a polygenic risk score below the 50th percentile and cases/controls as determined by the phenotype definition (Supp. Table 1-4). When examining the best respective score for each trait (Supp. Fig. 1-8), the various statistics used to assess disease prediction showed a wide range of relationships (Supp. Fig. 9-11). When examining the average score for each trait the AUC improvement was well correlated to both odds ratios (rho = 0.82, $p < 0.01$), and Nagelgerke's R^2 (rho = 0.80, $p < 0.01$), whereas correlation of AUC improvement to p-value was insignificant (rho = 0.26, $p = 0.20$), (Supp. Fig. 10). Alternatively, when examining the average score for each method, the various types of statistics all showed a similar relationship - uniform high correlations between all pairwise combinations of statistics ($p < 0.01$ for all metrics) (Supp. Fig. 11).

For each trait an ensemble score, created by taking an average of the top 5 polygenic risk scores weighted by their coefficients in a cross-validated logistic regression, generated a higher AUC improvement than any single method. On average the AUC improvement was 22% (95% CI 10%-35%), larger than that of the next best method (Fig. 1c, Supp. Fig. 12-14). However, when this ensemble technique was expanded to include far more than five scores, as implemented in the stackCT method, overfitting was observed. Therefore, the stackCT method was removed from all further analyses (Supp. Fig. 15).

Evaluation of Polygenic Risk Score Accuracy

For each trait, the method that corresponds to the best polygenic risk score, defined by producing the largest AUC improvement within the training data-set, was used to make predictions within the witheld data-set. Subsequent statistics revealed substantial predictive ability in several traits, and widespread variability in analyses of the consistency across statistics, phenotyping method, and number of variants in the component score. The AUC improvement produced among all 25 traits ranged from 0.002 to 0.15 (quantiles: 0.017, 0.029, 0.090) (Table 2). The traits at the top of this range often had poor total AUC, generated

from a model including age and sex covariates, as the overall correlation between total and improved AUC is 0.077 ($p = 0.72$). An example of this trend is coronary artery disease, which can be well predicted through age and sex, generating a covariate model AUC of 0.758, leaving the polygenic risk score only able to improve the AUC by 0.017. On the other hand, Crohn's Disease is less affected by age and sex, thereby generating a covariate AUC of 0.505, leaving the polygenic risk score to improve the AUC by 0.152. Out of 25 traits, genetic information improved AUC values by 0.05 or more in 9 traits.

Other statistics measuring the accuracy of polygenic risk scores demonstrated a similar variability among traits (Supp. Figs. 16-25). For example, when calculating odds ratios with a cut-off at the 1st percentile, the median value across all traits was 4.40, with three traits having an odds ratio above 10 (Supp. Fig. 18). The mean improvement in Nagelgerke's R^2 value was 0.055 (quantiles: 0.23, 0.044, 0.061) (Supp. Figs. 19, 20). The highest polygenic risk score p-value, generated from the logistic regression including age and sex covariates, was less than $5x10^{-8}$, indicating that all polygenic risk scores were significantly associated with their respective trait, even if they did not form a (Supp. Fig. 22). While these values were generated in a withheld testing group to prevent overfitting, a comparison between testing and training performance showed a slight drop in the testing group (linear slope = 0.64)(Supp. Fig. 26). This relative decrease in accuracy could be caused by slight overfitting in the training set, as 81 scores were compared to each trait.

Context of Prediction Analyses

Correlation analyses of the 25 polygenic risk scores evaluated within the withheld data-set indicated widespread pleiotropy. The first analysis correlated the 25 scores to each other in a pairwise fashion. There were potentially expected correlations between Crohn's Disease and ulcerative colitis, psoriaris and vitiligo, and amongst many autoimmune traits (Fig. 2a). Secondly, pair-wise correlations between the 25 previously analyzed scores to 23 other polygenic risk scores, including height, weight and blood protein levels, were calculated. Correlation p-values less than the Bonferroni Correction threshold were generated in 33% of pairs, and 23 of the 25 previously evaluated scores (Supp. Fig. 27). Thirdly, the 25 previously analyzed scores were used to predict in a pair-wise fashion 206 other traits (Supp. Fig. 28-31). A total of 141 predictions generated significant associations at the Bonferroni level (Supp. Fig. 32), according to the p-value of the score term in a logistic regression model which includes age and sex covariates. Most of these significant associations capture related disease etiology, such as the coronary artery disease polygenic risk score predicting angina, myocardial infarction and hypertension, all with a p-value less than 10^{-93}. Other associations have literature evidence, but their underlying mechanisms are less well understood, including how breast cancer is associated to meningitis[35] and schizophrenia is associated to irritable bowel syndrome[36]. The majority of comparisons however show insignificant or otherwise currently unexplainable relationships.

Beyond the assessment of trait prediction accuracy, additional meta-analyses revealed further variation and complexities within a typical polygenic risk score analysis. The first meta-analysis examined the sets of variants that generated the 25 polygenic risk scores evaluated within the withheld data-set (Supp. Fig. 33-37). The number of variants in each set ranged widely from 9 to 1,095,393 (mean 161,353; sd 322,134)(Table 2, Supp. Fig. 33). The predominant pattern among variants is to have many with low effect size and a few with high effect size (Supp. Fig. 34). The cumulative effect of the variants in the bottom three quartiles is larger than the cumulative effect of the those in the fourth quartile in all traits (Supp. Fig. 37). The second meta-analysis investigated associations between polygenic risk score accuracy and meta-statistics, including the number of variants and sample size in the GWAS. In these analyses, the only meta-statistic that consistently saw an association to the withheld data-set AUC improvement was the trait's heritability, producing a correlation of 0.53 ($p = 0.007$) (Fig. 2c, Supp. Fig. 38). A similar trend emerged when analyzing each method separately (Supp. Fig. 39). The third meta-analysis examined polygenic risk score performance across 4 different phenotype definitions (Supp. Table. 1-4, Supp. Fig. 40-46). In terms of the number of cases determined, the Dual-agreement and ICD methods, and the Self-report and Medication methods were similar (Supp. Fig. 40, 41). The definition used did not change the overall performance when comparing the AUC improvement for each trait in the withheld data-set ($p = 0.95$) (Fig. 2b). However, several individual traits did see large changes in prediction accuracy, including vitiligo and eczema whose AUC improvement values across the phenotyping methods ranged by more than 0.05. This clear lack of relationship between sample selectivity and accuracy highlights the difficulty in drawing simplistic

conclusions from these meta-analyses.

Discussion

Polygenic risk scores are easy to produce, simple to understand, and in the literature they are currently the most prominent metric for reporting disease risk attributable to genetics[37]. Unfortunately, their utility is often either convoluted by additional covariates[6], overstated due to overfitting[38,39], poorly constructed due to simplistic selection of underlying SNPs[40,41], or used in sensitive social issues[42,43]. This work aims to resolve these issues by providing a comprehensive yet straightforward analysis of polygenic risk score methods on a multitude of traits while providing extensive supplementary analyses. The associated findings lead to the following conclusions.

Out of the current polygenic risk scoring methods available, the simpler methods are superior. The low performance of sophisticated methods may not be due to incorrect genetic theory but rather caused by poor documentation, leading to improper implementation, and inferior results. One of the individual methods that was selected most often as the best polygenic risk scores is clumping. The combination of high accuracy and ease in application makes clumping a preferred first-line method. Only when maximal accuracy is absolutely demanded might complex methods warrant the effort of implementation. Alternatively, high accuracies can be obtained by creating an ensemble from the outputs of multiple simple methods. In the ensemble construction not only are labeled samples required, but also care must be taken to limit overfitting. In our work overfitting was limited by only including the best five individual scores in an ensemble. However, there was likely still some minor overfitting due to the large number of scores compared in the training data-set.

Polygenic risk score performance in the withheld data-set indicated that several could be considered clinically relevant, and many others could serve as valuable measurements in related analyses. However, this utility should be framed not only by many possible confounders[44], but also by the large variability found across all supplementary analyses. The sets of SNPs which produced the highest accuracy in the training set ranged in size from single digits to over a million. The withheld data-set AUC improvement ranged from values just above zero, indicating genetics provide negligible aid, to above 0.1, indicating genetics could have a beneficial place within the clinic[45]. The multiple statistics used to evaluate polygenic risk performance also often ranged widely in ranking the genetic predictability of a trait. Interestingly, this variability was also proven to be unrelated to many meta-statistics, including sample size and number of components SNPs. This consistency of variability in polygenic risk score analysis demonstrates that widespread assessments or declarations on polygenic risk scores require context on the trait, assessment statistic, SNPs analyzed and underlying GWAS utilized. Polygenic risk scores should not be assumed to be infallible representations of a trait's genetic components, but should also not be assumed to be inconsequential metrics.

When analyzing polygenic risk scores either through correlations to other scores or predicting additional traits, significant associations are often uncovered. These associations are usually within logical contexts, such as atrial fibrillation and coronary artery disease. Although, there are many other less obvious associations that are also significant, indicating widespread pleiotropy throughout the genome. This pleiotropy may be utilized to improve predictions by incorporating the genetic risk of a secondary trait into an analysis of the primary, related disease.

These conclusions are limited in their certainty because only a snapshot of a highly uniform prospective cohort was analyzed. The snapshot nature of the data could have been theoretically better modeled using a survival analysis, but a logistic regression model was chosen to follow precedent[6]. The uniformity is caused by the selection of only white, British individuals in this study. While narrowing the population follows literature precedence, and limits possible confounders, this practice creates many additional problems. One recently highlighted problem is participation bias[46], which may have caused starkly different results in certain traits, such as Schizophrenia, as compared to other studies. An additional large problem is that the portability of these results into other populations is unknown. On a wider scale, this problem will exacerbate health disparities[47]. Follow-up work should be performed to both test the polygenic risk score portability into multi-ethnic cohorts and to develop methods that can maximize the multi-ethnic utility of a single ancestry derived polygenic risk score.

This study produced polygenic risk scores across 25 traits and 15 methods. Simple polygenic risk score methods were

proven superior, as they were able to provide a sizeable accuracy boost for several traits when also considering basic covariates. These specific traits may be well suited for the consideration of genetics in the clinic - not as a solitary test, but rather as a supplementary measure to tailor screening schedules, prescription guidelines or lifestyle change recommendations. These changes may ultimately improve the health of an entire population.

Methods

Creation of the polygenic risk scores

All polygenic risk scores were derived from two sources of data: summary statistics provided in the GWAS Catalog[48] (Table 2), and genetic/phenotypic information within the UK Biobank[34]. The summary statistics, which were derived in genome wide association studies that did not contain the UK Biobank, were formatted to remove multi-allelic, multi-nucleotide, ambiguous, low minor allele frequency (< 0.01) or low imputation quality (INFO < 0.3) variants. The variants were further munged following Linkage Disequilibrium Score Regression (LDSC) methods[45]. The allele symbols were flipped when necessary to match the UK Biobank. Any reported odds ratios were logged to create a consistent effect size across traits.

Only individuals whose quality control reported white-British ancestry, not heterozygosity outliers, not holding sex chromosome aneuploidy, and matching reported and inferred sex were included. These individuals were split, based on precedence[6], into a training (N = 129,568) and testing (N = 278,685), following the phase 1 (batches UK BiLEVE 1-11 and batch 1-22) and phase 2 (batches 23-95) of release.

Once all summary statistics and genetic data were prepared, methods were applied to the training group following documentation outlined in the respective publication, code repository, or other online guide (Table 1). Some methods required external reference data, subset of the UK Biobank data, or computational output from external methods. Each method produced adjusted summary statistics, which were then used to generate the polygenic risk scores for both the training and testing groups. Adjustments for population structure, or additional covariates, were not made directly to the polygenic risk score but rather were included explicitly in the downstream analysis models. Traits in which only a portion of the population are most commonly susceptible were subset, namely breast and ovarian cancer were scored on females only and prostate cancer was scored on males only. Details on how each method was implemented can be found in the supplementary appendix II and Figure 3.

Comparison of Polygenic Risk Score Methods

Once all polygenic risk scores were generated, the best score for each trait was determined via a cross validation framework. In total, 81 scores were created for each trait, as multiple scores for each method were required to assess an array of possible parameters. Threefold cross validation produced logistic regression models within two thirds of the training group, which were then applied to the remaining one third of the training group for assessment. The logistic regression models were either termed covariate (including age, sex, first four principal components of the genetic variant matrix), score (including the polygenic risk score), or complete (including age, sex, first four principal components of the genetic variant matrix and polygenic risk score). The independent variable for all regressions was the phenotype, determined through ICD codes, OPCS codes and self-reported disease status (Supp. Tables 1-4). The primary statistic used to compare each score was AUC improvement, the difference between the AUCs generated from the complete and covariate models. Additional statistics such as odds ratio, p-value and Nagelkere's R^2 were computed. Odds ratios were calculated by considering individuals whose polygenic risk score was above the denoted cut off percentile to be in the exposure group and individuals with a polygenic risk score below the 50th percentile to be in the non-exposure group. p-values were calculated from the polygenic risk score term in the complete regression model. Nagelkerke's R^2 values were assessed upon either the score, covariate or complete logistic regression model. The rankings of the top method-parameter combinations according to these various statistics were compared. Finally, an ensemble score was created by assembling the best five polygenic risk scores and conducting threefold cross validation with each fold generating a logistic regression model. An average of each score's three logistic regression coefficients, weighted by the overall model's performance, became the final coefficients to define the simple linear combination ensemble. The motivation behind creating this ensemble score was to determine if a simple manipulation of scores, easily implementable with any group of scores at hand,

could improve prediction performance. Further, principal component analysis, statistic comparisons, phenotype comparison, SNP set analysis, ensemble weight analysis and method to meta-data associations were conducted to provide greater insight and verification of the comparison results. The principal components analysis is particularly noteworthy as the clusters which formed helped define the simple and complex methods (Supp. Fig. 47).

Generating Statistics from the Validation Data Set

For each trait, logistic regression models were fit with the best parameter-method score on the entire training set of data and assessed on the entire withheld testing set. This process was only carried out on one score per trait to best approximate the true extent of predictive performance without overfitting. Following the same analysis process as described in the methods comparison, a covariate, score, and complete model were fit to the entire training data-set and then used to predict the outcome of the withheld data-set. Producing three respective AUC measurements, these models led to the AUC improvement statistic. In addition, p-values, odds ratios, PR curves, prevalences, case/control distributions, significance of ROC curve difference and Nagelkerke's R^2 values were also all computed following the same procedures described in the previous section. The association between predictive accuracy and study statistics, such as sample size, were also computed, along with a diagnostic comparison of training and testing error.

Correlations of Scores

Several correlation analyses were completed, the first of which computed spearman correlation between pairwise combinations of the 25 evaluated polygenic risk scores. Pairs of traits which did not share any individuals in common, such as prostate and breast cancer, were left as NA. Corresponding empirical correlations between the numbers of cases were also computed using the Jaccard Similarity measure for comparison (Supp. Fig. 48). The next correlation analysis computed spearman correlation and the corresponding p-values between the 25 polygenic risk scores analyzed in the previous section and 23 additional polygenic risk scores generated using the clumping method. Lastly, unrelated traits were predicted by creating the three logistic regression models and calculating the corresponding AUC improvements, as described in the methods comparison section. This process was carried out over a grid of independent variables selected from the 25 best polygenic risk scores and the dependent variable determined by one of the unrelated traits. The exact codings of the unrelated traits were either determined by prevalent ICD-10 codes (n = 136) or self-reported status (n = 70). A final principal components analysis was applied to further visualize the relationship between multiple traits (Supp. Fig. 49).

All computational work was completed in R 3.5, or shell scripts. All code has been uploaded to GitHub (https://github.com /kulmsc). P-values derived from comparisons of groups were generated from ANOVA tests, P-values for individual terms were generated from coefficient terms in regression models, and P-values in relation to correlation analyses were derived from Pearson's product-moment correlation. Boxplots were produced with the default ggplot settings, the middle bar is the median, upper and lower bars are the 25^{the} and 75th percentiles respectively, the whiskers extend 1.5 times the interquartile range, and the notches extend 1.58 times the interquartile range divided by the square root of the sample. Additional notes on how methods were in implemented is in supplementary appendix II.

References

1. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590, DOI: 10.1038/s41576-018-0018-x (2018).
2. ho Lee, Y., Bang, H. & Kim, D. J. How to establish clinical prediction models. *Endocrinol. Metab.* **31**, 38, DOI: 10.3803/enm.2016.31.1.38 (2016).
3. Mars, N. J. *et al.* Polygenic and clinical risk scores and their impact on age at onset of cardiometabolic diseases and common cancers. DOI: 10.1101/727057 (2019).
4. Vreemann, S. *et al.* The added value of mammography in different age-groups of women with and without BRCA mutation screened with breast MRI. *Breast Cancer Res.* **20**, DOI: 10.1186/s13058-018-1019-6 (2018).
5. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci.* **111**, E5272–E5281, DOI: 10.1073/pnas.1419064111 (2014).
6. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224, DOI: 10.1038/s41588-018-0183-z (2018).
7. Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101, DOI: 10.1161/circulationaha.116.024436 (2017).
8. Khera, A. V. *et al.* Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New Engl. J. Medicine* **375**, 2349–2358, DOI: 10.1056/nejmoa1605086 (2016).
9. Verweij, K. J. *et al.* Short communication: Genetic association between schizophrenia and cannabis use. *Drug Alcohol Dependence* **171**, 117–121, DOI: 10.1016/j.drugalcdep.2016.09.022 (2017).
10. Läll, K. *et al.* Polygenic prediction of breast cancer: comparison of genetic predictors and implications for risk stratification. *BMC Cancer* **19**, DOI: 10.1186/s12885-019-5783-1 (2019).
11. Verweij, N. *et al.* The genetic makeup of the electrocardiogram. DOI: 10.1101/648527 (2019).
12. Ranlund, S. *et al.* Associations between polygenic risk scores for four psychiatric illnesses and brain structure using multivariate pattern recognition. *NeuroImage: Clin.* **20**, 1026–1036, DOI: 10.1016/j.nicl.2018.10.008 (2018).
13. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752, DOI: 10.1038/nature08185 (2009).
14. Gianola, D. & Hammond, K. *Advances in Statistical Methods for Genetic Improvement of Livestock* (Springer Science & Business Media, 2012). Google-Books-ID: UfDvCAAAQBAJ.
15. Malin, B., Karp, D. & Scheuermann, R. H. Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research. *J. Investig. Medicine* **58**, 11–18, DOI: 10.2310/JIM.0b013e3181c9b2ea (2010).
16. Dudbridge, F. & Newcombe, P. J. Accuracy of Gene Scores when Pruning Markers by Linkage Disequilibrium. *Hum. Hered.* **80**, 178–186, DOI: 10.1159/000446581 (2015).
17. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569, DOI: 10.1038/ng.608 (2010).
18. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7, DOI: 10.1186/s13742-015-0047-8 (2015).
19. Shi, J. *et al.* Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. *PLOS Genet.* **12**, e1006493, DOI: 10.1371/journal.pgen.1006493 (2016).

20. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the Most of Clumping and Thresholding for Polygenic Scores. *The Am. J. Hum. Genet.* **105**, 1213–1221, DOI: 10.1016/j.ajhg.2019.11.001 (2019).

21. So, H.-C. & Sham, P. C. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Sci. Reports* **7**, 1–11, DOI: 10.1038/srep41262 (2017).

22. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592, DOI: 10.1016/j.ajhg.2015.09.001 (2015).

23. Márquez-Luna, C. *et al.* Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* 375337, DOI: 10.1101/375337 (2019).

24. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1–10, DOI: 10.1038/s41467-019-09718-5 (2019).

25. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Comput. Biol.* **13**, e1005589, DOI: 10.1371/journal.pcbi.1005589 (2017).

26. Maier, R. M. *et al.* Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* **9**, 1–17, DOI: 10.1038/s41467-017-02769-6 (2018).

27. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 1–11, DOI: 10.1038/s41467-019-12653-0 (2019).

28. Paré, G., Mao, S. & Deng, W. Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci. Reports* **7**, 1–11, DOI: 10.1038/s41598-017-13056-1 (2017).

29. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480, DOI: 10.1002/gepi.22050 (2017).

30. Tesli, M. *et al.* Polygenic risk score and the psychosis continuum model. *Acta Psychiatr. Scand.* **130**, 311–317, DOI: 10.1111/acps.12307 (2014).

31. Reisberg, S., Iljasenko, T., Läll, K., Fischer, K. & Vilo, J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLOS ONE* **12**, e0179238, DOI: 10.1371/journal.pone.0179238 (2017).

32. Richardson, T. G., Harrison, S., Hemani, G. & Smith, G. D. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* **8**, DOI: 10.7554/elife.43657 (2019).

33. Woodfield, R., Sudlow, C. L. M. & and. Accuracy of patient self-report of stroke: A systematic review from the UK biobank stroke outcomes group. *PLOS ONE* **10**, e0137538, DOI: 10.1371/journal.pone.0137538 (2015).

34. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779, DOI: 10.1371/journal.pmed.1001779 (2015).

35. Jayson, G. C. *et al.* Carcinomatous meningitis in patients with breast cancer. An aggressive disease variant. *Cancer* **74**, 3135–3141, DOI: 10.1002/1097-0142(19941215)74:12<3135::AID-CNCR2820741212>3.0.CO;2-8 (1994). _eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0142%2819941215%2974%3A12%3C3135%3A%3AAID-CNCR2820741212%3E3.0.CO%3B2-8.

36. Gupta, S., Masand, P. S., Kaplan, D., Bhandary, A. & Hendricks, S. The relationship between schizophrenia and irritable bowel syndrome (IBS). *Schizophr. Res.* **23**, 265–268, DOI: 10.1016/S0920-9964(96)00099-0 (1997).

37. Warren, M. The approach to predictive medicine that is taking genomics research by storm. *Nature* **562**, 181–183, DOI: 10.1038/d41586-018-06956-3 (2018).

38. Derringer, J. *et al.* Predicting sensation seeking from dopamine genes. A candidate-system approach. *Psychol. Sci.* **21**, 1282–1290, DOI: 10.1177/0956797610380699 (2010).

39. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–178, DOI: 10.1038/nature10811 (2012).

40. Aly, M. *et al.* Polygenic risk score improves prostate cancer risk prediction: results from the Stockholm-1 cohort study. *Eur. Urol.* **60**, 21–28, DOI: 10.1016/j.eururo.2011.01.017 (2011).

41. Martin, J., O'Donovan, M. C., Thapar, A., Langley, K. & Williams, N. The relative contribution of common and rare genetic variants to ADHD. *Transl. Psychiatry* **5**, e506, DOI: 10.1038/tp.2015.5 (2015).

42. Abdellaoui, A. *et al.* Genetic Consequences of Social Stratification in Great Britain. preprint, Genetics (2018). DOI: 10.1101/457515.

43. Treff, N. R. *et al.* Utility and first clinical application of screening embryos for polygenic disease risk reduction. *Front. Endocrinol.* **10**, DOI: 10.3389/fendo.2019.00845 (2019).

44. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, DOI: 10.7554/elife.48376 (2020).

45. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295, DOI: 10.1038/ng.3211 (2015).

46. Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. DOI: 10.1101/2020.03.22.001453 (2020).

47. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591, DOI: 10.1038/s41588-019-0379-x (2019).

48. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012, DOI: 10.1093/nar/gky1120 (2019).

49. 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.* **47**, 1457–1464, DOI: 10.1038/ng.3434 (2015).

50. Christophersen, I. E. *et al.* Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 946–952, DOI: 10.1038/ng.3843 (2017).

51. Demenais, F. *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53, DOI: 10.1038/s41588-017-0014-7 (2018).

52. Dubois, P. C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. genetics* **42**, 295–302, DOI: 10.1038/ng.543 (2010).

53. International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol. Psychiatry* **23**, 1181–1188, DOI: 10.1038/mp.2017.154 (2018).

54. Jin, Y. *et al.* Early-onset autoimmune vitiligo associated with an enhancer variant haplotype that upregulates class II HLA expression. *Nat. Commun.* **10**, 391, DOI: 10.1038/s41467-019-08337-4 (2019).

55. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–154, DOI: 10.1038/ng.2500 (2013).

56. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458, DOI: 10.1038/ng.2802 (2013).

57. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986, DOI: 10.1038/ng.3359 (2015).

58. Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571, DOI: 10.1038/s41588-018-0084-1 (2018).

59. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537, DOI: 10.1038/s41588-018-0058-3 (2018).

60. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380, DOI: 10.1038/ng.3242 (2015).

61. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130, DOI: 10.1038/ng.3396 (2015).

62. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381, DOI: 10.1038/nature12873 (2014).

63. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386, DOI: 10.1038/ng.3245 (2015).

64. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. - PubMed - NCBI.

65. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. - PubMed - NCBI.

66. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691, DOI: 10.1038/ng.3826 (2017).

67. International Multiple Sclerosis Genetics Consortium *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219, DOI: 10.1038/nature10251 (2011).

68. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936, DOI: 10.1038/s41588-018-0142-8 (2018).

69. Tanikawa, C. *et al.* GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12. *Carcinogenesis* **39**, 652–660, DOI: 10.1093/carcin/bgy026 (2018).

70. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44**, 1341–1348, DOI: 10.1038/ng.2467 (2012).

71. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048, DOI: 10.1038/ng.3622 (2016).

Figure Captions

Table 1

The polygenic risk score methods implemented and attributes which define them. The parametric, external data and phenotyped data columns refer to whether the methods require the described entity in order to function.

Table 2

The diseases analyzed within the withheld data-set and corresponding statistics generated through polygenic risk score analyses.

Figure 1

Comparison of Polygenic Risk Scoring Methods. Panel A: Comparison of the performance of each polygenic risk scoring method, plotting a series of AUC improvement values for each polygenic risk scoring method. Each series is composed of one value for each trait, which corresponds to the best polygenic risk score generated by the method. Panel B: Examination of which type of model works best, indicating the number of times a trait chose a specific method as best, where best refers to the method which generated the highest AUC improvement. Panel C: Comparison of the performance of the ensemble and next best non-ensemble method by analyzing the best AUC improvement value for each trait in both categories of methods.

Figure 2

Correlations and Meta-Analyses of Polygenic Risk Scores. Panel A: Depicted as a heatmap, each square is the Spearman correlation value between one pairwise combination of the 25 evaluated polygenic risk scores. Values generated from scores that were either identical, or did not share any values (breast cancer and prostate cancer), were set to a null color. Panel B: The AUC improvement values for each trait, for the four different phenotyping methods. No phenotyping method produced significantly different values than any other. Panel C: The significance of correlations between meta-statistics and the four statistics used to evaluate the 25 traits in the withheld data-set. Only 3 associations surpass the Bonferonni -$\log_{10}$(P-value) level of 1.47.

Figure 3

Graphical depiction of methods. Through a flow chart, the raw inputs to the polygenic risk scores, the steps followed to pick the best polygenic risk score, and the final analyses conducted are depicted.

Figures

Table 1

Method	Complexity	Parametric	External Data	Phenotyped Data	Source of Software
basic	Simple	No	No	No	PLINK
Clump[18]	Simple	Yes	No	No	PLINK
WC-2d[19]	Simple	Yes	No	No	Custom
WC-likelihood[19]	Simple	No	No	No	Custom
WC-lasso[19]	Simple	Yes	No	No	Custom
Tweedie[21]	Simple	No	No	No	GitHub
stackCT[20]	Simple	Yes	No	Yes	bigsnpr Package
LDpred[22]	Simple	Yes	Yes	No	GitHub
LDpred-funct[23]	Compex	Yes	Yes	Yes	GitHub
AnnoPred[25]	Complex	Yes	Yes	Yes	GitHub
PRScs[24]	Complex	No	Yes	No	Personal Website
SBayesR[27]	Complex	Yes	Yes	No	GCTB Toolkit
sBLUP[26]	Complex	Yes	No	No	GCTA Toolkit
GraBLD[28]	Complex	Yes	No	Yes	GitHub
lassosum[29]	Complex	No	Yes	Yes	GitHub

Table 2

Disease Name	UKBB Cases	Best Score No. of SNPs	AUC Imp.	AUC Total
Systemic Lupus Erythematosus[49]	405	74111	0.06	0.718
Atrial Fibrillation[50]	13004	247	0.02	0.761
Asthma[51]	35941	34205	0.049	0.578
Celiac Disease[52]	1235	534535	0.094	0.681
Obessive Compulsive Disorder[53]	173	12114	0.002	0.547
Vitiligo[54]	145	199	0.11	0.658
Gout[55]	5562	9	0.022	0.8
Alzheimer's Disease[56]	602	663736	0.036	0.82
Ulcerative Colitis[57]	2909	1571	0.089	0.638
Crohn's Disease[57]	1455	19631	0.152	0.657
Type 2 Diabetes[58]	15014	5925	0.007	0.677
Stroke[59]	7668	1095393	0.004	0.695
Breast Cancer[60]	10315	1107	0.055	0.65
Coronary Artery Disease[61]	14912	150	0.017	0.775
Rheumatoid Arthritis[62]	5132	30362	0.029	0.658
Type 1 Diabetes[63]	2143	3429	0.02	0.621
Eczema[64]	7655	398	0.029	0.587
Chronic Kidney Disease[65]	6141	40467	0.004	0.705
Ovarian Cancer[66]	1141	1026573	0.008	0.605
Multiple Sclerosis[67]	1445	23309	0.102	0.726
Prostate Cancer[68]	6220	963	0.049	0.768
Schizophrenia[13]	704	383824	0.09	0.676
Colorectal Cancer[69]	5077	519	0.004	0.66
Psoriaris[70]	4368	26313	0.134	0.68
Amyotrophic Lateral Sclerosis[71]	206	54748	0.019	0.642

Figure 1

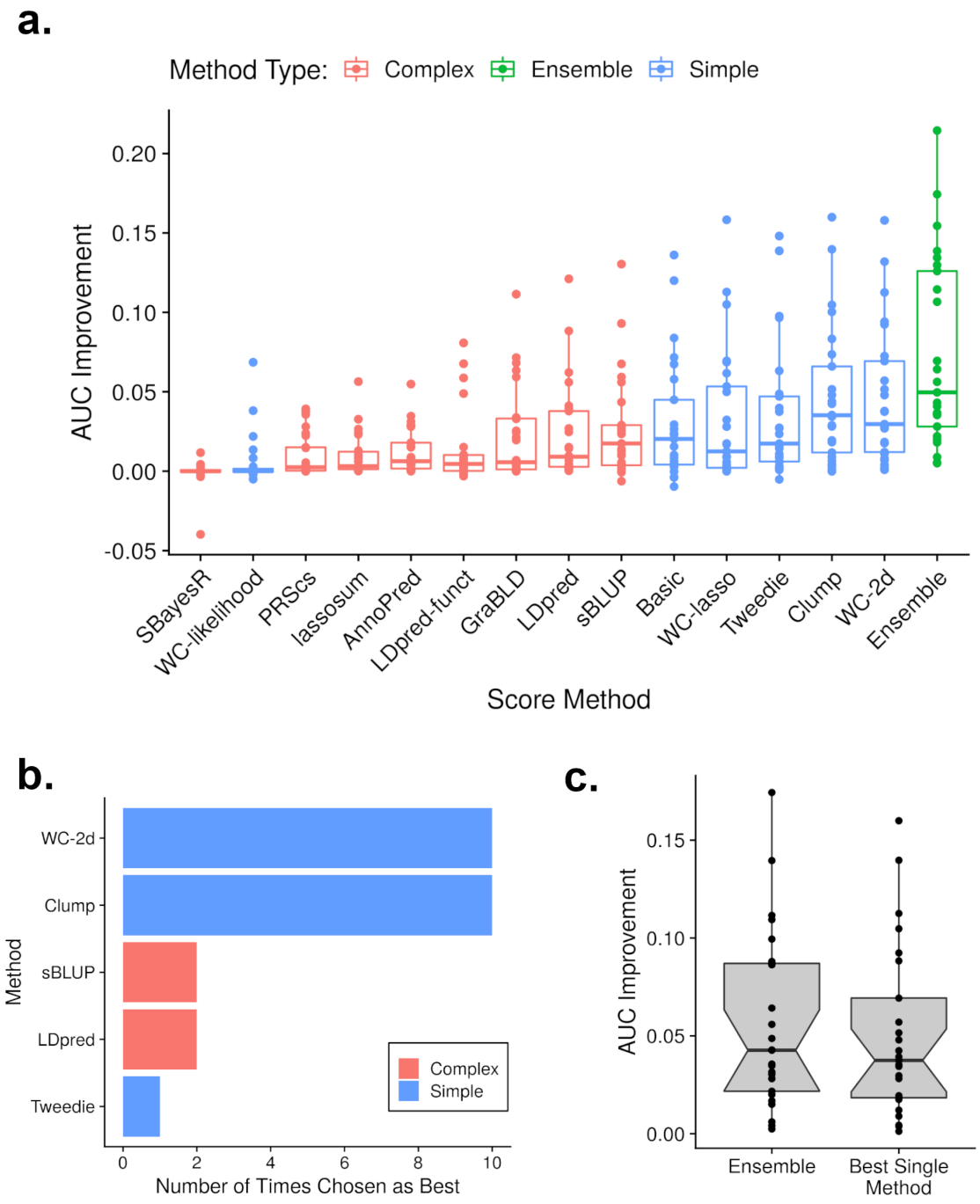

a.
Method Type: Complex Ensemble Simple
AUC Improvement
0.20
0.15
0.10
0.05
0.00
-0.05
SBayesR
WC-likelihood
PRScs
lassosum
AnnoPred
LDpred-funct
GraBLD
LDpred
sBLUP
Basic
WC-lasso
Tweedie
Clump
WC-2d
Ensemble
Score Method
b.
Method
WC-2d
Clump
sBLUP
LDpred
Tweedie
Complex
Simple
0
2
4
6
8
10
Number of Times Chosen as Best
c.
AUC Improvement
0.15
0.10
0.05
0.00
Ensemble
Best Single Method

Figure 2

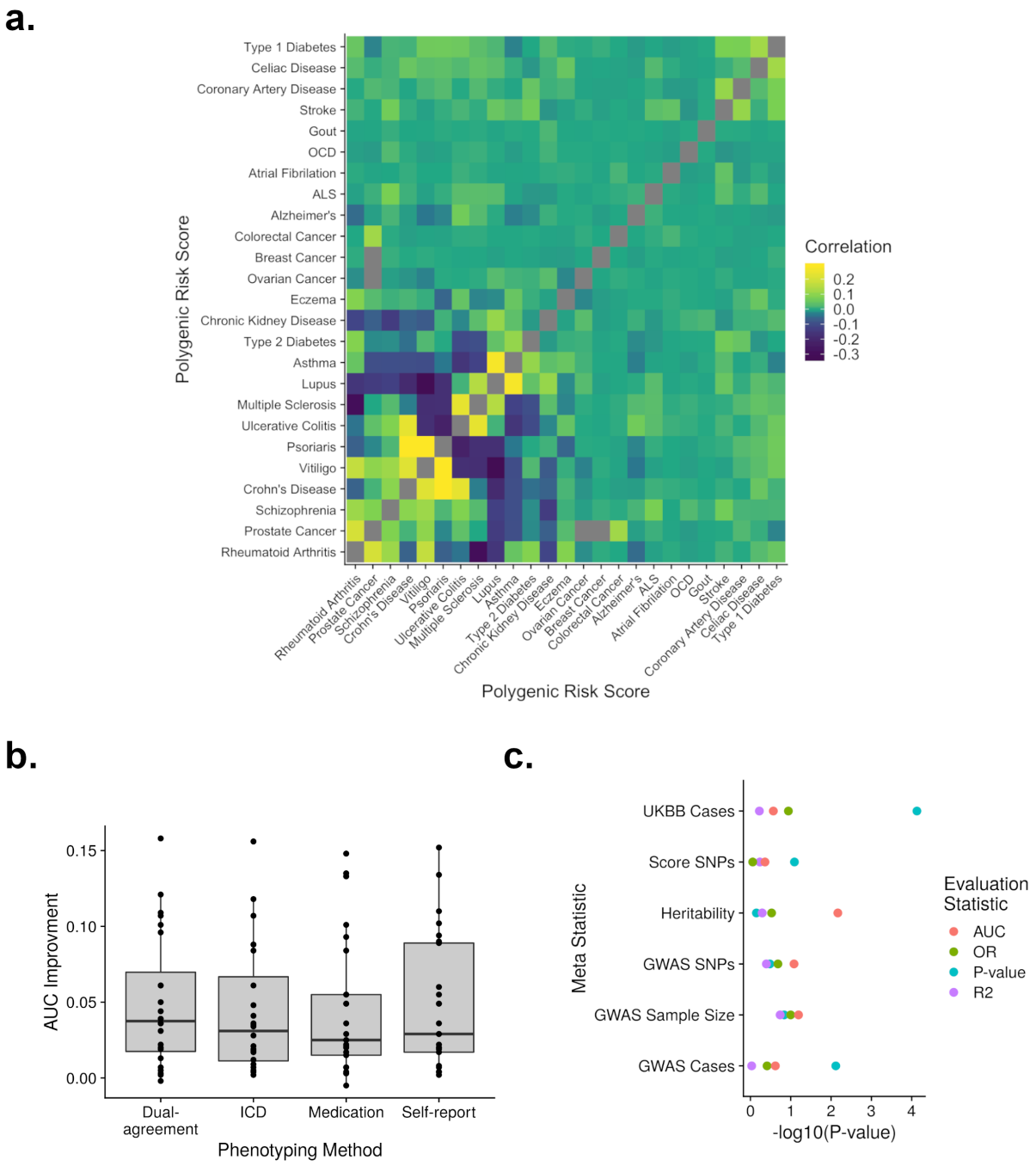
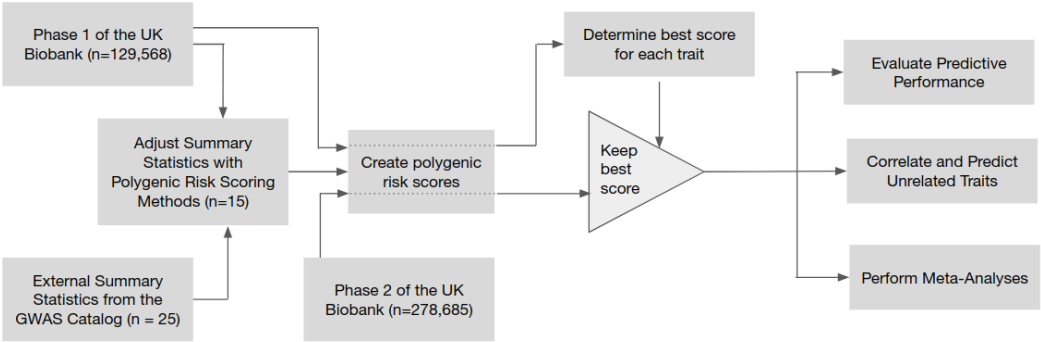

a.
Type 1 Diabetes
Celiac Disease
Coronary Artery Disease
Stroke
Gout
OCD
Atrial Fibrilation
ALS
Alzheimer's
Colorectal Cancer
Breast Cancer
Ovarian Cancer
Eczema
Chronic Kidney Disease
Type 2 Diabetes
Asthma
Lupus
Multiple Sclerosis
Ulcerative Colitis
Psoriaris
Vitiligo
Crohn's Disease
Schizophrenia
Prostate Cancer
Rheumatoid Arthritis
Polygenic Risk Score
Correlation
0.2
0.1
0.0
-0.1
-0.2
-0.3
b.
AUC Improvment
0.15
0.10
0.05
0.00
Dual-agreement
ICD
Medication
Self-report
Phenotyping Method
c.
Meta Statistic
UKBB Cases
Score SNPs
Heritability
GWAS SNPs
GWAS Sample Size
GWAS Cases
0
1
2
3
4
-log10(P-value)
Evaluation Statistic
AUC
OR
P-value
R2

Figure 3


Phase 1 of the UK Biobank (n=129,568)
Adjust Summary Statistics with Polygenic Risk Scoring Methods (n=15)
External Summary Statistics from the GWAS Catalog (n = 25)
Create polygenic risk scores
Phase 2 of the UK Biobank (n=278,685)
Determine best score for each trait
Keep best score
Evaluate Predictive Performance
Correlate and Predict Unrelated Traits
Perform Meta-Analyses