Assessing the performance of genome-wide association studies for predicting disease risk

Short title: Assessing the performance of genome-wide association studies for predicting disease risk

Authors:

Jonas Patron[1], Arnau Serra-Cayuela[1], Beomsoo Han[1], Carin Li[2], David S. Wishart[1,2]

Affiliations:

[1]Department of Biological Sciences, University of Alberta, Edmonton, Canada T6G 2E8

[2]Department of Computing Science, University of Alberta, Edmonton, Canada T6G 2E9

Corresponding author:

David S. Wishart, Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada, T6G 2E8

Tel: (780) 492-0383

Fax: (780) 492-1071

Email: dwishart@ualberta.ca

Abstract

To date more than 3700 genome-wide association studies (GWAS) have been published that look at the genetic contributions of single nucleotide polymorphisms (SNPs) to human conditions or human phenotypes. Through these studies many highly significant SNPs have been identified for hundreds of diseases or medical conditions. However, the extent to which GWAS-identified SNPs or combinations of SNP biomarkers can predict disease risk is not well known. One of the most commonly used approaches to assess the performance of predictive biomarkers is to determine the area under the receiver-operator characteristic curve (AUROC). We have developed an R package called G-WIZ to generate ROC curves and calculate the AUROC using summary-level GWAS data. We first tested the performance of G-WIZ by using AUROC values derived from patient-level SNP data, as well as literature-reported AUROC values. We found that G-WIZ predicts the AUROC with <3% error. Next, we used the summary level GWAS data from GWAS Central to determine the ROC curves and AUROC values for 569 different GWA studies spanning 219 different conditions. Using these data we found a small number of GWA studies with SNP-derived risk predictors having very high AUROCs (>0.75). On the other hand, the average GWA study produces a multi-SNP risk predictor with an AUROC of 0.55. Detailed AUROC comparisons indicate that most SNP-derived risk predictions are not as good as clinically based disease risk predictors. All our calculations (ROC curves, AUROCs, explained heritability) are in a publicly accessible database called GWAS-ROCS (http://gwasrocs.ca).

Introduction

A genome-wide association study (GWAS) is a comprehensive genetic analysis of the association between certain observable traits and specific genetic variations in the form of Single Nucleotide Polymorphisms (SNPs). The appeal of genome wide association (GWA) studies is that they provide a relatively facile approach for detecting potential genetic contributors to common, complex diseases (such as diabetes) or phenotypes (such as body mass index (BMI) or hair color) using a simple case-control study model. The first GWA study was performed in 2005 [1]. This early work explored the association between certain SNPs and age-related macular degeneration in a study population of 146 individuals (96 cases, 50 controls). To date, thousands of GWA studies looking at almost an equal number of conditions or phenotypes, with study populations as large as 1.3 million have been published [2]. Many of these GWA studies are now archived in public databases such as GWAS Central [3] and the NHGRI-EBI GWAS Catalog [4].

Public databases such as GWAS Central contain summary level findings from GWA studies collected on humans [3]. GWAS Central currently houses data from more than 3300 publications corresponding to over 6100 GWA studies, and lists more than 21 million p-values, ranging between 5×10^{-2} and 1×10^{-584}. The p-values in a GWA study report the likelihood of the odd-ratios between two different alleles being statistically different than one. The typical threshold of significance for most published GWA studies is $p = 5 \times 10^{-8}$. The average odds ratio for a statistically significant SNP is 1.33 with very few SNPs having an odds ratio above 3.0 [5].

GWA studies differ from most other ‘omics or clinical/epidemiological studies in their method of reporting of disease-associated or disease risk markers. In particular,

relatively strong emphasis is placed on the reporting of p-values of a single SNP marker (with p-values of < 10^{-50} often being achieved) along with its associated odds ratio [6]. Interestingly, very few GWA studies combine multiple SNPs together to produce a multi-marker risk predictor. As a result, it is rare to find a GWA study that reports on the sensitivity, specificity, receiver operating characteristic (ROC) curve or the area under the ROC curve (AUROC or C-statistic) associated with a given SNP or combination of SNPs for disease risk or trait prediction. In contrast to most GWA studies, clinical, proteomic and metabolomic studies involving marker identification rarely achieve p-values of < 10^{-6} and they infrequently report the performance of their markers in terms of odds ratios. Instead, most marker-based clinical, proteomic or metabolomic studies tend to combine multiple clinical or 'omic measures to generate multi-marker risk predictors. In these studies, multi-marker sensitivity, specificity, ROC curves and/or AUROCs are routinely reported [7].

One way for the different 'omics communities to better understand the predictive or discriminative ability of GWAS data is to convert the reported SNP data into a more conventional biomarker reporting format. In particular, if multi-SNP biomarker data could be combined and converted into ROC or AUROC data, then a more direct comparison could be performed regarding the performance of SNPs for predicting disease risks relative to clinical, metabolite or protein biomarkers for the same conditions. The only precise way of generating ROC curves or calculating AUROC values from GWA studies is to have the complete patient SNP data set. Unfortunately, obtaining ethics approvals and permissions for access to the full patient data sets from thousands of

individual GWA studies (found in GWAS Central, for example), conducted in multiple countries and spanning more than 15 years is currently impractical.

To overcome this problem we developed a novel technique to generate ROC curves and to calculate the AUROC for different SNP combinations using summary-level data. Summary-level data, which only contains study-wide averages, p-values, odds ratios, risk allele frequencies and other summary statistics for the entire study population, is the only data that GWAS data can be deposited in public repositories such as GWAS-Central. Our new method (called G-WIZ) combines population modeling with logistic regression (for risk prediction) to generate study-specific ROC curves and AUROC values. Our approach also enables the estimation of SNP heritability directly from summary level GWAS data. We have extensively tested G-WIZ by measuring its ability to predict ROC curves and AUROC values derived from a large collection of authentic individual patient SNP data (i.e. non-summary data), as well as to predict literature-reported AUROC values. We found that G-WIZ predicts the AUROC with <3% error. We then used G-WIZ to, calculate ROC curves and AUROC values for 569 GWA studies spanning 219 different conditions/phenotypes using summary level GWAS data from GWAS Central. Using these data we found at least five conditions/phenotypes that exhibit very high AUROCs (>0.75) and SNP-heritability values (>30%) using a multi-component SNP risk predictor. On the other hand, we found that the average AUROC value for GWAS risk predictors is only 0.55. In contrast, most predictive clinical, metabolite or protein-based risk predictors have AUROCs of >0.7 or 0.8 [8]. We also used the data from these calculations to derive a simple formula to estimate SNP heritability directly from the AUROC value. All of the G-WIZ calculations and accompanying analyses (ROC curves,

AUROCs, explained heritability, comparisons of AUROCs from other omics studies) have been deposited into a publicly accessible database called GWAS-ROCS (http://gwasrocs.ca). Having such a large, centralized database of SNP combinations, ROC curves, AUROC values and heritability estimates should open the door to performing more systematic comparisons of GWA studies or to identify new and unexpected trends or novel disease-gene relationships.

Methods

Ethics approvals

We received research ethics approval from the University of Alberta (REB - Pro00084706) and approval from the Wellcome Sanger Institute (request ID: 8303 and 7104) to obtain access to the Wellcome Trust Case Control Consortium (WTCCC) GWAS datasets [9].

Selection of GWAS Central studies

An in-house Python script was created to screen-scrape the summary data in GWAS Central. A total of 3307 GWAS publications corresponding to 6137 GWAS summaries were collected in this manner. These GWA studies were further filtered based on the inclusion of an odds ratio (OR) and a risk allele frequency (RAF) for each reported SNP. Additionally, to improve AUROC value prediction consistency we discarded GWA studies with a sample size of less than 1000 cases or 1000 controls. We thought this was reasonable since the WTCCC datasets all had sample sizes of more than 1000 cases and more than 1000 controls. Moreover, an analysis of every study in GWAS Central

indicated that roughly 70% of all published GWA studies (since 2009) have had sample sizes of greater than 1000 cases and 1000 controls. While choosing this threshold should have left us with about 4300 studies to analyze, we found that many studies in GWAS Central were missing either the SNP odds ratios or the risk allele frequencies – or both, which prevented their use in our calculations. Indeed, after applying these filters, we were left with a total of 569 GWAS Central studies, corresponding to 219 different conditions.

Population simulation for ROC and AUROC calculations

The calculation of sensitivity, specificity, ROC and AUROC data from case/control studies normally requires a data set where the variables of interest (SNPs, proteins, metabolites, clinical measures, etc.) are assigned to individual patients along with their health status. Unfortunately, the GWAS data as found in GWAS Central or other public databases, is only available as summary data. This means that this patient-specific information is not readily accessible. Indeed, the only information retrievable is the cohort size (cases and controls), the SNP identifiers, the p-values, ORs for each SNP and RAFs for each reported SNP. While it is possible to obtain detailed SNP profiles and health status information for each patient in each study, doing so for the entire GWAS Central Collection would have required extensive ethics reviews along with time and resources that were far beyond our means. This necessitated the development of a modeling program (called G-WIZ or Gwas WIZard) that would calculate ROC and AUROC data from summary-level information only. To do so we exploited that fact that almost no SNPs separated by more than 10 kb are in absolute linkage disequilibrium and that the vast majority of reported, disease-significant SNPs are in Hardy-Weinberg

equilibrium [10–12]. As a result, we assumed the independence of SNPs to create simulated patient populations with specific SNP profiles and assigned health conditions from the publicly available OR and RAF data. These synthetic populations were designed to be sufficiently large (typically >30,000 individuals) so that statistically anomalies would be averaged out. To assign a single SNP to an individual in the simulated population the following methodology was used:

Let H_s denote the number of risk alleles in the healthy group,

H_n denote the number of non-risk alleles in healthy group,

D_s denote the number of risk alleles in the diseased group,

D_n denote the number of non-risk alleles in diseased group,

RAF denote risk allele frequency,

OR denote odds ratio,

$N_{control}$ denote the number of controls in the simulated dataset,

N_{case} denote the number of cases in the simulated dataset,

$\lceil \cdot \rceil$ denote the ceiling function.

Using $N_{control}$, N_{case}, OR and RAF as the pre-specified input values, we can calculate H_s, H_n, D_s, D_n as follows:

$$H_s = \lceil 2 \cdot N_{control} \cdot RAF \rceil \quad (1)$$

$$H_n = 2 \cdot N_{control} - H_s \quad (2)$$

$$D_s = \left\lceil \frac{2 \cdot OR \cdot N_{case} \cdot H_s}{H_n + OR \cdot H_s} \right\rceil \quad (3)$$

$$D_n = 2 \cdot N_{case} - D_s \quad (4)$$

Under the assumption that each SNP is independent of the others we can repeat the above procedure to create a full SNP profile and an assigned health state for each member of the simulated population. The G-WIZ models were built using all the SNPs reported by each of the respective GWA studies. These SNPs from GWAS Central were reported on the basis of their significance as identified by the original depositors. However, we considered that it might still be possible that models created using only a subset (feature selection) of reported SNPs would perform better, as this would have controlled for over-parameterization. We tested for this by performing feature selection, using only the SNPs with lowest p-values, however no improvements to the models' performance were found. In the end our SNP profiles used every reported SNP. On average, each study had a SNP profile consisting of 6 significant SNPs. Moreover, the maximum SNP p-value was 9×10^{-6} and the minimum SNP p-value was 1×10^{-295}, indicating that the reported SNPs are all highly significant. Further, each G-WIZ model had on average 34,491 simulated patients (cases and controls). The largest number of SNPs used in any given SNP profile was 50, and the largest synthetic population generated by G-WIZ consisted of 808,380 individuals.

Statistical modeling for ROC curve generation

The creation of simulated populations consisting of full SNP profiles and assigned health states for each of the 569 condition/phenotype studies in GWAS Central allowed us to

calculate the corresponding ROC curves and AUROC values. A common modeling method used to generate ROC curves for multi-marker data is logistic regression. Logistic regression is a statistical method for modeling multiple independent variables (e.g. SNPs) to explain two possible outcomes (e.g. healthy or diseased). Once constructed, a logistic regression model will return a risk score between 0 and 1. A cut-off value can then be chosen (e.g. 0.5), and any individual that has a risk score above it is classified as 'diseased', and any individual below it is classified as 'healthy'. A plot of the sensitivity against 1-specificity for all possible cut-off values is known as a receiver-operating characteristic (ROC) curve. The classification accuracy of the logistic regression model can then be measured by calculating the area under the curve of the ROC curve (AUROC). A perfect model would have an AUROC of 1, while a model with no classification accuracy would have an AUROC of 0.5 [7].

We performed logistic regression analysis because it is easy to perform and interpret (13). Although, it is possible that better performing multi-SNP profiles could have been developed using advanced machine learning algorithms such as neural networks, decision trees, or support vector machines [14–20], it is also possible to over-train models with these very powerful pattern-finding tools. Indeed, it is not unusual, during validation studies, to see these models fall short when compared to logistic regression models [14–17]. These concerns regarding overfitting led us to limit our model complexity and to exclusively use logistic and ridge logistic regression to estimate the classification accuracy of these GWA studies.

Another common issue with regression models containing many explanatory variables is multicollinearity. Multicollinearity increases the variance of parameter

estimates, which will affect confidence intervals and hypothesis tests. This can lead to incorrect inferences about relationships between explanatory and response variables [21]. With these issues in mind we tested for multicollinearity by estimating the variance inflation factor (VIF) prior to building each regression model. Whenever multicollinearity was observed we used ridge logistic regression [22] to generate a biomarker model, otherwise we used a standard logistic regression model. In total 566 standard logistic regression models and 3 ridge logistic regression models were constructed for all 569 GWA studies.

To assess the performance of each biomarker or ROC-generative model, the simulated data was randomly split into training and testing sets. In the training set, nested cross-validation (outer 3-fold and inner 2-fold) was used to obtain an estimate of the classification accuracy [21]. Once the model was properly tuned, it was validated using the testing set.

The G-WIZ program

G-WIZ is written in the R programming language [23]. It consists of several modules including a custom-written tool to generate patient populations, the MLR package [24] to build the logistic regression models and perform cross-validation, as well as the pROC package [25] to generate ROC curves and to calculate the AUROCs. To analyze a study with G-WIZ the user must create a *.csv file with the study sample size, and the odds ratios and risk allele frequencies for each SNP in the study. A typical study (with a sample size of 50,000 individuals) can be simulated and analyzed via G-WIZ on a modern laptop computer in less than 30 seconds. For any given study, G-WIZ performs

population simulation, regression modelling, and then estimates ROC curves and AUROCs. The output consists of a *.csv file with the regression coefficients, a *.csv file with the sensitivity, specificity and AUROC of the study, a *.csv file with the simulated population, and a *.png file with the ROC curve. The code is freely available for download at https://bitbucket.org/wishartlab/gwiz-rscript/src/master/, and is compatible with a wide variety of UNIX platforms, Mac OS and Windows operating systems. The bitbucket download file also includes an example of the *.csv input file, as well as instructions on how to run the program. In addition, the complete set of simulated populations used in this study, along with their corresponding SNP profiles, disease status, calculated ROC curves and AUROC data is available for download at http://gwasrocs.ca.

SNP-derived heritability calculations

For each study in GWAS Central we also estimated the total variance in disease liability explained (often referred to as the SNP heritability) using the following formula described by Pawitan et al. [26].

Let $var(\cdot)$ denote the variance

OR denote the odds ratio

$log(\cdot)$ denote the natural logarithm

RAF denote the risk allele frequency

π denote the mathematical constant pi

h^2 denote the heritability

g denote the random genetic effect

Then $h^2 = \frac{var(g)}{\mathrm{var(g)} + {\pi^2}/{3}}$

$$var(g) = 2\sum_{k=1}^{n} RAF_k(1 - RAF_k)(logOR_k)^2$$

Where n is the number of SNPs in a particular study, and RAF_k and OR_k are the risk allele frequency and odds ratio of the kth SNP. We created an in-house R script to run this formula on all 569 GWA studies collected from GWAS Central. The results are shown in Supplementary Table S1.

Validation

To further ensure that our modeling methods and assumptions were correct, we validated our predictions in two different ways. In the first approach, we used real patient GWAS data from the Wellcome Trust Case Control Consortium (WTCCC). Using this data we ran logistic regression analyses for 7 different conditions on 2 different control datasets, and calculated the true ROCs and AUROCs. The 7 conditions were bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D)—spanning AUROC values from 0.51 to 0.72. The two control datasets were WTCCC1 project samples from UK National Blood Service and WTCCC1 project samples from 1958 British Birth Cohort. The SNP profiles and risk alleles we used are reported in Table 1, and are the same SNPs and alleles reported as being statistically significant by the WTCCC researchers [9]. For both control datasets we then applied the G-WIZ modelling method to the same set of SNPs and generated a synthetic population with SNP profiles by

directly calculating the RAFs and number of cases and controls for each disease study, and by estimating the ORs from the logistic regression coefficients. The true AUROCs were then compared to our G-WIZ calculated AUROCs. Additionally, the shape of the ROC curves was also compared using Delong's test [27].

Table 1: SNP profiles and risk alleles reported by the WTCCC publishers

WTCCC Disease	SNP	Risk Allele	Minor allele
Bipolar disorder	rs420259	A	G
Coronary artery disease	rs1333049	C	C
Crohn's disease	rs11805303	T	T
Crohn's disease	rs10210302	T	C
Crohn's disease	rs9858542	A	A
Crohn's disease	rs17234657	G	G
Crohn's disease	rs1000113	T	T
Crohn's disease	rs10761659	G	A
Crohn's disease	rs10883365	G	G
Crohn's disease	rs17221417	G	G
Crohn's disease	rs2542151	G	G
Hypertension	rs2820037	T	T
Hypertension	rs6997709	G	T
Hypertension	rs7961152	A	A
Hypertension	rs11110912	G	G
Hypertension	rs1937506	G	A
Hypertension	rs2398162	A	G
Rheumatoid arthritis	rs6679677	A	A
Rheumatoid arthritis	rs6457617	T	T
Type 1 diabetes	rs6679677	A	A
Type 1 diabetes	rs9272346	A	G
Type 1 diabetes	rs11171739	C	C
Type 1 diabetes	rs17696736	G	G
Type 1 diabetes	rs12708716	A	G
Type 2 diabetes	rs9465871	C	C
Type 2 diabetes	rs4506565	T	T
Type 2 diabetes	rs9939609	A	A

The second approach we used to validate our G-WIZ ROC and AUROC predictions involved comparing the G-WIZ predictions against 21 AUROC values provided by previously published GWA studies [28–48]. The GWA studies selected for this validation test covered a broad range of conditions/phenotypes, and a wide range of AUROC values (0.55 – 0.74). The same set of SNPs reported for these studies were used in the G-WIZ summary-level simulations. We used the published ORs, RAFs and the number of cases and controls reported in the 21 papers. Logistic regression or ridge regression (as required) was used to generate the ROC curves and to calculate the AUROCs for each of the 21 studies. The reported AUROCs were then compared to our calculated AUROCs. To compile these 21 studies, we created a list of all applicable papers in PubMed by searching the literature for all publications with genomic prediction models that reported an AUROC. This gave us a list of 112 studies (see Supplementary Table S2). Of these 112 studies, we further filtered the studies by study design, sample size requirements, and available SNP information. This left us with 21 studies that were both case-control studies, that had more than 1000 cases and 1000 controls, and that had the relevant SNP summary data to run G-WIZ.

Results

Validation testing

A crucial part of developing G-WIZ was to perform an extensive set of validation tests to ensure that our AUROC predictions were accurate. We tested the performance of G-WIZ with respect to four different SNP inheritance models (dominant, recessive, odd SNPs are

dominant, and odd SNPs are recessive). This was important because the SNP inheritance model in published GWA studies may vary from SNP to SNP, and is not typically disclosed. Thus, for each of the 7 WTCCC disease studies we calculated the true AUROC under these four SNP inheritance schemes. The results are outlined in Table 2. In total, we ran 56 different tests (spanning all 7 different disease studies, 2 control datasets and 4 SNP inheritance schemes). The average difference in our AUROC predictions was 0.026. The average p-value from the Delong's test was 0.43, indicating that the predicted ROC curves did not differ significantly from the true ROC curves. Indeed, 52 out of 56 tests had ROC curves that did not differ significantly when tested at the p = 0.05 level of significance.

Table 2: Comparison of true versus G-WIZ predicted AUROCs for the WTCCC data

	Controls Cohort: WTCCC1 UK National Blood Service		
Cases Cohort	True AUROC*	Predicted AUROC*	Difference
WTCCC1 Bipolar disorder	0.52	0.52	0.00
WTCCC1 Coronary artery disease	0.53	0.56	0.03
WTCCC1 Chron's disease	0.63	0.62	0.01
WTCCC1 Hypertension	0.57	0.57	0.00
WTCCC1 Rheumatoid arthritis	0.61	0.62	0.01
WTCCC1 Type 1 diabetes	0.68	0.70	0.02
WTCCC1 Type 2 diabetes	0.57	0.57	0.00
	Controls Cohort: WTCCC1 1958 British Birth Cohort		
Cases Cohort	True AUROC*	Predicted AUROC*	Difference
WTCCC1 Bipolar disorder	0.52	0.52	0.00
WTCCC1 Coronary artery disease	0.54	0.57	0.03
WTCCC1 Chron's disease	0.64	0.60	0.04
WTCCC1 Hypertension	0.55	0.55	0.00
WTCCC1 Rheumatoid arthritis	0.61	0.61	0.00
WTCCC1 Type 1 diabetes	0.66	0.71	0.05
WTCCC1 Type 2 diabetes	0.56	0.56	0.00

*Values are the average calculated over the 4 different SNP inheritance models
dominant, recessive, dominant (odd) - recessive (even), dominant (even) - recessive (odd)

Table 3 lists the AUROC values calculated by our logistic regression modeling technique along with the AUROCs from the 21 previously published GWAS papers. These published AUROC values were derived from GWA studies that used patient-identifiable, authentic clinical data. As can be seen from this table, our AUROC predictions had an average difference of 0.030 relative to the published AUROCs, with a maximum difference of 0.071 and a minimum difference of 0.004. This is remarkably good considering that our data consisted of simulated populations while the published studies were working with complete population data. Previously published studies on AUROC estimates have shown that the type of modeling method used to generate a multi-component biomarker signature can lead to AUROC variations of +/- 0.07 [14–17]. Similarly, we observed that the size of the sample population can also lead to differences in AUROC values (of +/- 0.04) with smaller populations or unbalanced numbers of cases and controls leading to larger differences.

Table 3: Comparison of published versus G-WIZ predicted AUROCs for 21 publications

Phenotype/Condition	Published AUROC [Reference)	Predicted AUROC	Difference
Type 2 diabetes	0.58 [30]	0.59	0.01
Type 2 diabetes	0.60 [28]	0.62	0.02
Type 2 diabetes	0.63 [31]	0.62	0.01
Type 2 diabetes	0.60 [29]	0.58	0.02
Type 2 diabetes	0.62 [33]	0.58	0.04

Type 2 diabetes	0.63 [34]	0.60	0.03
Coronary artery disease	0.61 [35]	0.58	0.02
Psoriasis	0.72 [36]	0.67	0.05
Rheumatoid arthritis	0.59 [37]	0.58	0.01
Breast cancer	0.58 [38]	0.63	0.05
Colorectal cancer	0.57 [32]	0.60	0.03
Breast cancer	0.58 [39]	0.61	0.03
Prostate cancer	0.66 [40]	0.65	0.01
Lung cancer	0.55 [41]	0.52	0.04
Esophageal squamous-cell carcinoma	0.63 [42]	0.63	0.00
Rheumatoid arthritis	0.66 [43]	0.69	0.03
Venous thrombosis	0.66 [44]	0.64	0.02
Alzheimer's disease	0.70 [45]	0.77	0.07
Colorectal cancer	0.56 [46]	0.59	0.03
Leprosy	0.74 [47]	0.80	0.06
Glaucoma	0.62 [48]	0.57	0.05

The excellent performance of our G-WIZ AUROC estimates in this validation study (along with the robustness shown in the cross-validation tests) gave us sufficient confidence to apply our logistic modeling method to all 569 GWA studies compiled from GWAS Central and calculate their ROC curves and AUROCs.

Analysis of GWAS Central studies

The complete set of ROCs, along with the calculated AUROCs for all 569 GWAS Central studies is available at http://gwasrocs.ca. These datasets were used to calculate the plot shown in Figure 1. This Figure displays a histogram of the calculated AUROC distribution for all 569 GWAS Central studies. The average AUROC was 0.55 while the median AUROC was 0.54. Ten GWA studies (2%) had an AUROC greater than 0.70 and 73 (13%) had an AUROC above 0.60. The GWA study with the highest AUROC was "Black vs. non-black hair color" [49,50], which had a predicted AUROC of 0.98. The

GWA study "Shingles" [51,52] had the lowest AUROC with a predicted AUROC of 0.50. The logistic regression models that we built used on average 6 SNPs. The largest number of SNPs used in a single model was 50. Six studies were modeled using 50 SNPs and 165 studies (29%) were modeled using only a single SNP (see Supplementary Table S3).

Fig 1: Histogram of the G-WIZ predicted AUROCs of the 569 studies from GWAS Central. Bin width of 0.01.

GWAS vs. non-GWAS risk prediction performance

One of the motivating factors behind this study was to compare the performance of GWAS-derived or SNP-derived biomarker profiles for disease prediction with other predictive biomarker profiles derived from clinical, metabolomic and/or proteomic (i.e. non-GWAS) data. These data are presented in Table 4 along with types of biomarkers used in each non-GWAS biomarker profile. In compiling this table, we performed an extensive literature survey (of 30 publications) to identify a number of validated or widely used clinical and/or 'omics' biomarker sets for disease risk prediction in 12 different conditions/phenotypes (Alzheimer's, type 2 diabetes, metabolic syndrome, prostate cancer, etc.) along with their reported AUROCs. The average GWAS AUROC was 0.64 whereas the average non-GWAS AUROC was 0.81.

Table 4: Comparison of typical metabolic, clinical, proteomic and genetic marker AUROCs for common conditions

Conditions/ Phenotypes	Metabolic Marker AUROC [Reference]	Clinical Factor AUROC [Reference]	Proteomic Marker AUROC [Reference]	Genetic Marker AUROC [Reference]

Alzheimer's Disease	0.770 [53]	0.770 [54]	---	0.640 [55]
Type 2 Diabetes	0.849 [56]	0.870 [57]	---	0.641 [58]
Age Related Macular Degeneration	---	0.780 [59]	---	0.820 [60]
Metabolic Syndrome	0.820 [61]	0.810 [62]	---	0.640 [63]
Colorectal Cancer	0.895 [64]	---	0.885 [65]	0.570 [32]
Prostate Cancer	---	---	0.690 [66]	0.610 [67]
Pulmonary Tuberculosis	---	0.910 [68]	---	0.640 [69]
Cardiovascular Disease	---	0.770 [70]	---	0.600 [71]
Breast Cancer	---	0.690 [72]	0.856 [73]	0.638 [74]
Esophageal Squamous-cell Carcinoma	---	0.639 [42]	---	0.632 [42]
Leprosy	0.862 [75]	---	---	0.707 [47]
Lung Cancer	0.800 [76]	0.903 [77]	---	0.551 (41)
Average:	0.83	0.79	0.81	0.64

Calculating SNP heritability using the AUROC

While comparing the AUROC values and heritability estimates plotted in the GWAS-ROCS website the values an interesting trend was noted. In particular, the heritability seemed to be well correlated with the square of the AUROC (r = 0.87, see Figure 2). This led to a more detailed investigation regarding the potential rational for this observation. Upon further reading we found that the $AUROC = (D + 1)/2$, where D is the Somers' rank correlation [78] between risk profile and disease status (1 = diseased, 0 = not diseased). Note that the squared Somers' D rank correlation is in fact the proportion of explained variance [79], and that the definition of heritability is precisely the proportion of explained variance. Thus, in the context of a SNP-only model trying to

predict disease status, the squared Somers' D rank correlation is, in fact, the SNP heritability. Rearranging for D and then squaring in the formula above, we find that $h^2 = D^2 = (2 \cdot AUROC - 1)^2$. This result highlights the utility of AUROC calculations for not only assessing the predictive performance of a multi-SNP panel but for also easily and rapidly calculating heritability of such a SNP panel.

Fig 2: Plot of the square of the AUROC versus the Pawitan et al. [26] SNP Heritability. Correlation coefficient, r = 0.867.

Comparison with competing methods

During the later stages of writing this paper, a similar paper by Song et al. [80] was published describing an R package (called SummaryAUC) that can be used to estimate AUROC values based on summary level GWAS data. The method is notable in that it incorporates data to handle SNPs in linkage disequilibrium. This makes it ideal for predicting AUROCs for models that include a large number of SNPs (100-2000 SNPs). However, the authors also note that the utility of their method may be diminished when only a few SNPs are included in the model or when cases and controls are from different population cohorts (which is typical of most GWA studies). To compare G-WIZ to SummaryAUC we tested the performance of both programs on the raw WTCCC1 data we had. The WTCCC1 data is characterized by a small number of SNPs in each model (typically less than 10), and sets of cases and controls which were taken from different cohorts. The results are shown in Figure 3. As can be seen in this figure, G-WIZ markedly outperforms SummaryAUC in this test set, with an average prediction

difference of 0.02 for G-WIZ, versus 0.13 for SummaryAUC. This comparison is of importance because GWA studies which report few significant SNPs and have cases and controls taken from separate population cohorts are typical of about 90% of the GWAS studies deposited in GWAS Central. As such, we believe that G-WIZ is much more suitable to perform GWAS ROC analyses on the summary GWAS data that is currently available. SummaryAUC may be of more utility in predicting the AUROCs for the rarer (or more specialized) GWA studies that report thousands of marginally significant SNPs. It is also worth noting that SummaryAUC only generates AUROC values for its output, whereas G-WIZ generates ROC curves, AUROC values, heritability estimates, predictive logistic regression models and simulated patient populations which could be used for a wide number of other research applications that go beyond merely knowing the AUROC.

Fig 3: Comparison of AUROC predictions accuracies made by G-WIZ and SummaryAUC where the difference between the actual AUROC and predicted AUROC for 7 different disorders is shown. Legend: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D)

The GWAS-ROCS database

Along with the creation of G-WIZ, we also created the GWAS-ROCS Database—a freely available electronic database containing SNP-derived AUROCs calculated via G-WIZ. GWAS-ROCS currently contains all the ROC curves, AUROCs, SNP heritability calculations, simulated populations, references, and SNP meta-data generated in the course of producing this paper. The database currently houses 569 simulated populations (corresponding to 219 different conditions) and SNP data (odds ratios, risk allele

frequencies, and p-values) for 2986 unique SNPs. Over the coming months, we intend to extend our web scraping to other databases such as NHGRI-EBI GWAS Catalog. In doing so we expect to be able to expand this database and add dozens more GWAS studies. All the data in the GWAS-ROCS database is downloadable, open-access and intended for applications in genomics, biomarker discovery, and general education.

The GWAS-ROCS database also contains numerous datasets with individual level GWAS simulated populations. A GWAS-ROCS simulated population is a *.csv file with computer-generated "patients" who are marked as either cases or controls. Each individual is assigned data about the presence of risk alleles (1 having the risk allele, or 0 not having the risk allele) at SNPs previously identified as being significant. Users can download these simulated patient populations to develop and test their own genetic risk prediction models or to perform other kinds of synthetic population experiments.

A screenshot montage illustrating the contents and design of the GWAS-ROCS database is shown in Figure 4. As can be seen from this figure the GWAS-ROCS website has a simple webpage layout (Figure 4a). There are four tabs at the top of the page: "Browse Study Simulations", "Downloads", "About" and "Contact Us". Clicking on the "Browse Study Simulations" tab allows users to view a scrollable series of images where they can easily browse through the 569 simulated GWA studies produced in this paper (Figure 4b). Users can sort the GWA studies according to their ID number, the condition/phenotype and the AUROC value. Clicking on a study sends users to a webpage with more information about that specific study (Figure 4c). This includes information such as hyperlinks to the reference GWAS Central study and the original PubMed publication, the number of control and case subjects, the SNP accession IDs, the

ORs and RAFs, and the simulated ROC plots, all of which can be found on this page. Additionally, a downloadable *.csv file with the simulated population for that specific GWA study can be found on this page too. The "Downloads" tab gives users a way to quickly and efficiently download the *.csv files with simulated populations and ROC plots, for every single study in GWAS-ROCS (Figure 4d). The "About" tab contains some documentation to help users navigate the site. And finally, the "Contact Us" tab gives users an easy way to contact the GWAS-ROCS team with any questions or concerns they may have.

Fig 4: A screenshot montage of the GWAS-ROCS database at http://gwasrocs.ca. **a)** The landing page. **b)** Browse Study Simulations tab. **c)** Study simulation page. **d)** Downloads tab.

Discussion

GWA studies have contributed significantly to our understanding of the genetic contributions to disease and disease risk. Hundreds of novel genes have been identified and implicated in various traits or conditions and many of these have led to new biological understandings and insights [81]. With continued improvements to GWA study designs (increased sample sizes, better population selection to remove confounders, more narrowly defined phenotypes) and GWAS analysis it is likely that many more important biological or genetic insights will be gained [81]. While the value of GWA studies is indisputable, there are still lingering concerns over the inability of SNPs to explain as much of the heritable variation as originally hoped (the missing heritability problem [82,83]) or as much of the disease risk as expected [83].

As remarked earlier, GWA studies that explore disease risk do not often adopt the convention used by most other multi-marker risk predictors to assess performance. In particular, the use of multi-component SNP models and the evaluation of ROC curves or AUROC values (C-statistics) is quite rare. Of the >3700 GWAS publications we evaluated, only 112 have published ROC curves or AUROC data. Of these, fewer than 30 provided sufficient data to independently validate their reported ROC or AUROC results. This has made it difficult to compare the performance of SNP-derived or GWAS-derived biomarkers in disease risk prediction with other types of disease-risk prediction biomarkers or models (clinical, metabolomic, proteomic, etc.). Furthermore, the difference in reporting methodologies between GWA studies (with an emphasis on p-values and odds ratios for individual SNPs) and non-GWA studies (with an emphasis on ROC curves and AUROCs calculated for multiple markers) has also led to an expectation by many non-GWAS specialists, or those with limited statistical training, that the predictive performance of GWAS-derived biomarkers should be much better than non-GWAS derived biomarkers.

Because of this "cultural" difference we undertook this study to help standardize biomarker reporting between GWAS derived and non-GWAS derived biomarker profiles. In particular, we used logistic regression modeling and simulated patient data to generate a comprehensive and publicly available database of GWAS ROC curves, AUROCs and SNP-heritability scores for a large number of conditions (219) and a large number of GWAS studies (569). These data were placed into an open-access database called GWAS-ROCS, which is publicly available at http://gwasrocs.ca.

In creating the GWAS-ROCS database we hoped to accomplish several objectives. First, we wanted to compile and consolidate an accurate and comprehensive set of SNP-derived AUROCs into a single, open-access site. Second, we wanted to use this consolidated data to systematically analyze interesting trends or features in GWAS AUROC data. One of the trends we wanted to explore in more detail concerned the performance of SNP biomarker panels in disease or phenotype prediction. Our results indicate that the average AUROC for a typical GWAS-derived biomarker profile is low, just 0.55 with a standard deviation of 0.05. This is significantly lower than what we expected given that (the few) published AUROCs typically report a range between 0.62-0.67 (Supplementary Table S2, [83]). The fact that published GWAS AUROCs tend to be high (~0.65) and unpublished GWAS AUROCs tend to be low (~0.55), suggests that one reason for the paucity of published GWAS AUROCs is that many AUROCs for SNP biomarker profiles are either uninterestingly low (<0.55), or not statistically different from those generated by a random predictor.

Another aspect that we wanted to explore in more detail was the performance of GWAS-derived SNP profiles for disease prediction compared to non-GWAS profiles for predicting identical diseases. As noted previously, the fact that GWAS disease risk assessments are not typically presented or measured in the same way as non-GWAS disease risk assessments, has made this kind of comparison difficult. As seen in Table 4, we found that non-genetic factors were generally better at predicting disease than genetic factors. In particular, for those conditions where GWAS-derived AUROCs and non-GWAS (clinical/proteomic/metabolomic) derived AUROCs could be compared, we found that a typical non-GWA study reported AUROCs closer to 0.81, which is

significantly more than the average GWAS-derived biomarker profile of 0.64 (see Table 4). On the other hand, it is important to note that the predictive ability or disease risk scores of SNP-derived biomarker profiles can effectively occur at birth (many decades prior to the onset of disease) while the non-SNP-derived biomarkers are generally only useful a few months or at most a few years prior to the onset of the disease. In this regard, the utility of SNP-biomarker profiles for long-term disease prevention or disease prophylaxis, even if modest compared to non-SNP profiles, is still quite significant.

A third objective of this study was to identify those conditions that appear to exhibit the best AUROC performance with multi-component SNP data. These high AUROC conditions would be expected to have a relatively high "explainable" genetic component with regard to disease risk. From the data compiled in GWAS Central we identified 5 conditions/phenotypes that had an AUROC greater than 0.75 and an estimated heritability of >25%. As can be seen in Supplementary Table S4, age-related macular degeneration (AUROC = 0.75), celiac disease (AUROC = 0.88), progressive supranuclear palsy (AUROC = 0.81), Craniofacial microsomia (AUROC = 0.76) and black hair colour (AUROC = 0.98) can be largely determined through SNP profiles. It is interesting to note that the very first SNP study ever recorded was one done on macular degeneration [1] and that macular degeneration has among the highest levels of heritability and among the highest AUROC values of all conditions we investigated. In many respects, macular degeneration was the equivalent of hitting the "mother lode" for GWA studies.

A fourth objective of this study was to identify those conditions where SNP information appears to be relatively uninformative with regard to disease risk prediction.

Our data indicates that 192 conditions (out of 219) have AUROCs <0.60, and 202 conditions (out of 219) have SNP heritability values of <15%. The majority of these conditions (such as type 1 diabetes, hypertension, amyotrophic lateral sclerosis and Parkinson's disease) are known to have a significant environmental or lifestyle contribution to disease risk. Indeed, non-GWAS derived risk scores (from clinical, proteomic or metabolomic studies) for several of these conditions have predictive AUROCs approaching 0.8 or 0.9. While many GWA studies have been undertaken due to moderately high heritability data estimated from twin studies, it appears that this heritability may be over-estimated due to small sample sizes or through undetected/unaccounted environmental (gut microbiome heritability) or epigenetic effects. We would suggest that prior to conducting large-scale GWA studies based on twin heritability estimates that some assessment regarding disease risk scores with (published) non-GWAS data should be conducted. Objectively assessing the contributions of clinical/phenotypic data, metabolomic data or proteomic data to disease risk prior to conducting a GWA study would certainly identify which kinds of conditions would most likely yield useful GWAS results or useful GWAS-derived biomarkers.

A fifth objective of this study was to demonstrate the utility of using simulated populations to model SNP distributions and to show how these populations, along with logistic regression modeling, could be used to create multi-marker SNP profiles from publicly available, summary-level GWAS data. Critical to this modeling is the assumption that all genotypes and allele proportions were in Hardy-Weinberg equilibrium with no linkage disequilibrium between genes. The method appears to be robust and particularly well-suited to handling most publicly available SNP data. It is also far faster

and less resource intensive than having to apply for study access and research ethics approvals (for non-public data) each time a new GWA study is released or a new GWA study is deposited in GWAS Central. However, simulated populations are not "real" and if the exact population and SNP structure is needed to answer a specific question or if absolute precision is needed in determining some SNP-derived model, then there is obviously no substitution for actual clinical data.

A sixth objective of this study was to explore whether certain trends in AUROCs, disease types or heritability estimates could be discerned by analyzing a large AUROC data set. As noted earlier, we found that the SNP-heritability as determined by Pawitan et al. [26], seemed to be well correlated with the square of the AUROC (r = 0.87, see Figure 2). This led to the discovery that $h^2 = (2 \cdot AUROC - 1)^2$. We used this newly derived formula to estimate the SNP-heritability for all 569 studies from GWAS Central. These heritability estimates can be found at http://gwasrocs.ca. Moreover, we compared our heritability estimates against the heritability values for 10 different conditions reported in the literature. Table 5 shows that on average our estimates differed from the true values by just 0.013. The formula we derived suggests that an AUROC of approximately 0.85 is needed to explain 50% of the heritability.

Table 5: Comparison of published heritability values versus those predicted using the AUROC for 10 conditions

Condition	Published SNP Heritability	AUROC derived SNP Heritability	Difference
Common psoriasis	0.280 [85]	0.270*	0.010
Crohn's disease	0.053 [86]	0.068**	0.015

Type 1 diabetes	0.127 [87]	0.130**	0.003
Coronary artery disease	0.106 [88]	0.073***	0.033
Age related macular degeneration	0.272 [89]	0.252***	0.020
Lupus	0.209 [90]	0.213***	0.004
Alzheimer's disease	0.059 [90]	0.055***	0.004
Type 2 diabetes	0.082 [90]	0.078***	0.004
Inflammatory bowel disease	0.117 [90]	0.091***	0.026
Rheumatoid arthritis	0.047 [91]	0.062***	0.015
Average:	0.135	0.129	0.013

* Using AUROC published in the same paper
** Using the true AUROCs calculated from raw WTCCC1 data
*** Using the highest predicted AUROC from GWAS Central

Much is often made of the difference in p-values between GWA studies and the p-values reported for non-GWA studies. As noted before, many GWA studies have SNPs with p-values < 1×10^{-50}, while it is rare for non-GWA studies to have clinical, protein or metabolite measures with p-values <1×10^{-6}. From our dataset one particular GWA study, "Alzheimer's disease" [92,93], had one SNP with the observed p-value (1×10^{-295}) of all SNPs in our GWAS Central dataset. This corresponds to a SNP located close to the well-known Alzheimer's disease-associated gene ApoE4 [94]. Using this single SNP alone it was possible to create an Alzheimer's disease risk predictor with an AUROC of 0.62. The addition of 3 more SNPs with p-values of 2×10^{-10}, 4×10^{-8}, 1×10^{-7} led to an increase in the Alzheimer's disease risk prediction AUROC to 0.65 [95]. So, despite the extremely low p-values for these Alzheimer's disease SNPs, the influence on the AUROC (and the heritability) was relatively modest. In another interesting example, the GWA study "Coronary Artery Disease" [96,97] had 50 SNPs, and a p-value of a staggering 1×10^{-101} for the most significant SNP. However, even with so many SNPs and the inclusion of

SNPs with remarkably low p-values, the AUROC of this SNP-panel reached just 0.58. Overall, our results indicate that it is not the p-value, but rather the odds ratio (OR), in conjunction with the risk allele frequency (RAF), that are most important for determining biomarker performance in disease risk prediction.

One criticism of our approach to calculating AUROCs from GWAS data is that is computationally inefficient. In particular, we construct large, simulated patient populations, and then used those simulated patient/SNP populations to estimate the AUROCs. A more efficient approach would have been to use machine learning methods or statistical techniques to predict the AUROC values directly. There are two reasons why we chose the population simulation approach. First, we believed it would be more useful for the scientific community to have access to simulated patient populations (with SNP data). This would allow others to perform their own statistical or modeling experiments. Furthermore, simulated SNP data can be used to create synthetic "patients" for electronic medical record (EMR) testing and training. Indeed, because of ethics and privacy restrictions, access to individual level GWAS data is often difficult, making generation of realistic genetic data for patients equally difficult. On the other hand, simulated individual level SNP data (and other 'omics' data) could be of great utility in the development and testing of "next-generation" EMR software and databases with realistic genetic data. As a result, G-WIZ was created as part of a separate EMR project to generate realistic "synthetic" patients with realistic conditions/phenotypes and correspondingly realistic genomic (SNP), metabolomic, proteomic and clinical profiles. In addition to the appeal of creating synthetic patient data, we also realized that by creating simulated populations we would be able to determine and plot ROC curves (with

which we could determine the AUROC values). Having a calculated ROC curve would give us another set of data with which to compare and validate our results. Indeed, we used the G-WIZ generated ROC curves to visually validate a number of the early ROC results during the testing phase of the program.

Conclusion

To summarize, we have created a software tool called G-WIZ to accurately predict GWAS ROC curves and AUROCs from summary level GWAS data. We subsequently compiled data from every sufficiently informative large-scale study in GWAS Central and calculated the corresponding ROC curves and AUROCs using G-WIZ. Using these calculated data, we conducted a number of comparisons to look for interesting results or unexpected trends. In particular, we compared these calculated GWAS AUROCs to typical AUROCs reported in other 'omics' studies and found some striking differences. We also derived a novel formula to calculate SNP-heritability and calculated the proportion of heritability explained by SNPs for all 569 GWAS Central studies that we analyzed. Through this assessment we were able to make some general suggestions regarding the evaluation and selection of medical conditions that should hopefully yield more significant and useful GWAS outcomes. The results of our G-WIZ calculations, along with other meta-data about each GWA study and the predicted heritability have been placed in an open-access database called GWAS-ROCS.

Acknowledgements

The authors thank Mr. John Bacon for his critical comments and suggested edits.

References

1. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 Apr 15;308(5720):385–9.

2. Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, et al. Genome-wide Analysis of Insomnia (N=1,331,010) Identifies Novel Loci and Functional Pathways. bioRxiv. 2018 Feb 1;214973.

3. Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur J Hum Genet.. 2014 Jul;22(7):949–52.

4. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.

5. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. N Engl J Med. 2010 Jul 8:363(2):166-76.

6. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Stat Med. 2015 Dec 10;34(28):3769–92.

7. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013 Apr;9(2):280–99.

8. Bitton A, Gaziano T. The Framingham Heart Study's Impact on Global Risk Assessment. Prog Cardiovasc Dis. 2010;53(1):68–78.

9. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun 7;447(7145):661–78.

10. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet. 2003 Apr 1;12(7):771–6.

11. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality Control Procedures for Genome Wide Association Studies. Curr Protoc Hum Genet. 2011 Jan;Chapter 1:Unit1.19. doi: 10.1002/0471142905.hg0119s68.

12. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2):e1608.

13. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol. 2012;8(12):e1002822.

14. Forberg JL, Green M, Björk J, Ohlsson M, Edenbrandt L, Öhlin H, et al. In search of the best method to predict acute coronary syndrome using only the electrocardiogram from the emergency department. J Electrocardiol. 2009 Jan 1;42(1):58–63.

15. Gulkesen KH, Koksal IT, Bilge U, Saka O. Comparison of methods for prediction of prostate cancer in Turkish men with PSA levels of 0-10 ng/mL. J Balk Union Oncol. 2010 Sep;15(3):537–42.

16. van der Ploeg T, Smits M, Dippel DW, Hunink M, Steyerberg EW. Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Med Res Methodol. 2011 Oct 25;11(1):143.

17. Wu J, Roy J, Stewart WF. Prediction Modeling Using EHR Data: Challenges, Strategies, and a Comparison of Machine Learning Approaches. Med Care. 2010 Jun;48:S106–13.

18. Lee HJ, Hwang SI, Han S, Park SH, Kim SH, Cho JY, et al. Image-based clinical decision support for transrectal ultrasound in the diagnosis of prostate cancer: comparison of multiple logistic regression, artificial neural network, and support vector machine. Eur Radiol. 2010 Jun 1;20(6):1476–84.

19. Muniz AMS, Liu H, Lyons KE, Pahwa R, Liu W, Nobre FF, et al. Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait. J Biomech. 2010 Mar 3;43(4):720–6.

20. Kim S, Kim W, Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. Healthc Inform Res. 2011 Dec 1;17(4):232–43.

21. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. J Interdiscip Math. 2010 Jun 1;13(3):253–67.

22. McDonald GC. Ridge regression. Wiley Interdiscip Rev Comput Stat. 2009 Jul 1;1(1):93–100.

23. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. 2018. Available from: https://www.R-project.org/.

24. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. J Mach Learn Res. 2016;17(170):1–5.

25. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77.

26. Pawitan Y, Seng KC, Magnusson PKE. How Many Genetic Variants Remain to Be Discovered? PLoS One. 2009 Dec 2;4(12):e7969.

27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837–45.

28. Lango H, Palmer CNA, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, et al. Assessing the Combined Impact of 18 Common Genetic Variants of Modest Effect Sizes on Type 2 Diabetes Risk. Diabetes. 2008 Nov;57(11):3129–35.

29. Sparsø T, Grarup N, Andreasen C, Albrechtsen A, Holmkvist J, Andersen G, et al. Combined analysis of 19 common validated type 2 diabetes susceptibility gene variants shows moderate discriminative value and no evidence of gene–gene interaction. Diabetologia. 2009 Jul 1;52(7):1308–14.

30. Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med. 2006 Oct;3(10):e374.

31. Miyake K, Yang W, Hara K, Yasuda K, Horikawa Y, Osawa H, et al. Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. J Hum Genet. 2009 Apr;54(4):236–41.

32. Dunlop MG, Tenesa A, Farrington SM, Ballereau S, Brewster DH, Koessler T, et al. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. Gut. 2013 Jun;62(6):871–81.

33. Hu C, Zhang R, Wang C, Wang J, Ma X, Lu J, et al. PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. PloS One. 2009 Oct 28;4(10):e7643.

34. Fontaine-Bisson B, Renström F, Rolandsson O, MAGIC, Payne F, Hallmans G, et al. Evaluating the discriminative power of multi-trait genetic risk scores for type 2 diabetes in a northern Swedish population. Diabetologia. 2010 Oct;53(10):2155–62.

35. Davies RW, Dandona S, Stewart AFR, Chen L, Ellis SG, Tang WHW, et al. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. Circ Cardiovasc Genet. 2010 Oct;3(5):468–74.

36. Chen H, Poon A, Yeung C, Helms C, Pons J, Bowcock AM, et al. A genetic risk score combining ten psoriasis risk loci improves disease prediction. PloS One. 2011 Apr 29;6(4):e19454.

37. Nakaoka H, Cui T, Tajima A, Oka A, Mitsunaga S, Kashiwase K, et al. A systems genetics approach provides a bridge from discovered genetic variants to biological pathways in rheumatoid arthritis. PloS One. 2011;6(9):e25389.

38. Darabi H, Czene K, Zhao W, Liu J, Hall P, Humphreys K. Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. Breast Cancer Res. 2012 Feb 7;14(1):R25.

39. Hüsing A, Canzian F, Beckmann L, Garcia-Closas M, Diver WR, Thun MJ, et al. Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. J Med Genet. 2012 Sep 1;49(9):601–8.

40. Akamatsu S, Takahashi A, Takata R, Kubo M, Inoue T, Morizono T, et al. Reproducibility, performance, and clinical utility of a genetic risk prediction model for prostate cancer in Japanese. PloS One. 2012;7(10):e46454.

41. Li H, Yang L, Zhao X, Wang J, Qian J, Chen H, et al. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. BMC Med Genet. 2012 Dec 10;13:118.

42. Chang J, Huang Y, Wei L, Ma B, Miao X, Li Y, et al. Risk prediction of esophageal squamous-cell carcinoma with common genetic variants and lifestyle factors in Chinese population. Carcinogenesis. 2013 Aug;34(8):1782–6.

43. Scott IC, Seegobin SD, Steer S, Tan R, Forabosco P, Hinks A, et al. Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking. PLoS Genet. 2013;9(9):e1003808.

44. Bruzelius M, Bottai M, Sabater-Lleal M, Strawbridge RJ, Bergendal A, Silveira A, et al. Predicting venous thrombosis in women using a combination of genetic markers and clinical risk factors. J Thromb Haemost. 2015 Feb;13(2):219–27.

45. Sleegers K, Bettens K, De Roeck A, Van Cauwenberghe C, Cuyvers E, Verheijen J, et al. A 22-single nucleotide polymorphism Alzheimer's disease risk score correlates with family history, onset age, and cerebrospinal fluid Aβ42. Alzheimers Dement J Alzheimers Assoc. 2015 Dec;11(12):1452–60.

46. Ibáñez-Sanz G, Díez-Villanueva A, Alonso MH, Rodríguez-Moranta F, Pérez-Gómez B, Bustamante M, et al. Risk Model for Colorectal Cancer in Spanish Population Using Environmental and Genetic Factors: Results from the MCC-Spain study. Sci Rep. 2017 24;7:43263.

47. Wang N, Wang Z, Wang C, Fu X, Yu G, Yue Z, et al. Prediction of leprosy in the Chinese population based on a weighted genetic risk score. PLoS Negl Trop Dis. 2018;12(9):e0006789.

48. Taylor KD, Guo X, Zangwill LM, Liebmann JM, Girkin CA, Feldman RM, et al. Genetic Architecture of Primary Open-Angle Glaucoma in Individuals of African Descent: The African Descent and Glaucoma Evaluation Study III. Ophthalmology. 2019 Jan;126(1):38–48.

49. Lin BD, Mbarek H, Willemsen G, Dolan CV, Fedko IO, Abdellaoui A, et al. Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. Genes. 2015 Jul 13;6(3):559–76.

50. GWAS-ROCS Database: Showing GR-Card for GR00070: Black vs. non-black hair color (HGVRS4156) [Internet]. [cited 2019 Jun 13]. Available from: http://gwasrocs.ca/study_simulations/GR00070

51. Crosslin DR, Carrell DS, Burt A, Kim DS, Underwood JG, Hanna DS, et al. Genetic variation in the HLA region is associated with susceptibility to herpes zoster. Genes Immun. 2015 Feb;16(1):1–7.

52. GWAS-ROCS Database: Showing GR-Card for GR00478: Shingles (HGVRS5860) [Internet]. [cited 2019 Jun 13]. Available from: http://gwasrocs.ca/study_simulations/GR00478

53. Mapstone M, Cheema AK, Fiandaca MS, Zhong X, Mhyre TR, MacArthur LH, et al. Plasma phospholipids identify antecedent memory impairment in older adults. Nat Med. 2014 Apr;20(4):415–8.

54. Barnes DE, Covinsky KE, Whitmer RA, Kuller LH, Lopez OL, Yaffe K. Dementia Risk Indices: A Framework for Identifying Individuals with a High Dementia Risk. Alzheimers Dement J Alzheimers Assoc. 2010 Mar;6(2):138–41.

55. Xiao Q, Liu Z-J, Tao S, Sun Y-M, Jiang D, Li H-L, et al. Risk prediction for sporadic Alzheimer's disease using genetic risk score in the Han Chinese population. Oncotarget. 2015 Nov 10;6(35):36955–64.

56. Floegel A, Stefan N, Yu Z, Mühlenbruch K, Drogan D, Joost H-G, et al. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. Diabetes. 2013 Feb 1;62(2):639–48.

57. Paprott R, Mühlenbruch K, Mensink GBM, Thiele S, Schulze MB, Scheidt-Nave C, et al. Validation of the German Diabetes Risk Score among the general adult population: findings from the German Health Interview and Examination Surveys. BMJ Open Diabetes Res Care. 2016 Nov 1;4(1):e000280.

58. Walford GA, Porneala BC, Dauriz M, Vassy JL, Cheng S, Rhee EP, et al. Metabolite Traits and Genetic Risk Provide Complementary Information for the Prediction of Future Type 2 Diabetes. Diabetes Care. 2014 Sep;37(9):2508–14.

59. Buitendijk GHS, Rochtchina E, Myers C, van Duijn CM, Lee KE, Klein BEK, et al. Prediction of age-related macular degeneration in the general population: the Three Continent AMD Consortium. Ophthalmology. 2013 Dec;120(12):2644–55.

60. Grassmann F, Fritsche LG, Keilhauer CN, Heid IM, Weber BHF. Modelling the Genetic Risk in Age-Related Macular Degeneration. PLoS One. 2012;7(5):e37979.

61. Pujos-Guillot E, Brandolini M, Pétéra M, Grissa D, Joly C, Lyan B, et al. Systems Metabolomics for Prediction of Metabolic Syndrome. J Proteome Res. 2017 02;16(6):2262–72.

62. Bener A, Yousafzai MT, Darwish S, Al-Hamaq AOAA, Nasralla EA, Abdul-Ghani M. Obesity index that better predict metabolic syndrome: body mass index, waist circumference, waist hip ratio, or waist height ratio. J Obes. 2013;2013:269038.

63. van Pelt ED, Mescheriakova JY, Makhani N, Ketelslegers IA, Neuteboom RF, Kundu S, et al. Risk genes associated with pediatric-onset MS but not with monophasic acquired CNS demyelination. Neurology. 2013 Dec 3;81(23):1996–2001.

64. Qiu Y, Cai G, Zhou B, Li D, Zhao A, Xie G, et al. A distinct metabolic signature of human colorectal cancer with prognostic potential. Clin Cancer Res Off J Am Assoc Cancer Res. 2014 Apr 15;20(8):2136–46.

65. Qiao J, Fang C-Y, Chen S-X, Wang X-Q, Cui S-J, Liu X-H, et al. Stroma derived COL6A3 is a potential prognosis marker of colorectal carcinoma revealed by quantitative proteomics. Oncotarget. 2015 Oct 6;6(30):29929–46.

66. Kim Y, Jeon J, Mejia S, Yao CQ, Ignatchenko V, Nyalwidhe JO, et al. Targeted proteomics identifies liquid-biopsy signatures for extracapsular prostate cancer. Nat Commun. 2016 28;7:11906.

67. Helfand BT, Fought AJ, Loeb S, Meeks JJ, Kan D, Catalona WJ. Genetic prostate cancer risk assessment: common variants in 9 genomic regions are associated with cumulative risk. J Urol. 2010 Aug;184(2):501–5.

68. Heffernan C, Doroshenko A, Egedahl ML, Barrie J, Senthilselvan A, Long R. Predicting pulmonary tuberculosis in immigrants: a retrospective cohort study. ERJ Open Res. 2018 Apr;4(2).

69. Hong EP, Go MJ, Kim H-L, Park JW. Risk prediction of pulmonary tuberculosis using genetic and conventional risk factors in adult Korean population. PloS One. 2017;12(3):e0174642.

70. Gander J. Factors Related to Coronary Heart Disease Risk Among Men: Validation of the Framingham Risk Score. Prev Chronic Dis. 2014 Aug 14;11:E140.

71. Kukava NG, Titov BV, Osmak GJ, Matveeva NA, Kulakova OG, Favorov AV, et al. Multilocus Analysis of Genetic Susceptibility to Myocardial Infarction in Russians: Replication Study. Acta Naturae. 2017;9(4):74–83.

72. Winkel RR, von Euler-Chelpin M, Nielsen M, Petersen K, Lillholm M, Nielsen MB, et al. Mammographic density and structural features can individually and jointly contribute to breast cancer risk assessment in mammography screening: a case–control study. BMC Cancer. 2016 Jul 7;16(1):414.

73. Chae YK, Gonzalez-Angulo AM. Implications of functional proteomics in breast cancer. The Oncologist. 2014 Apr;19(4):328–35.

74. Hsieh Y-C, Tu S-H, Su C-T, Cho E-C, Wu C-H, Hsieh M-C, et al. A polygenic risk score for breast cancer risk in a Taiwanese population. Breast Cancer Res Treat. 2017 May;163(1):131–8.

75. Al-Mubarak R, Vander Heiden J, Broeckling CD, Balagon M, Brennan PJ, Vissa VD. Serum metabolomics reveals higher levels of polyunsaturated fatty acids in lepromatous leprosy: potential markers for susceptibility and pathogenesis. PLoS Negl Trop Dis. 2011 Sep;5(9):e1303.

76. Haznadar M, Cai Q, Krausz KW, Bowman ED, Margono E, Noro R, et al. Urinary Metabolite Risk Biomarkers of Lung Cancer: A Prospective Cohort Study. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2016;25(6):978–86.

77. Raghu VK, Zhao W, Pu J, Leader JK, Wang R, Herman J, et al. Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models. Thorax. 2019 Jul 1;74(7):643–9.

78. Somers RH. A New Asymmetric Measure of Association for Ordinal Variables. Am Sociol Rev. 1962;27(6):799–811.

79. Mittlböck M, Schemper M. Explained variation for logistic regression. Stat Med. 1996 Oct 15;15(19):1987–97.

80. Song L, Liu A, Consortium MGOS, Shi J. SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. Bioinformatics 2019 Mar 26. pii: btz176. doi: 10.1093/bioinformatics/btz176.

81. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.

82. Sandoval-Motta S, Aldana M, Martínez-Romero E, Frank A. The Human Microbiome and the Missing Heritability Problem. Front Genet. 2017;8:80.

83. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012 Jan 13;90(1):7–24.

84. Kundu S, Mihaescu R, Meijer CMC, Bakker R, Janssens ACJW. Estimating the predictive ability of genetic risk models in simulated data based on published results from genome-wide association studies. Front Genet. 2014 Jun 13;5:179.

85. Tsoi LC, Stuart PE, Tian C, Gudjonsson JE, Das S, Zawistowski M, et al. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. Nat Commun. 2017 May 24;8:15382.

86. Yang S-K, Hong M, Zhao W, Jung Y, Baek J, Tayebi N, et al. Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut. 2014 Jan 1;63(1):80–7.

87. Fan Y, Song Y-Q. Finding the Missing Heritability of Genome-wide Association Study Using Genotype Imputation. Sci Matters [Internet]. 2016 May 4; Available from: https://www.sciencematters.io/articles/201604000013

88. The CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013 Jan;45(1):25–33.

89. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet. 2016 Feb;48(2):134–43.

90. Watanabe K, Stringer S, Frei O, Mirkov MU, Polderman TJC, Sluis S van der, et al. A global overview of pleiotropy and genetic architecture in complex traits. bioRxiv. 2018 Dec 19;500090.

91. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014 Feb;506(7488): 376–81.

92. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA. 2010 May 12;303(18):1832–40.

93. GWAS-ROCS Database: Showing GR-Card for GR00020: Alzheimer's disease (HGVRS1094) [Internet]. [cited 2019 Jun 13]. Available from: http://gwasrocs.ca/study_simulations/GR00020

94. Yu C-E, Seltman H, Peskind ER, Galloway N, Zhou PX, Rosenthal E, et al. Comprehensive Analysis of APOE and Selected Proximate Markers for Late-onset Alzheimer Disease: Pattern of Linkage Disequilibrium and Disease/Marker Association. Genomics. 2007 Jun;89(6):655–65.

95. GWAS-ROCS Database: Showing GR-Card for GR00299: Late onset Alzheimer's disease (HGVRS1241) [Internet]. [cited 2019 Jun 13]. Available from: http://gwasrocs.ca/study_simulations/GR00299

96. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet. 2017 Sep;49(9):1385–91.

97. GWAS-ROCS Database: Showing GR-Card for GR00171: Coronary artery disease (HGVRS4076) [Internet]. [cited 2019 Jun 13]. Available from: http://gwasrocs.ca/study_simulations/GR00171

Supporting Information

Supplementary Table S1. SNP heritability calculated using the method described by Pawitan et al. (2009) for every study in GWAS-ROCS.

Supplementary Table S2. List of every GWA study found in PubMed which reported an AUROC.

Supplementary Table S3. A breakdown of the number of SNPs reported by the 569 studies we scraped from GWAS Central.

Supplementary Table S4. List with the highest predicted AUROC for each of the 219 diseases analyzed from GWAS Central.

Figure 1

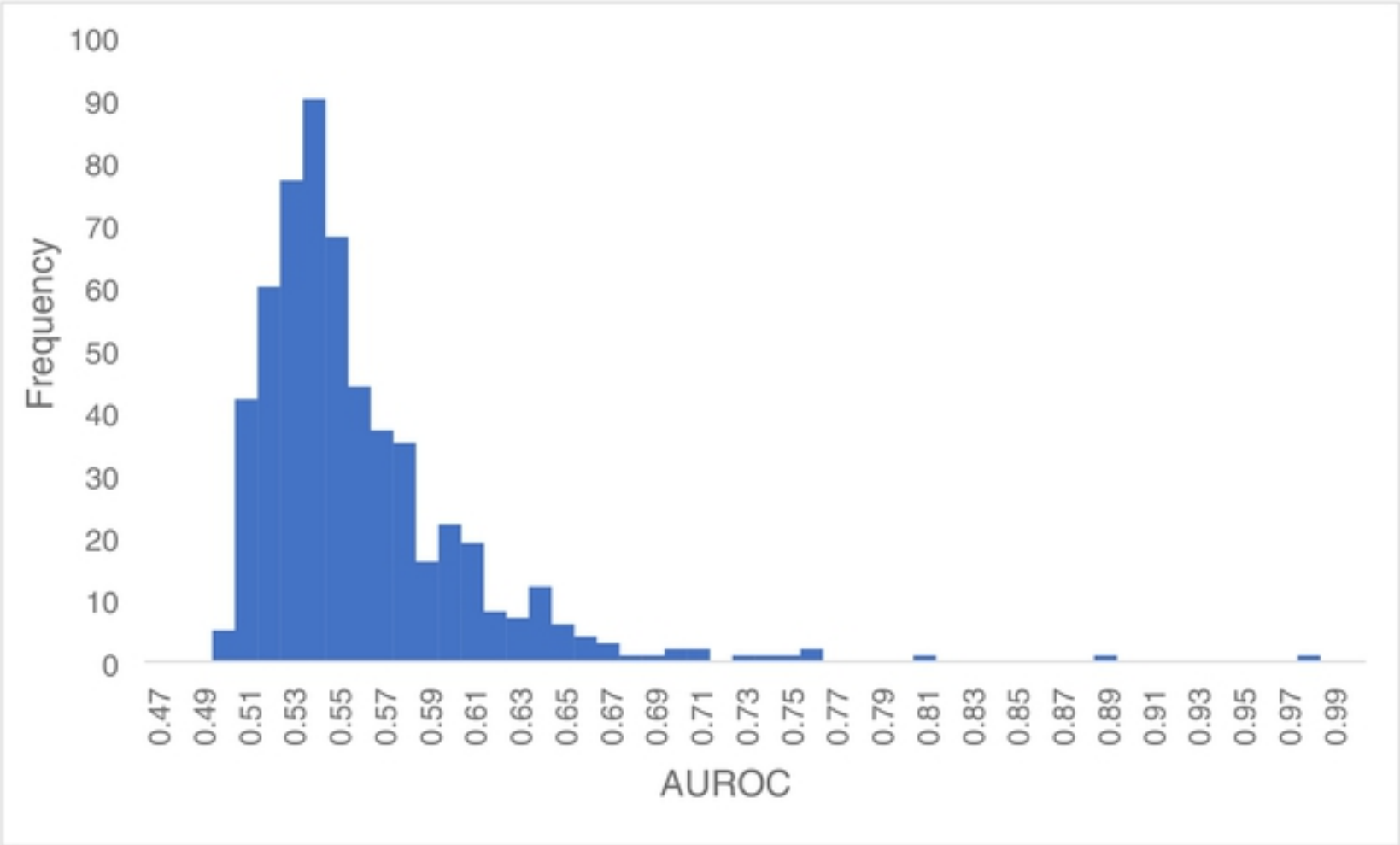

Figure 1

Figure 2

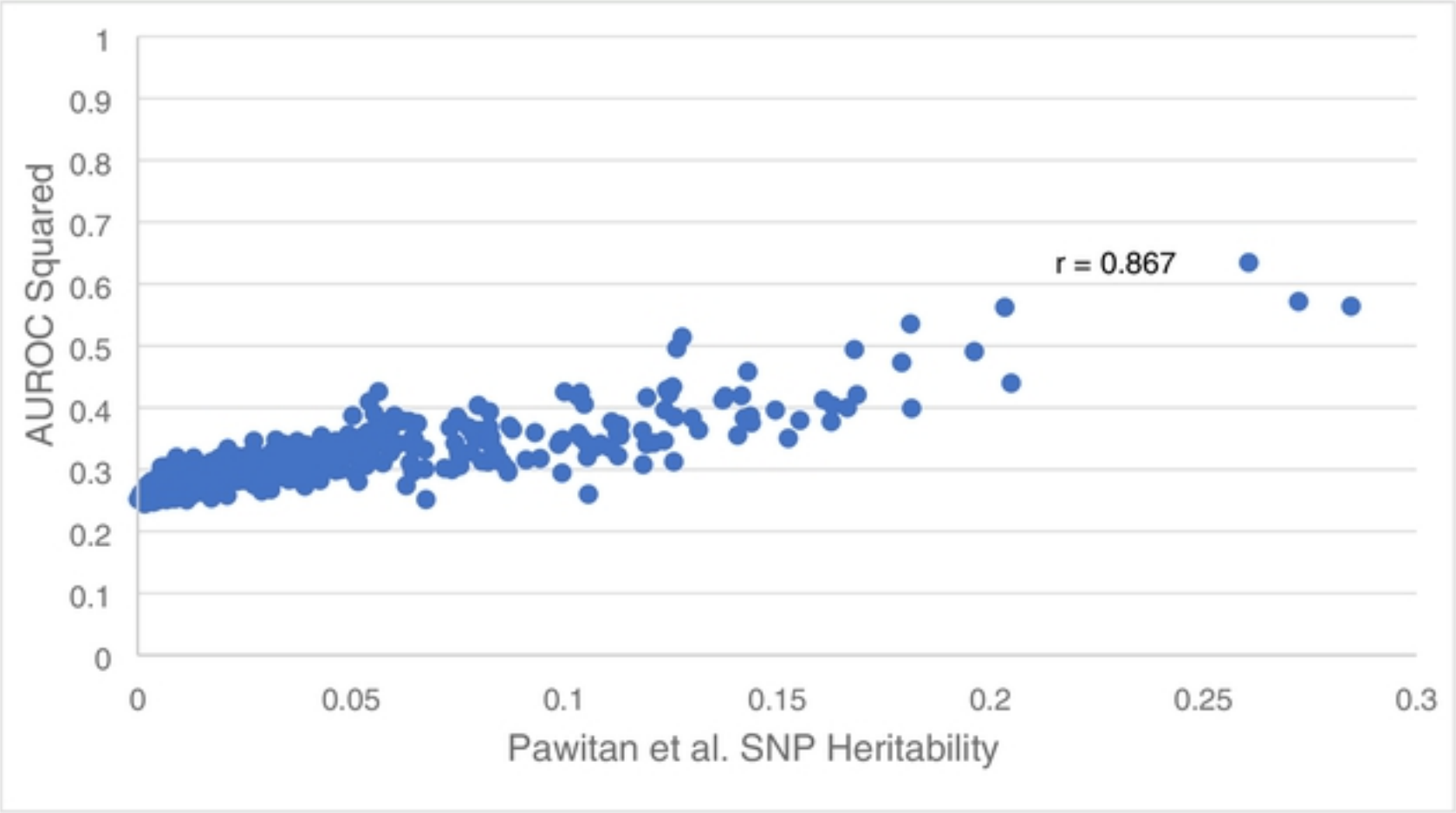

Figure 2

Figure 3

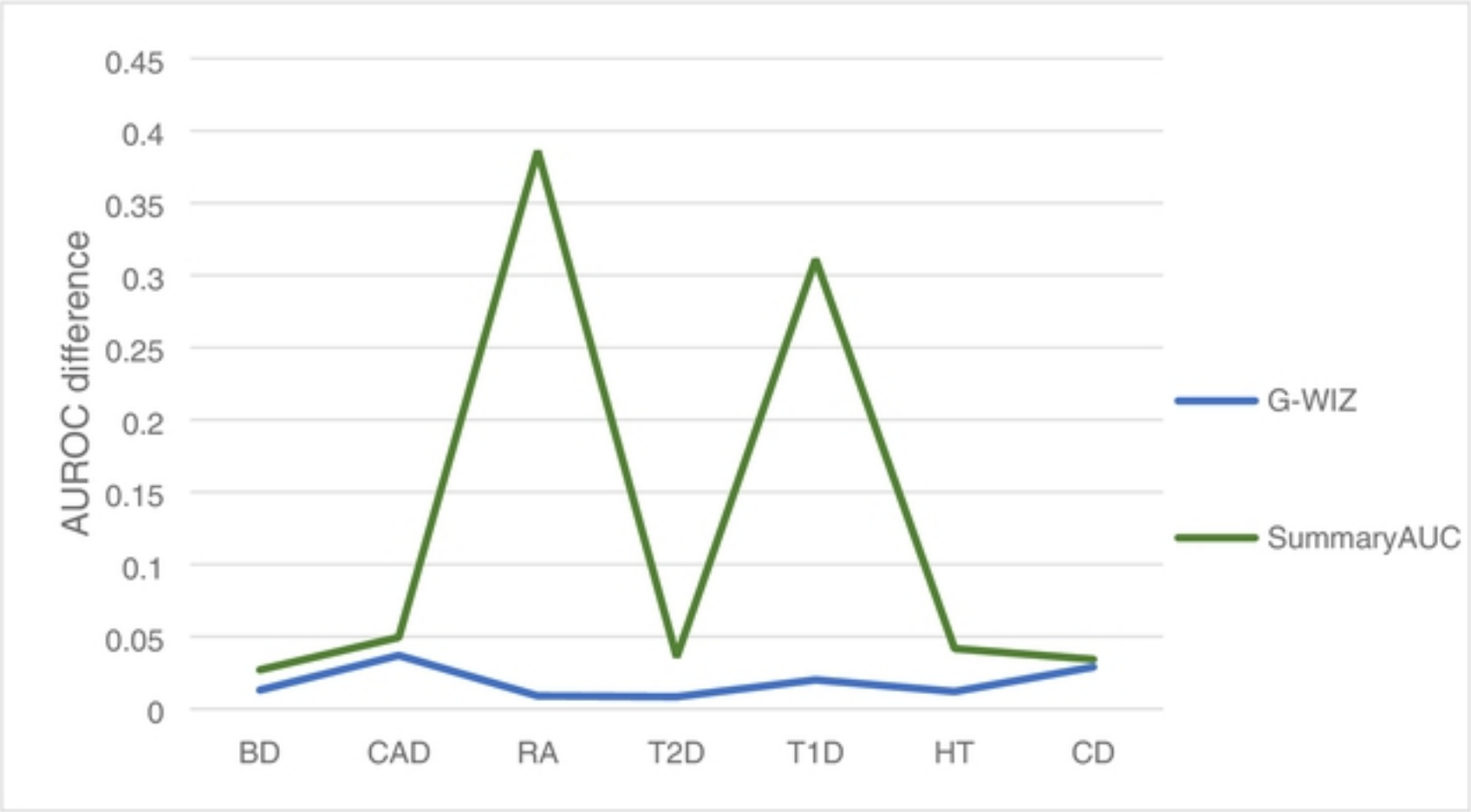

Figure 3

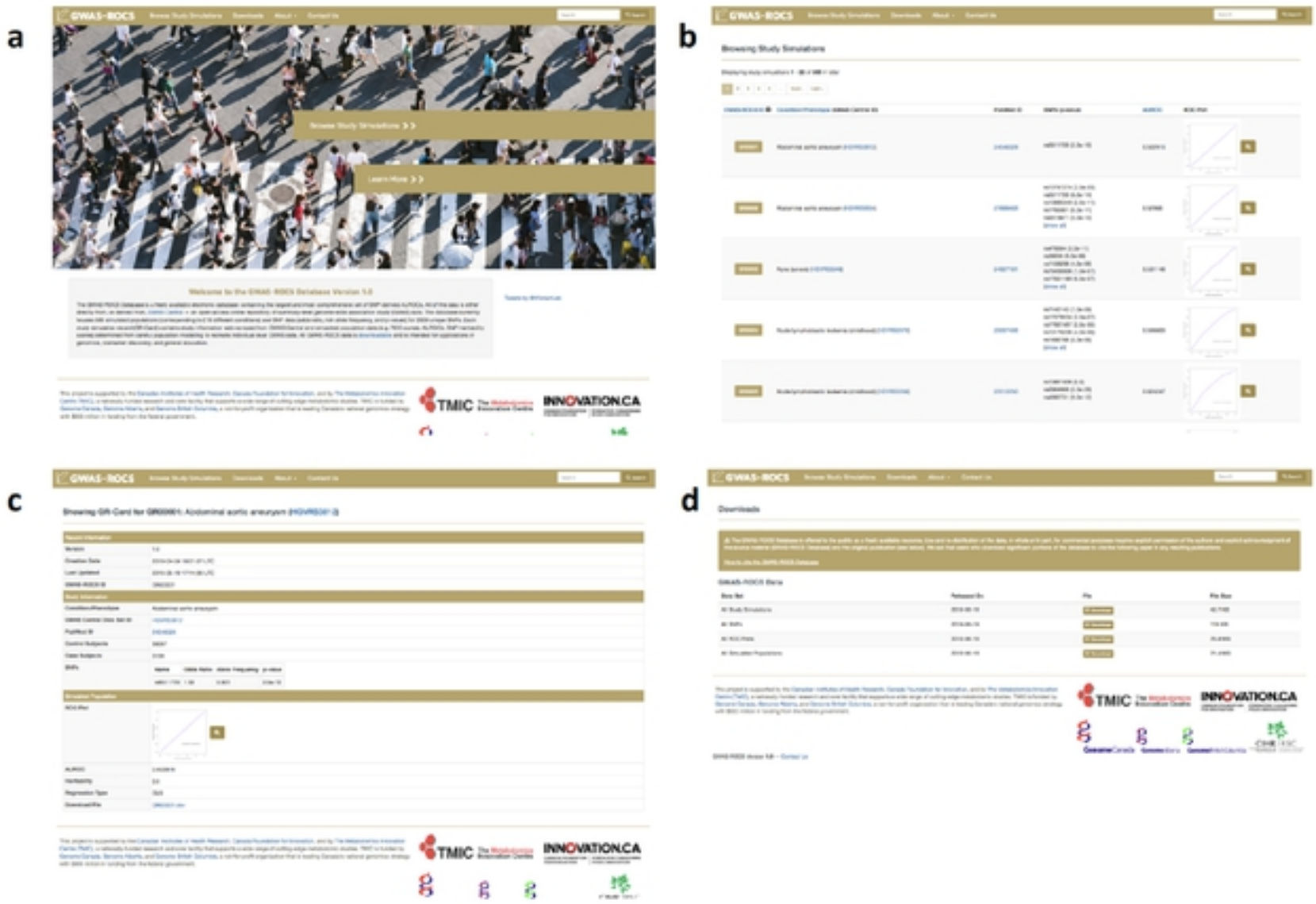

Figure 4

Figure 4